



CRC Press
Taylor & Francis Group

Tom Strachan
Anneke Lucassen

SECOND EDITION

Genetics and Genomics in Medicine



CRC Press
Taylor & Francis Group

Tom Strachan
Anneke Lucassen

SECOND EDITION

Genetics and Genomics in Medicine

GENETICS AND GENOMICS IN MEDICINE

SECOND EDITION

GENETICS AND GENOMICS IN MEDICINE

TOM STRACHAN AND ANNEKE LUCASSEN



CRC Press

Taylor & Francis Group

Boca Raton London New York

CRC Press is an imprint of the
Taylor & Francis Group, an **informa** business

Second edition published 2023

by CRC Press

6000 Broken Sound Parkway NW, Suite 300, Boca Raton, FL 33487-2742

and by CRC Press

4 Park Square, Milton Park, Abingdon, Oxon, OX14 4RN

CRC Press is an imprint of Taylor & Francis Group, LLC

© 2023 Taylor & Francis Group, LLC

This book contains information obtained from authentic and highly regarded sources. While all reasonable efforts have been made to publish reliable data and information, neither the author[s] nor the publisher can accept any legal responsibility or liability for any errors or omissions that may be made. The publishers wish to make clear that any views or opinions expressed in this book by individual editors, authors or contributors are personal to them and do not necessarily reflect the views/opinions of the publishers. The information or guidance contained in this book is intended for use by medical, scientific or healthcare professionals and is provided strictly as a supplement to the medical or other professional's own judgement, their knowledge of the patient's medical history, relevant manufacturer's instructions and the appropriate best practice guidelines. Because of the rapid advances in medical science, any information or advice on dosages, procedures or diagnoses should be independently verified. The reader is strongly urged to consult the relevant national drug formulary and the drug companies' and device or material manufacturers' printed instructions, and their websites, before administering or utilizing any of the drugs, devices or materials mentioned in this book. This book does not indicate whether a particular treatment is appropriate or suitable for a particular individual. Ultimately it is the sole responsibility of the medical professional to make his or her own professional judgements, so as to advise and treat patients appropriately. The authors and publishers have also attempted to trace the copyright holders of all material reproduced in this publication and apologize to copyright holders if permission to publish in this form has not been obtained. If any copyright material has not been acknowledged please write and let us know so we may rectify in any future reprint.

Except as permitted under U.S. Copyright Law, no part of this book may be reprinted, reproduced, transmitted, or utilized in any form by any electronic, mechanical, or other means, now known or hereafter invented, including photocopying, microfilming, and recording, or in any information storage or retrieval system, without written permission from the publishers.

For permission to photocopy or use material electronically from this work, access www.copyright.com or contact the Copyright Clearance Center, Inc. (CCC), 222 Rosewood Drive, Danvers, MA 01923, 978-750-8400. For works that are not available on CCC please contact mpkbookspermissions@tandf.co.uk

Trademark notice: Product or corporate names may be trademarks or registered trademarks and are used only for identification and explanation without intent to infringe.

ISBN: 978-0-367-49082-9 (hbk)

ISBN: 978-0-367-49081-2 (pbk)

ISBN: 978-1-003-04440-6 (ebk)

DOI: [10.1201/9781003044406](https://doi.org/10.1201/9781003044406)

Typeset in Utopia

by Apex CoVantage, LLC

Access the Support Material at: <https://www.routledge.com/9780367490812>

Contents

[Preface](#)

[Acknowledgements](#)

[1 FUNDAMENTALS OF DNA, CHROMOSOMES, AND CELLS](#)

[1.1 THE STRUCTURE AND FUNCTION OF NUCLEIC ACIDS](#)

[General concepts: the genetic material, genomes, and genes](#)

[The underlying chemistry of nucleic acids](#)

[Base pairing and the double helix](#)

[DNA replication and DNA polymerases](#)

[Genes, transcription, and the central dogma of molecular biology](#)

[1.2 THE STRUCTURE AND FUNCTION OF CHROMOSOMES](#)

[Why we need highly structured chromosomes, and how they are organized](#)

[Chromosome function: replication origins, centromeres, and telomeres](#)

[1.3 DNA AND CHROMOSOMES IN CELL DIVISION AND THE CELL CYCLE](#)

[Differences in DNA copy number between cells](#)

[The cell cycle and segregation of replicated chromosomes and DNA molecules](#)

[Mitosis: the usual form of cell division](#)

[Meiosis: a specialized reductive cell division giving rise to sperm and egg cells](#)

Why each of our gametes is unique

SUMMARY

QUESTIONS

FURTHER READING

2 FUNDAMENTALS OF GENE STRUCTURE, GENE EXPRESSION, AND HUMAN GENOME ORGANIZATION

2.1 PROTEIN-CODING GENES: STRUCTURE AND EXPRESSION

Gene organization: exons and introns

RNA splicing: stitching together the genetic information in exons

Translation: decoding messenger RNA to make a polypeptide

From newly synthesized polypeptide to mature protein

2.2 RNA GENES AND NONCODING RNA

The extraordinary secondary structure and versatility of RNA

RNAs that act as specific regulators: from quirky exceptions to the mainstream

2.3 WORKING OUT THE DETAILS OF OUR GENOME AND WHAT THEY MEAN

The Human Genome Project: working out the details of the nuclear genome

What the sequence didn't tell us and the goal of identifying all functional human DNA sequences

2.4 A QUICK TOUR OF SOME ELECTRONIC RESOURCES USED TO INTERROGATE THE HUMAN GENOME SEQUENCE AND GENE PRODUCTS

Gene nomenclature and the HGNC gateway

Databases storing nucleotide and protein sequences

Finding related nucleotide and protein sequences

Links to clinical databases

2.5 THE ORGANIZATION AND EVOLUTION OF THE HUMAN GENOME

A brief overview of the evolutionary mechanisms that shaped our genome

How much of our genome is functionally significant?

The mitochondrial genome: economical usage but limited autonomy.

Gene distribution in the human genome

The extent of repetitive DNA in the human genome

The organization of gene families

The significance of gene duplication and repetitive coding DNA

Highly repetitive noncoding DNA in the human genome

SUMMARY

QUESTIONS

FURTHER READING

3 PRINCIPLES UNDERLYING CORE DNA TECHNOLOGIES

3.1 AMPLIFYING DNA BY DNA CLONING

Amplifying desired DNA within bacterial cells

The need for vector DNA molecules

Physical clone separation

The need for restriction nucleases

DNA libraries and the uses and limitations of DNA cloning

3.2 AMPLIFYING DNA USING THE POLYMERASE CHAIN REACTION (PCR)

Basics of the polymerase chain reaction (PCR)

Quantitative PCR and real-time PCR

3.3 PRINCIPLES OF NUCLEIC ACID HYBRIDIZATION

Formation of artificial heteroduplexes

Hybridization assays: using known nucleic acids to find related sequences in a test nucleic acid population

Microarray hybridization: large-scale parallel hybridization to immobilized probes

3.4 PRINCIPLES OF DNA SEQUENCING

Dideoxy DNA sequencing

Massively parallel DNA sequencing (next-generation sequencing)

SUMMARY

QUESTIONS

FURTHER READING

4 PRINCIPLES OF GENETIC VARIATION

4.1 DNA SEQUENCE VARIATION ORIGINS AND DNA REPAIR

Genetic variation arising from errors in chromosome and DNA function

Various endogenous and exogenous sources can cause damage to DNA by altering its chemical structure

The wide range of DNA repair mechanisms

Repair of DNA damage or altered sequence on a single DNA strand

Repair of DNA lesions that affect both DNA strands

Undetected DNA damage, DNA damage tolerance, and translesion synthesis

4.2 POPULATION GENOMICS AND THE SCALE OF HUMAN GENETIC VARIATION

DNA variants, polymorphisms, and human population genomics

Small-scale variation: single nucleotide variants and small insertions and deletions

Microsatellites and other variable number of tandem repeat (VNTR) polymorphisms

Structural variation and low copy number variation

Taking stock of human genetic variation

4.3 FUNCTIONAL GENETIC VARIATION AND PROTEIN POLYMORPHISM

The vast majority of genetic variation has a neutral effect on the phenotype, but a small fraction is harmful

Different types of Darwinian natural selection operate in human lineages

Generating protein diversity by gene duplication: the example of olfactory receptor genes

4.4 EXTRAORDINARY GENETIC VARIATION IN THE IMMUNE SYSTEM

Pronounced genetic variation in four classes of immune system proteins

Programmed and random post-zygotic genetic variation

Somatic mechanisms allow cell-specific production of immunoglobulins and T-cell receptors

MHC (HLA) proteins: functions and polymorphism

The medical importance of the HLA system

SUMMARY

QUESTIONS

FURTHER READING

5 SINGLE-GENE DISORDERS: INHERITANCE PATTERNS, PHENOTYPE VARIABILITY, AND ALLELE FREQUENCIES

5.1 INTRODUCTION: TERMINOLOGY, ELECTRONIC RESOURCES, AND PEDIGREES

Background terminology and electronic resources with information on single-gene disorders

Investigating family history of disease and recording pedigrees

5.2 THE BASICS OF MENDELIAN AND MITOCHONDRIAL DNA INHERITANCE PATTERNS

Autosomal dominant inheritance

Autosomal recessive inheritance

Sex-linked inheritance

Matrilineal inheritance for mitochondrial DNA disorders

5.3 UNCERTAINTY, HETEROGENEITY, AND VARIABLE EXPRESSION OF MENDELIAN PHENOTYPES

Difficulties in defining the mode of inheritance in small pedigrees

Heterogeneity in the correspondence between phenotypes and the underlying genes and mutations

Nonpenetrance and age-related penetrance

5.4 ALLELE FREQUENCIES IN POPULATIONS

Allele frequencies and the Hardy-Weinberg law

Applications and limitations of the Hardy-Weinberg law

Ways in which allele frequencies change in populations

Population bottlenecks and founder effects

Mutation versus selection in determining allele frequencies

Heterozygote advantage: when natural selection favors carriers of recessive disease

SUMMARY

QUESTIONS

FURTHER READING

6 PRINCIPLES OF GENE REGULATION AND EPIGENETICS

The two fundamental types of gene regulation

Cis-acting and trans-acting effects in gene regulation

6.1 GENETIC REGULATION OF GENE EXPRESSION

Promoters: the major on-off switches in genes

Modulating transcription and tissue-specific regulation

Transcription factor binding and specificity

Genetic regulation during RNA processing: RNA splicing and RNA editing

Translational regulation by *trans*-acting regulatory proteins

Post-transcriptional gene silencing by microRNAs

Repressing the repressors: competing endogenous RNAs sequester miRNA

6.2 CHROMATIN MODIFICATION AND EPIGENETIC FACTORS IN GENE REGULATION

An overview of the molecular basis of epigenetic mechanisms

How changes in chromatin structure produce altered gene expression

Histone modification and histone substitution in nucleosomes

Modified histones and histone variants affect chromatin structure

The function of DNA methylation in mammalian cells

DNA methylation: mechanisms, heritability, and global roles during early development and gametogenesis

Long noncoding RNAs in mammalian epigenetic regulation

Genomic imprinting: differential expression of maternally and paternally inherited alleles

X-chromosome inactivation: compensating for sex differences in gene dosage

6.3 ABNORMAL EPIGENETIC REGULATION IN MENDELIAN DISORDERS AND UNIPARENTAL DISOMY

Principles of epigenetic dysregulation

“Chromatin diseases” due to mutations in genes specifying chromatin modifiers

Disease resulting from dysregulation of heterochromatin

Uniparental disomy and disorders of imprinting

Abnormal gene regulation at imprinted loci

SUMMARY

QUESTIONS

FURTHER READING

7 HOW GENETIC VARIATION IN DNA AND CHROMOSOMES CAUSES DISEASE

7.1 AN OVERVIEW OF HOW GENETIC VARIATION RESULTS IN DISEASE

The importance of repeat sequences in triggering pathogenesis

7.2 PATHOGENIC NUCLEOTIDE SUBSTITUTIONS AND TINY INSERTIONS AND DELETIONS

Pathogenic single nucleotide substitutions within coding sequences

Mutations that result in premature termination codons

Genesis and frequency of pathogenic point mutations

Surveying and curating point mutations that cause disease

7.3 PATHOGENESIS DUE TO VARIATION IN SHORT TANDEM REPEAT COPY NUMBER

The two main classes of pathogenic variation in short tandem repeat copy-number

Dynamic disease-causing mutations due to unstable expansion of short tandem repeats

Unstable expansion of short tandem repeats can cause disease in different ways

7.4 PATHOGENESIS TRIGGERED BY LONG TANDEM REPEATS AND INTERSPERSED REPEATS

Pathogenic exchanges between repeats occurs in both nuclear DNA and mtDNA

Nonallelic homologous recombination and transposition

Pathogenic sequence exchanges between chromatids at mispaired tandem repeats

Disease arising from sequence exchanges between distantly located repeats in nuclear DNA

7.5 CHROMOSOME ABNORMALITIES

Structural chromosomal abnormalities

Chromosomal abnormalities involving gain or loss of complete chromosomes

7.6 MOLECULAR PATHOLOGY OF MITOCHONDRIAL DISORDERS

Mitochondrial disorders due to mtDNA mutation show maternal inheritance and variable proportions of mutant genotypes

The two major classes of pathogenic DNA variant in mtDNA: large deletions and point mutations

7.7 EFFECTS ON THE PHENOTYPE OF PATHOGENIC VARIANTS IN NUCLEAR DNA

Mutations affecting how a single gene works: an overview of loss of function and gain of function

The effect of pathogenic variants depends on how the products of alleles interact: dominance and recessiveness revisited

Gain-of-function and loss-of-function mutations in the same gene can produce different phenotypes

Multiple gene dysregulation resulting from aneuploidies and mutations in regulatory genes

7.8 A PROTEIN STRUCTURE PERSPECTIVE OF MOLECULAR PATHOLOGY

Pathogenesis arising from protein misfolding

The many different ways in which protein aggregation can result in disease

7.9 GENOTYPE–PHENOTYPE CORRELATIONS AND WHY MONOGENIC DISORDERS ARE OFTEN NOT SIMPLE

The difficulty in getting reliable genotype–phenotype correlations

Modifier genes and environmental factors: common explanations for poor genotype–phenotype correlations

SUMMARY

QUESTIONS

FURTHER READING

8 IDENTIFYING DISEASE GENES AND GENETIC SUSCEPTIBILITY TO COMPLEX DISEASE

8.1 IDENTIFYING GENES IN MONOGENIC DISORDERS

A historical overview of identifying genes in monogenic disorders

Linkage analysis to map genes for monogenic disorders to defined subchromosomal regions

Chromosome abnormalities and other large-scale mutations as routes to identifying disease genes

Exome sequencing: let's not bother getting a position for disease genes!

8.2 APPROACHES TO MAPPING AND IDENTIFYING GENETIC SUSCEPTIBILITY TO COMPLEX DISEASE

The polygenic and multifactorial nature of common genetic disorders

Difficulties with lack of penetrance and phenotype classification in complex disease

Estimating heritability: the contribution made by genetic factors to the variance of complex diseases

The very limited success of linkage analyses in identifying genes underlying complex genetic diseases

The fundamentals of allelic association and the importance of HLA-disease associations

Linkage disequilibrium as the basis of allelic associations

How genomewide association studies are carried out

Moving from candidate subchromosomal region to identify causal genetic variants in complex disease can be challenging

The limitations of GWA studies and the issue of missing heritability

Alternative genome-wide studies and the role of rare variants and copy number variants in complex disease

The assessment and prediction of risk for common genetic diseases and the development of polygenic risk scores

8.3 ASPECTS OF THE GENETIC ARCHITECTURE OF COMPLEX DISEASE AND THE CONTRIBUTIONS OF ENVIRONMENTAL AND EPIGENETIC FACTORS

Common neurodegenerative disease: from monogenic to polygenic disease

The importance of immune system pathways in common genetic disease

The importance of protective factors and how a susceptibility factor for one complex disease may be a protective factor for another disease

Gene–environment interactions in complex disease

Epigenetics in complex disease and aging: significance and experimental approaches

SUMMARY

QUESTIONS

FURTHER READING

9 GENETIC APPROACHES TO TREATING DISEASE

9.1 AN OVERVIEW OF TREATING GENETIC DISEASE AND OF GENETIC TREATMENT OF DISEASE

Three different broad approaches to treating genetic disorders

Very different treatment options for different inborn errors of metabolism

Genetic treatment of disease may be conducted at many different levels

9.2 GENETIC INPUTS INTO TREATING DISEASE WITH SMALL MOLECULE DRUGS AND THERAPEUTIC

PROTEINS

An overview of how genetic differences affect the metabolism and performance of small molecule drugs

Phenotype differences arising from genetic variation in drug metabolism

Genetic variation in enzymes that work in phase II drug metabolism

Altered drug responses resulting from genetic variation in drug targets

When genotypes at multiple loci in patients are important in drug treatment: the example of warfarin

Translating genetic advances: from identifying novel disease genes to therapeutic small molecule drugs

Translating genomic advances and developing generic drugs as a way of overcoming the problem of too few drug targets

Developing biological drugs: therapeutic proteins produced by genetic engineering

Genetically engineered therapeutic antibodies with improved therapeutic potential

9.3 PRINCIPLES OF GENE AND CELL THERAPY

Two broad strategies in somatic gene therapy.

The delivery problem: designing optimal and safe strategies for getting genetic constructs into the cells of patients

Different ways of delivering therapeutic genetic constructs, and the advantages of *ex vivo* gene therapy.

Viral delivery of therapeutic gene constructs: relatively high efficiency but safety concerns

Virus vectors used in gene therapy.

The importance of disease models for testing potential therapies in humans

9.4 GENE THERAPY FOR INHERITED DISORDERS: PRACTICE AND FUTURE DIRECTIONS

Multiple successes for *ex vivo* gene supplementation therapy targeted at hematopoietic stem cells

In vivo gene therapy: approaches, barriers, and recent successes

An overview of RNA and oligonucleotide therapeutics

RNA interference therapy.

Future therapeutic prospects using CRISPR-Cas gene editing

Therapeutic applications of stem cells and cell reprogramming

Obstacles to overcome in cell therapy.

A special case: preventing transmission of severe mitochondrial DNA disorders by mitochondrial replacement

SUMMARY

QUESTIONS

FURTHER READING

10 CANCER GENETICS AND GENOMICS

10.1 FUNDAMENTAL CHARACTERISTICS AND EVOLUTION OF CANCER

The defining features of unregulated cell growth and cancer

Why cancers are different from other diseases: the contest between natural selection operating at the level of the cell and the level of the organism

Cancer cells acquire several distinguishing biological characteristics during their evolution

The initiation and multistage nature of cancer evolution and why most human cancers develop over many decades

Intratumor heterogeneity arises through cell infiltration, clonal evolution, and differentiation of cancer stem cells

10.2 ONCOGENES AND TUMOR SUPPRESSOR GENES

Two fundamental classes of cancer gene

Viral oncogenes and the natural roles of cellular oncogenes

How normal cellular proto-oncogenes are activated to become cancer genes

Tumor suppressor genes: normal functions, the two-hit paradigm, and loss of heterozygosity in linked markers

The key roles of gatekeeper tumor suppressor genes in suppressing G1-S transition in the cell cycle

The additional role of p53 in activating different apoptosis pathways to ensure that rogue cells are destroyed

Tumor suppressor involvement in rare familial cancers and non-classical tumor suppressors

The significance of miRNAs and long noncoding RNAs in cancer

10.3 GENOMIC INSTABILITY AND EPIGENETIC DYSREGULATION IN CANCER

Different types of chromosomal instability in cancer

Deficiency in mismatch repair results in unrepaired replication errors and global DNA instability

Different classes of cancer susceptibility gene according to epigenetic function, epigenetic dysregulation, and epigenome-genome interaction

10.4 NEW INSIGHTS FROM GENOME-WIDE STUDIES OF CANCERS

Genome sequencing has revealed extraordinary mutational diversity in tumors and insights into cancer evolution

Defining the landscape of driver mutations in cancer and establishing a complete inventory of cancer-susceptibility genes

Tracing the mutational history of cancers: just one of the diverse applications of single-cell genomics and transcriptomics in cancer

Genome-wide RNA sequencing enables insights into the link between cancer genomes and cancer biology and aids tumor classification

10.5 GENETIC INROADS INTO CANCER THERAPY

Targeted anticancer therapies are directed against key cancer cell proteins involved in oncogenesis or in escaping immunosurveillance

CAR-T Cell therapy and the use of genetically engineered T cells to treat cancer

The molecular basis of tumor recurrence and the evolution of drug resistance in cancers

The promise of combinatorial drug therapies

SUMMARY

QUESTIONS

FURTHER READING

11 GENETIC AND GENOMIC TESTING IN HEALTHCARE: PRACTICAL AND ETHICAL ASPECTS

11.1 AN OVERVIEW OF GENETIC TESTING

The different source materials and different levels of genetic testing

11.2 GENETIC TESTING FOR CHROMOSOME ABNORMALITIES AND PATHOGENIC STRUCTURAL VARIATION

Screening for aneuploidies using quantitative fluorescence PCR

Detecting large-scale copy number variants using chromosome SNP microarray analysis

Detecting and scanning for oncogenic fusion genes using, respectively, chromosome FISH and targeted RNA sequencing

Detecting pathogenic moderate- to small-scale deletions and duplications at defined loci is often achieved using the MLPA or ddPCR methods

Two very different routes towards universal genome-wide screens for structural variation: genome-wide sequencing and optical genome mapping

11.3 GENETIC AND GENOMIC TESTING FOR PATHOGENIC POINT MUTATIONS AND DNA METHYLATION TESTING

Diverse methods permit rapid genotyping of specific point mutations

The advantages of multiplex genotyping

Mutation scanning: from genes and gene panels to whole exome and whole genome sequencing

Interpreting and validating sequence variants can be aided by extensive online resources

Detecting aberrant DNA methylation profiles associated with disease

11.4 GENETIC AND GENOMIC TESTING: ORGANIZATION OF SERVICES AND PRACTICAL APPLICATIONS

The developing transformation of genetic services into mainstream genomic medicine

An overview of diagnostic and pre-symptomatic or predictive genetic testing

The different ways in which diagnosis of genetic conditions is carried out in the prenatal period

Preimplantation genetic testing is carried out to prevent the transmission of a harmful genetic defect using in vitro fertilization

Noninvasive prenatal testing (NIPT) and whole genome testing of the fetus

An overview of the different types of genetic screening

Pregnancy screening for fetal abnormalities

Newborn screening allows the possibility of early medical intervention

Different types of carrier screening can be carried out for autosomal recessive conditions

New genomic technologies are being exploited in cancer diagnostics

Bypassing healthcare services: the rise of direct-to-consumer (DTC) genetic testing

The downsides of improved sensitivity through whole genome sequencing: increased uncertainty about what variants mean

11.5 ETHICAL, LEGAL, AND SOCIETAL ISSUES (ELSI) IN GENETIC TESTING

Genetic information as family information

Consent issues in genetic testing

The generation of genetic data is outstripping the ability to provide clinical interpretation

New disease gene discovery and changing concepts of diagnosis

Complications in diagnosing mitochondrial disease

Complications arising from incidental, additional, secondary, or unexpected information

Consent issues in testing children

Ethical and societal issues in prenatal diagnosis and testing

Ethical and social issues in some emerging treatments for genetic disorders

The ethics of germline gene modification for gene therapy and genetic enhancement

SUMMARY

QUESTIONS

FURTHER READING

Glossary

Index

Preface

A rationale for establishing the first edition of *Genetics and Genomics in Medicine* was the suspicion that genomewide analyses might transform medicine. Using Sanger dideoxy sequencing the international Human Genome Project took about 13 years to deliver an almost complete genome sequence in 2003. Subsequent technological developments—first, genomewide microarray technologies and then massively parallel DNA sequencing—have certainly transformed genome analysis, permitting genome data in hours, not years.

The preface to the first edition of this book also included this question: might we soon live in societies where genome sequencing of citizens becomes the norm? Well, that day seems much closer now as millions of people have their genome sequenced, and debate has begun on whether population neonatal genome sequencing should be considered. The genome sequencing revolution found early major applications in medical genetics, then hematology and oncology, but is now being increasingly applied across multiple other medical disciplines. Various national genomic medicine initiatives have recently been established and, in 2020, NHS England became the first national health service to offer whole genome sequencing to patients as part of routine care.

In this book we try to summarize pertinent knowledge, and to structure it in the form of principles, rather than seek to compartmentalize information into chapters on topics such as epigenetics, evolutionary genetics, immunogenetics, pharmacogenetics, and so on. To help readers find broad topics that might be dealt with in two or more chapters, we provide a road

map on the inside front cover that charts how some broad themes are distributed between different chapters.

We start with three introductory chapters that provide basic background details. [Chapters 1](#) and [2](#) cover the fundamentals of DNA, chromosomes, the cell cycle, human genome organization and gene expression. [Chapter 3](#) introduces the basics of three core molecular genetic approaches used to manipulate DNA: DNA amplification (by DNA cloning or PCR), nucleic acid hybridization, and DNA sequencing, but we delay bringing in applications of these fundamental methods until later chapters, setting them against appropriate contexts that directly explain their relevance.

The next three chapters provide some background principles at a higher level. In [Chapter 4](#), we take a broad look at general principles of genetic variation, including DNA repair mechanisms and some detail on functional variation (but we consider how genetic variation contributes to disease in later chapters, notably [chapters 7](#), [8](#) and [10](#)). [Chapter 5](#) takes a look at how genes are transmitted in families and at allele frequencies in populations. [Chapter 6](#) moves from the basic principles of gene expression covered in [chapter 2](#) to explaining how genes are regulated by a wide range of protein and noncoding RNA regulators, and the central role of regulatory sequences in both DNA and RNA. In this chapter, too, we outline the principles of chromatin modification and epigenetic regulation and explain how aberrant chromatin structure underlies many single gene disorders.

The remainder of the book is largely devoted to clinical applications. We explain in [chapter 7](#) how chromosome abnormalities arise and their consequences, and how mutations and large-scale DNA changes can directly cause disease. In [chapter 8](#), we look at how genes underlying single gene disorders are identified, and also how genetic variants conferring susceptibility to complex diseases are identified. Then we consider the ways in which genetic variants, epigenetic dysregulation and environmental factors all make important contributions to complex diseases. [Chapter 9](#) briefly covers the wide the range of approaches for treating genetic disorders, before examining in detail how genetic approaches are used directly and indirectly in treating disease. In this chapter, too, we examine

how genetic variation affects how we respond to drug treatment. [Chapter 10](#) deals with cancer genetics and genomics and explains how cancers arise from a combination of abnormal genetic variants and epigenetic dysregulation. Finally, [Chapter 11](#) takes a broad look at diagnostic applications (and the exciting applications offered by new genome-wide technologies), plus ethical considerations in diagnosis and in some novel therapies.

Important recent advances have been made in applying genetic and genomic technologies to understanding pathogenesis, and in developing novel genetic testing methods, (including noninvasive ones), and novel treatments. There has been significant improvement, too, in pharmacogenomic approaches and in prenatal and preconception options to avoid serious genetic disease. Now we are no longer bound by the old approach of starting with a phenotype and then searching for a confirmatory genotype but can invert the process to predict phenotypes over a lifetime from a genotype. But challenges remain. Predicting phenotypes over a lifetime from a genotype, for example, is rarely clear-cut; the more we test without medical indications, the less likely we will predict diseases accurately. And, while acquiring genetic and genomic data is no longer the major rate-limiting step it was, data interpretation has become a huge challenge given the inherent complexities of interpreting the 4–5 million variants in a person’s genome and their implications for [ill] health.

Mainstreaming of genomic medicine—placing it at the center of healthcare — may be appealing, but its utility can be expected to be limited in the first instance to rare diseases and some easily studied cancers. Complex genetic disease is another matter. Genomewide association studies have undoubtedly been successful, especially in improving our understanding of the molecular pathways in a wide range of complex genetic diseases, but they have their limitations. Increasingly, attention has been devoted to finding rare variants by genomewide sequencing (with considerable recent success in some diseases, such as schizophrenia), and in investigating copy number variants. To properly appreciate the complexity of common genetic disease will require more information, too, from other

approaches, investigating modifier genes, environmental factors and so on, and reliance on phenotyping data from large population biobanks will be important.

The familial nature of much genetic information also poses challenges to many modern healthcare services for which there are no clear off-the-shelf solutions. Confidentiality in medicine remains important, yet shared familial inheritances may need disclosing at times, just as we attempted to trace contacts exposed to COVID-19. Sustainability aspects of long-term mass data storage are yet to be examined in any depth, and the lack of population diversity in most of the world's genomic repositories, and thus our understanding of genomic variation, needs urgent attention.

We have tried to convey the excitement of fast-moving research in genetics and genomics and their clinical applications, while explaining how the progress has been achieved. By weaving the ethical, legal and social aspects inherent in these developments throughout the text we hope to provide the reader with a realistic lens through which to view the promising developments in genetics and genomics. There is a long way to go, notably in understanding complex disease and in developing effective treatments for many disorders. But some impressive recent therapeutic advances, and new technological developments such as the prime editing and base editing refinements to CRISPR-Cas genome editing, have engendered an undeniable sense of excitement and optimism. How far will we move from the commonplace one-size-fits-all approach to disease treatment toward an era of personalized or precision medicine? At the very least, we might expect an era of stratified medicine where, according to the genetic variants exhibited by patients with a specific disease, different medical actions are taken.

We would like to thank the staff at CRC Press and Naughton Project Management Ltd: Jo Koster, Jordan Wearing and Nora Naughton, who have undertaken the job of converting our drafts into the finished product. We are also grateful to our family members: Meryl, Alex, James, Tim, Emily and Isobel for their steadfast support.

LITERATURE ACCESS

We live in a digital age and, accordingly, we have sought to provide electronic access to information. To help readers find references cited under Further Reading we provide the relevant PubMed identification (PMID) numbers for the individual articles—see also the PMID glossary item. We would like to take this opportunity to thank the US National Center for Biotechnology Information (NCBI) for their invaluable PubMed database that is freely available at: <http://www.ncbi.nlm.nih.gov/pubmed/>. Readers who are interested in new research articles that have emerged since publication of this book, or who might want to study certain areas in depth, may wish to take advantage of literature citation databases such as the freely available Google Scholar database (scholar.google.com).

For background information on single gene disorders, we often provide reference numbers to access OMIM, the Online Mendelian Inheritance in Man database (<http://www.omim.org>). For the more well-studied of these disorders, individual chapters in the University of Washington's GeneReviews series are highly recommended. They are electronically available at the NCBI's Bookshelf within its PubMed database. For convenience, we have given the PubMed Identifier (PMID) for individual articles that we refer to from the GeneReviews series. Note that all GeneReviews articles can be accessed through PubMed at PMID 20301295, where there is an alphabetic listing of all disorders covered by GeneReviews.

Tom Strachan and Anneke Lucassen

Acknowledgements

In writing this book, we have benefited greatly from the advice of many geneticists, biologists and clinicians. We are also grateful to various colleagues who contributed clinical profiles and/or laboratory data for case studies, or who advised on the contents of chapters and/or commented on some aspects of the text, notably the following: Chiara Bettolo, David Bourn, Gareth Breese, Heather Cordell, Jordi Diaz-Manera, Shaun Haigh, Rachel Horton, Majlinda Lako, Richard Martin, Ciaron McAnulty, Robert McFarland, Sabine Specht, Miranda Splitt, and Volker Straub.

1

Fundamentals of DNA, chromosomes, and cells

DOI: [10.1201/9781003044406-1](https://doi.org/10.1201/9781003044406-1)

CONTENTS

[1.1 THE STRUCTURE AND FUNCTION OF NUCLEIC ACIDS](#)

[1.2 THE STRUCTURE AND FUNCTION OF CHROMOSOMES](#)

[1.3 DNA AND CHROMOSOMES IN CELL DIVISION AND THE CELL CYCLE](#)

[SUMMARY](#)

[QUESTIONS](#)

[FURTHER READING](#)

Three structures are the essence of life: cells, chromosomes, and nucleic acids. Cells receive basic sets of instructions from DNA molecules that must also be transmitted to successive generations. And DNA molecules work in the context of larger structures: chromosomes.

Many organisms consist of single cells that can multiply quickly. They are genetically relatively stable, but through changes in their DNA they can adapt rapidly to changes in environmental conditions. Others, including ourselves, animals, plants, and some types of fungi, are multicellular.

Multicellularity offers specialization and complexity: individual cells can be assigned different functions, becoming muscle cells, neurons, or lymphocytes, for example. All the different cells in an individual arise originally from a single cell, and so all nucleated cells carry the same DNA sequences. During development, however, the DNA structure within chromosomes is changed to allow specific changes in gene expression that determine a cell's identity, whether it be a muscle cell or a neuron, for example.

Growth during development and tissue maintenance requires cell division. When a cell divides to produce daughter cells, our chromosomes and the underlying DNA sequences must undergo coordinated duplication and then be carefully segregated to the daughter cells.

Some of our cells can carry our DNA to the next generation. When that happens, chromosomes swap segments and DNA molecules undergo significant changes that make us different from our parents and from other individuals.

1.1 THE STRUCTURE AND FUNCTION OF NUCLEIC ACIDS

General concepts: the genetic material, genomes, and genes

Nucleic acids provide the *genetic material* of cells and viruses. They carry the instructions that enable cells to function in the way that they do and to divide, allowing the growth and reproduction of living organisms. Nucleic acids also control how viruses function and replicate. As we describe later, viruses can be highly efficient at inserting genes into human cells, and modified viruses are widely used in gene therapy.

Nucleic acids are susceptible to small changes in their structure (**mutations**). Occasionally, that can change the instructions that a nucleic acid gives out. The resulting genetic variation, plus mechanisms for shuffling the genetic material from one generation to the next, explains why individual organisms of the same species are nevertheless different from

each other. And genetic variation is the substrate that evolutionary forces work on to produce different species. (But note that the different types of cell in a single multicellular organism cannot be explained by genetic variation—the cells each contain the same DNA and the differences in cell types must arise instead by **epigenetic** mechanisms.)

In all cells the genetic material consists of double-stranded DNA in the form of a double helix. (Viruses are different. Depending on the type of virus, the genetic material may be double-stranded DNA, single-stranded DNA, double-stranded RNA, or single-stranded RNA.) As we describe below, DNA and RNA are highly related nucleic acids. RNA is functionally more versatile than DNA (it is capable of self-replication and individual RNA sequences can also serve as templates to make a protein, or act as regulators of gene expression). RNA is widely believed to have developed at a very early stage in evolution. Subsequently, DNA evolved; being chemically much more stable than RNA, it was more suited to being the store of genetic information in cells.

Genome is the collective term for all the *different* DNA molecules within a cell or organism. In prokaryotes—simple unicellular organisms, such as bacteria, that lack organelles—the genome usually consists of just one type of circular double-stranded DNA molecule that can be quite large and has a small amount of protein attached to it. A very large DNA-protein complex such as this is traditionally described as a **chromosome**.

Eukaryotic cells are more complex and more compartmentalized (containing multiple organelles that serve different functions), and they have multiple different DNA molecules. As we will see below, for example, the cells of a man have 25 different DNA molecules but a woman's cells have a genome made up of 24 types of DNA molecule.

In our cells—and in those of all animals and fungi—the genome is partitioned between the nucleus and the mitochondria. Most of the DNA is found in the nucleus, existing as extremely long linear DNA molecules complexed with a variety of different proteins and some types of RNA to form highly organized chromosomes. However, in mitochondria there is just one type of small circular DNA molecule that is largely devoid of

protein. (In plant cells, chloroplasts also have their own type of small circular DNA molecule.)

Genes are the DNA segments that carry the genetic information to make proteins or functional noncoding RNA molecules within cells. The great bulk of the genes in a eukaryotic cell are found in the chromosomes of the nucleus; just a few genes are found in the small mitochondrial or chloroplast DNA molecules.

The underlying chemistry of nucleic acids

Each nucleic acid strand is a polymer, a long chain containing many sequential copies of a simple repeating unit, a **nucleotide**. Each nucleotide in turn consists of a sugar molecule, to which is attached a nitrogenous base and a phosphate group. In DNA the sugar is deoxyribose, which has five carbon atoms that are labeled 1¢ (one *prime*) to 5¢. It is very closely related to ribose, the sugar molecule found in RNA—the only difference is that a hydroxyl (-OH) group at carbon 2¢ of ribose is replaced by a hydrogen atom in deoxyribose ([Figure 1.1](#)).

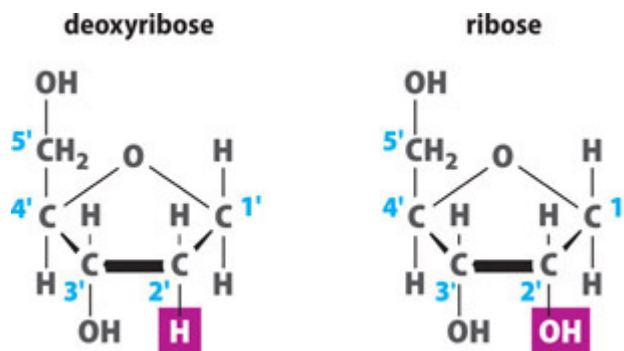


Figure 1.1 Structure of deoxyribose (left) and ribose (right). The five carbon atoms are numbered 1¢ (one *prime*) to 5¢ (five *prime*). The magenta shading is meant to signify the only structural difference between deoxyribose (the sugar found in DNA) and ribose (the sugar found in RNA): ribose has a hydroxyl (-OH) group in place of the highlighted hydrogen atom attached to carbon 2' of deoxyribose. The more precise name for deoxyribose is therefore 2¢-deoxyribose.

Individual nucleotides are joined to their neighbors by a negatively charged phosphate group that links the sugar components of the neighboring nucleotides. As a result, nucleic acids are polyanions, and have a *sugar-phosphate backbone* with bases bonded to the sugars. As explained in [Box 1.1](#), the sugar-phosphate backbone of each nucleic acid strand is asymmetric and the ends of each strand are asymmetric, giving direction to each strand.

BOX 1.1 5' AND 3' ENDS, AND STRAND ASYMMETRY OF NUCLEIC ACIDS

In a nucleic acid strand each phosphate group links carbon atom 3' from the sugar on one nucleotide to a carbon 5' on the sugar of a neighboring nucleotide. Internal nucleotides will therefore be linked through both carbon 5' and carbon 3' of the sugar to the neighboring nucleotides on opposing sides. However, the nucleotides at the extreme ends of a DNA or RNA strand will have different functional groups. At one end, the **5' end**, the nucleotide has a terminal sugar with a carbon 5' that is not linked to another nucleotide and is capped by a phosphate group; at the other end, the **3' end**, the terminal nucleotide has a sugar with a carbon 3' that is capped by a hydroxyl group ([Figure 1](#)).

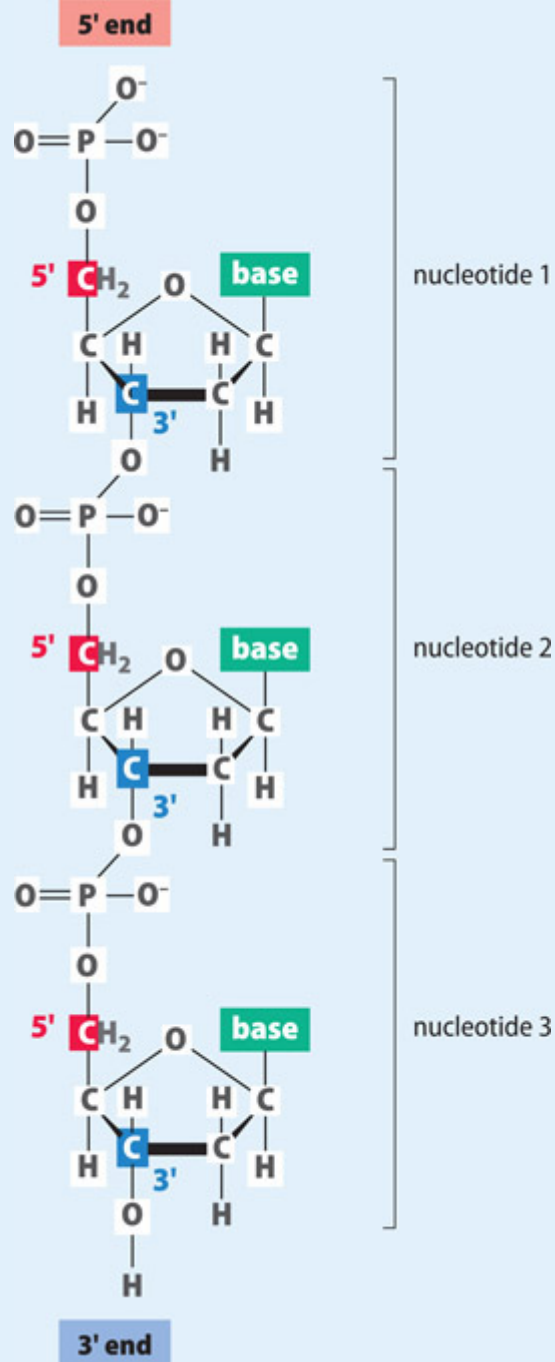


Figure 1 Repeating structure and asymmetric 5' and 3' ends in nucleic acids.

The resulting asymmetry between the two ends of a nucleic acid give it a direction. That is important in packing a nucleic acid because when two single nucleic acid strands pair up to make a stable duplex, they must be *anti-parallel*: the 5' @ 3' direction of one strand must be opposite to that

of its partner strand. And direction is important for synthesis of a nucleic acid: a growing nucleic acid strand always extends in a 5' → 3' direction.

Unlike the sugar molecules, the nitrogenous bases come in four different types, and it is the sequence of different bases that identifies the nucleic acid and its function. Two of the bases have a single ring based on carbon and nitrogen atoms (a **pyrimidine**) and two have a double ring structure (a **purine**). In DNA the two purines are adenine (A) and guanine (G), and the two pyrimidines are cytosine (C) and thymine (T). The bases of RNA are very similar; the only difference is that in place of thymine there is a very closely related base, uracil (U) ([Figure 1.2](#)).

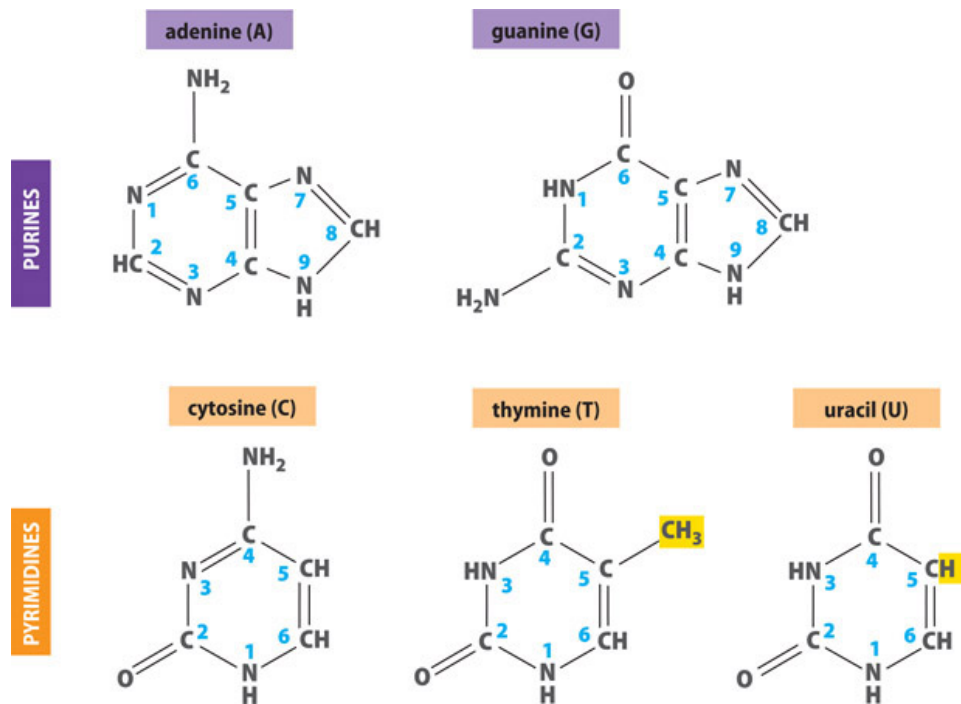


Figure 1.2 Structure of the bases found in nucleic acids. Adenine and guanine are purines with two interlocking rings based on nitrogen and carbon atoms (numbered 1 to 9 as shown). Cytosine and thymine are pyrimidines with a single ring. Adenine, cytosine, and guanine are found in both DNA and RNA, but the fourth base is thymine in DNA and uracil in RNA (they are closely related bases—carbon atom 5 in thymine has an attached methyl group, but in uracil the methyl group is replaced by a hydrogen atom).

Base pairing and the double helix

Cellular DNA exists in a double-stranded (or *duplex*) form, in which the two very long single DNA strands are wrapped round each other. In the resulting double helix each base on one DNA strand is noncovalently linked (by hydrogen bonding) to an opposing base on the opposite DNA strand, forming a **base pair**. However, the two DNA strands fit together correctly only if opposite every A on one strand is a T on the other strand, and opposite every G is a C. (Only two types of base pairs are normally tolerated in double-stranded DNA: A–T and G–C base pairs.) G–C base pairs, which are held together by three hydrogen bonds, are stronger than A–T base pairs, which are held together by two base pairs; see [Figure 1.3](#).

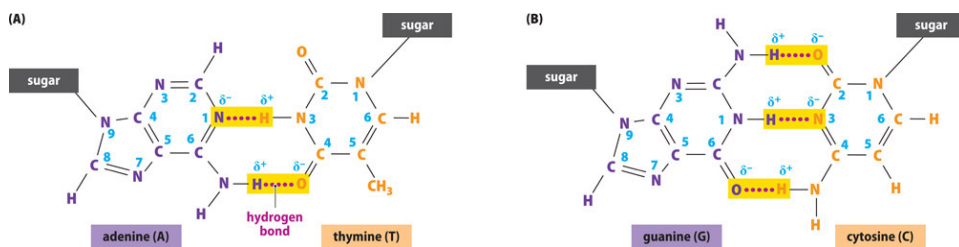


Figure 1.3 Structure of base pairs. In the A–T base pair shown in (A), the adenine is connected to the thymine by two hydrogen bonds. In the G–C base pair shown in (B), three hydrogen bonds link the guanine to the cytosine; a G–C base pair is therefore stronger than an A–T base pair. δ^+ and δ^- indicate fractional positive charges and fractional negative charges.

There is one additional restriction on how two single-stranded nucleic acids form a double-stranded nucleic acid. In addition to a sufficient degree of base pairing, for a duplex to form, the two single strands must be anti-parallel; that is, the 5' \rightarrow 3' direction of one strand is the opposite of the 5' \rightarrow 3' direction of the other strand.

Two single nucleic acid strands that can form a double helix with perfect base matching (according to the base pairing rules given above) are said to have **complementary sequences**. As a result of base pairing rules, the sequence of one DNA strand in a double helix can immediately be used to predict the base sequence of the complementary strand ([Box 1.2](#)). Note that

base pairing can also occur in RNA; when an RNA strand participates in base pairing, the base pairing rules are more relaxed (see [Box 1.2](#)).

DNA replication and DNA polymerases

Base pairing rules also explain the mechanism of DNA replication. In preparation for new DNA synthesis before cell division, each DNA double helix must be unwound using a helicase. During the unwinding process the two individual single DNA strands become available as templates for making complementary DNA strands that are synthesized in the 5' → 3' direction ([Figure 1.4](#)).

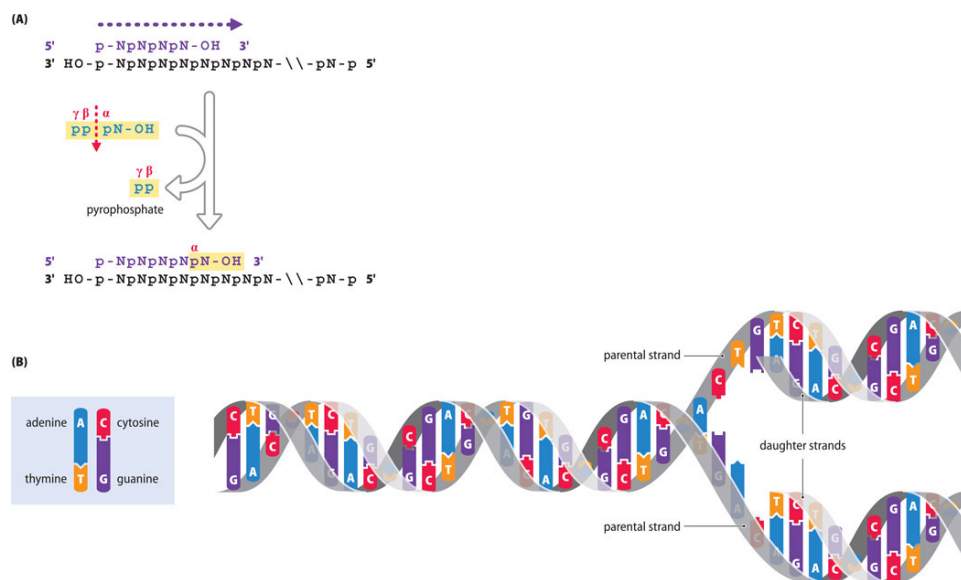


Figure 1.4 DNA synthesis and replication. (A) DNA synthesis. Using a pre-existing DNA strand (black) as a template, a new DNA strand (purple) is synthesized in a 5' → 3' direction (dashed arrow) using a DNA polymerase to insert successive dNMPs obtained by cleaving the two external phosphates (α and β), to give pyrophosphate residue (which is discarded). (B) DNA replication. The parental DNA duplex consists of two complementary DNA strands that unwind to serve as templates for the synthesis of new complementary DNA strands (daughter strands). Each completed daughter DNA duplex contains one of the two parental DNA strands plus one newly synthesized DNA strand and is structurally identical to the original parental DNA duplex.

BOX 1.2 BASE PAIRING PREVALENCE, SEQUENCE COMPLEMENTARITY, AND SEQUENCE NOTATION FOR NUCLEIC ACIDS

THE PREVALENCE OF BASE PAIRING

The DNA of cells—and of viruses that have a double-stranded DNA genome—occurs naturally as double helices in which base pairing is restricted to A–T and C–G base pairs.

Double-stranded RNA also occurs naturally in the genomes of some kinds of RNA viruses. Although cellular RNA is often single-stranded, it can also participate in base pairing in different ways. Many single-stranded RNAs have sequences that allow intramolecular base pairing—the RNA bends back upon itself to form local double-stranded regions for structural stability and/or for functional reasons. Different RNA molecules can also transiently base pair with each other over short to moderately long regions, allowing functionally important interactions (such as base pairing between messenger RNA and transfer RNA during translation, for example; see [Section 2.1](#)). G–U base pairs are allowed in RNA–RNA base pairing, in addition to the standard A–U and C–G base pairs.

RNA–DNA hybrids also form transiently in different circumstances. They occur when a DNA strand is transcribed to give an RNA copy, for example, and when an RNA is reverse transcribed to give a DNA copy.

SEQUENCE COMPLEMENTARITY

Double-helical DNA within cells shows perfect base matching over extremely long distances, and the two

DNA strands within a double helix are said to exhibit **base complementarity** and to have *complementary sequences*. Because of the strict base pairing rules, knowing the base sequence of just one DNA strand is sufficient to immediately predict the sequence of the complementary strand, as illustrated below.

SEQUENCE NOTATION

Because the base sequence of a nucleic acid governs its biological properties it is customary to define a nucleic acid by its base sequence, which is always written in the 5' → 3' direction. While a single-stranded oligonucleotide sequence might be written accurately as 5' p-C-p-G-p-A-p-C-p-C-p-A-p-T-OH 3', where p = phosphate, it is simpler to write it just as CGACCAT.

For a double-stranded DNA the sequence of just one of the two strands is needed (the sequence of the complementary strand can immediately be predicted by the base pairing rules given above). If a given DNA strand has the sequence CGACCAT, the sequence of the complementary strand can easily be predicted to be ATGGTCG (in the 5' → 3' direction as shown below, where A–T base pairs are shown in green and C–G base pairs in blue).



DNA replication therefore uses one double helix to make two double helices, each containing one strand from the parental double helix and one newly synthesized strand (semi-conservative DNA replication). Because DNA synthesis occurs only in the 5' → 3' direction, one new strand (the *leading strand*) can be synthesized continuously; the other strand (the *lagging strand*) needs to be synthesized in pieces, known as Okazaki fragments ([Figure 1.5](#)).

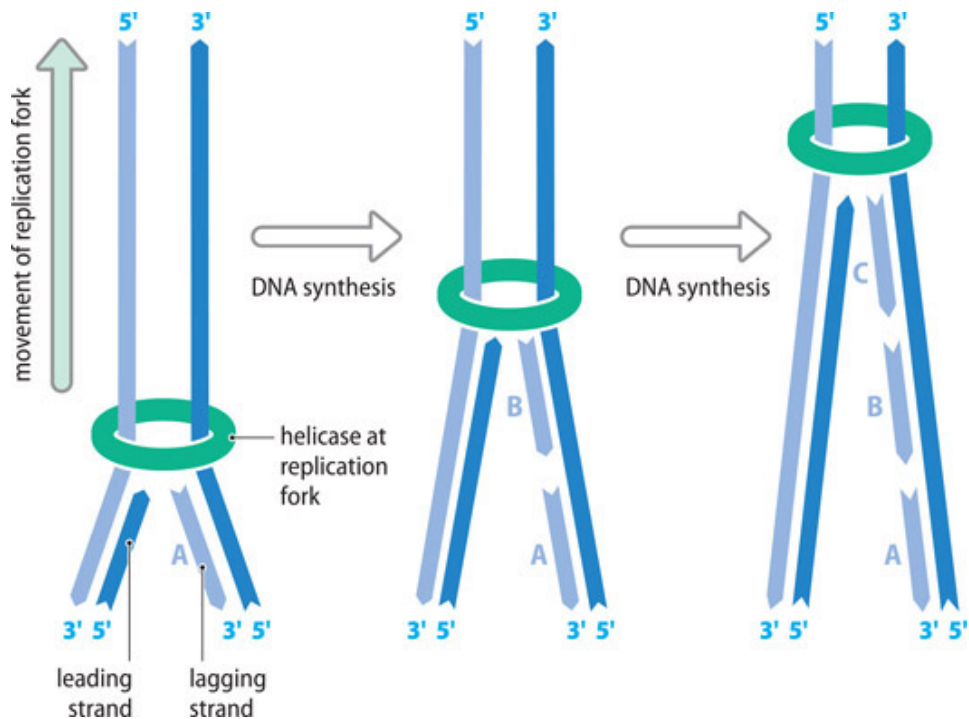


Figure 1.5 Semi-discontinuous DNA replication. The enzyme DNA helicase opens up a **replication fork**, where synthesis of new daughter DNA strands can begin. The overall direction of movement of the replication fork matches that of the continuous 5' to 3' synthesis of one daughter DNA strand, the *leading strand*. Replication is semi-discontinuous because the *lagging strand*, which is synthesized in the opposite direction, is built up in pieces (Okazaki fragments, shown here as fragments A, B, and C) that will later be stitched together by a DNA ligase.

Mammalian cells have very many kinds of DNA-dependent DNA polymerases that serve a variety of roles, including DNA replication initiation, synthesis of the leading and lagging strands, and also, as described in [Section 4.2](#), multiple roles in DNA repair. Our cells also contain specialized DNA polymerases that use RNA as a template to synthesize a complementary DNA; see [Table 1.1](#).

TABLE 1.1 CLASSICAL DNA-DEPENDENT AND RNA-DEPENDENT DNA POLYMERASES OF MAMMALIAN CELLS	
DNA polymerases	Roles

DNA polymerases	Roles
Classical DNA-dependent DNA polymerases α (alpha) δ (delta) and ϵ (epsilon) β (beta) γ (gamma)	Standard DNA replication and/or DNA repair initiates DNA synthesis (at replication origins, and also when priming the synthesis of Okazaki fragments on the lagging strand) major nuclear DNA polymerases and multiple roles in DNA repair base excision repair (repair of deleted bases and simply modified bases) dedicated to mitochondrial DNA synthesis and mitochondrial DNA repair
RNA-dependent DNA polymerases Retrosposon reverse transcriptase TERT (telomerase reverse transcriptase)	Genome evolution and telomere function occasionally converts mRNA and other RNA into complementary DNA, which can integrate elsewhere into the genome; can occasionally give rise to new genes and new exons, and soon. replicates DNA at ends of linear chromosomes, using an RNA template

Note: The classical DNA-dependent DNA polymerases are high-fidelity polymerases—they insert the correct base with high accuracy; however, our cells also have many non-classical DNA-dependent DNA polymerases that exhibit comparatively low fidelity of DNA replication. We will consider the non-classical DNA polymerases in [Chapter 4](#), because of their roles in certain types of DNA repair and in maximizing the variability of immunoglobulins and T-cell receptors.

Genes, transcription, and the central dogma of molecular biology

As a repository of genetic information, DNA must be stably *transmitted* from mother cell to daughter cells, and from individuals to their progeny; DNA replication provides the necessary mechanism. But within the context of individual cells, the genetic information must also be *interpreted* to dictate how cells work. **Genes** are discrete segments of the DNA whose sequences are selected for this purpose, and gene expression is the

mechanism whereby genes are used to direct the synthesis of two kinds of product: RNA and proteins.

The first step of gene expression is to use one of the two DNA strands as a template for synthesizing an RNA copy whose sequence is complementary to the selected template DNA strand. This process is called *transcription*, and the initial RNA copy is known as the primary transcript ([Figure 1.6](#)). Subsequently, the primary transcript undergoes different processing steps, eventually giving a mature RNA that belongs to one of two broad RNA classes:

- *Coding RNA*. RNAs in this class contain sequences that direct the synthesis of polypeptides (the major component of proteins) in a process called translation. This type of RNA has traditionally been called a messenger RNA (mRNA) because it carries genetic instructions to be decoded by the protein synthesis machinery.
- *Noncoding RNA*. All other mature functional RNAs fall into this class, and here the RNAs, not proteins, are the functional endpoint of gene expression. Noncoding RNAs have a variety of roles in cells, as described in later chapters.

In all forms of life, genetic information is interpreted in what initially seemed to be one direction only: DNA → RNA → protein, a principle that became known as the central dogma of molecular biology. However, certain DNA polymerases, known as reverse transcriptases, found initially in certain types of viruses, can reverse the flow of genetic information by making a DNA copy of an RNA molecule. The cells of complex organisms also have their reverse transcriptases, as described below. In addition, RNA can sometimes also be used as a template to make a complementary RNA copy. So, although genetic information in cells mostly flows from DNA to RNA to protein, the central dogma is no longer strictly valid.

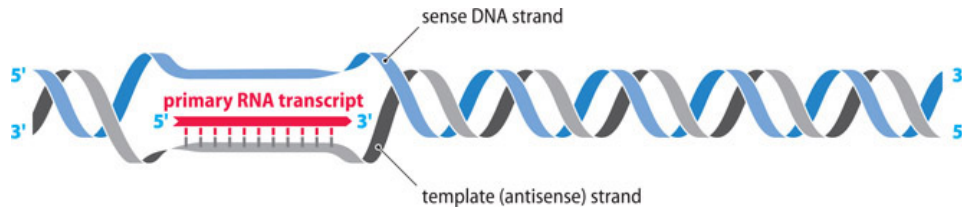


Figure 1.6 Transcription. Transcription results in the synthesis of an RNA transcript in the 5' → 3' direction. The nucleotide sequence of the primary RNA transcript is complementary to that of the *template strand* and so is identical to that of the *sense strand*, except that U replaces T. Note: for simplification, the diagram does not show the coiling of the RNA transcript around the template DNA strand to form a double helix.

We will explore gene expression (including protein synthesis) in greater detail in [Chapter 2](#). And in [Chapter 6](#) we will focus on both genetic and epigenetic regulation of gene expression.

1.2 THE STRUCTURE AND FUNCTION OF CHROMOSOMES

In this section we consider general aspects of the structure and function of our chromosomes that are largely shared by the chromosomes of other complex multicellular organisms. We will touch on human chromosomes when we consider aspects of the human genome in [Chapter 2](#), when we first introduce the banded pattern of human chromosomes. In [Chapter 7](#) we consider how disease-causing chromosome abnormalities arise. We describe the methodology and the terminology of human chromosome banding in [Box 7.2](#), and diagnostic chromosome analyses in [Chapter 11](#).

Why we need highly structured chromosomes, and how they are organized

Before replication, each chromosome in the cells of complex multicellular organisms normally contains a single, immensely long DNA double helix. For example, an average-sized human chromosome contains a single DNA

double helix that is about 4.8 cm long with 140 million nucleotides on each strand; that is, 140 million base pairs (140 megabases (Mb)) of DNA.

To appreciate the difficulty in dealing with molecules this long in a cell only about 10 μm across, imagine a model of a human cell 1 meter across (a 10^5 -fold increase in diameter). Now imagine the problem of fitting into this 1-meter-wide cell 46 DNA double helices that when scaled up by the same factor would each be just 0.2 mm thick but on average 4.8 km (about three miles) long. Then there is the challenge of replicating each of the DNA molecules and arranging for the cell to divide in such a way that the replicated DNA molecules are segregated equally into the two daughter cells. All this must be done in a way that avoids any tangling of the long DNA molecules.

To manage nuclear DNA molecules efficiently and avoid any tangling, they are complexed with various proteins and sometimes noncoding structural RNAs to form **chromatin** that undergoes different levels of coiling and compaction to form chromosomes. In interphase—the stages of the cell cycle other than mitosis (see [Section 1.3](#))—the nuclear DNA molecules are still in a very highly extended form and normally the very long slender interphase chromosomes remain invisible under the light microscope. But even in interphase cells, the 2 nm-thick double helix is subject to at least two levels of coiling. First, the double helix is periodically wound round a specialized complex of positively charged histone proteins to form a 10 nm nucleosome filament. The nucleosome filament is then coiled into a 30 nm chromatin fiber that undergoes looping and is supported by a scaffold of nonhistone proteins ([Figure 1.7](#)).

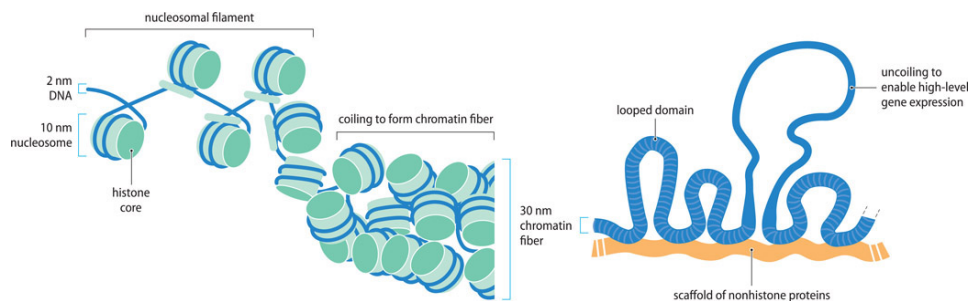


Figure 1.7 From DNA double helix to interphase chromatin. Binding of basic histone proteins causes the 2 nm DNA double helix to undergo coiling, forming first a 10 nm filament studded with nucleosomes that is further coiled to give a 30 nm chromatin fiber. In interphase, the chromatin fiber is organized in looped domains, each containing about 50–200 kilobases of DNA, that are attached to a central scaffold of nonhistone proteins. High levels of gene expression require local uncoiling of the chromatin fiber to give the 10 nm nucleosomal filaments. The diagram does not show structural RNAs that can be important in chromatin. (Adapted from Grunstein M [1992] *SciAm* 267:68–74; PMID 1411455. With permission from Macmillan Publishers Ltd; and Alberts B, Johnson A, Lewis J et al. [2008] *Molecular Biology of the Cell*, 5th ed. Garland Science.)

During interphase most chromatin exists in an extended state (**euchromatin**) that is dispersed through the nucleus. Euchromatin is not uniform, however—some euchromatic regions are more condensed than others, and genes may or may not be expressed, depending on the cell type and its functional requirements. Some chromatin, however, remains highly condensed throughout the cell cycle and is generally genetically inactive (**heterochromatin**).

As cells prepare to divide, the chromosomes need to be compacted much further to maximize the chances of correct pairing and segregation of chromosomes into daughter cells. Packaging of DNA into **nucleosomes** and then the 30 nm chromatin fiber results in a linear condensation of about 50-fold. During the M (mitosis) phase, higher-order coiling occurs (see [Figure 1.7](#)), so that DNA in a human metaphase chromosome is compacted to about 1/10 000 of its stretched-out length. As a result, the short, stubby metaphase chromosomes are readily visible under light microscopes.

Chromosome function: replication origins, centromeres, and telomeres

The DNA within a chromosome contains genes that are expressed according to the needs of a cell. But it also contains specialized sequences

that are needed for chromosome function. Three major classes are described below.

Centromeres

When a cell divides the chromosomes must be correctly segregated to the two daughter cells. This requires a **centromere**, a region to which a pair of large protein complexes called kinetochores will bind just before the preparation for cell division ([Figure 1.8](#)). Centromeres can be seen at metaphase as the primary constriction that separates the short and long chromosome arms. Microtubules attached to each kinetochore are responsible for positioning the chromosomes correctly at metaphase and then pulling the separated chromosomes to opposite poles of the mitotic spindle.

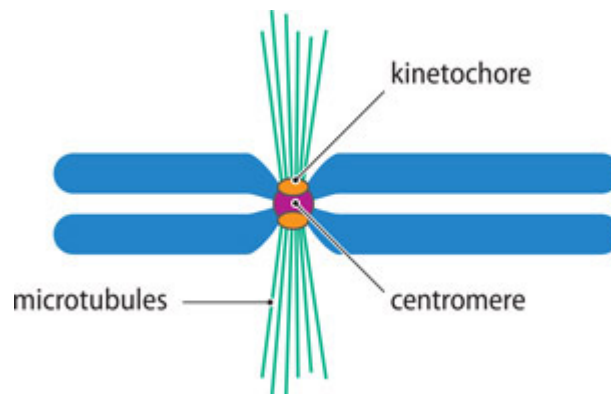


Figure 1.8 Centromere function relies on the assembly of kinetochores and attached microtubules.

The DNA sequences at centromeres are very different in different organisms. In a mammalian chromosome, the centromeric DNA is a heterochromatic region dominated by highly repetitive DNA sequences that often extend over megabases of DNA.

Replication origins

For a chromosome to be replicated, it needs one or more replication origins—DNA sequence components to which protein factors bind in preparation for initiating DNA replication. The chromosomes of budding yeast can be replicated using a single very short highly defined DNA sequence, but in the cells of complex organisms, such as mammals, DNA is replicated at multiple initiation sites along each chromosome; the replication origins are quite long and do not have a common base sequence.

Telomeres

Telomeres are specialized structures at the ends of chromosomes that are necessary for the maintenance of chromosome integrity (if a telomere is lost after chromosome breakage, the resulting chromosome end is unstable; it tends to fuse with the ends of other broken chromosomes, or to be involved in recombination events, or to be degraded).

Unlike centromeric DNA, telomeric DNA has been well conserved during evolution. In vertebrates, the DNA of telomeres consists of many tandem (sequential) copies of the sequence TTAGGG to which certain telomeric proteins bind. Most of the telomere DNA is double-stranded with one strand containing TTAGGG repeats (the G-rich strand) and the complementary strand containing CCCTAA repeats (the C-rich strand). However, at its 3' end, the G-rich strand has an overhang (with about 30 TTAGGG repeats) that folds back and base pairs with the C-rich strand. The resulting T-loop is thought to protect the telomere DNA from natural cellular exonucleases that repair double-strand DNA breaks ([Figure 1.9](#)).

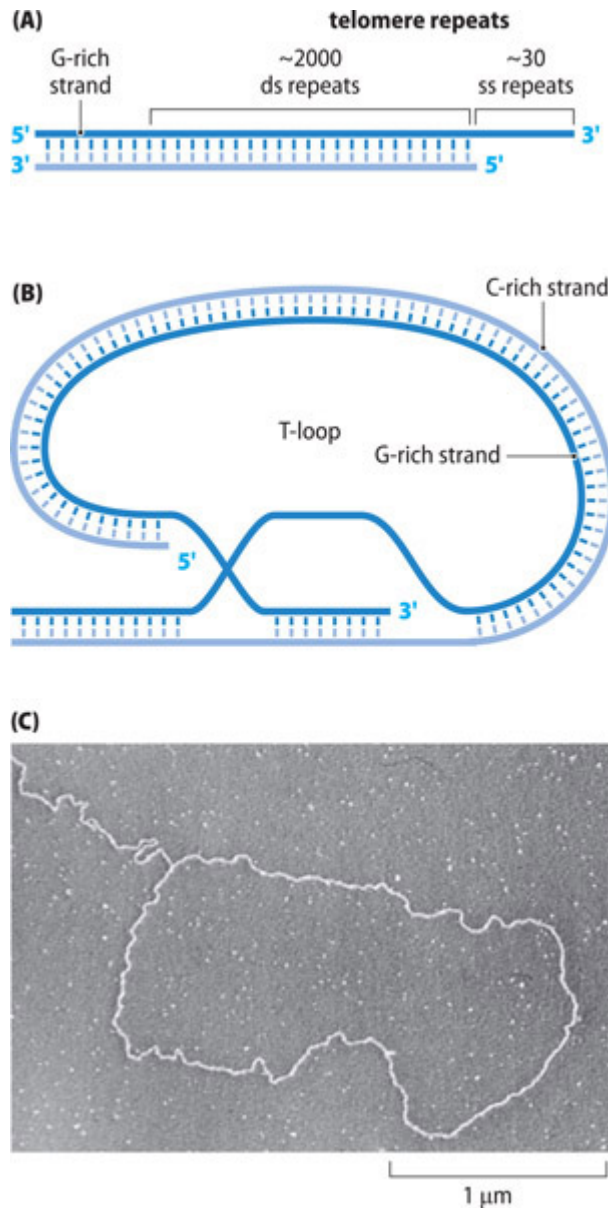


Figure 1.9 Telomere structure and T-loop formation. (A) Human telomere structure. A tandem array of roughly 2000 copies of the double-stranded hexanucleotide (TTAGGG/CCCTAA) repeat followed by a protrusion of about 30 single-stranded TTAGGG repeats. Abbreviations: ss, single-strand; ds, double-strand. (B) T-loop formation. The single-stranded terminus can loop back and invade the double-stranded region by base pairing with the complementary C-rich strand. (C) Electron micrograph showing formation of a roughly 15-kilobase T-loop at the end of an interphase human chromosome. (From Griffith JD et al. [1999] *Cell* 97:503–514; PMID 10338214. With permission from Elsevier.)

1.3 DNA AND CHROMOSOMES IN CELL DIVISION AND THE CELL CYCLE

Differences in DNA copy number between cells

Like other multicellular organisms, we have cells that are structurally and functionally diverse. In each individual the different cell types have the same genetic information, but only a subset of genes is expressed in each cell. What determines the identity of a cell—whether a cell is a B lymphocyte or a hepatocyte, for example—is the pattern of expression of the different genes across the genome.

As well as differences in gene expression, different cells can vary in the number of copies of each DNA molecule. The term *ploidy* describes the number of copies (n) of the basic chromosome set (the collective term for the different chromosomes in a cell) and also describes the copy number of each of the different nuclear DNA molecules.

The DNA content of a single chromosome set is represented as C . Human cells—and the cells of other mammals—are mostly **diploid** ($2C$), with nuclei containing two copies of each type of chromosome, one paternally inherited and one maternally inherited. Sperm and egg cells are **haploid** cells that contain only one of each kind of chromosome ($1C$). Human sperm and eggs each have 23 different types of chromosomes and so $n = 23$ in humans.

Some specialized human cells are nulliploid ($0C$) because they lack a nucleus—examples include erythrocytes, platelets, and terminally differentiated keratinocytes. Others are naturally polyploid (more than $2C$). Polyploidy can occur by two mechanisms. The DNA might undergo multiple rounds of replication without cell division, as when the large megakaryocytes in blood are formed (they have from 16 to 64 copies of each chromosome, and the nucleus is large and multilobed). Alternatively, polyploid cells originate by cell fusion to give cells with multiple nuclei, as in the case of muscle fiber cells.

Mitochondrial DNA copy number

The great majority of our cells are diploid and contain two copies of each nuclear DNA molecule. In stark contrast, the number of copies of the mitochondrial DNA (mtDNA) can vary from hundreds to many thousands according to the cell type, and can even vary over time in some cells. The two types of haploid cells show very large differences in mtDNA copy number: a human sperm typically has about 100 mtDNA copies, but a human egg cell usually has about 250 000 mtDNA molecules.

The cell cycle and segregation of replicated chromosomes and DNA molecules

Cells also differ according to whether they actively participate in the cell cycle and undergo successive rounds of cell division. Each time a cell divides, it gives rise to two daughter cells. To keep the number of chromosomes constant there needs to be a tight regulation of chromosome replication and chromosome segregation. Each chromosome needs to be replicated just once to give rise to two daughter chromosomes, which must then segregate equally so that one passes to each daughter cell.

During normal periods of growth there is a need to expand cell number. In the fully grown adult, the majority of cells are terminally differentiated and do not divide, but stem cells and progenitor cells continue to divide to replace cells that have a high turnover, notably blood, skin, sperm, and intestinal epithelial cells.

Each round of the cell cycle involves a phase in which the DNA replicates—S phase (synthesis of DNA)—and a phase where the cell divides—M phase. Note that M phase involves both nuclear division (mitosis) and cell division (cytokinesis). In the intervals between these two phases are two gap phases—G₁ phase (gap between M phase and S phase) and G₂ phase (gap between S phase and M phase)—see [Figure 1.10](#).

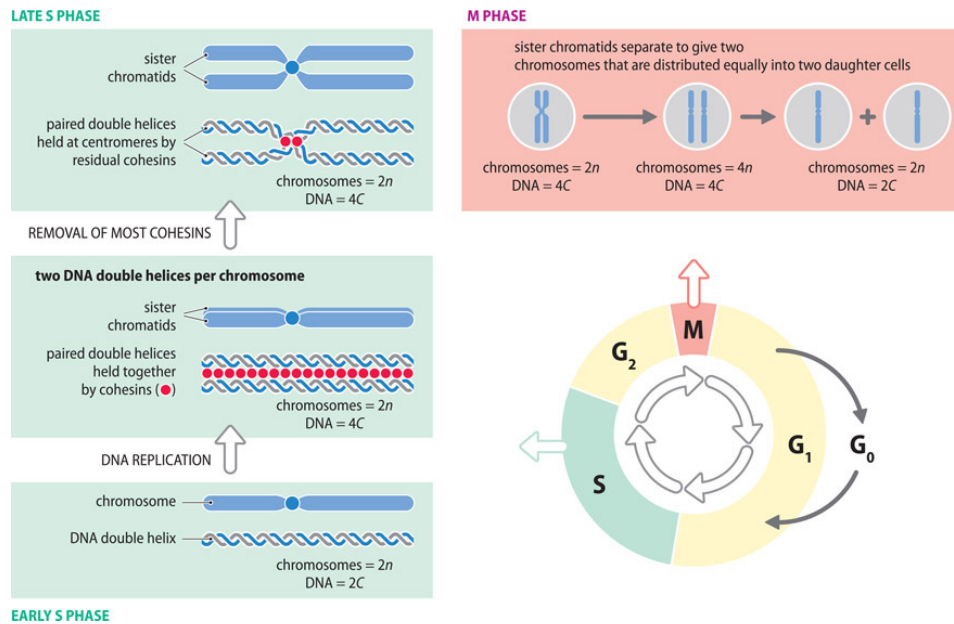


Figure 1.10 Changes in chromosomes and DNA content during the cell cycle. The cell cycle consists of four major phases as shown at the bottom right (in the additional G₀ phase a cell exits from the cell cycle and remains suspended in a stationary phase that resembles G₁ but can subsequently rejoin the cell cycle under certain conditions). In the expanded panels for M and S phases we show for convenience just a single chromosome, and we illustrate in each of the three boxes at left, representing S phase at different stages, how a single chromosome (top) relates to its DNA molecule (bottom). Chromosomes contain one DNA double helix from the end of M phase right through until just before the DNA duplicates in S phase. After duplication, the two double helices are held tightly together along their lengths by binding proteins called cohesins (red circles), and the chromosome now consists of two sister chromatids each having a DNA double helix. The sister chromatids becomes more obvious in late S phase when most of the cohesins are removed except for some at the centromere, which continue to hold the two sister chromatids together. The sister chromatids finally separate in M phase to form two independent chromosomes that are then segregated into the daughter cells. Note that the S-phase chromosomes in the boxes at left are shown, purely for convenience, in a compact form; in reality they are enormously extended.

Cell division takes up only a brief part of the cell cycle. For actively dividing human cells, a single turn of the cell cycle might take about 24 hours; M phase often occupies about 1 hour. During the short M phase, the

chromosomes become extremely highly condensed in preparation for nuclear and cell division. After M phase, cells enter a long growth period called **interphase** (= G₁ + S + G₂ phases), during which chromosomes are enormously extended, allowing genes to be expressed.

G₁ is the long-term end state of terminally differentiated nondividing cells. For dividing cells, the cells will enter S phase only if they are committed to mitosis; if not, they are induced to leave the cell cycle to enter a resting phase, the G₀ phase (a modified G₁ stage). When conditions become suitable later on, cells may subsequently move from G₀ to re-enter the cell cycle.

Changes in cell chromosome number and DNA content

During the cell cycle, the amount of DNA in a cell and the number of chromosomes change. In the box panels in [Figure 1.10](#) we follow the fate of a single chromosome through M phase and then through S phase. If we were to consider a diploid human cell this would be one chromosome out of the 46 ($2n$) chromosomes present after daughter cells are first formed. We also show in this figure how a single chromosome (top) relates to its DNA double helix content at different stages in S phase. The progressive changes in the number of chromosomes and the DNA content of cells at different stages of the cell cycle are listed below.

- From the end of the M phase right through until DNA duplication in S phase, each chromosome of a diploid ($2n$) cell contains a single DNA double helix; the total DNA content is therefore $2C$.
- After DNA duplication, the total DNA content per cell is $4C$, but specialized binding proteins called cohesins hold the duplicated double helices together as **sister chromatids** within a single chromosome. The chromosome number remains the same ($2n$), but each chromosome now has double the DNA content of a chromosome in early S phase. In late S phase, most of the cohesins

are removed but cohesins at the centromere are retained to keep the sister chromatids together.

- During M phase, the residual cohesins are removed and the duplicated double helices finally separate. That allows sister chromatids to separate to form two daughter chromosomes, giving $4n$ chromosomes. The duplicated chromosomes segregate equally to the two daughter cells so that each will have $2n$ chromosomes and a DNA content of $2C$.

[Figure 1.10](#) can give the misleading impression that all the interesting action happens in S and M phases. That is quite wrong—a cell spends most of its life in the G_0 or G_1 phases, and that is where the genome does most of its work, issuing the required instructions to make the diverse protein and RNA products needed for cells to function.

Mitochondrial DNA replication and segregation

In advance of cell division, mitochondria increase in mass and mtDNA molecules replicate before being segregated into daughter mitochondria that then need to segregate into daughter cells. Whereas the replication of nuclear DNA molecules is tightly controlled, the replication of mtDNA molecules is not directly linked to the cell cycle.

Replication of mtDNA molecules simply involves increasing the number of DNA copies in the cell, without requiring equal replication of individual mtDNAs. That can mean that some individual mtDNAs might not be replicated and other mtDNA molecules might be replicated several times ([Figure 1.11](#)).

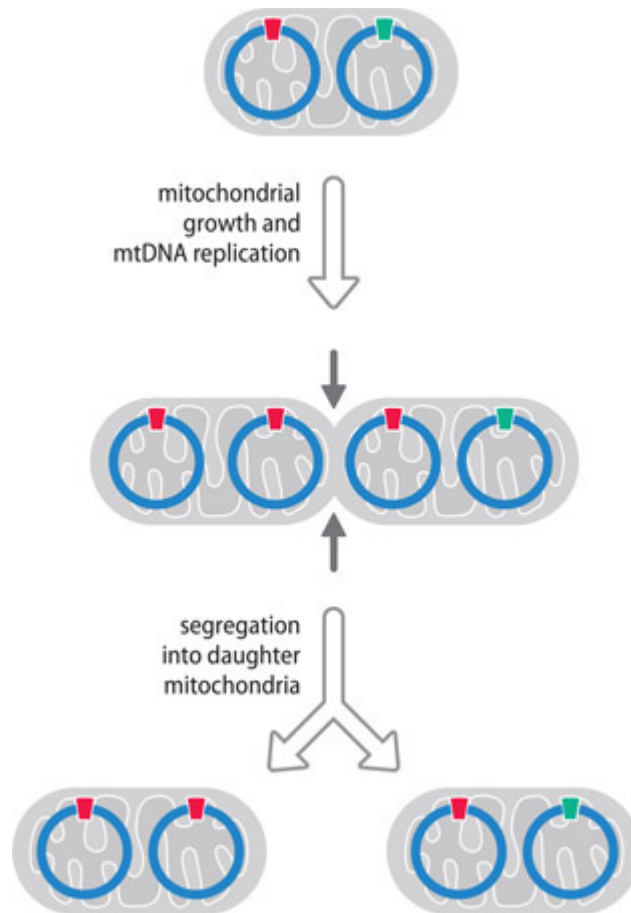


Figure 1.11 Unequal replication of individual mitochondrial DNAs. Unlike in the nucleus, where replication of a chromosomal DNA molecule is tightly controlled and normally produces two copies, mitochondrial DNA (mtDNA) replication is stochastic. When a mitochondrion increases in mass in preparation for cell division, the overall amount of mitochondrial DNA increases in proportion, but individual mtDNAs replicate unequally. In this example, the mtDNA with the green tag fails to replicate and the one with the red tag replicates to give three copies. Variants of mtDNA can arise through mutation so that a person can inherit a mixed population of mtDNAs (heteroplasmy). Unequal replication of pathogenic and nonpathogenic mtDNA variants can have important consequences, as described in [Chapter 5](#).

Whereas the segregation of nuclear DNA molecules into daughter cells needs to be equal and is tightly controlled, segregation of mtDNA molecules into daughter cells can be unequal. Even if the segregation of mtDNA molecules into daughter mitochondria is equal (as shown in [Figure](#)

[1.11](#)), the segregation of the mitochondria into daughter cells is thought to be stochastic.

Mitosis: the usual form of cell division

Most cells divide by a process known as mitosis. In the human life cycle, mitosis is used to generate extra cells that are needed for periods of growth and to replace various types of short-lived cells. Mitosis ensures that a single parent cell gives rise to two daughter cells that are both genetically identical to the parent cell (barring any errors that might have occurred during DNA replication). During a human lifetime, there may be something like 10^{17} mitotic divisions.

The M phase of the cell cycle includes both nuclear division (mitosis, which is divided into the stages of prophase, prometaphase, metaphase, anaphase, and telophase), and also cell division (cytokinesis), which overlaps the final stages of mitosis ([Figure 1.12](#)). In preparation for cell division, the previously highly extended duplicated chromosomes contract and condense so that, by the metaphase stage of mitosis, they are readily visible when viewed under the microscope.

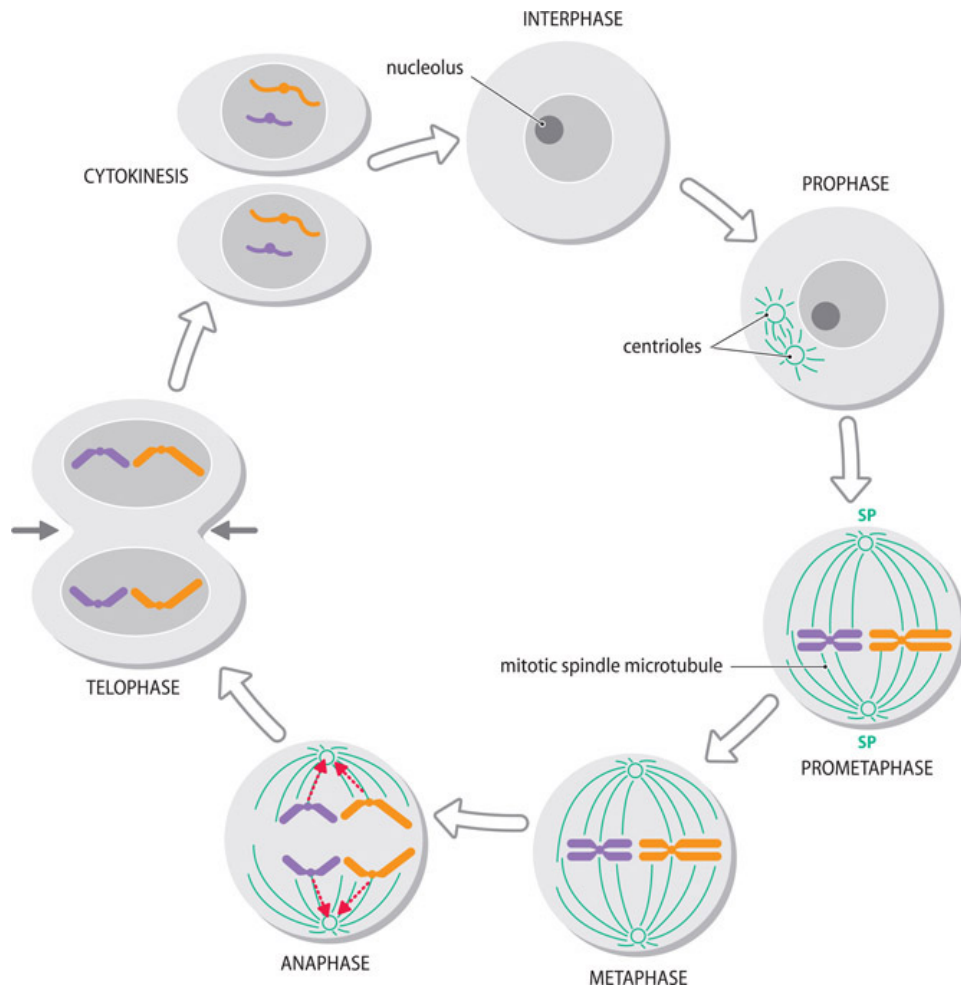


Figure 1.12 Mitosis (nuclear division) and cytokinesis (cell division). Early in prophase, centrioles (short cylindrical structures made up of microtubules and associated proteins) begin to separate and migrate to opposite poles of the cell. They give rise to the spindle poles (SP) from which microtubules will extend to the center of the cell to form the mitotic spindle. In prometaphase, the nuclear envelope breaks down, and the now highly condensed chromosomes become attached at their centromeres to the array of mitotic spindle microtubules. At metaphase, the chromosomes all lie along the middle of the mitotic spindle, still with the sister chromatids bound together (because of residual cohesins at the centromere that hold the duplicated DNA helices together). Removal of the residual cohesins allows the onset of anaphase: the sister chromatids separate and begin to migrate toward opposite poles of the cell (shown by dashed red arrows). The nuclear envelope forms again around the daughter nuclei during telophase, and the chromosomes decondense, completing mitosis. Before the final stages of mitosis, and most obviously at telophase, cytokinesis begins. The cell

becomes progressively constricted at its middle (shown at telophase by converging horizontal arrows), eventually resulting in full cytokinesis to produce two daughter cells.

The chromosomes of early S phase have one DNA double helix; however, after DNA replication, two identical DNA double helices are produced and held together by cohesins. Later, when the chromosomes undergo compaction in preparation for cell division, the cohesins are removed from all parts of the chromosomes apart from the centromeres. As a result, as early as prometaphase (when the chromosomes are now visible under the light microscope), individual chromosomes can be seen to comprise two **sister chromatids** that remain attached at the centromere (bound by some residual cohesins).

Later, at the start of anaphase, the remaining cohesins are removed and the two sister chromatids can now disengage to become independent chromosomes that will be pulled to opposite poles of the cell and then distributed equally to the daughter cells (see [Figure 1.12](#)).

Meiosis: a specialized reductive cell division giving rise to sperm and egg cells

The **germ line** is the collective term for cells that can pass genetic material to the next generation. It includes haploid sperm and egg cells (the **gametes**) and all the diploid precursor cells from which they arise by cell division, going all the way back to the zygote. The nongermline cells are known as **somatic cells**.

In humans, where $n = 23$, each gamete contains one sex chromosome plus 22 nonsex chromosomes (**autosomes**). In eggs the sex chromosome is always an X; in sperm it may be either an X or a Y. After a haploid sperm fertilizes a haploid egg, the resulting diploid **zygote** and almost all of its descendant cells have the chromosome constitution 46,XX (female) or 46,XY (male).

Diploid primordial germ cells migrate into the embryonic gonad and engage in repeated rounds of mitosis, to generate spermatogonia in males

and oogonia in females. Further growth and differentiation produce primary spermatocytes in the testis and primary oocytes in the ovary. The diploid spermatocytes and oocytes can then undergo **meiosis**, the cell division process that produces haploid gametes.

Meiosis is a *reductive* division because it involves two successive cell divisions (meiosis I and meiosis II) but only one round of DNA replication ([Figures 1.13](#) and [1.14](#)). As a result, it gives rise to four haploid cells. In males, the two meiotic cell divisions are each symmetric, producing four functionally equivalent spermatozoa. Huge numbers of sperm are produced, and spermatogenesis is a continuous process from puberty onward.

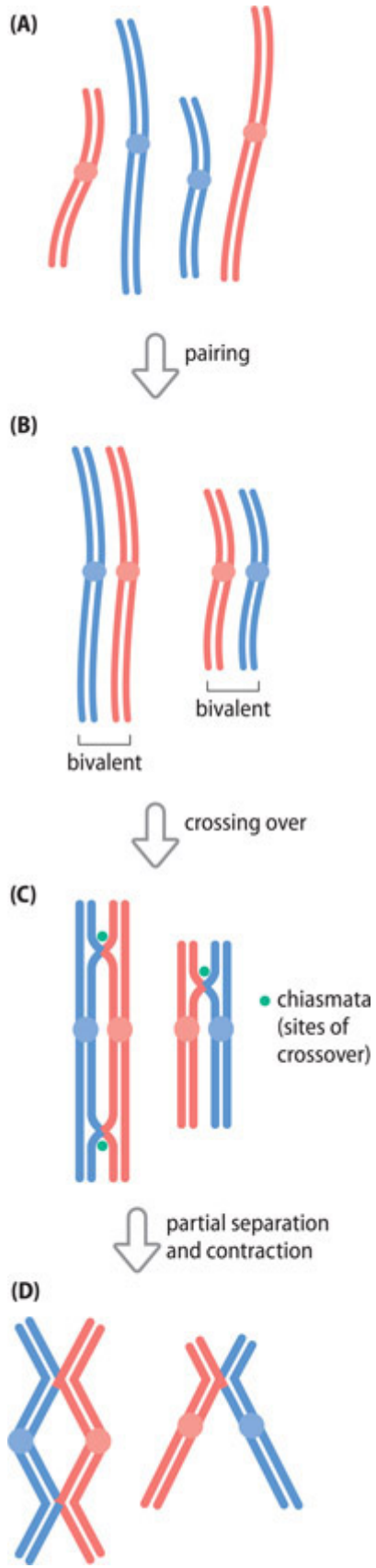


Figure 1.13 Prophase stages in meiosis I. (A) In leptotene, the duplicated chromosomes (each with a pair of sister chromatids) begin to condense but remain unpaired. (B) In zygotene, pairing of maternal and paternal homologous chromosomes (**homologs**) occurs, to form bivalents with four chromatids. (C) In pachytene, recombination (crossing over) occurs through the physical breakage and subsequent rejoining of maternal and paternal chromosome fragments. There are two chiasmata (crossovers) in the bivalent on the left, and one in the bivalent on the right. For simplicity, both chiasmata on the left involve the same two chromatids. In reality, more chiasmata may occur, involving three or even all four chromatids in a bivalent. (D) During diplotene, the homologous chromosomes may separate slightly, except at the chiasmata. A further stage, diakinesis, is marked by contraction of the bivalents and is the transition to metaphase I. In this figure, only 2 of 23 possible pairs of homologs are illustrated (with the maternal homolog colored pink, and the paternal homolog blue).

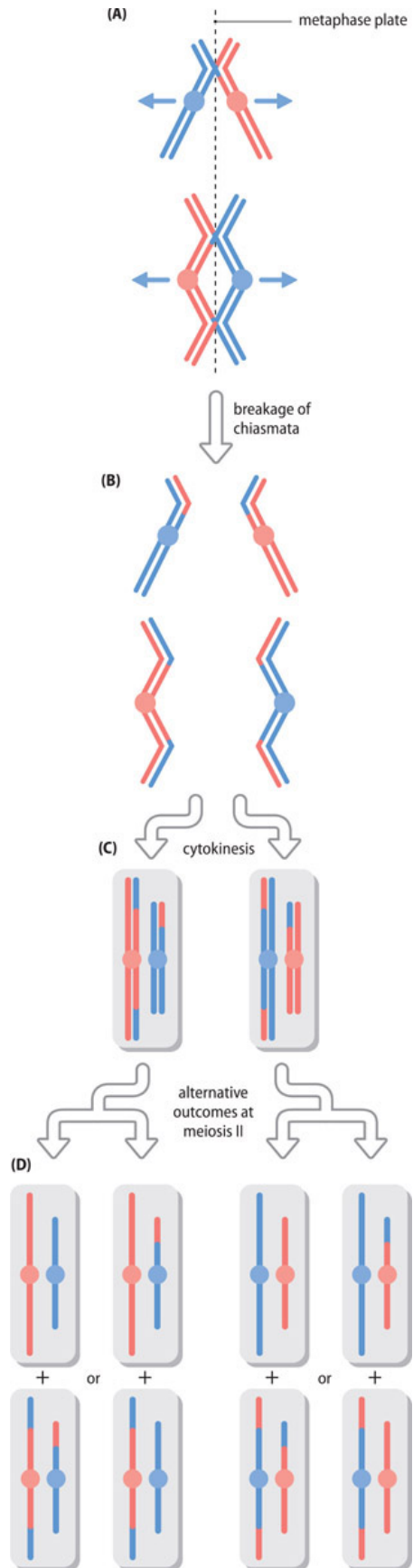


Figure 1.14 Metaphase I to production of gametes. (A) At metaphase I, the bivalents align on the metaphase plate, at the center of the spindle apparatus. Contraction of spindle fibers draws the chromosomes in the direction of the spindle poles (arrows). (B) The transition to anaphase I occurs at the consequent rupture of the chiasmata. (C) Cytokinesis segregates the two chromosome sets, each to a different primary spermatocyte. Note that, after recombination during prophase I (see [Figure 1.13C](#)), the chromatids share a single centromere but are no longer identical. (D) Meiosis II in each primary spermatocyte, which does not include DNA replication, generates unique genetic combinations in the haploid secondary spermatocytes. Only 2 of the possible 23 different human chromosomes are depicted, for clarity, so only 2^2 (that is, 4) of the possible 2^{23} (8 388 608) possible combinations are illustrated. Although oogenesis can produce only one functional haploid gamete per meiotic division, the processes by which genetic diversity arises are the same as in spermatogenesis.

Female meiosis is different: cell division is asymmetric, resulting in unequal division of the cytoplasm. The products of female meiosis I (the first meiotic cell division) are a large secondary oocyte and a small cell, the *polar body*, which is discarded. During meiosis II the secondary oocyte then gives rise to the large mature egg cell and a second polar body (which again is discarded).

In humans, primary oocytes enter meiosis I during fetal development but are then all arrested at prophase until after the onset of puberty. After puberty in females, one primary oocyte completes meiosis with each menstrual cycle. Because ovulation can continue up to the fifth and sometimes sixth decades, this means that meiosis can be arrested for many decades in those primary oocytes that are not used in ovulation until late in life.

Pairing of paternal and maternal homologs (synapsis)

Each of our diploid cells contains two copies (**homologs**) of each type of chromosome, one maternal copy and one paternal copy. So, for example,

paternal chromosome 1 and maternal chromosome 1 are homologs. The exception, of course, are the X and Y chromosomes in males.

A special feature of meiosis I—that distinguishes it from mitosis and meiosis II—is the pairing (*synapsis*) of paternal and maternal homologs. Then, the maternal and paternal homologs, each with sister chromatids following DNA replication, align along their lengths and become bound together. The resulting *bivalent* has four strands: two paternally inherited sister chromatids and two maternally inherited sister chromatids (see Figures 1.13B–D and 1.14).

The pairing of homologs is required for recombination to occur (as described in the next subsection). It must ultimately be dictated by high levels of DNA sequence identity between the homologs. The high sequence matching between homologs required for pairing does not need to be complete, however: when there is some kind of chromosome abnormality so that the homologs do not completely match, the matching segments usually manage to pair up.

Pairing of maternal and paternal sex chromosomes is straightforward in female meiosis, but in male meiosis there is the challenge of pairing a maternally inherited X chromosome with a paternally inherited Y. The human X chromosome is very much larger than the Y, and their DNA sequences are very different. However, they do have some sequences in common, notably a major *pseudoautosomal region* located close to the short-arm telomeres. The X and Y chromosomes cannot pair up along their lengths, but because they have some sequences in common, they can always pair up along these regions. We will explore this in greater detail in [Chapter 5](#) when we consider pseudoautosomal inheritance.

Recombination

The prophase of meiosis I begins during fetal life and, in human females, can last for decades. During this extended process, paternal and maternal chromatids within each bivalent normally exchange segments of DNA at

randomly positioned but matching locations. This process—called **recombination** (or crossover)—involves physical breakage of the DNA in one paternal and one maternal chromatid, and the subsequent joining of maternal and paternal fragments.

Recombined homologs seem to be physically connected at specific points. Each such connection marks the point of crossover and is known as a chiasma (plural chiasmata—see [Figure 1.13C](#)). The distribution of chiasmata across chromosomes is nonrandom. The number of chiasmata per meiosis shows significant sex differences, and there are very significant differences between individuals of the same sex (and even between individual meioses from a single individual). In a large recent study of human meiosis, an average of 38 recombinations were detected per female meiosis, while 24 meioses occurred on average in male meiosis but with very significant variation (shown in [Figure 8.3](#) on page 244). In addition to their role in recombination, chiasmata are thought to be essential for correct chromosome segregation during meiosis I.

There are hotspot regions where recombination is more likely to occur. For example, recombination is more common in subtelomeric regions. In the case of X–Y crossover there is an obligate crossover within a short 2.6 Mb pseudoautosomal region located at the tips of the short arms of the X and Y. This region is so called because it is regularly swapped between the X and Y chromosomes and so the inheritance pattern for any DNA variant here is not X-linked or Y-linked but instead resembles autosomal inheritance.

Why each of our gametes is unique

The sole purpose of sex in biology is to produce novel combinations of gene variants, and the instrument for achieving that aim is meiosis. The whole point of meiosis is to produce *genetically unique* gametes by selecting different combinations of DNA sequences on maternal and paternal homologs.

Although a single ejaculate may contain hundreds of millions of sperm, meiosis ensures that no two sperm will be genetically identical. Equally, no two eggs are genetically identical. Each zygote must also be unique because at fertilization a unique sperm combines with a unique egg. However, a unique fertilization event can occasionally give rise to two genetically identical (**monozygotic**) twins if the embryo divides into two at a very early stage in development (monozygotic twins are nevertheless unique individuals—genetics is not everything in life!).

The second division of meiosis is identical in form to mitosis; meiosis I is where the genetic diversity originates, and that involves two mechanisms. First, there is independent assortment of paternal and maternal homologs. After DNA replication, the homologous chromosomes each comprise two sister chromatids, so each bivalent is a four-stranded structure at the metaphase plate. Spindle fibers then pull one complete chromosome (two chromatids) to either pole. In humans, for each of the 23 homologous pairs, the choice of which daughter cell each homolog will enter is independent. This allows 2^{23} or about 8.4×10^6 different possible combinations of parental chromosomes in the gametes that might arise from a single meiotic division ([Figure 1.15](#)).

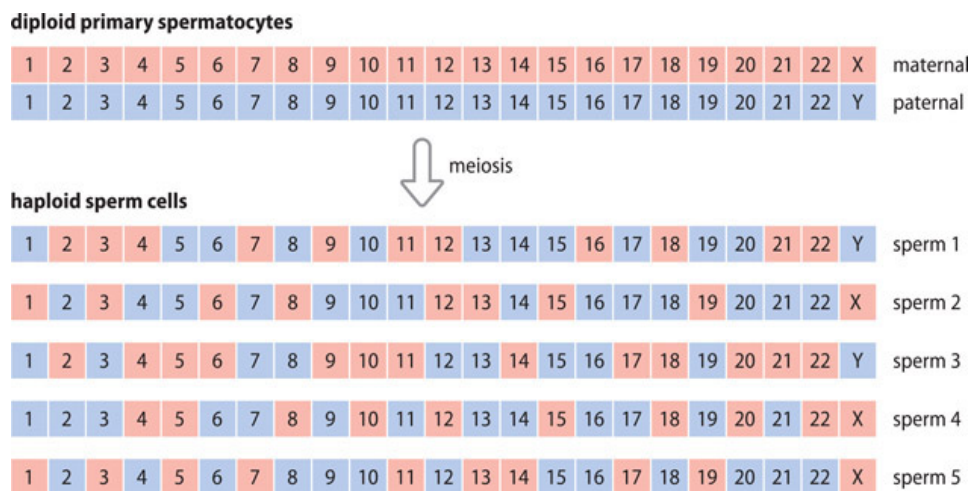


Figure 1.15 Independent assortment of maternal and paternal homologs during meiosis. The figure shows a random selection of just 5 of the 8 388 608 (2^{23}) theoretically possible combinations of homologs that might occur in haploid human

spermatozoa after meiosis in a diploid primary spermatocyte. Maternally derived homologs are represented by pink boxes, and paternally derived homologs by blue boxes. For simplicity, the diagram ignores recombination—but see [Figure 1.16](#).

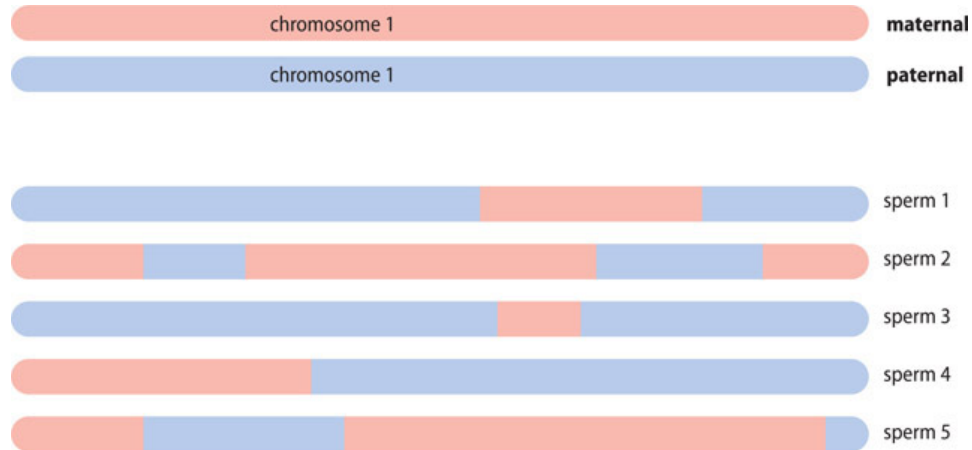


Figure 1.16 Recombination superimposes additional genetic variation at meiosis I.

[Figure 1.15](#) illustrates the contribution to genetic variation at meiosis I made by independent assortment of homologs, but for simplicity it ignores the contribution made by recombination. In reality each transmitted chromosome is a mosaic of paternal and maternal DNA sequences, as shown here. See [Figure 8.1](#) on page 243 for a real-life example.

The second mechanism that contributes to genetic diversity is recombination. Whereas sister chromatids within a bivalent are genetically identical, the paternal and maternal chromatids are not. On average their DNA will differ at roughly 1 in every 1000 nucleotides. Swapping maternal and paternal sequences by recombination will therefore produce an extra level of genetic diversity ([Figure 1.16](#)). It raises the number of permutations from the 8.4 million that are possible just from the independent assortment of maternal and paternal homologs alone, to a virtually infinite number.

SUMMARY

- Nucleic acids are negatively charged long polymers composed of sequential nucleotides that each consist of a sugar, a nitrogenous base, and a phosphate group. They have a sugar-phosphate backbone with bases projecting from the sugars.
- Nucleic acids have four types of bases: adenine (A), cytosine (C), guanine (G), and either thymine (T) in DNA or uracil (U) in RNA. The sequence of bases determines the identity of a nucleic acid and its function.
- RNA normally consists of a single nucleic acid chain, but in cells a DNA molecule has two chains (strands) that form a stable duplex (in the form of a double helix). Duplex formation requires hydrogen bonding between matched bases (base pairs) on the two strands.
- In DNA two types of base pairing exist: A pairs with T, and C pairs with G. According to these rules, the two strands of a DNA double helix are said to have complementary base sequences.
- Base pairing also occurs in RNA and includes G–U base pairs, as well as G–C and A–U base pairs. Two different RNA molecules with partly complementary sequences can associate by forming hydrogen bonds. Intramolecular hydrogen bonding also allows a single RNA chain to form a complex three-dimensional structure.
- DNA carries primary instructions that determine how cells work and how an individual is formed. Defined segments of DNA called genes are used to make a single-stranded RNA copy that is complementary in sequence to one of the DNA strands (transcription).
- DNA is propagated from one cell to daughter cells by replicating itself. The two strands of the double helix are unwound, and each strand is used to make a new complementary DNA copy. The two new nuclear DNA double

helices (each with one parental DNA strand and one new DNA strand) are segregated so that each daughter cell receives one DNA double helix.

- RNA molecules function in cells either as a mature noncoding RNA, or as a messenger RNA with a coding sequence used to make the polypeptide chain of a protein (translation).
- Each nuclear DNA molecule is complexed with different proteins and some noncoding RNAs to form a chromosome that condenses the DNA and protects it.
- Packaging DNA into chromosomes stops the long DNA chains from getting entangled within cells, and by greatly condensing the DNA in preparation for cell division it allows the DNA to be segregated correctly to daughter cells and to offspring.
- Our sperm and egg cells are haploid cells with a set of 23 different chromosomes (each with a single distinctive DNA molecule). There is one sex chromosome (an X chromosome in eggs; either an X or Y in sperm) and 22 different autosomes (nonsex chromosomes).
- Most of our cells are diploid with two copies of the haploid chromosome set, one set inherited from the mother and one from the father. Maternal and paternal copies of the same chromosome are known as homologs.
- There is one type of mitochondrial DNA (mtDNA); it is present in many copies with wide variation in copy number between different cell types. Both the replication of mtDNA and its segregation to daughter cells occur stochastically.
- Cells need to divide as we grow. In fully formed adults, most of our cells are specialized, nondividing cells, but some cells are required to keep on dividing to replace short-lived cells, such as blood, skin, and intestinal epithelial cells.
- Mitosis is the normal form of cell division. Each chromosome (and chromosomal DNA) replicates once and the duplicated

chromosomes are segregated equally into the two daughter cells.

- Meiosis is a specialized form of cell division required to produce haploid sperm and egg cells. The chromosomes in a diploid spermatogonium or oogonium replicate once, but there are two successive cell divisions to reduce the number of chromosomes in each cell.
- Each sperm cell produced by a man is unique, as is each egg cell that a woman produces. During the first cell division in meiosis, maternal and paternal homologs associate and exchange sequences by recombination. Largely random recombination results in unpredictable new DNA sequence combinations in each sperm and in each egg.

QUESTIONS

Questions can be downloaded by visiting the following link, under Support Materials: www.routledge.com/9780367490812.

FURTHER READING

More detailed treatment of the subject matter in this chapter can be found in more comprehensive genetics and cell biology textbooks such as:

Alberts B, Johnson A, Lewis J, Morgan D, Raff M, Roberts K & Walter P (2015) *Molecular Biology of the Cell*, 6th ed. Garland Science.

Strachan T & Read AP (2019) *Human Molecular Genetics*, 5th ed. CRC Press, Taylor & Francis.

2

Fundamentals of gene structure, gene expression, and human genome organization

DOI: [10.1201/9781003044406-2](https://doi.org/10.1201/9781003044406-2)

CONTENTS

[2.1 PROTEIN-CODING GENES: STRUCTURE AND EXPRESSION](#)

[2.2 RNA GENES AND NONCODING RNA](#)

[2.3 WORKING OUT THE DETAILS OF OUR GENOME AND WHAT THEY MEAN](#)

[2.4 A QUICK TOUR OF SOME ELECTRONIC RESOURCES USED TO INTERROGATE THE HUMAN GENOME SEQUENCE AND GENE PRODUCTS](#)

[2.5 THE ORGANIZATION AND EVOLUTION OF THE HUMAN GENOME](#)

[SUMMARY](#)

[QUESTIONS](#)

[FURTHER READING](#)

Our genome is complex, comprising about 3.2 Gb (3.2×10^9 base pairs; Gb = giga-base) of DNA. One of its main tasks is to produce a huge variety of different proteins that dictate how our cells work. Surprisingly, however, **coding DNA**—DNA sequences that specify the polypeptides of our proteins—account for just about 1.5 % of our DNA.

The remainder of our genome is noncoding DNA that does not make any protein. A significant fraction of the noncoding DNA is functionally important, including many different classes of DNA regulatory sequences that control how our genes work (such as promoters and enhancers), and DNA sequences that specify short regulatory sequence elements that work at the RNA level.

Additionally, we have many thousands of genes that do not make polypeptides; instead they make different classes of functional noncoding RNA. Some of these **RNA genes**—such as genes encoding ribosomal RNA and transfer RNA needed for protein synthesis—have been known for decades, but one of the big surprises in recent years has been the sheer number and variety of noncoding RNAs in our cells. In addition to the RNA genes, our protein-coding genes frequently make noncoding RNA transcripts as well as messenger RNAs (mRNAs).

Like other complex genomes, our genome has a large proportion of moderately to highly repetitive DNA sequences. Some of these are important in centromere and telomere function; others are

important in genome evolution.

By 2003, the Human Genome Project (HGP) provided the first comprehensive insights into our genome, delivering an essentially complete nucleotide sequence of the gene-rich euchromatic component of the genome. Follow-up studies have compared our genome with other genomes, helping us to understand how our genome evolved. The comparative genomics studies, together with genome-wide functional and bioinformatic analyses, are providing major insights into how our genome works.

2.1 PROTEIN-CODING GENES: STRUCTURE AND EXPRESSION

Proteins are the main functional endpoints of gene expression and perform a huge diversity of roles that govern how cells work (acting as structural components, enzymes, carrier proteins, ion channels, signaling molecules, gene regulators, and so on). They each consist of one or more polypeptides, long sequences of amino acids that are encoded by a coding DNA. In many cases a protein also contains carbohydrate or lipid components (which are not genetically determined).

Protein-coding genes come in a startling variety of organizations, as described below, and synthesize one or more polypeptides. Polypeptide synthesis is not the endpoint, however. A newly synthesized polypeptide must undergo multiple different maturation steps, usually involving chemical modification and cleavage events, and often then associates with other polypeptides to form a working protein.

Gene organization: exons and introns

The protein-coding genes of bacteria are small (on average about 1000 bp long) and simple. The gene is transcribed to give an mRNA with a continuous coding sequence that is then translated to give a linear sequence of about 300 amino acids on average. Unexpectedly, the genes of eukaryotes turned out to be much bigger and much more complex than anticipated. And, as we will see, our protein-coding genes often contain a rather small amount of coding DNA.

For most eukaryotic protein-coding genes, the coding DNA is split into segments (**exons**) separated by noncoding DNA sequences (**introns**). The number of exons and introns in a gene varies considerably (there seems little logic about precisely where introns insert within genes).

Excluding single-exon genes (some genes lack introns), average exon lengths show moderate variation from gene to gene, but introns can show extraordinary size differences. Our genes are therefore often large, sometimes extending over more than a megabase of DNA ([Table 2.1](#)).

TABLE 2.1 EXAMPLES OF DIFFERENTIAL GENE ORGANIZATION FOR HUMAN PROTEIN-CODING GENES

Human gene	Size in genome (kb)	No. of exons	Average size of exon (bp) *	Average size of intron (bp) **
<i>SRY</i> (male sex-determinant)	0.9	1	850	-

Items in brackets show the protein name. kb, kilobases (= 1000 bp).

* Note that the shortest human exon is just two nucleotides long, and final exons can quite often be long, the record being 27 303 bp.

** The shortest human intron is 26bp, and the longest is 1160 411 bp—see PMID31164174

Human gene	Size in genome (kb)	No. of exons	Average size of exon (bp) *	Average size of intron (bp) **
<i>HBB</i> (-globin)	1.6	3	150	490
<i>TP53</i> (p53)	39	10	236	3076
<i>F8</i> (factor VIII)	186	26	375	7100
<i>CFTR</i> (cystic fibrosis transmembrane regulator)	250	27	227	9100
<i>DMD</i> (dystrophin)	2400	79	180	30 770

Items in brackets show the protein name. kb, kilobases (= 1000 bp).

* Note that the shortest human exon is just two nucleotides long, and final exons can quite often be long, the record being 27 303 bp.

** The shortest human intron is 26bp, and the longest is 1160 411 bp—see PMID31164174

RNA splicing: stitching together the genetic information in exons

Like all genes, genes that are split into exons are initially transcribed by an RNA polymerase to give a long RNA transcript. This primary transcript is identical in base sequence to the transcribed region of the sense DNA strand, except that U replaces T (the transcribed region of DNA is called a **transcription unit**). Thereafter, the primary RNA transcript undergoes a form of processing called **RNA splicing**.

RNA splicing involves first cleaving the RNA transcript at the junctions between transcribed exons and introns. The individual transcribed intron sequences are often degraded, but the transcribed exon sequences are then covalently linked (spliced) in turn to make a mature RNA ([Figure 2.1](#)). RNA splicing is performed within the nucleus by spliceosomes, complex assemblies of protein factors and small nuclear RNA (snRNA) molecules.

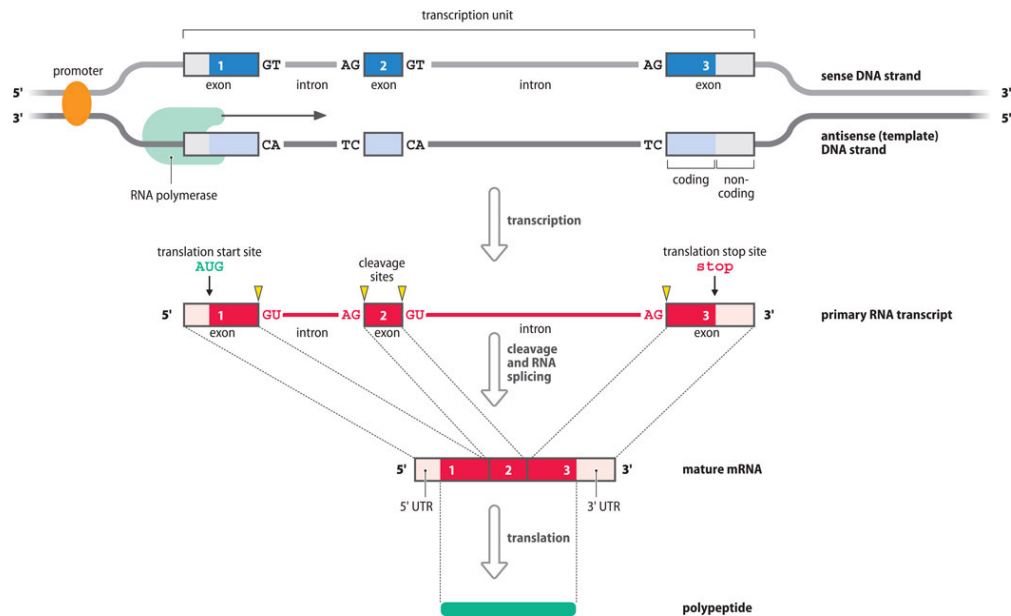


Figure 2.1 RNA splicing brings transcribed exon sequences together. Most of our protein-coding genes (and many RNA genes) undergo RNA splicing. In this generalized example a protein-coding gene is illustrated with an upstream promoter and three exons separated by two introns that each begin with the dinucleotide GT and end in the dinucleotide

AG. The central exon (exon 2) is composed entirely of coding DNA (deep red), but exons 1 and 3 have both coding DNA and also noncoding DNA sequences (shown in pink; they will eventually be used to make untranslated sequences in the mRNA). The three exons and the two separating introns are transcribed together to give a large primary RNA transcript. The RNA transcript is cleaved at positions corresponding to exon-intron boundaries. The two transcribed intron sequences that are excised are each degraded, but the transcribed exon sequences are joined (*spliced*) together to form a contiguous mature RNA that has noncoding sequences at both the 5' and 3' ends. In the mature mRNA these terminal sequences will not be translated and so are known as *untranslated regions* (UTRs). The central coding sequence of the mRNA is defined by a translation start site (which is almost always the trinucleotide AUG) and a translation stop site, and is read (*translated*) to produce a polypeptide.

We do not fully understand how the spliceosome is able to recognize and cut the primary transcript at precise positions marking the start and end of introns. However, we do know that certain sequences are important in signaling the splice sites that define exon-intron boundaries. For example, very nearly all introns begin with a GT dinucleotide on the sense DNA strand and end with an AG so that the transcribed intron sequence begins with a GU (that marks the **splice donor site**) and ends in an AG (marking the **splice acceptor site**). The GT (GU) and AG end sequences need to be embedded in broader splice site consensus sequences that we will describe in [Section 6.1](#) when we consider how gene expression is regulated. As we will see in [Chapter 7](#), mutations at splice sites are important causes of disease.

[Figure 2.1](#) might give the erroneous impression that all protein-coding genes undergo a specific, single type of RNA splicing. However, close to 10 % of our protein-coding genes have a single, uninterrupted exon and do not undergo RNA splicing at all—notable examples include histone genes. And most of the genes that go through RNA splicing undergo alternative RNA splicing patterns; a single gene can therefore produce different gene products that may be functionally different. We consider the concept of alternative splicing in greater detail in [Chapter 6](#), in the context of gene regulation.

The evolutionary value of RNA splicing

As we will see in [Section 2.2](#), many RNA genes also undergo RNA splicing. At this stage, one might reasonably wonder why RNA splicing is so important in eukaryotic cells, and so especially prevalent in complex multicellular organisms. Why do we need to split the genetic information in genes into sometimes so many different little exons? The answer is to help stimulate the formation of novel genes and novel gene products that can permit greater functional complexity during evolution.

The huge complexity of humans and other multicellular organisms has been driven by genome evolution. In addition to periodic gene duplication, various genetic mechanisms allow individual exons to be duplicated or swapped from one gene to another on an evolutionary timescale. That allows different ways of combining exons to produce novel hybrid genes. An additional source of complexity comes from using different combinations of exons to make alternative transcripts from the same gene (alternative splicing).

Translation: decoding messenger RNA to make a polypeptide

Messenger RNA (mRNA) molecules produced by RNA splicing in the nucleus are exported to the cytoplasm. Here they are bound by ribosomes, very large complexes consisting of four types of ribosomal RNA (rRNA) and many different proteins.

Although an mRNA is formed from exons only, it has sequences at its 5' and 3' ends that are noncoding. Having bound to mRNA, the job of the ribosomes is to scan the mRNA sequence to find and interpret a central coding sequence that will be translated to make a polypeptide. The noncoding sequences at the ends are known as **untranslated regions** (UTRs; as shown in [Figure 2.1](#)), and contain sequences that are important in regulating gene expression.

A polypeptide is a polymer made up of a linear sequence of **amino acids** ([Figure 2.2A](#)). Amino acids have the general formula $\text{NH}_2\text{-CH(R)-COOH}$, where R is a variable side chain that defines the chemical identity of the amino acid and is connected to the central (alpha) carbon of the NH-CH-CO framework sequence. There are 20 common amino acids ([Figure 2.2C](#)). Polypeptides are made by a condensation reaction between the carboxyl (COOH) group of one amino acid and the amino (NH₂) group of another amino acid, forming a peptide bond (see [Figure 2.2B](#)).

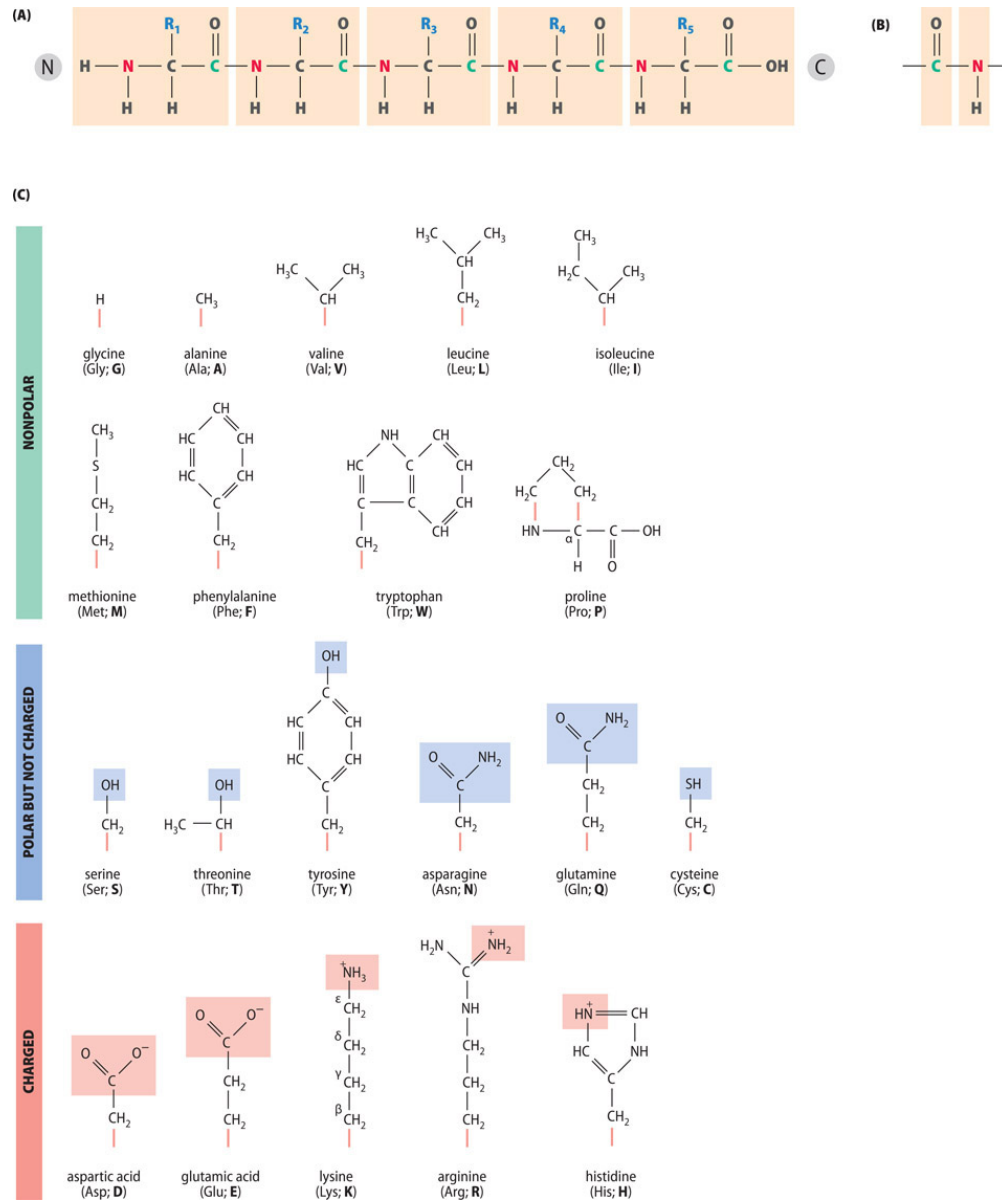


Figure 2.2 Polypeptide and amino acid structure. (A) Polypeptide primary structure. A pentapeptide is shown with its five amino acids highlighted. Here, the left end is called the N-terminal end because the amino acid has a free amino (NH_2) group; the right end is the C-terminal end because the last amino acid has a free carboxyl (COOH) group. The side chains (R_1 to R_5) are variable and determine the identity of the amino acid. They are joined to the central carbon atom of the repeating framework sequence: $-\text{NH}-\text{CH}-\text{CO}-$. Note that at physiological pH the free amino and carboxyl groups will be charged: NH_3^+ and COO^- respectively. (B) Neighboring amino acids are joined by a peptide bond. A peptide bond is formed by a condensation reaction between the end carboxyl group of one amino acid and the end amino group of another: $-\text{COOH} + \text{NH}_2 \rightarrow -\text{CONH}- + \text{H}_2\text{O}$. (C) Side chains of the 20 principal amino acids. Red lines represent the covalent bond attaching the side chain to the framework protein structure. Note that the structure of proline is unusual and its full structure is given here because its side chain connects to the nitrogen atom of the framework amino group as well as to the alpha carbon, thereby forming a five-membered ring.

To make a polypeptide, the coding sequence within an mRNA is translated in groups of three nucleotides at a time, called **codons**. There are 64 possible codons (four possible bases at each of three nucleotide positions makes $4 \times 4 \times 4$ permutations). Of these, 61 are used to specify an amino acid; three others signal an end to protein synthesis. The universal **genetic code**, the set of rules that dictate how codons are interpreted, therefore has some redundancy built into it. For example, the amino acid serine can be specified by any of six codons (UCA, UCC, UCG, UCU, AGU, and AGG), and, on average, an amino acid is specified by any of three codons. As a result, nucleotide substitutions within coding DNA quite often do not cause a change of amino acid. We discuss the genetic code in some detail in [Section 7.2](#) when we consider the effects of single nucleotide substitutions.

The process of translation

Translation begins when ribosomes bind to the 5' end of an mRNA and then move along the RNA to find a translational start site, the initiation codon—an AUG trinucleotide embedded within the broader, less well defined Kozak consensus sequence (GCC**Pu**CCAUGG; the most conserved bases are shown in bold, and Pu represents purine).

The initiation codon is the start of an **open reading frame** of codons that specify successive amino acids in the polypeptide chain (see [Box 2.1](#) for the concept of translational reading frames). As described below, a family of transfer RNAs (tRNAs) is responsible for transporting the correct amino acids to be inserted in the required position of the growing polypeptide chain. Individual types of tRNA carry a specific amino acid; they can recognize and bind to a specific codon, and when they do so they unload their amino acid cargo.

BOX 2.1 TRANSLATIONAL READING FRAMES AND SPLITTING OF CODING SEQUENCES BY INTRONS

TRANSLATIONAL READING FRAMES

In the examples of different translational **reading frames** below, we use sequences of words containing three letters to represent the triplet nature of the genetic code. We designate the reading frames (RF) as 1, 2, or 3 depending on whether the reading frame starts before the first, second, or third nucleotide in the sequence.

Reading frame 1 (RF1) in [Figure 1](#) makes sense, but a shift to another reading frame produces nonsense. The same principle generally applies to coding sequences. So, for example, if one or two nucleotides are deleted from a coding sequence or there is an insertion of one or two nucleotides, the effect is to produce a **frame-shift** (a change of reading frame) that will result in nonsense.

sequence: THEOLDMANGOTOFFTHEBUSANDSAWTHEBIGREDDOGANDHERPUP

RF1: THE OLD MAN GOT OFF THE BUS AND SAW THE BIG RED DOG AND HER PUP

RF2: T HEO LDM ANG OTO FFT HEB USA NDS AWT HEB IGR EDD OGA NDH ERP UP

RF3: TH EOL DMA NGO TOF FTH EBU SAN DSA WTH EBI GRE DDO GAN DHE RPU P

Figure 1 The importance of using the correct translational reading frame. The sequence of letters at the top can be grouped into sets of three (codons) that make sense in reading frame 1 (RF1) but make no sense when using reading frame 2 (RF2) or reading frame 3 (RF3).

SPLITTING OF CODING SEQUENCES BY INTRONS

At the DNA level, introns may interrupt a coding sequence at one of three types of position: at a point precisely between two codons (a *phase 0 intron*), after the first nucleotide of a codon (a *phase 1 intron*), or after the second nucleotide of a codon (a *phase 2 intron*).

An internal exon may be flanked by introns of the same phase; in an exon like this the number of nucleotides is always exactly divisible by three. Where an exon is flanked by two introns of a different phase, the exon will have a number of nucleotides that is not exactly divisible by three. That can have important consequences when deletions occur within genes (see [Figure 2](#)).

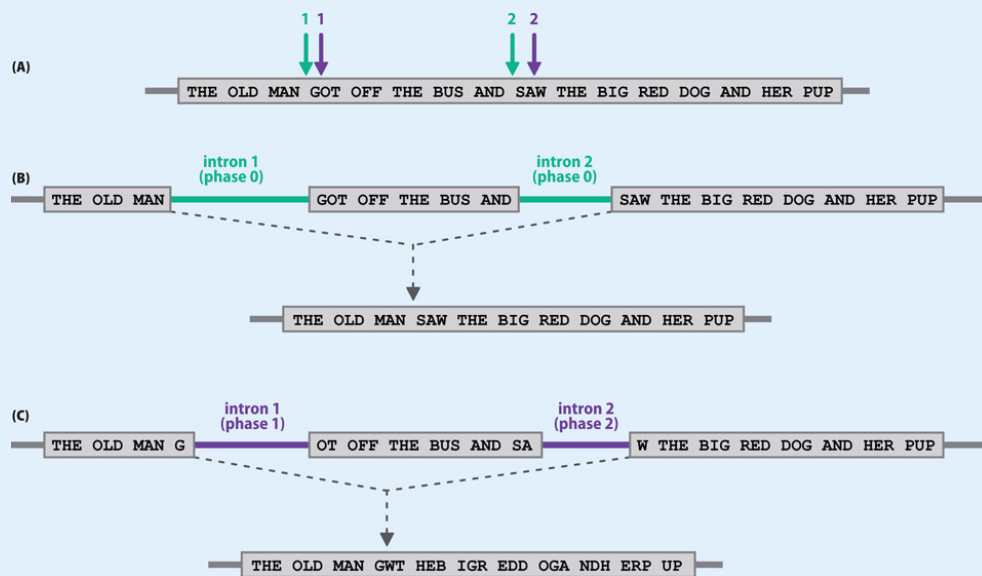


Figure 2 Effects on the translational reading frame caused by the deletion of coding exons. (A) Here we show a coding sequence split at the DNA level by a pair of introns, 1 and 2. Now imagine two alternative possibilities, shown by green and purple arrows. Green arrows indicate two flanking introns of the same phase: in this case both are phase 0 introns, having inserted at analogous positions *between* codons. Purple arrows indicate an alternative where the introns are of different phases, respectively phase 1 for intron 1 and phase 2 for intron 2. B) The green introns result in the central exon having a number of nucleotides that is exactly divisible by three; it can be deleted without an effect on the downstream reading frame. If the exon does not encode a critical component of the protein, the functional consequences may not be too grave. (C) If instead introns 1 and 2 are located as shown in purple, the central exon has a number of nucleotides that cannot be divided exactly by three. If it were to be deleted, the downstream reading frame would be scrambled with a high chance of a premature termination codon, frequently resulting in lack of function.

As each new amino acid is unloaded it is bonded to the previous amino acid so that a polypeptide chain is formed ([Figure 2.3](#)). The first amino acid has a free NH₂ (amino) group and marks the N-terminal end (N) of the polypeptide. The polypeptide chain terminates after the ribosome encounters a **stop codon** (which signifies that the ribosome should disengage from the mRNA, releasing the

polypeptide; for translation on cytoplasmic ribosomes, there are three choices of stop codon: UAA, UAG, or UGA). The last amino acid that was incorporated in the polypeptide chain has a free COOH (carboxyl group) and marks the C-terminal end (C) of the polypeptide.

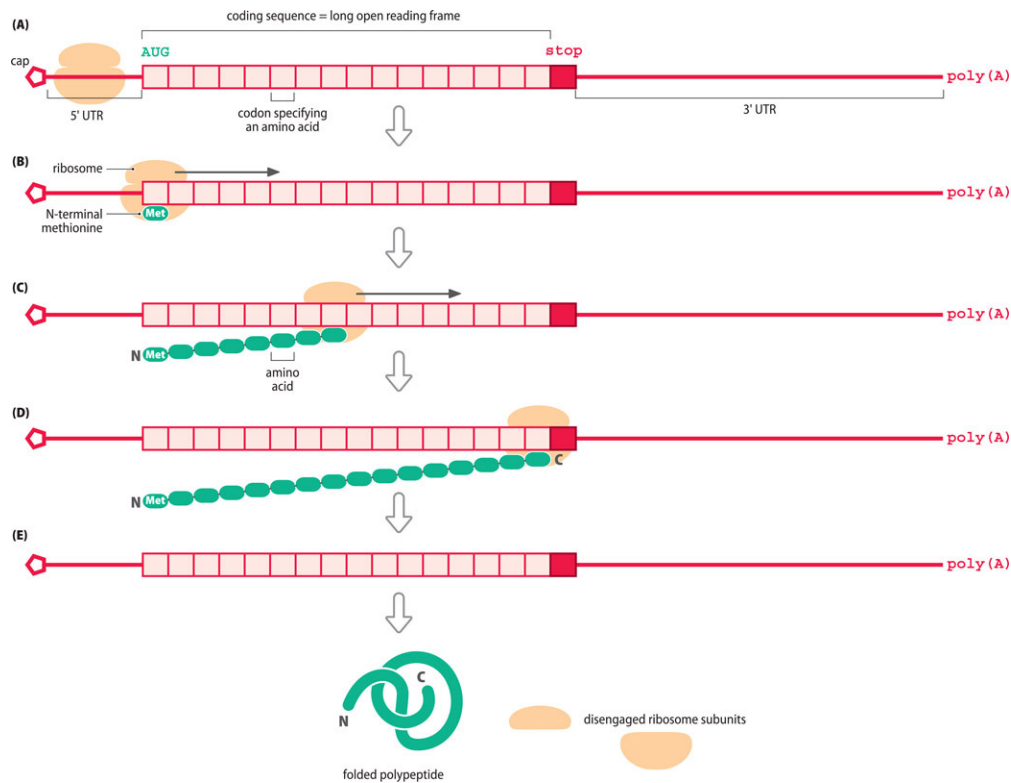


Figure 2.3 The basics of translation. (A) A ribosome attaches to the 5' untranslated region (5' UTR) of the mRNA and then slides along until (B) it encounters the initiation codon AUG, at which point a methionine-bearing transfer RNA (not shown) engages with the AUG codon and deposits its methionine cargo (green bar, labeled MET). (C) The ribosome continues to move along the mRNA and as it encounters each codon in turn a specific amino-acid-bearing tRNA is recruited to recognize the codon and to deposit its amino acid, according to the *genetic code*. The ribosome catalyses the formation of a *peptide bond* (Figure 2.2B) between each new amino acid and the last amino acid, forming a polypeptide chain (shown here, for convenience, as a series of joined green ovals). (D) Finally, the ribosome encounters a stop codon, at which point (E) the ribosome falls off the mRNA and dissociates into its two subunits, releasing the completed polypeptide. The polypeptide undergoes *post-translational modification* as described in the text, which may sometimes involve cleavage at the N-terminal end so that methionine may not be the N-terminal amino acid in the mature polypeptide.

Transfer RNA as an adaptor RNA

Transfer RNAs have a classic cloverleaf structure resulting from intramolecular hydrogen bonding (Figure 2.4A). They serve as adaptor RNAs because their job is to base pair with mRNAs and help decode the coding sequence messages carried by mRNAs. The base pairing is confined to a three-nucleotide sequence in the tRNA called an *anticodon*, which is complementary in sequence to a codon. According to the identity of their anticodons, different tRNAs carry different amino acids covalently linked to their 3' ends. Through base pairing between codon and anticodon, individual

amino acids can be sequentially ordered according to the sequence of codons in an mRNA, and sequentially linked together to form a polypeptide chain (see [Figure 2.4B](#)).

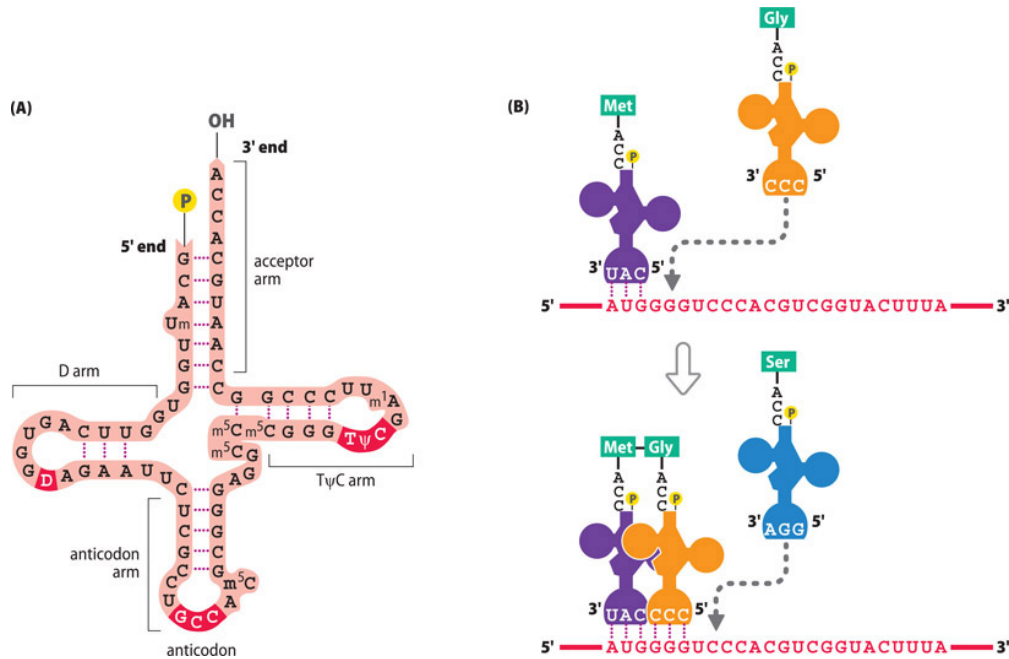


Figure 2.4 Transfer RNA structure, and its role as an adaptor RNA in translation. (A) Transfer RNA structure. The tRNA^{Gly} shown here illustrates the classical cloverleaf tRNA structure. Intramolecular base pairing produces three arms terminating in a loop plus an acceptor arm formed by pairing of 5' and 3' end sequences. The latter ends in a -CAA trinucleotide and covalently binds a specific amino acid. The three nucleotides at the centre of the middle loop form the *anticodon*, which identifies the tRNA according to the amino acid it will bear. Minor nucleotides are: D, 5,6-dihydrouridine; Y, pseudouridine (5-ribosyluracil); m⁵C, 5-methylcytidine; m¹A, 1-methyladenosine; Um, 2'-O-methyluridine. (B) Role of adaptor RNA. Different tRNAs carry different amino acids, according to the type of anticodon they bear. As a ribosome traverses an mRNA it identifies the AUG initiation codon. A methionine-bearing tRNA with the complementary CAU anticodon sequences then engages with the ribosome so that the CAU anticodon base pairs with the AUG codon. Note: for ease of illustration we show the tRNAs in the opposite orientation to the standard form shown in (A), with the acceptor arm on the left, not the right. Thereafter, a tRNA bearing glycine engages the second codon, GGG, by base pairing with its CCC anticodon. The ribosome's peptidyltransferase then forms a peptide bond between the N-terminal methionine and glycine. The ribosome moves along by one codon and the tRNA^{Met} is cleaved so that it can be reused and the process continues with an incoming tRNA carrying a serine and an anticodon GGA to bind to the third codon UCC, after which the incoming serine will be covalently bonded to the glycine by the ribosome's peptidyltransferase.

Untranslated regions and 5' cap and 3' poly(A) termini

As illustrated in [Figure 2.3](#), each mature mRNA has a large central coding DNA sequence flanked by two **untranslated regions**, a short 5' untranslated region (5' UTR) and a rather longer 3' untranslated region (3' UTR). The untranslated regions regulate mRNA stability and contain regulatory sequences that are important in determining how genes are expressed.

As well as sequences copied from the gene sequence, mRNA molecules usually also have end sequences added post-transcriptionally to the pre-mRNA. At the 5' end a specialized cap is added: 7-methylguanosine linked to the first nucleotide by a distinctive 5'-5' phosphodiester bond (instead of a normal 5'-3' phosphodiester bond). The cap protects the transcripts against 5' → 3' exonuclease attack and facilitates transport to the cytoplasm and ribosome attachment. At the 3' end a dedicated poly(A) polymerase sequentially adds adenylate (AMP) residues to give a poly(A) tail, about 150–200 nucleotides long. The poly(A) helps in transporting mRNA to the cytoplasm, facilitates binding to ribosomes, and is also important in stabilizing mRNAs.

From newly synthesized polypeptide to mature protein

The journey from newly synthesized polypeptide released from the ribosome to fully mature protein requires several steps. The polypeptide typically undergoes post-translational cleavage and chemical modification. Polypeptides also need to fold properly, and they often bind to other polypeptides as part of a multisub-unit protein. And then there is a need to be transported to the correct intracellular or extracellular location.

Chemical modification

We describe below one type of chemical modification that involves cross-linking between two cysteine residues within the same polypeptide or on different polypeptides. Often, however, chemical modification involves the simple covalent addition of chemical groups to polypeptides or proteins. Sometimes small chemical groups are attached to the side chains of specific amino acids ([Table 2.2](#)). These groups can sometimes be particularly important in the structure of a protein (as in the case of collagens, which have high levels of hydroxyproline and hydroxylysine).

TABLE 2.2 COMMON TYPES OF CHEMICAL MODIFICATION OF PROTEINS BY COVALENT ADDITION OF CHEMICAL GROUPS TO A SIDE CHAIN

Type of chemical modification	Target amino acids	Comments
ADDITION OF SMALL CHEMICAL GROUP		
Hydroxylation	Pro; Lys; Asp	can play important structural roles
Carboxylation	Glu	especially in some blood clotting factors
Methylation	Lys	specialized enzymes can add or remove the methyl, acetyl, or phosphate group, causing the protein to switch between two states, with functional consequences
Acetylation	Lys	
Phosphorylation	Tyr;Ser;Thr	
ADDITION OF COMPLEX CARBOHYDRATE OF LIPID GROUP		
N-glycosylation	Asn	added to the amino group of Asn in endoplasmic reticulum and Golgi apparatus
O-glycosylation	Ser;Thr; Hydroxylysine	added to the side-chain hydroxyl group; takes place in Golgi apparatus

Type of chemical modification	Target amino acids	Comments
N-lipidation	Gly	added to the amino group of an N-terminal glycine; promotes protein-membrane interactions
S-lipidation	Cys	a palmitoyl or prenyl group is added to the thiol of the cysteine. Often helps anchor proteins in a membrane

In other cases, dedicated enzymes add or remove small chemical groups to act as switches that convert a protein from one functional state to another. Thus, specific kinases can add a phosphate group that can be subsequently removed by a dedicated phosphatase. The change between phosphorylated and dephosphorylated states can result in a major conformational change that affects how the protein functions. Similarly, methyltransferases and acetyltransferases add methyl or acetyl groups that can be removed by the respective demethylases and deacetylases. As we will see in [Chapter 6](#), they are particularly important in modifying histone proteins to change the conformation of chromatin and thereby alter gene expression.

In yet other cases, proteins can be modified by covalently attaching complex carbohydrates or lipids to a polypeptide backbone. Thus, for example, secreted proteins and proteins destined to be part of the excretory process of cells routinely have oligosaccharides attached to the side chains of specific amino acids. Different types of lipids are also often added to membrane proteins (see [Table 2.2](#)).

Folding

The amino acid sequence, the primary structure, dictates the pattern of folding, but certain regions of polypeptides adopt types of secondary structure important in protein folding ([Box 2.2](#) gives an outline of protein structure). Until correct folding has been achieved, a protein is unstable; different chaperone molecules help with the folding process (careful supervision is needed because partly folded or misfolded proteins can be toxic to cells).

BOX 2.2 A BRIEF OUTLINE OF PROTEIN STRUCTURE

Four different levels of structure are recognized:

- primary structure—the linear sequence of amino acids in constituent polypeptides
- secondary structure—the path that a polypeptide backbone follows within local regions of the primary structure
- tertiary structure—the overall three-dimensional structure of a polypeptide
- quaternary structure—the aggregate structure of a multimeric protein (composed of two or more polypeptide subunits that may be of more than one type).

ELEMENTS OF SECONDARY STRUCTURE

Secondary structure is notably shaped by intramolecular hydrogen bonding. The α -helix, for example, is a rigid cylinder stabilized by hydrogen bonding between the carbonyl oxygen of a

peptide bond and the hydrogen atom of the amino nitrogen of a peptide bond located four amino acids away (**Figure 1A**). α -Helices often occur in transcription factors and other proteins that perform key cellular functions.

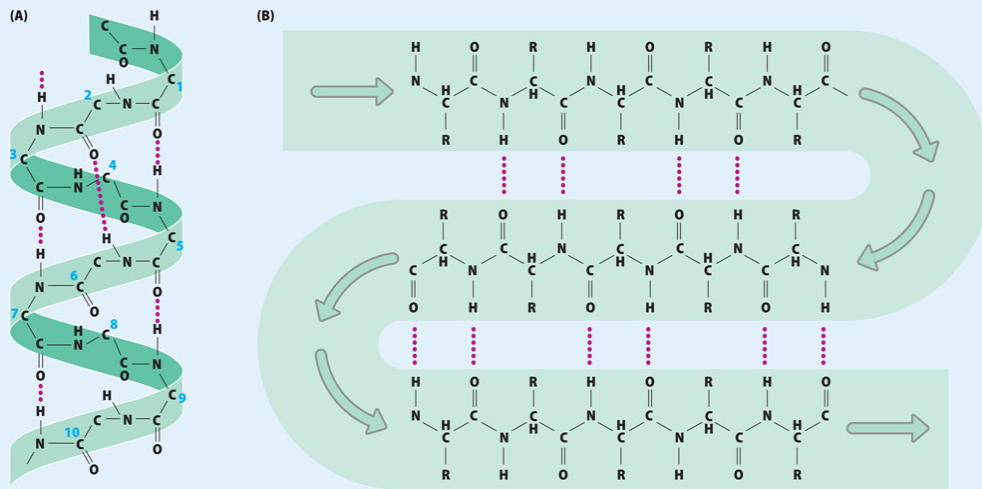


Figure 1 Elements of protein secondary structure. (A) An α -helix. The solid rod structure is stabilized by hydrogen bonding between the oxygen of the carbonyl group ($C = O$) of each peptide bond and the hydrogen on the peptide bond amide group (NH) of the fourth amino acid away, yielding 3.6 amino acids per turn of the helix. The side chains of each amino acid are located on the outside of the helix; there is almost no free space within the helix. Note: only the backbone of the polypeptide is shown, and some bonds have been omitted for clarity. (B) A β -sheet (also called a β -pleated sheet). Here, hydrogen bonding occurs between the carbonyl oxygens and amide hydrogens on adjacent segments of a sheet that may be composed either of parallel segments of the polypeptide chain or, as shown here, of antiparallel segments (arrows mark the direction of travel from N-terminus to C-terminus), hydrogen bond.

In the β -sheet (also called the β -pleated sheet), the hydrogen bonds occur between opposed peptide bonds in parallel or antiparallel segments of the same polypeptide chain (**Figure 1B**). β -Sheets occur—often together with α -helices—at the core of most globular proteins.

The β -turn involves hydrogen bonding between the peptide-bond carbonyl ($C = O$) of one amino acid and the peptide-bond NH group of an amino acid located only three places farther along. The resulting hairpin turn allows an abrupt change in the direction of a polypeptide, enabling compact globular shapes to be achieved. β -Turns can connect neighboring segments in a β -sheet, when the polypeptide strand has to undergo a sharp turn.

When placed in an aqueous environment, proteins are stabilized by having amino acids with hydrophobic side chains located in the interior of the protein, whereas hydrophilic amino acids tend to be located toward the surface. For many proteins, notably globular proteins, the folding pattern is also stabilized by a form of covalent cross-linking that can occur between certain distantly located cysteine residues—the sulfhydryl groups of the cysteine side chains interact to form a disulfide bond (alternatively called a disulfide bridge—see **Figure 2.5**).

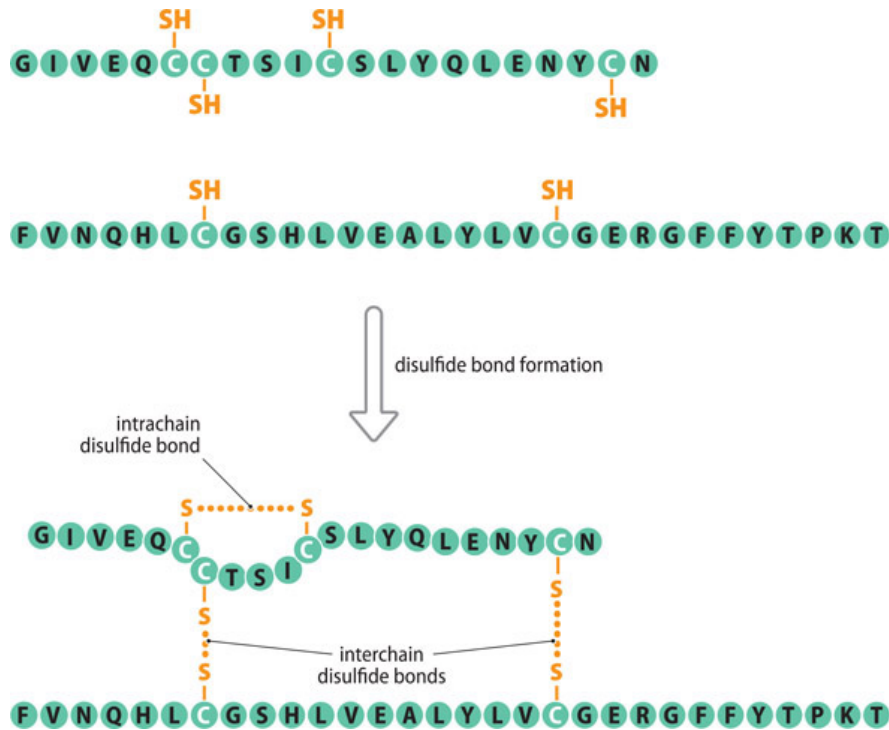


Figure 2.5 Intrachain and interchain disulfide bridges in human insulin. Human insulin is composed of two peptide chains, an A chain with 21 amino acids, and a B chain with 30 amino acids. Disulfide bridges (–S–S–) form by a condensation reaction between the sulfhydryl (–SH) groups on the side chains of cysteine residues. They form between the side chains of cysteines at positions 6 and 11 within the insulin A chain, and also between cysteine side chains on the insulin A and B chains. Note that here all the cysteines participate in disulfide bonding, which is unusual. When disulfide bonding occurs in large proteins, only certain cysteine residues are involved.

Cleavage and transport

The initial polypeptide normally undergoes some type of N-terminal cleavage. Sometimes just the N-terminal methionine is removed. But for proteins secreted from cells, the polypeptide precursor carries an N-terminal leader sequence (signal peptide) that is required to assist the protein to cross the plasma membrane, after which the signal peptide is cleaved at the membrane, releasing the mature protein. (The signal peptide, often 10–30 amino acids in length, carries multiple hydrophobic amino acids.)

Other short internal peptide sequences can act simply as address labels for transporting proteins to the nucleus, mitochondria, plasma membrane, and so on. They are retained in the mature protein.

Binding of multiple polypeptide chains

Proteins are often made of two or more polypeptide subunits. Occasionally, constituent polypeptides are covalently linked with disulfide bridges (as in the case of joining the different chains of immunoglobulins; see [Figure 4.10](#) on page 99). Often, however, constituent polypeptides are held together mainly by noncovalent bonds, including nonpolar interactions and hydrogen bonds. For example, hemoglobins are tetramers, composed of two copies each of two different globin chains that

associate in this way. Collagens provide a good example of very intimate structural association between polypeptides, consisting of three chains (two of one type; one of another) wrapped round each other to form a triple helix.

2.2 RNA GENES AND NONCODING RNA

The majority of our genes are **RNA genes**, genes devoted to making functional **noncoding RNA (ncRNA)** as their end product. (The latest GENCODE data – RELEASE 40, April 2022 – revealed a total of 26 372 human RNA genes and 19 988 protein-coding genes). The vast majority of the RNA genes regulate gene expression in some way, or directly assist in the expression of protein-coding genes, and proteins remain the main functional endpoint in cells.

Like proteins (and mRNA), noncoding RNAs are made as precursors that often undergo enzymatic cleavage to become mature gene expression products. They are also subject to chemical modification: minority bases such as dihydrouridine or pseudouridine and various methylated bases are quite common—see [Figure 2.4A](#) for some examples in a tRNA.

Until quite recently, ncRNAs were largely viewed as having important but rather dull functions. For the most part, they seemed to act as ubiquitous accessory molecules that worked directly or indirectly in protein production. After ribosomal and transfer RNAs, we came to know about various other ubiquitous ncRNAs that mostly work in RNA maturation: spliceosomal small nuclear RNAs (snRNAs); small nucleolar RNAs (snoRNAs) that chemically modify specific bases in rRNA; small Cajal-body RNAs (scaRNAs) that chemically modify spliceosomal snRNA; and certain RNA enzymes (ribozymes) that cleave tRNA and rRNA precursors. All of these types of RNA can be viewed as accessory molecules needed, like rRNA and tRNA, to support protein synthesis in general. In stark contrast to RNA, proteins were viewed as the functionally important endpoints of genetic information, the exciting pacesetters that performed myriad roles in cells.

The view that noncoding RNAs (ncRNAs) are mostly ubiquitous accessory molecules that assist general protein synthesis is no longer tenable. Over the past two decades we have become progressively more aware of the functional diversity of ncRNA and of the many thousands of ncRNA genes in our genome. Multiple new classes of regulatory RNAs have very recently been discovered to be expressed in certain cell types only, or at certain stages of development. Working out what they do has become an exciting area of research.

With hindsight, perhaps we should not be so surprised at the functional diversity of RNA. DNA is simply a self-replicating repository of genetic information, but RNA can serve this function (in the case of RNA viruses) and can also have catalytic functions. In the “RNA world” hypothesis RNA is viewed as the original genetic material and as also being capable of executive functions before DNA and proteins developed. That is possible because, unlike naked double-stranded DNA (which has a comparatively rigid structure), single-stranded RNA has a very flexible structure and can form complex shapes by intramolecular hydrogen bonding, as described below. As will be described in later chapters, the relatively recent understanding of just what RNA does in cells and how it can be manipulated is driving some important advances in medicine. Mutations in certain RNA genes are now known to underlie some genetic disorders and cancers, and RNA therapeutics offers important new approaches to treating disease.

The extraordinary secondary structure and versatility of RNA

The primary structure of nucleic acids and proteins is the sequence of nucleotides or amino acids that defines their identity; however, higher levels of structure determine how they work in cells. Single-stranded RNA molecules are much more flexible than naked double-stranded DNA, and like proteins they have a very high degree of secondary structure where intramolecular hydrogen bonding causes local alterations in structure.

The secondary structure of single-stranded RNA depends on base pairing between complementary sequences on the same RNA strand. Intervening sequences that do not engage in base pairing will loop out, producing stem-loop structures (called hairpins when the loop is short)—see [Figure 2.6](#). Higher-level structures can form when, for example, a sequence within the stem of one loop base pairs with another sequence, and extraordinarily intricate structures can develop. Note that base pairing in RNA includes G–U base pairs as well as more stable A–U and G–C base pairs.

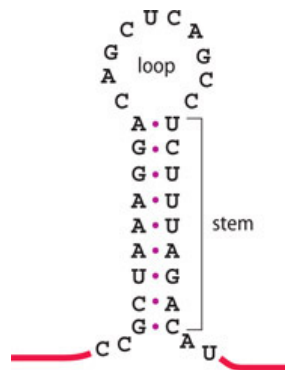


Figure 2.6 A stem-loop structure. This is formed when the RNA folds back on itself so that two short regions can base pair to form the stem while a small intervening sequence loops out. Note that G–U base pairs form in RNA, in addition to G–C and A–U base pairs. Related structures but with shorter stems are important in tRNA structure as shown in [Figure 2.4A](#).

Stem-loop structures in RNA have different functions. As described in [Chapter 6](#), they can serve as recognition elements for binding regulatory proteins, and they are crucially important in determining the overall structure of an RNA that can be important for function.

In general, because of the flexible structure of single-stranded RNA, different RNAs can adopt different shapes according to the base sequence; this enables them to do different jobs, such as working as enzymes. Many different classes of RNA enzyme (ribozyme) are known in nature, and some originated very early in evolution. For example, the catalytic activity of the ribosome (the peptidyltransferase responsible for adding amino acids to the growing poly-peptide chain) is due solely to the large RNA (28S rRNA) present in the large subunit. In recent years RNAs have been found to work in a large variety of roles ([Figure 2.7](#)).

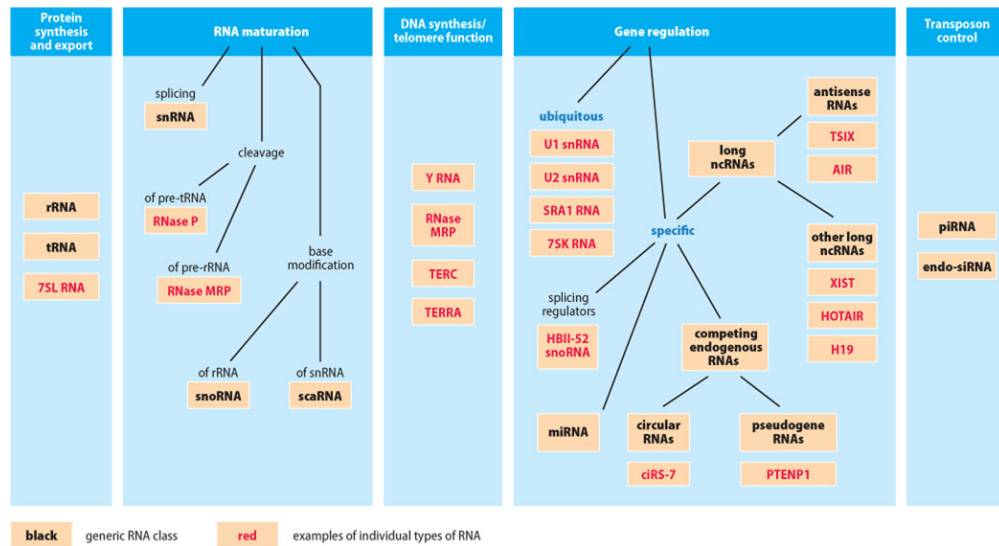


Figure 2.7 The versatility of noncoding RNA. The two panels on the left show ubiquitously expressed RNAs that are important in generally assisting protein production and export, including RNA families that supervise the maturation of other RNAs, notably: small nuclear RNA (snRNA); small nucleolar RNA (snoRNA); and small Cajal-body RNA (scaRNA). The central panel includes RNAs involved in DNA replication (the ribozyme RNase MRP has a crucial role in initiating mtDNA replication, as well as in cleaving pre-rRNA), and developmentally regulated telomere regulators (TERC is the RNA component of telomerase; TERRA is telomere RNA). Diverse classes of noncoding RNA regulate gene expression. In addition to the listed ubiquitous RNAs that have general roles in transcription, many classes of RNA regulate *specific* target genes and are typically restricted in expression. They work at different levels: transcription (such as antisense RNAs), splicing, and translation (notably miRNAs, which bind to certain regulatory sequences in the untranslated regions of target mRNAs). Some RNAs, notably the highly prevalent class of circular RNAs, regulate the interaction between miRNAs and their targets. piRNAs, and to a smaller extent endogenous short interfering RNAs (endo-siRNA), are responsible for silencing transposable elements in germline cells. We describe how RNAs regulate gene expression in detail in [Chapter 6](#).

RNAs that act as specific regulators: from quirky exceptions to the mainstream

The first examples of more specific regulatory RNAs were discovered more than 20 years ago. For a long time they were considered interesting but exceptional cases. They included RNAs working in epigenetic regulation to produce monoallelic gene expression. For most genes both the maternal and paternal allele are normally expressed, but for a few genes it is *normal* that only one of the two parental alleles is expressed. Some of our genes are *imprinted* so that, according to the specific gene, either the maternal allele or the paternal is consistently expressed, while the other allele is silenced. And in women (and female mammals), genes on one of the two X chromosomes, either the maternal or the paternal X chromosome chosen at random, are *normally* silenced (X-chromosome inactivation). We describe the underlying mechanisms in [Chapter 6](#).

We now know that there are many thousands of different RNA genes in our genome. Many of these genes make regulatory ncRNAs that are expressed in certain cell types only, including some large families of long noncoding RNAs and tiny noncoding RNAs.

Long noncoding RNAs

In addition to the very few ribosomal RNAs, there are a very large number of long noncoding RNAs, the great majority of which are associated with chromatin and act as regulators of gene expression. They come in two broad classes. **Antisense RNAs** are transcribed using the *sense* strand of a gene as a template and are not subject to cleavage and RNA splicing. As a result they can be quite large, often many thousands of nucleotides long. They work by binding to the complementary sense RNA produced from the gene, downregulating gene expression.

A second class of long regulatory RNAs are formed from primary transcripts that are typically processed like the primary transcripts of protein-coding genes (and so normally undergo RNA splicing). Many of these RNAs regulate neighboring genes, but some control the expression of genes on other chromosomes. We consider the details of how they work in [Chapter 6](#).

Tiny noncoding RNAs

Thousands of tiny noncoding RNAs (less than 35 nucleotides long) also work in human cells. They include many microRNAs (miRNAs) that are usually 20–22 nucleotides long and are expressed in defined cell types or at specific stages of early development. As described in [Chapter 6](#), a miRNA works by recognizing and binding to defined target regulatory sequences present in specific mRNAs in order to downregulate their expression. MicroRNAs are important in a wide variety of different cellular processes.

Human germ cells also make many thousands of different 26–32-nucleotide Piwi protein-interacting RNAs (piRNAs). The piRNAs work in germ cells to damp down excess activity of **transposons** (mobile DNA elements). Active mobile elements in the human genome can make a copy that migrates to a new location in our genome and can be harmful (by disrupting genes or inappropriately activating some types of cancer gene).

2.3 WORKING OUT THE DETAILS OF OUR GENOME AND WHAT THEY MEAN

The human genome consists of 25 different DNA molecules partitioned between two physically separate genomes, one in the nucleus and one in the mitochondria. In the nucleus there are either 23 or 24 different types of linear DNA molecule (one each for the *different* types of chromosome: 23 in female cells or 24 in male cells). The chromosomal DNA molecules are immensely long (ranging in size from 48 Mb to 249 Mb). In the mitochondria there is just one type of DNA molecule: a comparatively tiny circular DNA just 16.6 kilobases (kb) long, roughly 1/10 000 of the size of an average nuclear DNA molecule. Unlike the chromosomal DNA molecules (each present in only two copies in diploid cells), there are many mitochondrial DNA copies in a cell, and the copy number can vary very significantly according to the type of cell.

In what was a heroic effort at the time, the mitochondrial DNA (often called the mitochondrial genome) was sequenced by a single research team in Cambridge, UK, as far back as 1981. Despite its small size, it is packed with genes. The complexity of the nuclear genome—roughly 200 000 times the size of the mitochondrial genome—posed a much more difficult challenge. That would require an international collaboration between many research teams, as described below.

Working out the nucleotide sequence was only the first step. The next challenge, which is still continuing and may take decades, is to work out the details of how our genome functions and what all the component sequences do.

The Human Genome Project: working out the details of the nuclear genome

For decades, the only available map of the nuclear genome was a low-resolution physical map based on chromosome banding. Chromosomes can be stained with certain dyes, such as Giemsa, to reveal an alternating pattern of dark and light bands for each chromosome, as represented by the image shown in [Figure 2.8](#).

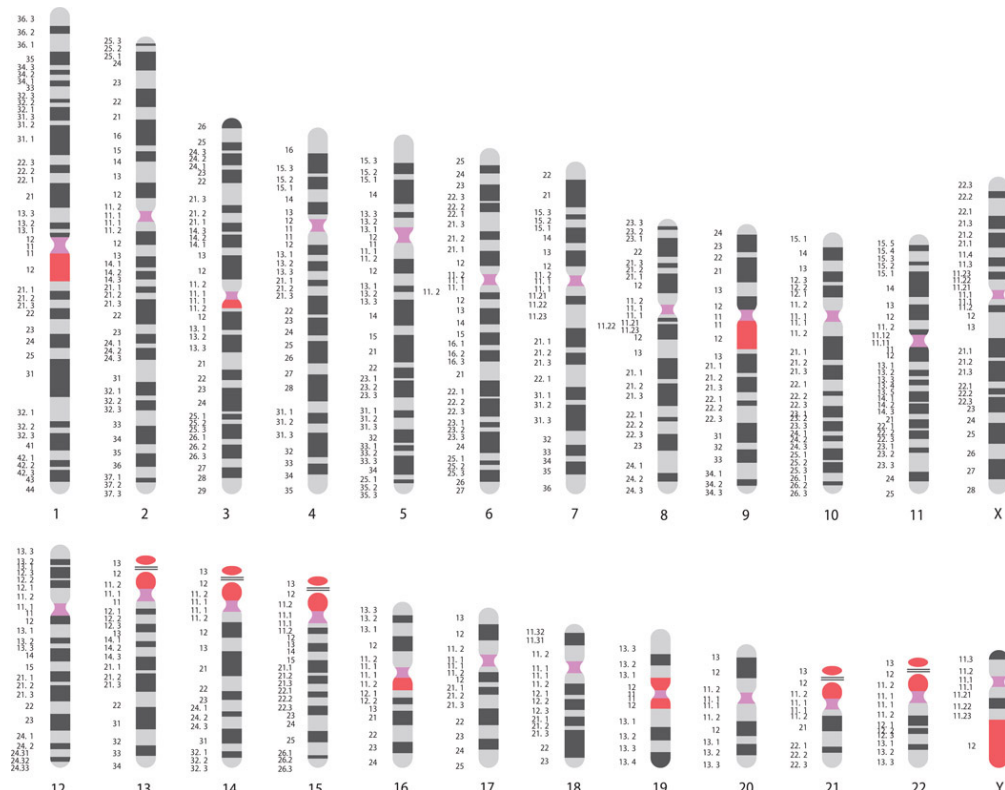


Figure 2.8 Ideogram showing a 550-band Giemsa banding pattern and constitutive heterochromatin within human metaphase chromosomes. Dark bands represent DNA regions where there is a low density of G–C base pairs and a generally low density of exons and genes. Pale bands represent DNA regions where there is a high density of G–C base pairs and a generally high density of exons and genes. Centromeric heterochromatin is illustrated by mauve blocks; non-centromeric and non-telomeric constitutive heterochromatin is shown as bright red blocks. Note the large amounts of non-centromeric heterochromatin on the Y chromosome, the short arms of the acrocentric chromosomes (13, 14, 15, 21, and 22), and on chromosomes 1, 9, and 16. Numbers to the left are the numbers of individual chromosome bands; for the nomenclature of chromosome banding, see [Box 7.2](#) on pages 204–5.

We describe the methodology and terminology of human chromosome banding in [Box 7.2](#). For now, there are two salient points to note. First, the alternating pattern of bands reflects different staining intensities. That in turn reflects differences in chromatin organization along chromosomes (as a result of differences in base composition), and differences in gene and exon density (see the legend to [Figure 2.8](#)). Secondly, the resolution of the map is low—in even a high-resolution chromosome

map the average size of a band is several megabases of DNA. What was needed was a map with a 1 bp resolution, a DNA sequence map.

The principal objective of the international Human Genome Project (HGP) was to obtain a *reference sequence* for our nuclear genome, that is, the aggregated DNA sequences of each of the 24 different human chromosomes. It is necessarily a reference sequence: if we assume 8 billion people on the planet, and without even considering somatic variation, there are at least 16 billion different human genomes (each of us has inherited two different genomes: a maternal genome and a paternal genome that also show differences from the DNA of our parents because of meiosis). The HGP recruited a small number of individuals who donated blood cells to provide genomic DNA for the project. Ultimately, for each chromosome, a map was constructed based on many DNA clones with long inserts that could be ordered as a series of DNA clones with partly overlapping DNA sequences. Finally, the DNAs from selected clones were sequenced and used to build chromosome-wide DNA sequences.

Not all regions of DNA were sequenced: a low priority was given to long regions composed largely of highly repetitive DNA repeats that were technically difficult to sequence. They notably included regions of **heterochromatin**, the parts of the genome where the chromatin of cells is highly condensed throughout the cell cycle—see [Box 2.3](#) for an overview of heterochromatin.

Instead, the focus was on sequencing the gene-rich **euchromatin** component of our genome. A draft sequence for the human nuclear euchromatin genome was published in 2001 (after collation of all the different chromosome DNA sequences). Thereafter, almost complete sequences of the euchromatin region of all 24 nuclear DNA molecules were obtained and published by 2003–2004. The DNA of the euchromatin component was found to represent about 93 % of the nuclear genome. Subsequently, significant components of heterochromatin DNA have been sequenced.

BOX 2.3 AN OVERVIEW OF EUCHROMATIN AND HETEROCHROMATIN

Like other complex genomes, the human nuclear genome is composed of regions of gene-poor **heterochromatin** (where the DNA has a very highly condensed structure that acts as a barrier to transcription factors), and gene-rich **euchromatin** (where the DNA is more open and generally more accessible to transcription factors).

There is some variability in both cases. For euchromatin, the transcriptional activity of euchromatin can vary between cells, notably between cells of different types. In addition to genes expressed in essentially all nucleated cells, many genes in euchromatin show restricted expression; to allow tissue-specific expression, specific regions of euchromatin are induced to be more condensed and transcriptionally inactive in some cells but in other cells the equivalent gene has an open structure accessible to transcription factors.

For heterochromatin, the variability depends on the extent to which the highly condensed structure is consistently maintained or is prone to being altered by an epigenetic mechanism to allow transcription under certain circumstances. As a result, two types of heterochromatin have been distinguished, as listed below.

- **Facultative heterochromatin.** In this case, the same specific regions of DNA can be very highly condensed in some circumstances but be induced to undergo a change in structure to behave as euchromatin in other circumstances. For example, because of an epigenetic

phenomenon known as X-inactivation, the two X chromosomes in a woman are very different: one where most of the X is highly condensed and transcriptionally inactive, and the other where almost all of the X has a relaxed euchromatin structure.

- **Constitutive heterochromatin.** This type of gene-poor chromatin is consistently condensed, at least in somatic cells, and accounts for ~7 % of the nuclear genome. It is found at the centromeres and telomeres, and also over significant other regions of certain autosomes (notably human chromosomes 1, 9, 13–16, 19, 21, 22) and the Y chromosome (see [Figure 2.8](#)). Note: constitutive heterochromatin does contain some genes that are expressed in germline cells. An example is the *DUX4* gene located in telomeric heterochromatin at 4q35.2. If inappropriately expressed in muscle cells, it can cause facioscapulohumeral muscular dystrophy, as described in [Section 6.3](#).

What the sequence didn't tell us and the goal of identifying all functional human DNA sequences

With hindsight, we can appreciate that obtaining the human genome sequence was the easy part. The hard part is to work out what the sequence means. We now know that much of our genome is composed of repetitive sequences. Many genes, for example, are members of families of closely related genes that also often contain *pseudogenes*, inactive copies of functional genes. And a good deal of the genome is composed of highly repetitive noncoding DNA sequences. Because much of the genome sequence does not appear to be critically important, a priority in the “post-genome era” has been to identify all the functionally important DNA sequences and understand how they work.

Protein-coding genes

When our genome sequence was first obtained, it was widely anticipated that new (previously unstudied) genes would be revealed. And indeed, many new human protein-coding genes were soon discovered using computer-based analyses (coding DNA sequences are usually easy to identify because they have long open reading frames and are highly conserved during evolution). Various programs can scan the genome for longer-than-expected open reading frames (as expected of coding DNAs). BLAST programs (described below) can then compare candidate human coding DNAs to seek related sequences in the genomes of other mammals (such as mouse). BLAST programs that rely on translating the putative coding DNA in all six reading frames (three each for the two DNA strands) and then comparing them with similarly translated mammalian genomes, were especially effective (protein sequences are even more highly conserved in evolution than coding DNA—see [Figure 2.9](#) for the example of human and mouse p53 protein sequences).

score	expect	method	identities	positives	gaps
574 bits(1480)	0.0	Compositional matrix adjust.	304/393(77%)	326/393(82%)	6/393(1%)
Query 1	MEEPQSDPSVEPPLSQETFSDLWKLLENNVLSPLPSQAMDDLMLSPDDIEQWFTEDPGP				60
Sbjct 4	MEE QSD S+E PLSQETFS LWKLLP ++L P P MDDL+L P D+E++F GP				57
Query 61	DEAPRMPEAAPVAPAPAAPTPAAPAPAPSWPLSSSVPSQKTYQGSYGFRGLFGLHSGTAK				120
Sbjct 58	EA R+ A P P P APAPA WPLSS VPSQKTYQG+YGF LGFL SGTAK				117
Query 121	SVTCTYSPALNKMFCQLAKTQCPVQLWVDSTPPPGTRVRAMAIYKQSQHMTEVVRRCPHHE				180
Sbjct 118	SV CTYSP LNK+FCQLAKTQCPVQLWV +TPP G+RVRAMAIYK+SQHMTEVVRRCPHHE				177
Query 181	RCSDSDGLAPPQHLIRVEGNLRVEYLDRNTFRHSVVVPYEPPEVGSDCCTTIHNYMCNS				240
Sbjct 178	RCSD DGLAPPQHLIRVEGNL EYL+DR TFRHSVVVPYEPPE GS+ TTIHY YMCNS				237
Query 241	SCMGMNRRPILTIITLEDSSGNLLGRNSFEVRVCACPRDRRTEENLRKKGEPHHELP				300
Sbjct 238	SCMGMNRRPILTIITLEDSSGNLLGR+SFEVRVCACPRDRRTEEN RKK ELP				297
Query 301	PGSTKRALPNNTSSSPQKKKPLDGEYFTLQIRGRERFEMFRELNEALELKDAQAKEPG				360
Sbjct 298	PGS KRALP TS+SP KKKPLDGEYFTL+IRGR+RFEMFRELNEALELKDA A +E G				357
Query 361	GSAHSSHLKSKKGQSTSRHKKLMFKTEGPDS		393		
Sbjct 358	SRAHSS+LK+KKGQSTSRHKK M K GPDS		390		

Figure 2.9 Unlike RNA sequences, protein sequences are often highly conserved in evolution: the example of human and mouse p53 proteins. The alignment was generated by the BLASTP program that compared a query sequence, the human p53 protein, against the mouse p53 sequence, the subject (Sbjct) sequence. The alignment is slightly skewed because the mouse p53 sequence has an extra three amino acids at its N-terminus, but lacks some amino acids found in the human p53 sequence, such as the EDP at positions 56–58 in human p53). The middle lines between the query and subject lines show identical residues (at 302 out of the 393 matched positions). The + symbol in the middle lines denote functionally similar amino acids; they occur here at 22 positions, which when added to 304 identities gives a total of 326 positive matches out of 393.

RNA genes and regulatory elements

The vast majority of known human RNA genes and regulatory sequences (such as promoters, enhancers, and so on) were initially not identified from the genome sequence. RNA genes lack open reading frames, some of them are tiny sequences, and unlike the polypeptide sequences of proteins, RNA sequences are often poorly conserved during evolution (in RNA molecules the shape is important, but the sequence less so); but nucleotides are important in maintaining the RNA shape or in ensuring correct binding to interacting molecules and so are often well-conserved. Regulatory elements are clusters of tiny sequences that are often not as well-conserved as coding DNA, making them often difficult to identify by computer programs in the past. Our knowledge of human regulatory elements was therefore also limited at the time when the human genome sequence was reported in 2003.

For follow-up projects to hunt down all functional human DNA sequences, the major priority became to identify and catalog all RNA transcripts and also all regulatory sequences (working at either the DNA level, or at the RNA level). The projects used genome-wide transcription analyses,

evolutionary sequence comparisons and functional assays, all underpinned by bioinformatic analyses. The most prominent international study—the ENCODE (**Encyclopedia of DNA Elements**) Project reported its findings in 2012, and one of its most important findings was that RNA transcription is pervasive. And, as detailed in [Section 2.5](#), it has become clear that there are significantly more RNA genes in the human genome than protein-coding genes.

2.4 A QUICK TOUR OF SOME ELECTRONIC RESOURCES USED TO INTERROGATE THE HUMAN GENOME SEQUENCE AND GENE PRODUCTS

A wide variety of databases and computer programs currently provide a wealth of information on the human genome, human genes, and gene products ([Table 2.3](#)). Genome browsers help users navigate a sequenced genome by programs that employ graphical user interfaces to portray genome information for selected chromosomes and subchromosomal regions. The characteristics (genes, exons, transcripts, and so on) of a selected human chromosome or chromosome region can be tracked, moving from large scale to nucleotide scale, with click-over facilities to identify the characteristics and download the sequences of genes and associated exons, RNAs, and proteins. The principal browsers are listed in [Table 2.3](#).

TABLE 2.3 SOME OF THE PRINCIPAL ELECTRONIC RESOURCES FOR INTERROGATING THE HUMAN GENOME, HUMAN GENES AND GENE PRODUCTS

Resource use	Popular resources	Website address
Gateways to multiple electronic resources	Human Genome Resources HGNC portal NIH National Library of Medicine	http://www.ncbi.nlm.nih.gov/genome/guide/human/ http://www.genenames.org/ http://www.ncbi.nlm.nih.gov/gquery/
Reference nucleotide and protein sequences	RefSeq (for mRNAs, ncRNAs and proteins) RefSeqGene (for genes)	http://www.ncbi.nlm.nih.gov/refseq/ http://www.ncbi.nlm.nih.gov/refseq/rsg/
Identifying related	BLAST programs	http://blast.ncbi.nlm.nih.gov/Blast.cgi https://genome.ucsc.edu/cgi-bin/hgBlat?command=start

For further descriptions of individual resources, see main text. HGNC: the HUGO (human genome organization) Gene Nomenclature Committee; NCBI: the US National Center for Biotechnology Information; UCSC: University of California at Santa Cruz.

Resource use	Popular resources	Website address
sequences and homologs	BLAT HomoloGene HCOP (orthology predictions)	http://www.ncbi.nlm.nih.gov/homologene http://www.genenames.org/tools/hcop
Protein sequence analysis	UniProt InterPro	http://www.uniprot.org https://www.ebi.ac.uk/interpro/
Genome browsers	Ensembl UCSC Genome Browser	http://www.ensembl.org/ https://genome.ucsc.edu/
Genome annotation	GENCODE NCBI Annotation (release 109)	https://www.encodegenes.org/ https://www.ncbi.nlm.nih.gov/genome/annotation_euk/Homo_sapiens/109/

For further descriptions of individual resources, see main text. HGNC: the HUGO (human genome organization) Gene Nomenclature Committee; NCBI: the US National Center for Biotechnology Information; UCSC: University of California at Santa Cruz.

Specialized electronic gateways with electronic links to numerous such electronic resources are especially useful. If the starting point of interest is a gene, gene product or genetic disorder, the HGNC portal organized by the HUGO Gene Nomenclature Committee has a simple, user-friendly architecture. A comprehensive coverage of electronic resources is also available through the Human Genome Resources section of the US National Center for Biotechnology Information (NCBI)—see [Table 2.3](#)).

Gene nomenclature and the HGNC gateway

Gene symbols for human genes are allocated by the HUGO Gene Nomenclature Committee (HGNC). They typically have between three and seven characters, and are displayed in italicized uppercase, such as *HBB* (hemoglobin beta subunit), *CFTR* (cystic fibrosis transmembrane regulator), or *RN7SL* (7SL RNA). Mitochondrial genes are prefixed by *MT-* (for example, *MT-RNR1* is the mitochondrial 12S rRNA gene). Pseudogenes, naturally occurring but defective copies of a normally functional gene, usually have a symbol that is the same as a related functional gene, but followed by a P, or by a P followed by a number (for example, *CFTRP3* is one of three pseudogenes related to the *CFTR* gene). Note that the format for gene symbols for other species is often different. For example, the mouse and rat orthologs of the human *CFTR* gene are each given the symbol *Cftr*.

The HGNC portal at www.genenames.org has links to many databases and browsers. It can be interrogated by using as a query a gene symbol, if known, or descriptive text for an associated gene product or disease. For example, entering hemoglobin as the search term returns a list of genes encoding the different subunits of all forms of human hemoglobin, and entering cystic fibrosis yields related results including the cystic fibrosis gene *CFTR*. Selecting a gene symbol such as *CFTR* opens

the way to an extraordinary amount of related information through linked databases and genome browsers ([Figure 2.10](#)). It can also be used to identify groups of related genes under the section entitled Gene groups.

The screenshot displays the HGNC portal interface for the CFTR gene. At the top, there is a search bar and navigation links. The main content is divided into several sections:

- Symbol report for CFTR:** A central section with a yellow border containing basic information such as the approved symbol (CFTR), approved name (CF transmembrane conductance regulator), locus type (gene with protein product), HGNC ID (HGNC:1884), symbol status (Approved), previous symbols (CF, ABCC7), previous names (cystic fibrosis transmembrane conductance regulator, ATP-binding cassette (sub-family C, member 7)), alias symbols (MRP7, ABC35, TNR-CFTR, dJ760C5.1, CFTR/MRP), alias names (ATP-binding cassette sub-family C, member 7), chromosomal location (7q31.2), and gene groups (Chloride channels, ATP-gated CFTR, ATP binding cassette subfamily C).
- Gene resources:** Links to Ensembl (ENSG000001626), NCBI Gene (1080), and UCSC (uc003yqj.4).
- Nucleotide resources:** Links to INSDC (M28668), RefSeq (NM_000492), and CCDS (CCDS5773).
- Protein resources:** Links to UniProt/Swiss-Prot (P13569) and InterPro (IP089, PDBRef, Reactome).
- Orthologs from selected species:** A list of orthologous genes in other species, including Bos taurus, Canis familiaris, Equus caballus, Felis catus, Macaca mulatta, Mus musculus, Rattus norvegicus, and Sus scrofa.
- Specialist resources:** Links to IUPHAR/BPS Guide to PHARMACOLOGY (707).
- Clinical resources:** Links to OMIM (602421), LRG (LRG_663), and LSDB (Cystic Fibrosis).
- Genetics Home Reference:** Links to DECIPHER, ClinGen, and Genetic Testing Registry.
- Other resources:** Links to AmiGO, QuickGO, BioGPS, GeneCards, Monarch, and WikiGenes.
- References:** A list of scientific references, including Rommens JM et al. (Science 1989) and PMID: 2772657.

Figure 2.10 Getting a wealth of information on a selected human gene starting from the HGNC portal at www.genenames.org. The figure shows the output displayed after using *CFTR* (the human cystic fibrosis gene) as a search query. Under the Report tab, the box at top left outlined in yellow shows basic items of information maintained by HGNC, including Gene groups at bottom. The other sections, with titles highlighted in gray, provide links to numerous external databases, with information on associated gene, RNA and protein sequences, databases with clinical information and mutant sequences, and information on closely related orthologs in other species. Pressing on the second tab at top left, marked, HCOP homology predictions, identifies related sequences in a very wide range of organisms.

Databases storing nucleotide and protein sequences

In the example of [Figure 2.10](#), the link for “Nucleotide resources” begins with *general* nucleotide sequence databases that are part of the International Nucleotide Sequence Database Collaboration (INSDC), comprising the European Nucleotide Archive (ENA), GenBank and the DNA Database of Japan (DDBJ). The ID number for the *CFTR* mRNA sequence in these databases is M28668, and the 6129-nucleotide sequence is presented. Under “Protein resources” is the useful UniProt/Swiss-Prot database where extensive information on the CFTR protein (ID number: P13569) can be found.

The problem with general nucleotide and protein sequence databases is that they contain *redundant* sequences (sometimes with many entries for the same sequence from independent DNA clones, some having partial sequences, some full length). To make it much easier to find a complete sequence of interest, the RefSeq databases were established at the NCBI to provide a comprehensive, nonredundant, and well-annotated set of reference sequences for different species. The standard

RefSeq database has non-redundant reference sequences for mRNAs (ID numbers prefixed by NM_); noncoding RNAs (ID numbers prefixed by NR_); and proteins inferred from a mRNA (ID numbers prefixed by NP_). The separate RefSeqGene database stores gene reference sequences.

Finding related nucleotide and protein sequences

Sequences evolutionarily related to a query nucleic acid or protein sequence are most often identified using one of the BLAST programs hosted at the NCBI. BLASTN uses a nucleotide query to compare with other nucleotide sequences, and BLASTP compares a protein sequence against other protein sequences (see [Figure 2.9](#) for an example output). TBLASTN is powerful because it can compare a protein sequence against all possible translations of all sequences in nucleotide databases. Significant homology may be apparent across the length of the query sequence or be limited to a region, such as a conserved protein domain or some other shared sequence.

The BLAT program allows rapid sequence searching across whole genomes. Query nucleotide sequences (up to a total of 25 000 nucleotides) and query protein sequences (up to 10 000 amino acids) can be entered to search for homologous sequences across the human genome, or the genome of any of multiple model organisms. The output lists significant hits, with given chromosome coordinates and sequence alignments.

The sequence comparisons include searching to find equivalent sequences (*orthologs*) in other species, using programs such as Homologene and HCOP. The HGNC reports for genes show multiple orthologs from other species, such as shown in the *CFTR* gene report in [Figure 2.10](#).

Links to clinical databases

As shown in [Figure 2.10](#) HGNC reports also provide links to a variety of clinical databases under the section entitled “Clinical resources”. They include the On-line Mendelian Inheritance in Man (OMIM), Genetic Home Reference and various databases recording disease-associated mutations, including COSMIC (focusing on cancer) and ClinVar (which documents relationships between human DNA variants and phenotypes).

2.5 THE ORGANIZATION AND EVOLUTION OF THE HUMAN GENOME

Our genome has some curious characteristics, such as: division into a massive complex genome in the nucleus and a tiny simple genome in our mitochondria; a huge number of repetitive DNA sequences; a profusion of pseudogenes; a vast range in gene sizes; and a small proportion of functionally important nucleotides. In order to make sense of how our genome is organized, we begin this section by examining important evolutionary forces that shaped the genome.

A brief overview of the evolutionary mechanisms that shaped our genome

The widely accepted endosymbiont hypothesis proposes that our two physically separated genomes, the nuclear and mitochondrial genomes, originated when a type of aerobic prokaryotic cell was endocytosed (engulfed) by an anaerobic eukaryotic precursor cell, at a distant time in the past when oxygen started to accumulate in significant quantities in the Earth’s atmosphere. Over a long period,

much of the original prokaryote genome was excised, causing a large decrease in its size, and this much reduced prokaryotic genome gave rise to the mitochondrial genome. DNA fragments excised from the aerobic prokaryote cell were transferred to the genome of the engulfing cell. The latter genome increased in size as a result, but then progressively went on to undergo further very significant changes in both size and form during evolution, developing into our nuclear genome.

The theory explains why mitochondria have their own ribosomes and their own protein-synthesizing machinery and why our mitochondrial DNA closely resembles in form a reduced (stripped-down) bacterial genome. But how did the genome of the engulfing cell become so large and complex? That largely happened by a series of different mechanisms that copied existing DNA sequences and added them to the genome. After some considerable time, the copies can acquire mutations that make them different from the parent sequences; ultimately new genes, new exons, and so on can be formed in this way.

Whole genome duplication is a quick way of increasing genome size, and comparative genomics has provided very strong evidence that this mechanism has occurred from time to time in different evolutionary lineages. There is compelling evidence, for example, that whole genome duplication occurred in the early evolution of our chordate ancestors just before the appearance of vertebrates.

Additional duplications of moderately large to small regions of DNA occur comparatively frequently on an evolutionary timescale, and they also give rise ultimately to novel genes, novel exons, and so on. They occur by copying mechanisms that work at the level of genomic DNA, or at the RNA level by using *reverse transcriptases* (RNA-dependent DNA polymerases) to make DNA copies of RNA transcripts that then insert into the genome.

Comparative genomics can reveal when new genes were formed in evolution by screening the genomes of multiple species to identify those that possess versions of the same gene (the different versions, such as the human *CFTR* gene and the mouse *Cftr* gene are said to be *orthologs*). In the examples given in [Section 2.3](#), the gene encoding the p53 tumor suppressor first appeared at the time of bony vertebrates (it is not present in invertebrates or in non-bony vertebrates), but others have evolved very recently (the *TCP10L* gene, for example, is found only in Old World monkeys, apes, and humans).

The duplication mechanisms that led to a progressive increase in genome size and to the formation of novel genes and other novel functional sequences are, to a limited extent, offset by occasional evolutionary loss of functional DNA sequences, including genes. After whole genome duplication, for example, many of the new gene copies pick up mutations that cause them to be silenced, and they are eventually lost. And the Y chromosome is believed to have shed many genes over hundreds of million years. Gene loss can happen on a smaller scale, too.

Gene birth and gene loss are comparatively infrequent events, and even though humans and mice diverged from a common evolutionary ancestor about 80 million years ago, our gene repertoire is extremely similar to that of the mouse. However, *cis*-acting regulatory sequences such as enhancers often evolve rapidly, and although we have much the same set of genes as a mouse, they are often expressed in different ways. Differential gene regulation is a primary explanation for the differences between species that are evolutionarily closely related.

How much of our genome is functionally significant?

We will end this chapter by looking in some detail at different facets of our genome. But first, let us step back and take a broad look at its design. Here is one perspective attributed to the evolutionary

biologist David Penny: “I would be quite proud to have served on the committee that designed the [*Escherichia coli*] genome. There is, however, no way that I would admit to serving on a committee that designed the human genome. Not even a university committee could botch something that badly.”

The *E. coli* genome is a sleek genome, packed with gene sequences (90 % of the genome is made up of coding DNA sequences). By contrast, like the genome of many complex organisms, our genome seems rather flabby: coding DNA accounts for just 1.2 % of the genome, and the majority of the DNA is made up of highly repetitive noncoding DNA sequences of questionable functional value (see [Figure 2.11](#)). For decades much of our genome had largely been regarded as “junk DNA,” an idea supported by the lack of correlation between genome size and organism complexity (the genome of the diploid onion, for example, is more than five times the size of our genome).

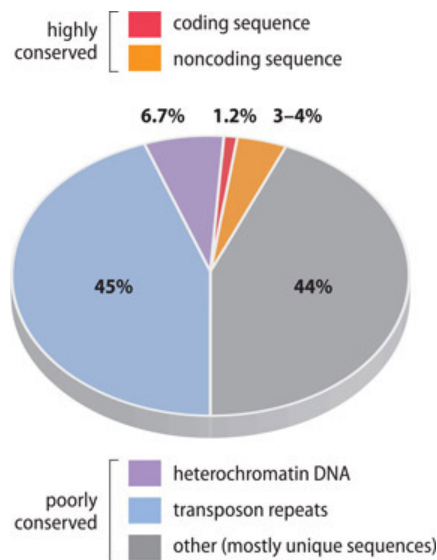


Figure 2.11 Human genome organization: extent of evolutionary conservation and repetitive sequences. Only just over 1.2 % of our genome is coding DNA that specifies protein sequences, and another roughly 3–4 % or so of our genome is made up of noncoding DNA sequences that have been highly or moderately conserved during evolution (as determined by looking at nucleotide substitutions in mammalian sequence alignments). Some of this conserved sequence is present in multiple copies and includes different types of repeated genes (gene families). The 6.7 % of our genome that is located in constitutive heterochromatin is very largely made up of poorly conserved repetitive DNA sequences that include sequences responsible for centromere function. Transposon repeats include highly repetitive interspersed repeats such as Alu and LINE-1 repeats; it is thought that during evolution some of these repeats contributed to the formation of new exons and regulatory elements, including some long ncRNAs.

The ENCODE Project seemed to offer a different perspective. The data suggested that much of the genome was transcribed, often from both DNA strands in specific regions, and 80.4 % of the human genome was claimed to participate in at least one RNA-associated or chromatin-structure-associated event in at least one cell type. However, the possible conclusion that much of our genome might be functionally significant has been strongly resisted by evolutionary biologists. Part of the difficulty in interpreting the ENCODE data is that much of the 80.4 % figure comes from the observed representation of RNA transcripts, but many RNAs are produced at very low levels and their functional status is uncertain.

Sequence conservation due to selection and estimating functional constraint

Although mutation can potentially change any nucleotide in DNA, there must be constraints on changing a functionally important sequence during evolution. If we take a protein-coding sequence, for example, even a single amino acid change might quite often result in loss of the protein's function or produce an aberrant protein that might contribute to disease.

Mutations that result in adverse changes to the phenotype are effectively removed from populations over generations. That happens by a type of Darwinian natural selection called **purifying selection**. Compared with normal alleles, the mutant allele is not efficiently transmitted to subsequent generations (some people that carry it will not reproduce as well as people with normal alleles). Over long periods of evolutionary time, therefore, protein-coding sequences are constrained by the need to maintain function (*functional constraint*), and they change slowly in comparison with most other DNA sequences.

The amount of the human genome that is highly conserved (under functional constraint as a result of purifying selection) was initially estimated to be about 5 % (see [Figure 2.11](#)). However, that figure came from comparisons with many different mammals. Additional functional constraint became apparent in the 1000 Genomes Project, in which multiple human genomes were compared.

Functionally important DNA sequences that are rapidly evolving might not be seen to be conserved (and therefore constrained) when comparisons are made between a broad range of mammalian species. Although some rare non-coding DNA sequences are very strongly conserved—sometimes more than coding DNA—it is clear that many regulatory DNA sequences and RNA genes are rapidly evolving. They do, however, make important contributions to functional constraint in narrower evolutionary lineages, including primate and then human lineages in our case.

Taking that into account, the proportion of our genome that is subject to purifying selection is now thought to be of the order of at most 10 %; on that basis, most of our genome does not seem to have a valuable function. However, there is evolutionary value in having a large genome with surplus non-functional DNA because the non-functional DNA component can, through successive mutations, provide new functional sequences in the future, as described below.

The mitochondrial genome: economical usage but limited autonomy

David Penny's comment about the human genome in general (see above) certainly does not apply to the mitochondrial genome. Our mitochondrial DNA closely resembles in form a reduced (stripped-down) bacterial genome. Mitochondria, like chloroplasts, have their own ribosomes and their own protein-synthesizing machinery and almost certainly originated when a prokaryotic cell was engulfed by an anaerobic eukaryote precursor cell, allowing aerobic eukaryotes to develop.

The human mitochondrial genome has a total of 37 genes. Of these, 24 are RNA genes that make all the RNA required for protein synthesis in the mitochondrial ribosomes: the two rRNAs and 22 tRNAs ([Figure 2.12](#)). The remaining genes make 13 out of the 89 polypeptide subunits of the oxidative phosphorylation system (OXPHOS). (The other 76 OXPHOS subunits, like all other mitochondrial proteins, are encoded by nuclear genes and synthesized on cytoplasmic ribosomes before being imported into the mitochondria.)

[Section 7.1](#); the differences between the nuclear and mitochondrial genetic codes are given in [Figure 7.2](#).

None of the mitochondrial genes is interrupted by introns, and the genome is a model of economical DNA usage: close to 95 % of the genome (all except 1 kb out of the 16.6 kb of DNA) makes functional gene products. Note that transcription of the two DNA strands occurs using one promoter each to generate large multigenic transcripts that are subsequently cleaved to generate individual mRNAs and ncRNAs.

Gene distribution in the human genome

More than 90 % of the mitochondrial DNA sequence directly specifies a protein or functional ncRNA, and there is one intronless gene every 450 bp on average. The nuclear genome is very different: the gene density is much lower, genes are frequently interrupted by introns as listed in [Table 2.1](#) above, and a sizable fraction of the genome is made up of repetitive DNA, notably highly repetitive non-coding DNA.

Close to 7 % of the nuclear genome is located in constitutive **heterochromatin** that remains highly condensed throughout the cell cycle (the chromosomal locations of human heterochromatin are given in [Figure 2.8](#) above). Although almost devoid of genes, the constitutive heterochromatin of somatic cells does contain a tiny number of genes that are inactive in somatic tissues but expressed in germ cells (where chromatin has a more open structure). The remaining 93 % of our genome is accommodated in less-condensed, gene-rich **euchromatin**.

Protein-coding genes have been comparatively easy to identify. GENCODE data estimate close to 20 000 human protein-coding genes, but the number can never be exact because of variation between individuals (and sometimes between maternally and paternally inherited genomes) in copy number for some repeated genes).

Identifying RNA genes is much more problematic. Three characteristics make it difficult to do so: the lack of a sizable open reading frame, lack of evolutionary sequence conservation, and in some cases extremely small sizes (making them easily overlooked). Establishing the functional significance of many transcribed noncoding DNA sequences can therefore be difficult. The most recent GENCODE data identify more RNA genes than protein-coding genes ([Table 2.4](#)), but the functional status of some putative RNA genes remains unproven.

TABLE 2.4 A SNAPSHOT OF THE NUMBERS OF HUMAN GENES AND PSEUDOGENES LISTED BY GENCODE
VERSION 40 (RELEASED IN APRIL 2022)

Class	Number
PROTEIN-CODING GENES	19988
RNA GENES	26372
making long ncRNA	18805
making short ncRNA	7567
PSEUDOGENES	14774
processed	10661
unprocessed	3566

Obtained at <http://genecodegenes.org/human/stats.html>

Class	Number
other	547

Obtained at <http://genecodegenes.org/human/stats.html>

Genome sequencing showed that gene and exon density in the euchromatic regions can vary enormously. Some chromosomes are gene-rich, such as chromosomes 19 and 22; others are gene-poor, notably the Y chromosome (which makes only 31 different proteins that mostly function in male determination). Within a chromosome, the pattern of alternating dark and light bands reflects different base compositions, and also differences in gene and exon density (as described in the legend to [Figure 2.8](#)).

The extent of repetitive DNA in the human genome

Our large nuclear genome is the outcome of periodic changes that have occurred over very long timescales during evolution, including rare whole genome duplication and intermittent chromosome rearrangements, localized DNA duplications, DNA duplication followed by dispersal to other genome locations, and loss of DNA sequences. The net result has been a gradual increase in DNA content and gene number through evolution.

Previous whole genome duplications were followed by a gradual loss of most of the duplicated sequences; accordingly, unique sequences make up quite a sizable fraction of our genome. Nevertheless, highly repetitive DNA sequences originating from transposons (mobile DNA elements; see below) plus the repetitive DNA families found in heterochromatin account for more than 50 % of our genome (described in [Figure 2.11](#) above).

In addition to transposon-derived repeats, our euchromatin contains clear evidence of localized DNA duplications. In some cases, the repeats have diverged considerably in sequence—the duplication occurred many tens or hundreds of millions of years ago in evolution, and subsequent mutations have led to divergence in sequence between the repeats. But other localized duplications are quite striking because they have occurred very recently in evolution. For example, about 5 % of our euchromatin DNA consists of neighboring duplicated segments that are more than 1 kb long and show more than 90 % sequence identity. Many of these **segmental duplications** are primate-specific, and they are particularly common close to telomeres and centromeres (about 40 % occur in subtelomeric regions; about 33 % occur at pericentromeric regions).

There is a significant amount of repetitive coding DNA within genes and also many repeated genes. Within a gene, repetitive coding DNA may be found in an individual exon (usually as a tandem duplication of one or more nucleotides), or one or more exons has been repeated ([Figure 2.13A](#) and [2.13B](#)).

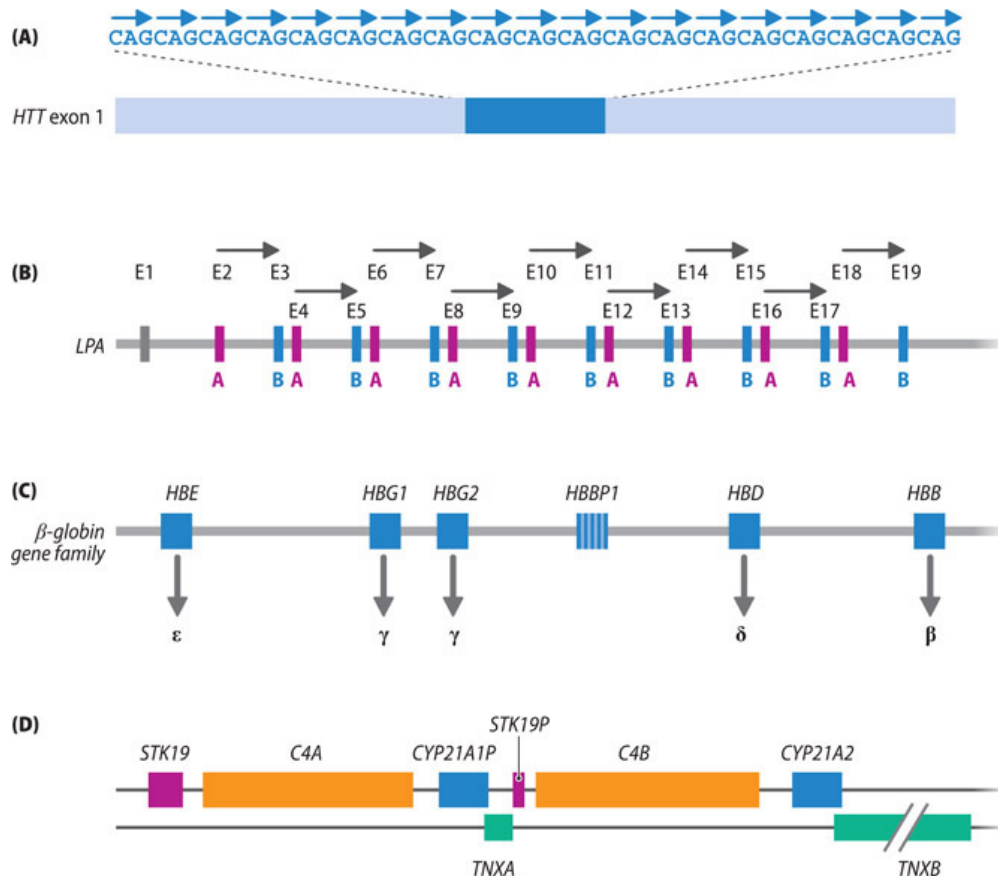


Figure 2.13 Examples of tandemly repetitive coding DNA and clustered gene families. (A) Normal alleles of the *HTT* huntingtin gene have an array of tandemly repeated CAG codons in exon 1 that varies in number up to 35 repeats (having more than 36 repeats results in Huntington disease). (B) The *LPA* gene encodes lipoprotein Lp(a), a protein with multiple kringle domains that are each 114 amino acids long and extremely similar in sequence. Each kringle repeat is encoded by a tandemly repeated pair of exons (here labeled A and B) that encode two adjoining parts of the kringle domain and that can be present in different copy numbers. The example shown here has nine pairs of A and B exons, starting with exons 2 and 3 (E2 and E3) and continuing through to exons 18 and 19 (E18 and E19). (C) The β -globin gene family has six highly related genes. Four genes make alternative globins used in hemoglobin, but the status of *HBD* is uncertain (q-globin is never incorporated into a hemoglobin protein) and *HBBP1* is a pseudogene. (D) The HLA (human leukocyte antigen) region of normal individuals has a tandemly duplicated unit containing four gene sequences, encoding serine threonine kinase 19, complement C4, cytochrome P450 21-hydroxylase, and tenascin-X (transcribed from the opposite strand). Subsequently, three of the genes became pseudogenes, having acquired inactivating mutations (*CYP21A1P*) or having also lost significant amounts of sequence (*STK19P* and *TNXA*).

On a larger scale, the repeated unit can consist of a whole gene or occasionally two or more unrelated genes (Figure 2.13C and 2.13D). The resulting multi-gene families contain two or more genes that produce related or even identical gene products. The more recently duplicated genes are readily apparent because they make very closely related or identical products. Genes originating from more evolutionarily ancient duplications make more distantly related products.

The organization of gene families

Different classes of multigene families exist in the human genome, and the number of genes in a gene family can range from two to many hundreds ([Table 2.5](#)). Some are clustered genes confined to one subchromosomal region. They typically arise by tandem gene duplication events in which chromatids first pair up unequally so that they become aligned out of register over short regions. The mispaired chromatids then exchange segments at a common breakpoint ([Figure 2.14](#)).

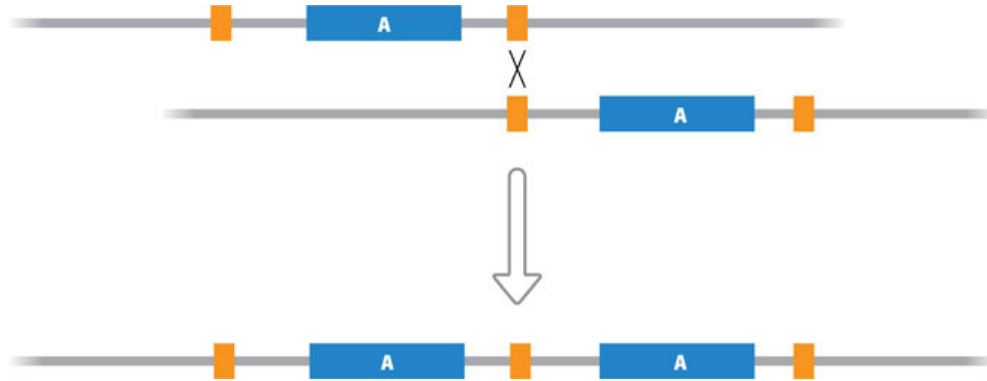


Figure 2.14 Tandem gene duplication. Gene duplication can occur after sister chromatids or non-sister chromatids of homologous chromosomes mispair so that they are slightly out of alignment. In this example, the result is that gene A on one chromatid is out of register with gene A on the opposing chromatid. Subsequent breakage of both chromatids at the position marked by the cross (X) and swapping of fragments between chromatids can result in a chromatid that has two copies of gene A (the left fragment of the top chromatid joins to the right fragment of the bottom chromatid). The exchange may be facilitated by base pairing between highly related noncoding repetitive DNA sequences (orange boxes).

TABLE 2.5 EXAMPLES OF MULTIGENE FAMILIES IN THE HUMAN GENOME

Gene family	Copy number	Genome organization
β -Globin	6 (includes one pseudogene)	clustered within 50 kb at chromosome 11 p15 (Figure 2.13C)
Class I human leukocyte antigen (HLA)	17 (includes many pseudogenes and gene fragments)	clustered over 1.8Mb at 6p21.3
Neurofibromatosis type I	1 functional gene; 8 unprocessed pseudogenes	functional gene, NF1, at 17q11.2; pseudogenes dispersed over pericentromeric regions on several other chromosomes
Ferritin heavy chain	1 functional gene; 27 processed pseudogenes ^a	functional gene, FTH1, at 11q13; pseudogenes dispersed over multiple chromosome locations
U6 snRNA	49 genes; 800 processed pseudogenes ^a	scattered on many chromosomes

^a A processed pseudogene (retro pseudogene) is a copy of a gene transcript and so has counterparts of exon sequences only. By contrast, an unprocessed pseudogene (which is a copy of the genomic sequence) also has sequences corresponding to introns and upstream promoters); see [Box 2.4](#).

The α - and β -globin gene clusters on chromosomes 16 and 11, respectively, arose by a series of tandem gene duplications. Some of the duplicated genes (such as the *HBA1* and *HBA2* genes, which

make identical α -globins, or the *HBG1* and *HBG2* genes, which make γ -globins that differ at a single amino acid) are the outcome of very recent gene duplication. Other globin genes are clearly related to each other but have more divergent sequences. The different globin classes have slightly different properties, an advantage conferred by gene duplication (see below).

Other gene families are distributed over two or more different chromosomal regions. In some cases they originally arose from duplicated genes in gene clusters that were then separated by chromosome rearrangements. In other cases, cellular reverse transcriptases were used to make natural complementary DNA (cDNA) copies of the mRNA produced by a gene, and the cDNA copies were able to insert successfully elsewhere in the genome of germline cells. Because the cDNA copies of mRNA lacked promoters, as well as intron sequences, they very frequently degenerated into nonfunctional **pseudogenes**. For some genes, however, cDNA copies have integrated during evolution at other chromosomal locations to produce functional genes (**Box 2.4**).

BOX 2.4 PSEUDOGENES

One common consequence of gene duplication is that a gene copy diverges in sequence but instead of producing a variant gene product it gradually accumulates deleterious mutations to become a **pseudogene**. A pseudogene copy of a protein-coding gene can usually be detected by identifying deleterious mutations in the sequence that corresponds to the coding DNA sequence; RNA pseudogenes are less easy to identify as pseudogenes. GENCODE lists more than 14 000 pseudogenes in the human genome ([Table 2.4](#)). According to their origin, pseudogenes can be divided into two major classes: unprocessed pseudogenes and processed pseudogenes. Despite their name, some pseudogenes are known to have functionally important roles, as described below.

UNPROCESSED PSEUDOGENES

An unprocessed pseudogene arises from a gene copy made at the level of genomic DNA, for example after tandem gene duplication ([Figure 2.14](#)). Initially, the copied gene would have copies of all exons and introns of the parental gene plus neighboring regulatory sequences including any upstream promoter. Acquisition of deleterious mutations could lead to gene inactivation ('silencing') and subsequent decay, and sometimes instability (substantial amounts of the DNA sequence can be lost, leaving just a fragment of the parental gene). Unprocessed pseudogenes are typically found in the immediate chromosomal vicinity of the parental functional gene (see the example of *HBBP1* in [Figure 2.13C](#)). Sometimes, however, they are transposed to other locations because of instability of pericentromeric or subtelomeric regions. For example, the *NF1* neurofibromatosis type I gene is located at 17q11.2, and eight highly related unprocessed *NF1* pseudogenes are found (with one exception) in pericentromeric regions of other chromosomes as a result of comparatively frequent interchromosomal exchanges at pericentromeric regions.

PROCESSED PSEUDOGENES (RETROPSEUDOGENES)

A processed pseudogene arises by reverse transcription of an RNA from a parental gene followed by random integration of the resulting cDNA copy elsewhere in the genome ([Figure 1](#)). The cDNA copy lacks any sequences corresponding to introns and regulatory sequences occurring outside exons, such as upstream promoters. Integration of a cDNA copy of a protein-coding gene will

usually mean that the cDNA is not expressed, and it will acquire deleterious mutations to become a retropseudogene. If, however, the cDNA integrates at a position adjacent to an existing promoter it may be expressed and acquire some useful function to become a **retrogene** (see [Figure 1](#)).

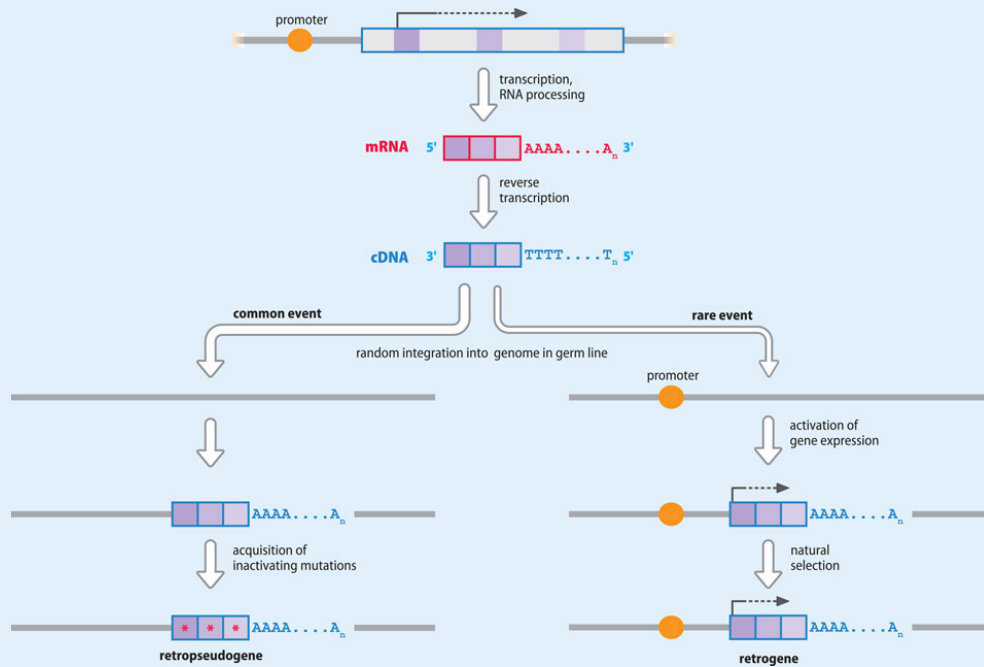


Figure 1 Retrogenes and retropseudogenes originate following reverse transcription from RNA transcripts. In this example, a protein-coding gene with three exons is transcribed from an upstream promoter, and introns are excised from the transcript to yield an mRNA. The mRNA can then be converted naturally into an antisense single-stranded cDNA by using a cellular reverse transcriptase. If this occurs in the germline and the resulting transcript then integrates randomly into the genome, it will most probably be incapable of expression (it will lack a promoter) and will degenerate into a retropseudogene, after acquiring harmful inactivating mutations (red asterisks, bottom left). If, however, the transcript inserts next to an endogenous promoter it may be expressed. Very occasionally, and according to need, the expressed gene may be useful and be preserved by natural selection as a functional retrogene (bottom right).

Several protein-coding genes have associated retropseudogenes ([Table 2.5](#)), but the highest numbers of retropseudogenes derive from RNA genes. RNA genes transcribed by RNA polymerase III have *internal promoters*. That is, the sequences needed to attract the transcriptional activation complexes (which then go on to recruit an RNA polymerase) are located within the transcription unit itself, instead of being located upstream like the promoters of protein-coding genes. When transcripts of these genes are copied into cDNA, the DNA copies of their transcripts also have promoter sequences, giving the potential to reach very high copy number. Human Alu repeats, for example, originated as cDNA copies of 7SL RNA, and the high copy number mouse B1 and B2 repeats are diverged cDNA copies of tRNAs.

FUNCTIONAL PSEUDOGENES

Comparative genomic studies indicate that some pseudogene sequences have evolved under purifying selection (they are more evolutionarily conserved than would have been expected for a functionless DNA sequence). Many pseudogenes are known to be transcribed, and there is good

evidence that the transcripts of some pseudogenes have important regulatory roles. For example, the *PTEN* gene (which is mutated in multiple advanced cancers) is located on chromosome 10 and is regulated by an RNA transcript from a closely related processed pseudogene *PTENP1* located on chromosome 9. As described in [Chapter 6](#), *PTENP1* regulates cellular levels of *PTEN*, and thereby acts as a tumor suppressor).

Many RNA genes are also members of large gene families. For example, the short arms of chromosomes 13, 14, 15, 21, and 22 each have 30–40 tandem repeats of a 45 kb DNA sequence that specifies 28S, 18S, and 5.8S ribosomal RNA. During mitosis, the megabase-sized clusters of ribosomal DNA on the different chromosomes can pair up and exchange segments, a type of interchromosomal recombination.

Noncoding RNAs can also be reverse transcribed to give cDNA copies that can integrate elsewhere in the genome, like the mRNA-derived cDNAs in [Figure 1](#) of [Box 2.4](#). During evolution, cDNA copies of certain classes of RNA genes that have internal promoters and are transcribed by RNA polymerase III were particularly successful in integrating into the genome because the copies carried with them promoter sequences. In some cases this resulted in a huge increase in copy number, and as explained below it gave rise to the most commonly occurring repetitive DNA sequence in the human genome, the Alu repeat.

The significance of gene duplication and repetitive coding DNA

Over long periods of evolutionary time there seems to have been a relentless drive to duplicate DNA in complex genomes. That has meant that whole genes have been duplicated to give gene families as described above. Tandem duplication of exons (which occurs by the same mechanisms that produce tandem gene duplication) is also evident in about 10 % of human protein-coding genes. There are several advantages of DNA sequence duplication, as listed below.

Gene dosage

Duplication of genes can be advantageous simply because it allows more gene product to be made. Increased gene dosage is an advantage for genes that make products needed in very large amounts in cells—we have hundreds of virtually identical copies of genes that make individual ribosomal RNAs and individual histone proteins, for example. Exon duplication might also be an advantage when an exon (or group of exons) encodes a structural motif that can be repeated, allowing proteins such as collagens to extend the size of structural domains during evolution.

Novel genetic variants

Once a gene or exon has duplicated, there are initially two copies with identical sequences. When that happens, the constraints on changing the sequence imposed by Darwinian natural selection may be applied to one of the two sequences only. The other sequence is free from normal constraints to maintain the original function; it can diverge in sequence over many millions of years to produce a different but related genetic variant. Divergent exons allowed the formation of different but related

protein domains and the possibility of alternative splicing to produce transcripts with different exon combinations. Additionally, as described below, certain types of mobile element allow the copying of exons from one gene to another (exon shuffling) to produce novel combinations of exons.

Divergent genes produced by tandem gene duplication allow the production of variant but related proteins. The vertebrate globin superfamily provides illustrative examples. Over a period of 800 million years or so, a single ancestral globin gene gave rise to all existing globin genes by a series of periodic gene duplications. Early duplications led to diverged gene copies that ultimately came to be expressed in different cell types, producing globins that were adapted to work in blood (hemoglobins), in muscle (myoglobin), in the nervous system (neuroglobin), or in multiple cell types (cytoglobin).

More recent duplications in the α - and β -globin gene clusters (see [Figure 2.13C](#) for the latter) led to different varieties of hemoglobin being produced at different stages of development. Thus in early development, zeta (ζ)-globin is used in place of α -globin, while epsilon (ϵ)-globin is used instead of β -globin in the embryonic period, and γ -globin is used instead of β -globin in the fetal period. The globins incorporated into hemoglobin in the embryonic and fetal periods have been considered to be better adapted to the more hypoxic environment at these stages.

There are, however, disadvantages to DNA sequence duplication. One consequence of repetitive coding DNA and tandemly repeated gene sequences is that the repeated DNA sequences can be prone to genetic instability, causing disease in different ways. We examine this in detail in [Chapter 7](#).

Highly repetitive noncoding DNA in the human genome

Just over half of the human genome is made up of highly repetitive noncoding DNA sequences, of which a minority (about 14 %) is found in constitutive heterochromatin (which accounts for a total of about 7 % of our DNA). Euchromatin accounts for about 93 % of our DNA, of which just under half is made up of highly repetitive noncoding DNA (accounting for 45 % of the total genome).

The repetitive noncoding DNA in heterochromatin is a mixture of repetitive DNA sequences that are found in both heterochromatin and euchromatin (see examples below) plus DNA repeats that are characteristic of heterochromatin. The latter include different satellite DNA families of highly repetitive tandem repeats. Satellite DNAs are common at centromeres and include: alphoid DNA, with a 171 bp repeat unit (found at all human centromeres); a 68 bp β repeat unit at the centromeres of the acrocentric chromosomes plus chromosomes 1, 9, and Y; plus different other satellite DNAs with comparatively small repeat units.

Like most heterochromatin DNA, the DNA of centromeric heterochromatin is very poorly conserved between species. Telomeric heterochromatin is the exception. It is based on TTAGGG repeats (that extend over lengths of 5–15 kb at the chromosome ends); the TTAGGG telomere repeat sequence is conserved throughout vertebrates and is highly similar to the telomere repeats of many invertebrates and plants.

Transposon-derived repeats in the human genome

Different classes of highly repetitive DNA sequences occur in an interspersed fashion (rather than as tandem repeats) and are commonly found within genes (usually in introns, but sometimes in exons). They arose from *transposons*, mobile elements that were able to migrate from one location in the genome to another.

The vast majority of transposon-derived repeats in the human genome can no longer transpose and are considered “transposon fossils”. They are either truncated, having lost key sequences, or have picked up inactivating mutations during evolution. As a result, only a very small number of human transposon repeats are now capable of transposing autonomously (but other transposon-derived repeats can sometimes transpose by hitching a ride when located physically close to an autonomously transposing repeat).

Only about 6 % of the highly repetitive interspersed repeats in euchromatin originated from DNA transposon families that transpose by a cut-and-paste mechanism. The great majority, accounting for at least 40 % of the genome, originated from **retrotransposons** that transpose through an RNA intermediate. Here an RNA is copied by cellular reverse transcriptases to make a cDNA copy that integrates elsewhere in the genome (the same principle as shown in the figure in [Box 2.4](#)). There are three major classes and one minor class of retrotransposon-derived repeats in the human genome, as listed below and in [Table 2.6](#).

TABLE 2.6 HUMAN TRANSPOSON REPEAT CLASSES AND FAMILIES

Transposon repeats by origin	Repeat class	Full length	% of genome	Examples (*see Fig. 2.15)
RETROTRANSPOSON REPEATS (via RNA intermediate)	LINEs	6–8 kb	21%	LINE-1*
	SINEs	100–300bp	13%	Alu repeat*
	Retrovirus-like elements	1.5–11kb	8%	HERV family
				LTR element
SVA elements	~2kb	0.1 %	SVA element*	
DNA TRANSPOSON REPEATS (via cut and paste)	Various	2–3 kb	3%	

Note: the great majority of repeats are truncated, having lost sequence components during evolution. Abbreviations are: LINEs—Long Interspersed Nuclear Elements; SINEs—Short Interspersed Nuclear Elements; HERV—human endogenous retrovirus; LTR—long terminal repeat; SVA—Sine-R-VNTR-Alu.

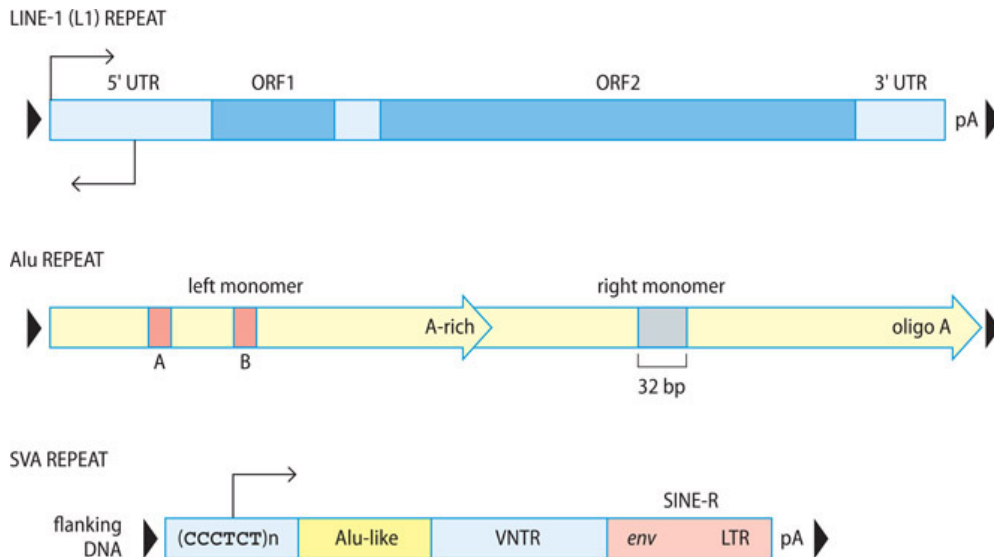


Figure 2.15 Structure of three types of commonly transposing human transposon-derived repeats. Some full-length LINE-1 repeats can transpose autonomously. The ORF2 open reading frame makes an endonuclease that can cut DNA (preferentially at AT-rich sequences) and a reverse transcriptase that uses the released 3'-OH end to prime cDNA synthesis. New insertion sites are flanked by a small target-site duplication (flanking black arrowheads). Alu repeats often consist of two monomer repeats with similar sequences terminating in an A-rich or oligo A sequences. Nonautonomous SVA repeats have both an Alu-like sequence and a 3' HERV fragment (SINE-R), separated by a sequence with variable number of tandem repeats (VNTR). They may often be transcribed from promoters in flanking DNA sequence. Arrows indicate propensity for transcription.

- *LINEs (long interspersed nuclear elements)*. Of the three LINE families, the most numerous is the LINE-1 (also called L1) family. The only human LINE elements currently capable of transposition are a small subset (about 80–100 copies) of full-length LINE-1 repeats.
- *SINEs (short interspersed nuclear elements)*. SINEs are non-autonomous: they need a reverse transcriptase to be supplied (for example by a neighboring LINE-1 repeat). The primate-specific Alu repeat family, with close to 1.5 million copies, is the dominant family and is preferentially located in euchromatic regions of the genome (unlike LINEs that tend to be located in heterochromatin). Alu repeats have evolved from cDNA copies of 7SL RNA (a component of the signal recognition particle that regulates transport of proteins out of cells). The other two SINE families have evolved respectively from reverse transcription of a tRNA and a 5S rRNA.
- *Retrovirus-like LTR elements*. The human endogenous retroviruses (HERVs) contain sequences resembling the key retroviral genes, *gag*, *pol* (encoding reverse transcriptase) and *env*, flanked by *long terminal repeats (LTRs)*, but there is little evidence of actively transposing human HERVs. Truncated versions with remnants of the *gag* gene are also common.
- *SVA (SINE-R-VNTR-Alu) elements*. This very small family has compound repeats with a mix of an Alu-like sequence and a HERV-like sequence (confusingly known as SINE-R)—see [Figure 2.15](#).

The evolutionary value of transposon repeats

At this stage, one might wonder why so much—at least 50 %—of our genome is composed of transposon-derived repeats, the great majority of which are now incapable of transposition. A likely answer is the drive to increase genetic novelty. In complex genomes a relentless evolutionary drive to duplicate DNA sequences has resulted in duplicated genes, duplicated regulatory sequences, duplicated exons, and so on, allowing the freedom to alter duplicates, while conserving the original functions of duplicated elements. Transposition can carry regulatory elements, and even exons, to different parts of the genome where they can change the functions of genes (for an example, see [Figure 2.16](#) for how a LINE-1 repeat can cause *exon shuffling* between genes). In very rare cases, transposable elements even appear to have given rise to new genes in evolution.

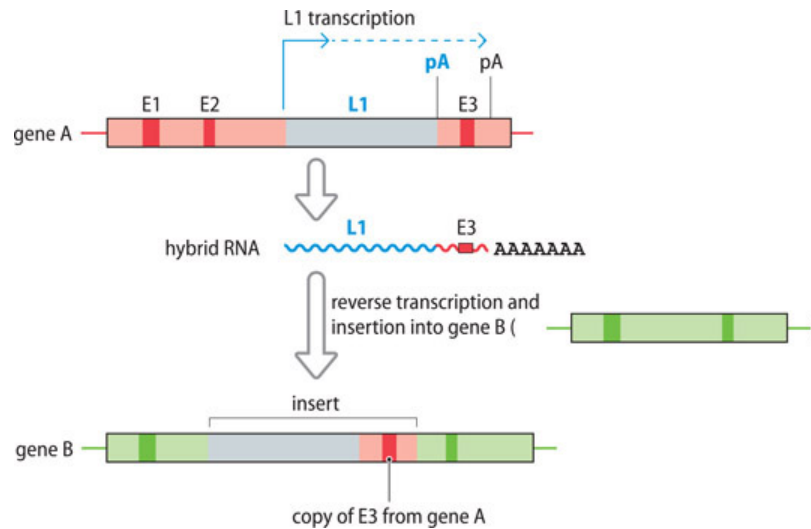


Figure 2.16 Retrotransposons can mediate exon shuffling. Exon shuffling can be carried out using retrotransposons such as actively transposing members of the LINE1 (L1) sequence family, as shown here. LINE-1 elements have weak poly(A) signals (pA) and so transcription often continues past such a signal until another downstream poly(A) signal is reached (for example after exon 3 (E3) in gene A). The resulting RNA copy contains a transcript not just of L1 sequences but also of a downstream exon (in this case E3). The L1 reverse transcriptase machinery can then act on the extended poly(A) sequence to produce a hybrid cDNA copy that contains both L1 and E3 sequences. Subsequent transposition into a new chromosomal location may lead to the insertion of exon 3 into a different gene (gene B)—[Figure 1](#) Box 9.3.

SUMMARY

- Genes are transcribed to make RNA. Protein-coding genes make an mRNA that is decoded to make a poly-peptide. RNA genes make a functional noncoding RNA.
- An mRNA contains a central coding sequence, flanked by noncoding untranslated regions that contain regulatory sequences.
- In eukaryotes, the DNA sequence corresponding to the coding sequence of an mRNA is often divided into exons that are separated by intervening introns.
- The primary RNA transcript, an RNA copy of both exons and introns, is cleaved at exon-intron boundaries. Most of the transcribed intron sequences are discarded, and transcribed exon sequences are spliced together. Specialized end sequences—a 5' cap sequence and a 3' poly(A) sequence—protect the ends of the mRNA and assist in transfer to the cytoplasm to engage with ribosomes.
- The coding sequence of an mRNA is translated using groups of three nucleotides (codons) to specify individual amino acids that are bonded together to make a polypeptide.
- Introns are also found in noncoding DNA, both in some RNA genes and in many DNA sequences that make untranslated regions in mRNAs.
- Noncoding RNAs perform many different functions in cells, but most regulate gene expression, either ubiquitously (to assist general protein synthesis), or by controlling certain target genes in selected cell types.

- The human nuclear genome is composed of 24 different types of very long linear DNA molecules, one each for the 24 different types of human chromosomes (1–22, X, and Y). It has about 20 000 protein-coding genes and over 26 000 RNA genes (genes that make noncoding RNAs).
- Our mitochondrial genome consists of one type of small circular DNA molecule that is present in many copies per cell. It has 37 genes that make all the rRNAs and tRNAs needed for protein synthesis on mitochondrial ribosomes plus a few of the proteins involved in oxidative phosphorylation.
- The great majority of the genome consists of poorly conserved DNA sequences and only about 10 % of genome sequences are thought to be under selective constraint to maintain function.
- About 1.2 % of the nuclear genome is decoded to make proteins. These coding DNA sequences have mostly been highly conserved during evolution—for each human protein, recognizably similar proteins exist in many other organisms.
- RNA genes are more rapidly evolving than coding DNA, and while the shape of a noncoding RNA is particularly important the linear sequence is less important than that of polypeptides.
- Regulatory sequences are generally less conserved than coding DNA and polypeptide sequences. Humans and mice have a very similar set of genes but they are expressed differently because of differences in regulatory elements.
- Repetitive DNA sequences are very common in the human genome. They include both tandem repeats (often sequential head-to-tail repeats) and dispersed repeats.
- Tandem repeats may be found within genes and coding sequences, and whole genes can be duplicated several times to produce a clustered gene family. Other gene families are made up of gene copies that are dispersed across two or more chromosomes.
- Gene families often contain defective gene copies (pseudogenes and gene fragments) in addition to functional genes.
- Dispersed gene copies often arise in evolution from RNA transcripts that are copied by a reverse transcriptase to make a complementary DNA that integrates randomly into chromosomal DNA (retrotransposition).
- DNA sequence lying outside exons is largely composed of repetitive sequences, including highly repetitive interspersed repeats such as Alu repeats. They originated by retrotransposition (DNA copies were made of RNA transcripts that then integrated into the genome). Very few of the repeats are currently able to transpose.
- The DNA of centromeres and telomeres is largely composed of very many tandemly repeated copies of short sequences.
- Gene and exon duplication has been a driving force during genome evolution. Novel genes and exons are occasionally produced by tandem duplication events. Novel exons and novel regulatory sequences can also be formed by retrotransposition.

QUESTIONS

Questions can be downloaded by visiting the following link, under Support Materials: www.routledge.com/9780367490812.

FURTHER READING

More detailed treatment of much of the subject matter in this chapter, including a detailed account of human genome organization, gene evolution, and the Human Genome Project, can be found in the following:

Strachan T & Read AP (2019) *Human Molecular Genetics*, 5th ed. Garland Science.

Protein-coding genes and protein structure

Agris PF (2007) tRNA's wobble decoding of the genome: 40 years of modification. *J Mol Biol* 366:1–13; PMID 17187822.

Piovesan A (2019) Human protein-coding genes and gene feature statistics in 2019. *BMC Res Notes*: 12: 315; PMID 31164174

Preiss T & Hentze MW (2003) Starting the protein synthesis machine: eukaryotic translation initiation. *BioEssays* 25:1201–1211; PMID 14635255.

Whitford D (2005) *Protein Structure and Function*. John Wiley & Sons.

RNA genes and regulatory RNA

Amaral PP (2008) The eukaryotic genome as an RNA machine. *Science* 319:1787–1789; PMID 18369136.

Ponting CP (2009) Evolution and functions of long noncoding RNAs. *Cell* 136:629–641; PMID 19239885

Human genome: analysis and internet resources

Djebali S (2012) Landscape of transcription in human cells. *Nature* 489:101–108; PMID 22955620.

ENCODE Project Consortium (2012) An integrated encyclopedia of DNA elements in the human genome. *Nature* 489:57–74; PMID 22955616.

GENCODE Project statistics (with useful statistics on the numbers of human genes and transcripts) are available at <http://www.encodegenes.org/human/stats.html>

Genome Browsers and Internet Resources – see [Table 2.3](#).

User's Guide to the Human Genome Nat Genet 35 supplement no. 1, September 2003. Available at <http://www.nature.com/ng/journal/v35/n1s/index.html>

Human genome: organization and evolution

- Bailey JA (2002) Recent segmental duplications in the human genome. *Science* 297:1003–1007; PMID 12169732.
- Conrad B & Antonorakis SE (2007) Gene duplication: a drive for phenotypic diversity and cause of human disease. *Annu Rev Genomics Hum Genet* 8:17–35; PMID 17386002.
- Konkel MK & Batzer MA (2010) A mobile threat to genome stability: the impact of non-LTR retrotransposons upon the human genome. *Semin Cancer Biol* 20:211–221; PMID 20307669.
- Long M (2013) New gene evolution: little did we know. *Annu Rev Genet* 47:307–333; PMID 24050177.
- Mills RE (2007) Which transposable elements are active in the human genome? *Trends Genet* 23:183–191; PMID 17331616.
- Muotri AR (2007) The necessary junk: new functions for transposable elements. *Hum Molec Genet* 16:R159-R167; PMID 17911158.
- Pink RC (2011) Pseudogenes: pseudo-functional or key regulators in health and disease. *RNA* 17:792–798; PMID 21398401.
- Ponting CP & Hardison RC (2011) What fraction of the human genome is functional? *Genome Res* 21:1769–1776; PMID 21875934.
- Vinckenbosch N (2006) Evolutionary fate of retroposed gene copies in the human genome. *Proc Natl Acad Sci USA* 103:3220–3225; PMID 16492757.

3

Principles underlying core DNA technologies

DOI: [10.1201/9781003044406-3](https://doi.org/10.1201/9781003044406-3)

CONTENTS

[3.1 AMPLIFYING DNA BY DNA CLONING](#)

[3.2 AMPLIFYING DNA USING THE POLYMERASE CHAIN REACTION \(PCR\)](#)

[3.3 PRINCIPLES OF NUCLEIC ACID HYBRIDIZATION](#)

[3.4 PRINCIPLES OF DNA SEQUENCING](#)

[SUMMARY](#)

[QUESTIONS](#)

[FURTHER READING](#)

Defining the genetic basis of disease requires the analysis of DNA and sometimes chromosomes. That can be challenging because the vast majority of our genetic material is organized as immensely long DNA molecules, and disease may result from a mutation that changes just a single nucleotide out of the more than six billion nucleotides in the combined maternal and paternal genomes that we inherit. Sophisticated DNA technologies allow us

to study and analyze genes, enabling diagnostic, predictive, and therapeutic applications that we describe in later chapters. Here we confine ourselves to describing the *principles* of four core technologies for purifying and analyzing DNA sequences.

Three of the four core DNA technologies began to be employed in the first efforts to study and identify individual human genes. The problem here was that human genes are often composed of very small exons separated by very large introns, and individual exons and genes represent a tiny fraction of our genome. Although we can readily isolate DNA from human cells, a single coding sequence exon, averaging just 150 bp, is a tiny fraction of the DNA (just 1/20 000 000 of the genome). Many full-length genes are also extremely small components of the genome. To allow us to focus on just a single exon or single gene, two quite different approaches can be employed. Either the DNA of interest must be purified by selectively increasing its copy number (DNA amplification), or it must be specifically recognized in some way (see [Table 3.1](#)).

TABLE 3.1 TWO VERY DIFFERENT APPROACHES USED TO ENABLE DETAILED STUDY OF A SHORT DNA OF INTEREST (“TARGET DNA”) IN A COMPLEX GENOME

General approach	Subapproaches	Core technology
1. Purify the target DNA by selective amplification*	Selective amplification <i>within cells</i> (using a cellular DNA polymerase)	DNA cloning (Section 3.1)
	Selective amplification <i>in vitro</i> (using a purified DNA polymerase)	PCR** (Section 3.2)
2. Specifically recognize the target	Various (Section 3.3)	Nucleic acid hybridization

* That is, selectively increase the number of copies of the target DNA.

** PCR (the polymerase chain reaction) is the most widely used way to amplify a target DNA *in vitro*, but alternative methods exist.

General approach	Subapproaches	Core technology
DNA		(Section 3.3)

* That is, selectively increase the number of copies of the target DNA.

** PCR (the polymerase chain reaction) is the most widely used way to amplify a target DNA *in vitro*, but alternative methods exist.

Table 3.1 shows three core DNA technologies. There is of course, a fourth. DNA sequencing is the ultimate way of tracking changes in genes and DNA sequences. It used to be expensive, time-consuming, and restricted in scope. All that has changed. Now, almost two decades into the “post-genome era” (after the human genome sequence was obtained), DNA sequencing is such a hugely efficient and dominant DNA technology that we can sequence whole human genomes quickly and cheaply. Accordingly, we can now *simultaneously* analyze all known genes across our genome. We consider the general principles, and the most basic of the commonly used DNA sequencing methods in [Section 3.4](#), but we introduce the most recent DNA sequencing techniques in [Chapter 11](#), and some medical applications in various other chapters.

3.1 AMPLIFYING DNA BY DNA CLONING

Cloning DNA in cells is a way of purifying DNA sequences and is usually carried out in bacterial cells. It can allow very many identical copies of a desired DNA sequence to be produced, enabling it to be studied or put to some use. To do that, cells are first treated to optimize the transfer of DNA molecules to be cloned into the cells, a process known as **transformation**. In each case the DNA to be cloned is first covalently joined to some *vector* DNA molecule that will help it replicate within the host cells, as detailed below. The joining of DNA fragments to vector molecules results in the formation of an artificial **recombinant DNA**. There is normally some kind of selection or screening system that helps identify those cells that have been successfully transformed and that contain recombinant DNA.

Transformation, a key step in DNA cloning, is highly selective: when foreign DNA does get into a cell, just a *single* DNA molecule is usually taken up by a cell. A population of cells therefore serves as a sorting office that can efficiently fractionate a complex mixture of DNA fragments ([Figure 3.1](#)).

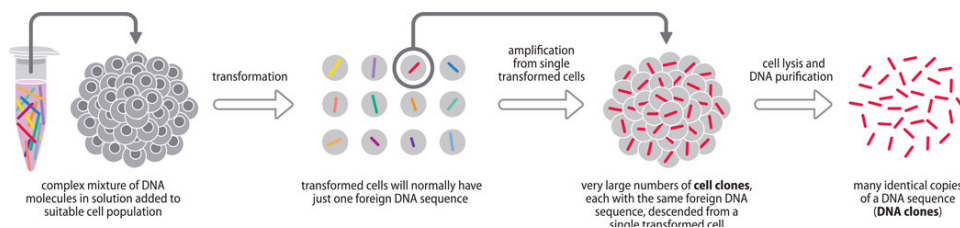


Figure 3.1 Transformation as a way of fractionating a complex sample of DNA fragments. The key point is that transformation is selective: when a cell is transformed, it usually picks up a *single* DNA molecule from the environment, and so different fragments are taken up by different cells. (For simplicity, the figure shows only the DNA sequences that are to be cloned—in practice they would be joined to a *vector DNA* molecule to make a recombinant DNA that is often circular.) Cell clones can form by cell division from a single transformed cell. Thereafter, they are propagated to produce a large number of cells with an identical foreign DNA sequence that can be purified after the cells have been broken open.

Amplifying desired DNA within bacterial cells

Large quantities of a cloned DNA can be obtained using bacterial cells because the inserted DNA can be amplified to very high copy numbers ([Figure 3.1](#)). That is possible for two reasons. First, a single bacterium containing a cloned DNA can rapidly divide and eventually produce a huge number of identical bacterial cell clones, each with the same foreign DNA sequence. Secondly, some vector molecules can replicate *within* a bacterial cell to reach quite high copy numbers; if they have a foreign DNA sequence covalently linked to them, it too will be amplified within the cell ([Figure 3.2](#)).

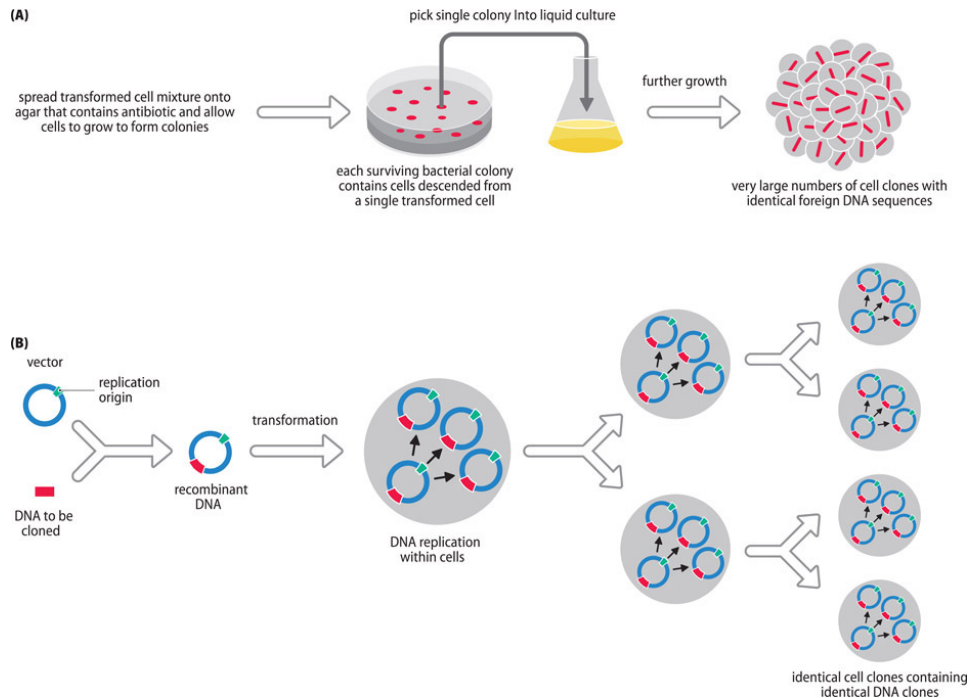


Figure 3.2 DNA cloning in bacterial cells: DNA copy number amplification and separation of clones from different transformed cells. (A) For a cell transformed by a recombinant DNA, an increase in cell number leads to a proportional expansion in copy number of that recombinant DNA. Growth occurs initially in a solid medium (after the transformed cells have been spread out on a plate of agar containing antibiotics). Individual cells will be physically dispersed on the plate, and then go through several rounds of cell division *in situ* to form *separate* visible colonies. An individual colony can be picked and allowed to go through a second round of amplification by growth in liquid culture. For simplicity, the cloned DNA fragments are shown in the absence of the vector molecule. (B) Vectors have their own replication origin allowing them to replicate *independently* of, and much more frequently than, the bacterial chromosome.

The need for vector DNA molecules

Fragments of human DNA cannot normally replicate after being transferred into bacterial cells: they lack a special DNA sequence capable of initiating DNA replication in that cell type. Such sequences are called replication origins; molecules containing a replication origin are known as **replicons**.

To permit replication within a bacterial cell the human DNA fragment must first be covalently joined (*ligated*) to a suitable replicon, forming a

recombinant DNA. Extrachromosomal replicons are typically used for this purpose, either *plasmids* (small circular double-stranded DNAs that can replicate in bacteria) or *bacteriophages* (bacterial viruses). When used to ferry desired DNA fragments into bacterial cells, the plasmid or bacteriophage DNA is known as a **vector DNA**.

To be useful as a cloning vector the original plasmid or bacteriophage needs to be genetically modified so that we can efficiently join a foreign DNA to it (described below) and so that transformed cells can easily be recognized. The vector will often have been genetically engineered to contain a gene conferring resistance to some antibiotic to which the host bacterial cells are sensitive. After transformation, the cells are grown on agar containing the antibiotic; untrans-formed cells die, but transformed cells survive. Because some cells are transformed by naked vector DNA (lacking other DNA), screening systems are often also devised to ensure that cells with recombinant DNA can be identified.

Physical clone separation

How can cells that have taken up different DNA fragments be separated from each other? That relies on the formation of *physically separated* cell colonies. After transformation of bacterial cells, for example, aliquots of the cell mixture are spread over the surface of antibiotic-containing agar in Petri dishes (plating out); successfully transformed cells should grow and multiply; if the plating density is optimal, they form well-separated cell colonies (see [Figure 3.2A](#)). Each colony consists of identical descendant cells (cell clones) that originate from a single transformed cell and so the cell clones each contain the same single foreign DNA molecule.

An individual well-separated cell colony can then be physically picked and used to start the growth of a large culture of identical cells all containing the same foreign DNA molecule, resulting in very large amplification of a single DNA sequence of interest ([Figure 3.2A](#)). Thereafter, the cloned foreign DNA can be purified from the bacterial cells.

The need for restriction nucleases

DNA cloning in bacterial cells is most efficient when transferring relatively small DNA fragments. However, when DNA is isolated from the cells of complex organisms, the immensely long nuclear DNA molecules are fragmented by physical shearing forces to give an extremely heterogeneous collection of still rather long fragments with heterogeneous ends. The long fragments need to be reduced to pieces of a much smaller, manageable size with more uniform end sequences to facilitate ligation.

Recombinant DNA technology was first developed in the 1970s. The crucial breakthrough was to exploit the ability of restriction endonucleases to cut the DNA at *defined* places. As a result, the DNA could be reduced to small well-defined fragments with uniform end sequences that could be easily joined by a DNA ligase to similarly cut vector molecules ([Box 3.1](#)).

BOX 3.1 RESTRICTION ENDONUCLEASES: FROM BACTERIAL GUARDIANS TO GENETIC TOOLS

THE NATURAL ROLE OF RESTRICTION ENDONUCLEASES: HOST CELL DEFENSE

Restriction endonucleases are bacterial enzymes that recognize specific short sequence elements within a double-stranded DNA molecule, and then cleave the DNA on both strands, within, or close to, the recognition sequences. They provide a form of self-defense against bacteriophages: the restriction nuclease produced by a bacterium is designed to selectively cleave the DNA of the invading bacteriophage into small pieces, while leaving the bacterial genome intact. Different types and strains of bacteria produce restriction endonucleases of different sequence specificity.

For example, restriction nuclease *EcoRI* from the *Escherichia coli* strain RY13 specifically recognizes the sequence GAATTC and cleaves DNA strands within this recognition sequence (called a **restriction site**) – see [Figure 1A](#). The same bacterial strain also initially produces an *EcoRI* methyltransferase to modify its own genome: it recognizes the same sequence GAATTC and methylates the central adenosine on both DNA

strands. The *EcoRI* restriction nuclease cannot cleave at previously methylated GAATTC sequences within the bacterial genome but will cleave at unmethylated GAATTC sequences in the DNA of invading pathogens, cutting it up into small pieces that are degraded.

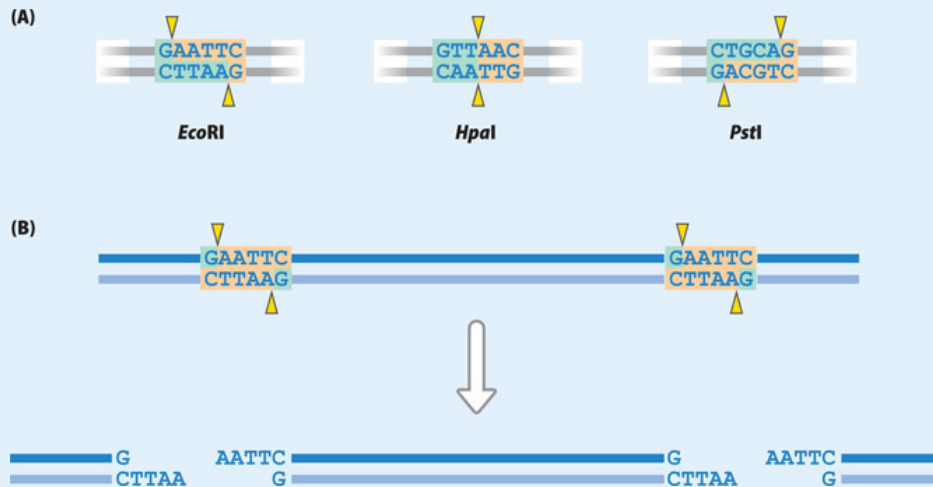


Figure 1 Sequence specificity of type II restriction nucleases. (A) Shown are the sequence specificities of three type II restriction nucleases that cleave within palindromic recognition sites. *EcoRI* (from *Escherichia coli* strain RY13) and *PstI* from *Providencia stuartii* cut asymmetrically within their recognition sequences to produce four-nucleotide overhanging ends (5' AATT overhangs for *EcoRI*; 3' TGCA for *PstI*). *HpaI* from *Haemophilus parainfluenzae* cuts symmetrically at the middle of its recognition sequence to leave blunt-ended fragments. (B) Cleavage of genomic DNA with *EcoRI* produces fragments with two 5' AATT overhangs.

RESTRICTION NUCLEASES AS MOLECULAR GENETIC TOOLS

There are different classes of restriction nucleases but type II restriction nucleases are widely used in manipulating and analyzing DNA. They recognize short sequence elements that are typically *palindromes* (the 5' 3' sequence is the same on both strands, as in the sequence GAATTC); they then cleave the DNA either within, or very close to, the recognition sequence. Cleavage often occurs at asymmetric positions within the two strands to produce fragments with overhanging 5' ends or overhanging 3'

ends, but sometimes a restriction enzyme cuts symmetrically to produce “blunt ends” (see [Figure 1](#)).

Under appropriate conditions, it is possible to use a restriction nuclease to cut complex genomic DNA into thousands or millions of fragments that can then be individually joined using a DNA ligase to a similarly cut vector molecule to produce recombinant DNA molecules ([Figure 2](#)). For cloning DNA in bacterial cells, vector molecules are often based on circular plasmids that have been artificially engineered so that they contain unique restriction sites for certain restriction nucleases. The recombinant DNA molecules can then be transferred into suitable host cells and amplified.

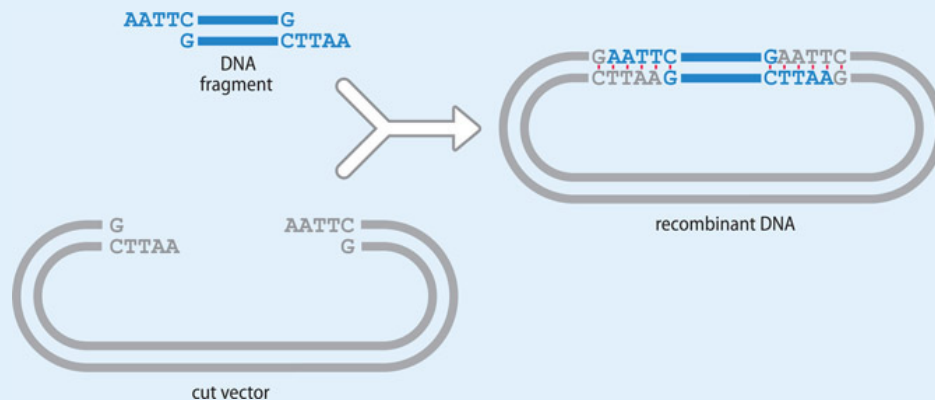


Figure 2 Formation of recombinant DNA. In this example, the vector has been cut at a unique *EcoRI* site to produce 5' ends with an overhanging AATT sequence, and the DNA fragment to be cloned has the same 5' AATT overhangs, having also been produced by cutting with *EcoRI*. The AATT overhangs are examples of *sticky ends* because they can hydrogen bond to other fragments with the same overhang thus facilitating intermolecular interactions. (Vertical red lines in the recombinant DNA represent hydrogen bonds between paired 5' AATT overhangs in the vector and in the DNA to be cloned.)

DNA libraries and the uses and limitations of DNA cloning

Once DNA cloning was established it was soon used to make **DNA libraries**; that is, collections of DNA clones representing all types of DNA sequence in a complex starting material.

DNA isolated from white blood cells, for example, provides a complex genomic DNA that can be cut into many pieces and attached to vector DNA molecules. The resulting mixture of different recombinant DNA molecules is used to transform bacteria to produce very many different clones, a *genomic DNA library*. A good genomic DNA library would have so many different DNA clones that there was a good chance that the library would include just about all the different DNA sequences in the genome.

An alternative was to make gene-centred *cDNA libraries* starting with mRNA. Because RNA cannot be cloned, DNA copies of the RNA were made using a specialized reverse transcriptase that naturally copies a single-stranded RNA template to make a **complementary DNA (cDNA)** copy. Once the cDNA strand has been made, the original RNA is destroyed by treatment with ribonuclease and the remaining DNA strand is copied in turn to give a complementary DNA, thereby making double-stranded cDNA that can be cloned like any other DNA.

DNA cloning started a revolution in genetics. It prepared the way for obtaining panels of DNA clones representing all the sequences in the genome of organisms, and that in turn made genome projects possible to obtain the complete sequence of genomic DNA in a variety of organisms. Once that was done, the structure of genes could be determined, paving the way for comprehensive studies to analyze gene expression and to determine how individual genes work.

There is a drawback: cloning DNA in cells is laborious and time-consuming. It is also not suited to performing rapid parallel amplifications of the same DNA sequence in multiple different samples of DNA. That required a new technology, as described in the next section.

3.2 AMPLIFYING DNA USING THE POLYMERASE CHAIN REACTION (PCR)

PCR, a cell-free method for amplifying DNA, was first developed in the mid-1980s and revolutionized genetics. It was both very fast and readily allowed parallel amplifications of DNA sequences from multiple starting DNA samples. If you wanted to amplify each exon of the b-globin gene from blood DNA samples from 100 different individuals with b-thalassemia, a single person could now do that in a very short time.

Basics of the polymerase chain reaction (PCR)

PCR relies on using a heat-stable DNA polymerase to synthesize copies of a small, predetermined DNA segment of interest within a complex starting DNA (such as total genomic DNA from easily accessed blood or skin cells). To initiate the synthesis of a new DNA strand, a DNA polymerase needs a single-stranded oligonucleotide **primer** that is designed to bind to a *specific* complementary sequence within the starting DNA.

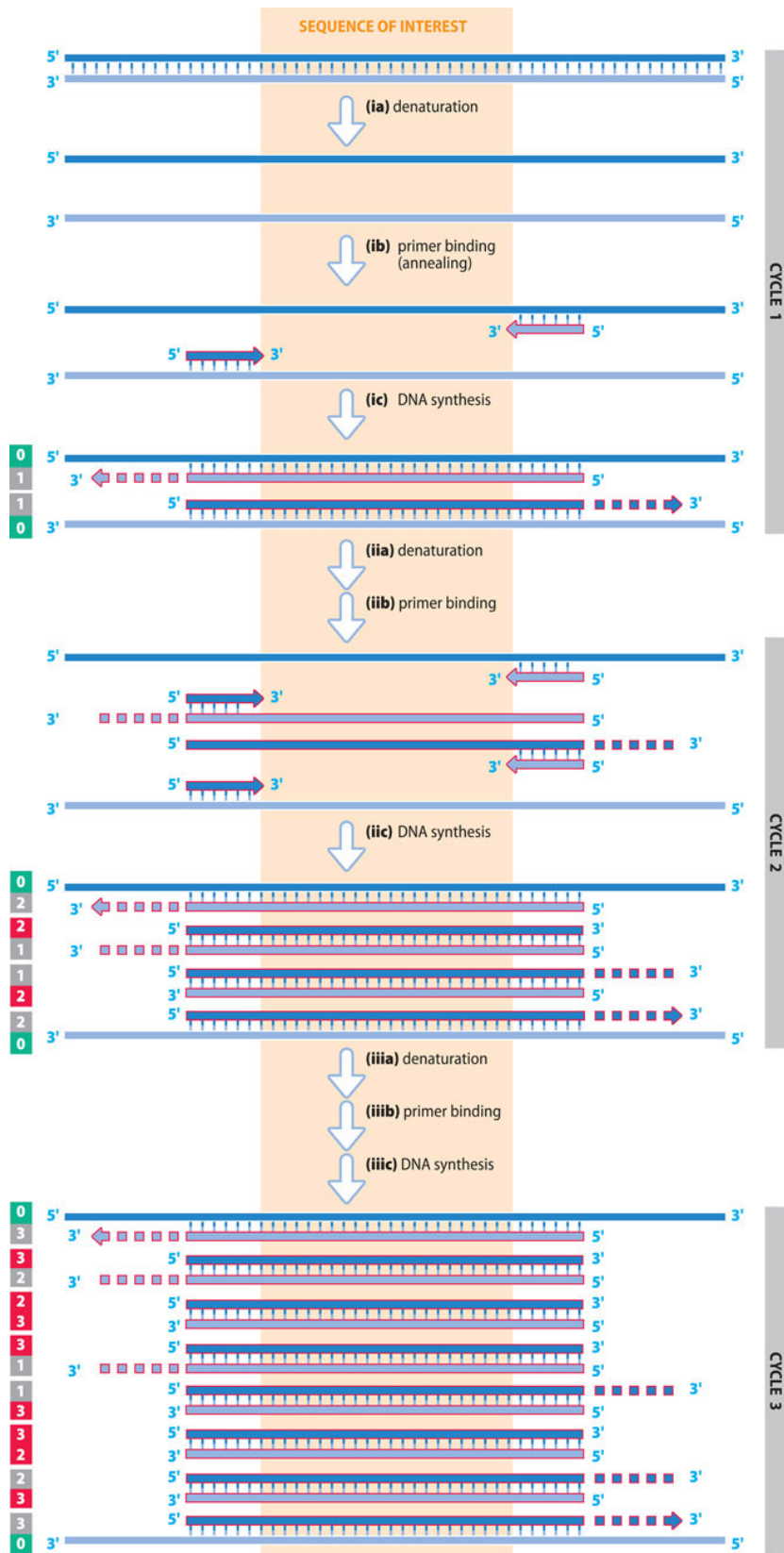
For the primer to bind preferentially at just one desired location in a complex genome, the oligonucleotide often needs to be about 20 nucleotides long or more and is designed to be able to base pair perfectly to its intended target sequence (the strength of binding depends on the number of base pairs formed and the degree of base matching).

To allow the primer to bind, the DNA needs to be heated. At a high enough temperature, the hydrogen bonds holding complementary DNA stands together are broken, causing the DNA to become single stranded. Subsequent cooling allows the oligonucleotide primer to bind to its perfect complementary sequence in the DNA sample (**annealing** or **hybridization**). Once bound, the primer can be used by a suitably heat-stable DNA polymerase to synthesize a complementary DNA strand.

In PCR, two primers are designed to bind to complementary sequences on *opposing* DNA strands, so that copies are made of both DNA strands. The primers are designed to be long enough for them to bind specifically to sequences that closely flank the DNA sequence of interest in such a way that the direction of synthesis of each new DNA stand is toward the sequence that is bound by the other primer. In further cycles of DNA

denaturation, primer binding, and DNA synthesis, the previously synthesized DNA strands become targets for binding by the other primer, causing a chain reaction to occur.

Synthesis of new DNA strands continues until the end of the template DNA is reached or until the polymerase disengages from the template DNA. The initial template DNA strands are often very long and on different copies of the template DNA the polymerase may disengage at variable places, thereby producing strands with variable 3' ends. However, increasingly, as the PCR reaction proceeds, template strands with fixed ends terminating in a primer sequence begin to predominate, and as a result, a product with fixed 5' and 3' ends is hugely predominant ([Figure 3.3](#)).



0
1
1
0

0
2
2
1
1
2
2
0

0
3
3
2
2
3
3
1
1
3
3
2
2
3
3
0

CYCLE 1

CYCLE 2

CYCLE 3

Figure 3.3 The polymerase chain reaction (PCR). The reaction usually consists of about 25–30 cycles of (a) DNA denaturation, (b) binding of oligonucleotide primers flanking the desired sequence, and (c) new DNA synthesis in which the desired DNA sequence is copied and primers are incorporated into the newly synthesized DNA strands. Numbers in the vertical strips to the left indicate the origin of the DNA strands, with original DNA strands represented by 0 and PCR products by 1 (made during first cycle), 2 (second cycle), or 3 (third cycle). The first cycle will result in new types of DNA product with a fixed 5′ end (determined by the primer) and variable 3′ ends (extending past the other primer). After the second cycle, there will be two more products with variable 3′ ends but also two desired products of fixed length (shown at the left by filled red squares) with both 5′ and 3′ ends defined by the primer sequences. Whereas the products with variable 3′ ends increase arithmetically (amount = $2n$, where n is the number of cycles), the desired products initially increase exponentially until the reaction reaches a stationary phase as the number of reactants becomes depleted (see [Figure 3.4](#)). After 25 or so cycles, the desired product accounts for the vast majority of the DNA strands.

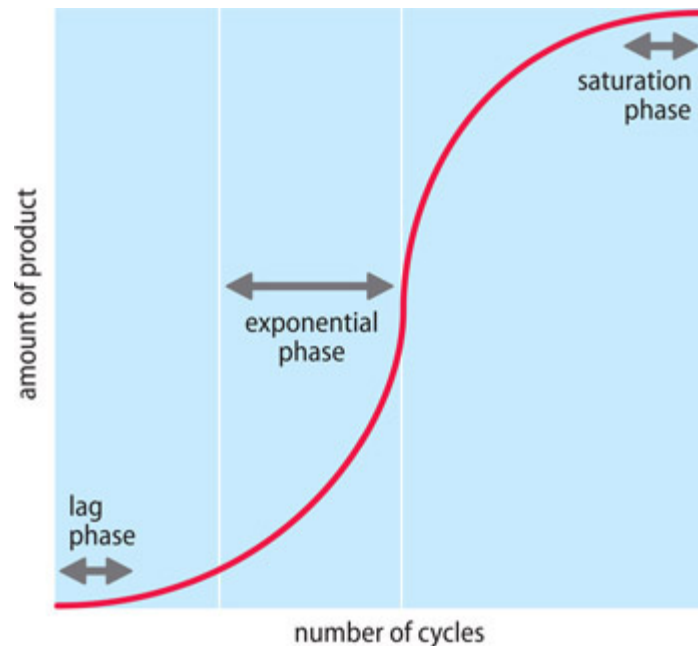


Figure 3.4 Different phases in a PCR reaction. After a lag phase, the amount of PCR product increases gradually at first. In the exponential phase, beginning after about 16–18 cycles and continuing to approximately the 25th cycle, the amount of PCR product is

taken to be proportional to the amount of input DNA; quantitative PCR measurements are made on this basis. With further cycles, the amount of product increases at first but then tails off as the saturation phase approaches, when the reaction efficiency diminishes as reaction products increasingly compete with the remaining primer molecules for template DNA.

The end result is that millions of copies can be made of just the desired DNA sequence of interest within the complex starting DNA. By amplifying the desired sequence we can now study it in different ways—by directly sequencing the amplified DNA, for example.

PCR is very sensitive and can successfully amplify DNA fragments from tiny amounts of tissue samples and even from single cells. And it is robust enough to work on badly degraded tissue samples (and sometimes even samples fixed in formalin). As a result, there have been numerous applications in forensic and archaeological studies.

PCR can also be used to analyze RNA transcripts. In that case the RNA transcripts are first converted into complementary DNA (cDNA) with a reverse transcriptase (the process is called reverse transcription-PCR or RT-PCR).

Quantitative PCR and real-time PCR

In routine PCR, all that is required is to generate a detectable or usable amount of product. However, for some purposes there is a need to quantitate the amount of product. Some **quantitative PCR** methods give a *relative quantitation* of a sequence of interest within test samples and controls, and in [Chapter 11](#) we describe different diagnostic DNA screening methods that use PCR to get relative quantitation. Fluorescently labeled PCR products from the exponential phase of the PCR reaction ([Figure 3.4](#)) are removed and analyzed to measure the ratio of the fluorescence exhibited by the PCR product from a test sample (one that is associated with disease or is suspected as being abnormal) and the fluorescence exhibited by the PCR product from a control sample. The basis of the quantitation is that

during the exponential phase the amount of PCR product is proportional to the amount of target DNA sequence in the input DNA.

Real-time PCR is a form of quantitative PCR that can provide absolute quantitation (the absolute number of copies), as well as relative quantitation, and is performed in specialized PCR machines. Instead of waiting for the end of the reaction, the quantitation is performed while the PCR reaction is still progressing: the amplified DNA is detected as the PCR reaction proceeds in real time within the PCR machine. Important applications are found in profiling gene expression (using RT-PCR) and also in assays for altered nucleotides in DNA, as detailed in [Chapter 11](#).

3.3 PRINCIPLES OF NUCLEIC ACID HYBRIDIZATION

In a double-stranded DNA molecule, the hydrogen bonds between paired bases act as a fastening system that holds the two complementary DNA strands together. Two hydrogen bonds form between A and T in each A–T base pair, and three hydrogen bonds hold G and C together in each G–C base pair (see [Figure 1.3](#)). A region of DNA that is GC-rich (having a high proportion of G–C base pairs) is therefore more stable than a region that is AT-rich.

Each hydrogen bond is individually weak, but when base matching extends over many base pairs, the cumulative strength of the hydrogen bonds becomes quite strong. (As an analogy, think of Velcro®: a single Velcro hook and loop attachment is very weak, but thousands of them make for a strong fastening system.)

Double-stranded DNA can be manipulated in different ways to break the hydrogen bonds so that the two DNA strands are separated (**denaturation**). For example, if we heat the DNA to a high enough temperature (or expose it to strong concentrations of a highly polar molecule such as formamide or urea), the hydrogen bonds break and the two complementary DNA strands separate. Subsequent gradual cooling of heat-denatured DNA allows the

separated DNA strands to come together again, re-forming the base pairs in the correct order to restore the original double-stranded DNA ([Figure 3.5A](#)).

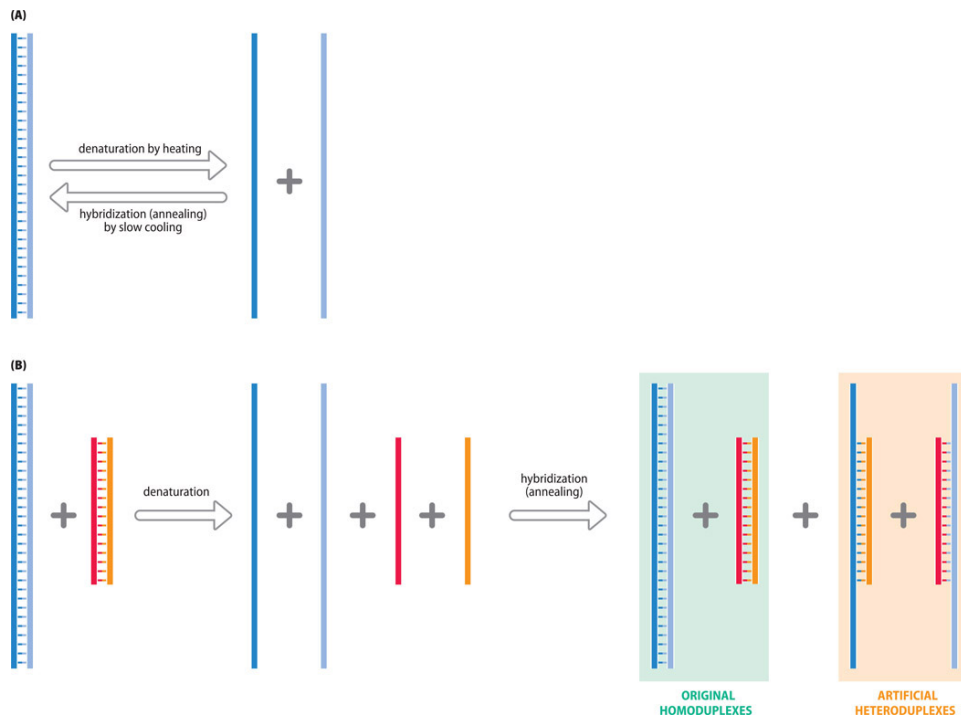


Figure 3.5 Denaturation and annealing of homologous DNA molecules to form artificial heteroduplexes and natural homoduplexes. (A) Denaturation means breaking of the hydrogen bonds in a double-stranded (duplex) nucleic acid and can be achieved by heating (or by exposure to highly polar chemicals such as urea and formamide). Under certain conditions, the separated strands can reassociate (hybridize or re-anneal) to reconstitute the original double-stranded DNA. (B) Artificial duplex formation between *homologous sequences* (in strands that have very similar nucleotide sequences) from two different DNA sources that have been denatured and mixed. For a proportion of the denatured DNA molecules, the original double-stranded DNAs re-form (homoduplexes), but in other cases artificial duplexes form between the partly complementary sequences.

Formation of artificial heteroduplexes

The association of any two complementary nucleic acid strands to form a double-stranded nucleic acid is known as nucleic acid **hybridization** (or

annealing). Under experimental conditions, two single nucleic acid strands with a high degree of base complementarity can be allowed to hybridize to form an artificial duplex. For example, if we mix cloned double-stranded DNA fragments that come from two different sources but have high levels of sequence identity, and then heat the mixture to disrupt all hydrogen bonding, all the millions of molecules of double-stranded DNA in the mixed samples from the two sources will be made single-stranded ([Figure 3.5B](#)).

Now imagine allowing the mixture to cool slowly: two different types of DNA duplex can form. First, a proportion of the single-stranded DNA molecules will base pair to their original partner to reconstitute the original DNA strands (homoduplexes). But in addition, sometimes a single-stranded DNA molecule will base pair to a complementary DNA strand in the DNA from the other source to form an artificial **heteroduplex** (see [Figure 3.5B](#)). (Note that we will use the term heteroduplex to cover all artificial duplexes in which base pairing is not perfect across the lengths of the two complementary strands. In the example in [Figure 3.5B](#) there is perfect base matching over the length of the small DNA strands but much of the blue strands remains unpaired. Very rarely, complementary DNA strands from two different sources might be generated that have both identical lengths and perfect base matching—if so, they could form artificial homoduplexes.)

The formation of artificial duplexes, almost always heteroduplexes, is the essence of the nucleic hybridization assays that are widely used in molecular genetics. For convenience we have illustrated cloned double-stranded DNAs in [Figure 3.5B](#). But as we will see below, the starting nucleic acids may sometimes include RNA (usually already single-stranded) or synthetic oligonucleotides as well as DNA. Often, too, one or both starting nucleic acids are complex mixtures of fragments, such as total RNA from cells or fragments of total genomic DNA. Like cloned DNA, the starting nucleic acids are usually isolated from millions of cells, and so individual sequences are normally present in many copies, often millions of copies.

Hybridization assays: using known nucleic acids to find related sequences in a test nucleic acid population

The object of a hybridization assay is to use a known nucleic acid population (the **probe**) to find related nucleic acid sequences within a poorly understood test sample. Such assays exploit the specificity of hybridization. Two single poly-nucleotide (DNA or RNA) or oligonucleotide strands will form a *stable* double-stranded hybrid (duplex) only if there is a significant amount of base pairing between them. The stability of the resulting duplex depends on the extent of base matching, and assay conditions can be chosen to allow perfectly matched duplexes only, or to allow degrees of base mismatching.

Hybridization assays can be performed in many different ways, with multiple applications in both research and diagnostics. But there is a common underlying principle: a *known*, well-characterized population of nucleic acid molecules or synthetic oligonucleotides (the *probe population*) is used to interrogate an imperfectly understood population of nucleic acids (the test sample). To do that, both nucleic acid populations must be separated into single strands and then mixed so that single probe strands can form artificial duplexes with complementary strands in the test sample.

After the probe has bound to complementary nucleic acid strands in the test sample, the resulting probe–test-sample heteroduplexes need to be identified in some way. To do that, two conditions are needed. First, either the probe or the test-sample nucleic acid population needs to be *labeled* at the outset with modified nucleotides containing some distinctive chemical group (such as one that can emit fluorescence—we describe how nucleic acids are labeled later on in [Box 3.2](#)). [Figure 3.6](#) gives one approach where the probe molecules are labeled. Secondly, there must be some way of separating the probe–test-sample heteroduplexes from labeled nucleic acid homoduplexes. As described below, that usually requires that the unlabeled nucleic acids be bound to some type of solid support.

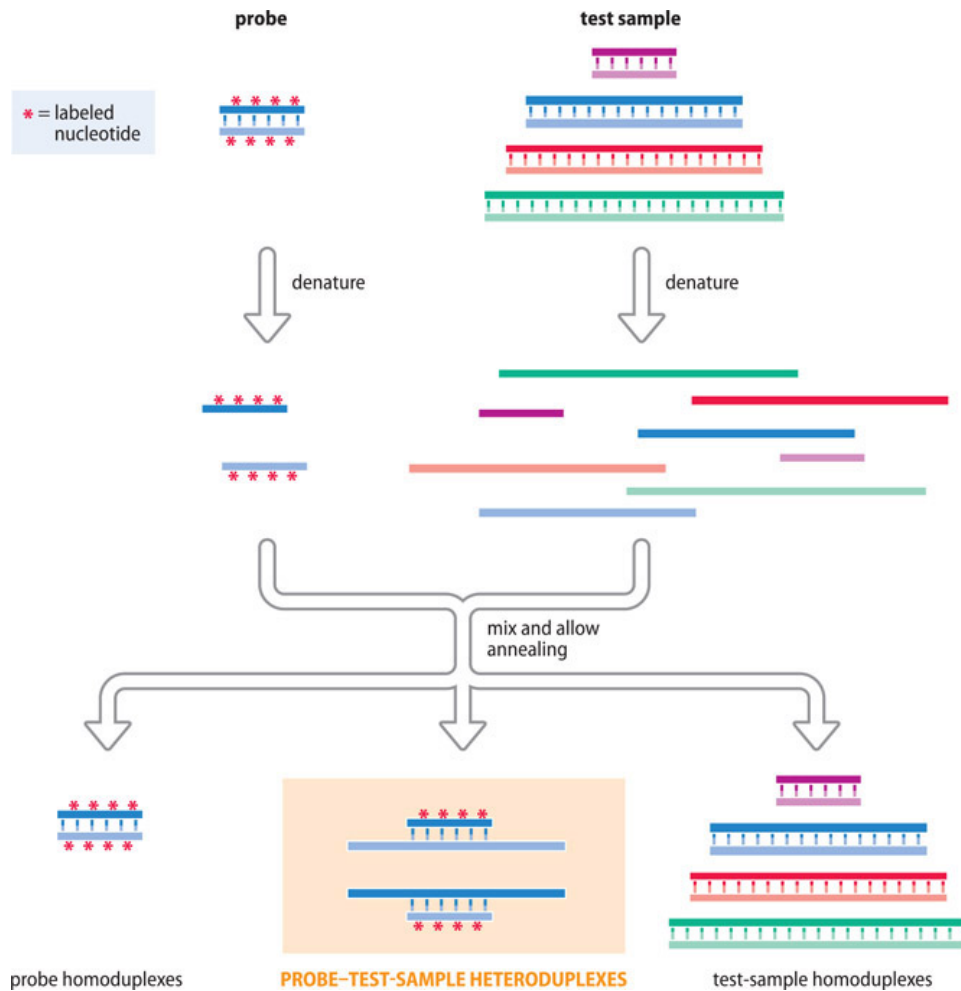


Figure 3.6 Heteroduplex formation in a nucleic acid hybridization assay. A defined probe population of known nucleic acid or oligonucleotide sequences and a test nucleic acid sample population are both made single-stranded (as required), then mixed and allowed to anneal. Many of the fragments that had previously been base paired in the two populations will reanneal to reconstitute original homoduplexes (bottom left and bottom right). In addition, new artificial duplexes will be formed between (usually) partly complementary probe and test-sample sequences (bottom center). The hybridization conditions can be adjusted to favor formation of the novel duplexes. In this way, probes can selectively bind to and identify closely related nucleic acids within a complex nucleic acid population. In this example, some kind of labeled nucleotide (*) has been introduced into the probe, but in some hybridization assays it is the test sample nucleic acid that is labeled.

Using high and low hybridization stringency

A hybridization assay can be used to identify nucleic acid sequences that are distantly related from a given nucleic acid probe. We might want to start with a DNA clone from a human gene and use that to identify the corresponding mouse gene. The human and mouse genes might be significantly different in sequence, but if we choose a long DNA probe and reduce the stringency of hybridization, stable heteroduplexes can be allowed to form even though there might be significant base mismatches ([Figure 3.7A](#)).

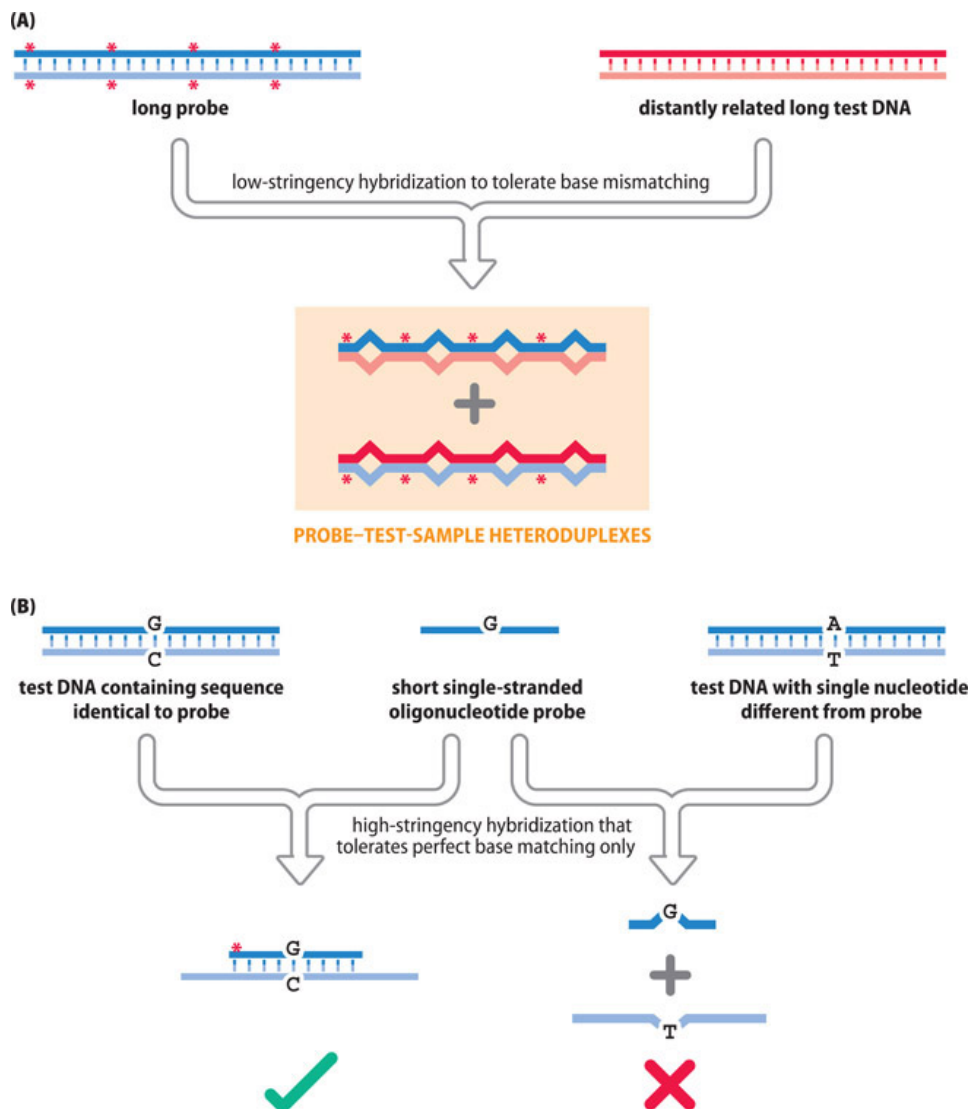


Figure 3.7 Using low or high hybridization stringency to detect nucleic acid sequences that are distantly related or show perfect base matching with a given probe. In any hybridization assay we can control the degree of base matching between complementary strands in the probe and test sample. If, for example, we increase salt concentrations and/or reduce the temperature, we lower hybridization stringency. (A) In some circumstances a long probe strand can form a thermodynamically stable duplex with a comparable but distantly related strand within the test DNA (or RNA), even though there might be significant base mismatching. (B) Alternatively, we can use high temperatures and low salt concentrations to achieve high hybridization stringency that might allow only perfect base matching. That is most easily achieved with a short oligonucleotide probe and allows assays to discriminate between alleles that differ at a single nucleotide position.

Conversely, we can choose hybridization conditions to accept only perfect base matching. If we choose an oligonucleotide probe, we can use a high hybridization stringency so that the only probe-test duplexes that can form are ones that contain exactly the same sequence as the probe ([Figure 3.7B](#)). That can happen because a single mismatch out of, say, 18 base pairs can make the duplex thermodynamically unstable. Oligonucleotides can therefore be used to identify alleles that differ by a single nucleotide (*allele-specific oligonucleotides*).

Two classes of hybridization assay

There are many types of hybridization assay, but they all fall into two broad classes. In one case the probe molecules are labeled and the test-sample molecules are unlabeled, as in [Figure 3.6](#). In that case, the probe is often a single type of cloned DNA and it is usually labeled by using a polymerase to synthesize complementary DNA or RNA strands in the presence of one or more fluorescently labeled nucleotides (**Box 3.2**). The alternative type uses unlabeled probe molecules and it is the test-sample molecules that are labeled (see below).

BOX 3.2 LABELING OF NUCLEIC ACIDS AND OLIGONUCLEOTIDES

Hybridization assays involve the labeling of either the probe or the test-sample population. Usually this involves making labeled DNA copies of a starting DNA or RNA with a suitable DNA polymerase in the presence of the four precursor deoxynucleotides (dATP, dCTP, dGTP, and dTTP). In the case of a starting RNA, a specialized DNA polymerase, a reverse transcriptase, uses the RNA as a template for making a complementary DNA copy. For some purposes, labeled RNA copies are made of a starting DNA using an RNA polymerase and the four precursor ribonucleotides (ATP, CTP, GTP, and UTP).

Whichever procedure is used, particular chemical groups (labels) are introduced into the DNA or RNA copies and can be specifically detected in some way. Often, at least one of the four nucleotide precursors has been modified so that it has a label attached to the base; alternatively, labeled oligonucleotide primers are incorporated.

Unlike DNA or RNA, oligonucleotides are chemically synthesized by the sequential addition of nucleotide residues to a starting nucleotide that will be the 3' terminal nucleotide. Amine or sulfhydryl groups can be incorporated into the oligonucleotide and can then be conjugated with amine-reactive or sulfhydryl-reactive labels.

Different labeling systems can be used ([Table 1](#)). Fluorescent dyes—such as derivatives of fluorescein—are popular; they can be detected readily because they emit fluorescent light of a defined wavelength when suitably stimulated. Some other labels are detected by specific binding to an antibody or to a very strongly interacting protein (see [Table 1](#)). In these cases, the detecting protein is usually conjugated to a fluorescent group (**fluorophore** or **fluorochrome**) or to an enzyme, such as alkaline phosphatase or peroxidase, which can permit detection via colorimetric assays or chemical luminescence assays.

TABLE 1 POPULAR SYSTEMS FOR LABELING NUCLEIC ACIDS

Labeling system	Examples of labels	Label detection
Fluorescence	FITC (fluorescein isothiocyanate)	using laser scanners/fluorescence microscopy
Antibody detection	digoxigenin (a steroid found in Digitalis plants)	via a digoxigenin-specific antibody that is coupled to a fluorophore or suitable enzyme
Specific protein interaction	biotin (= vitamin B7)	via streptavidin (a bacterial protein with an extraordinarily high affinity for biotin) that has been conjugated to a fluorophore or enzyme

The point of using labeled nucleic acids in a hybridization assay is to allow probe-test sample heteroduplexes to be identified. But how can we distinguish between the label in these duplexes and the label in the original labeled probe or labeled test-sample DNA? The answer is to immobilize the unlabeled nucleic acid population on a solid support (often plastic, glass, or quartz) and expose it to an aqueous solution of the labeled nucleic acid population. When labeled nucleic acid strands hybridize to complementary sequences on the solid support, they will be physically bound to the support, but labeled molecules that do not find a partner on the support or that stick nonspecifically can be washed off. That leaves behind the complementary partners that the assay is designed to find ([Figure 3.8](#)).

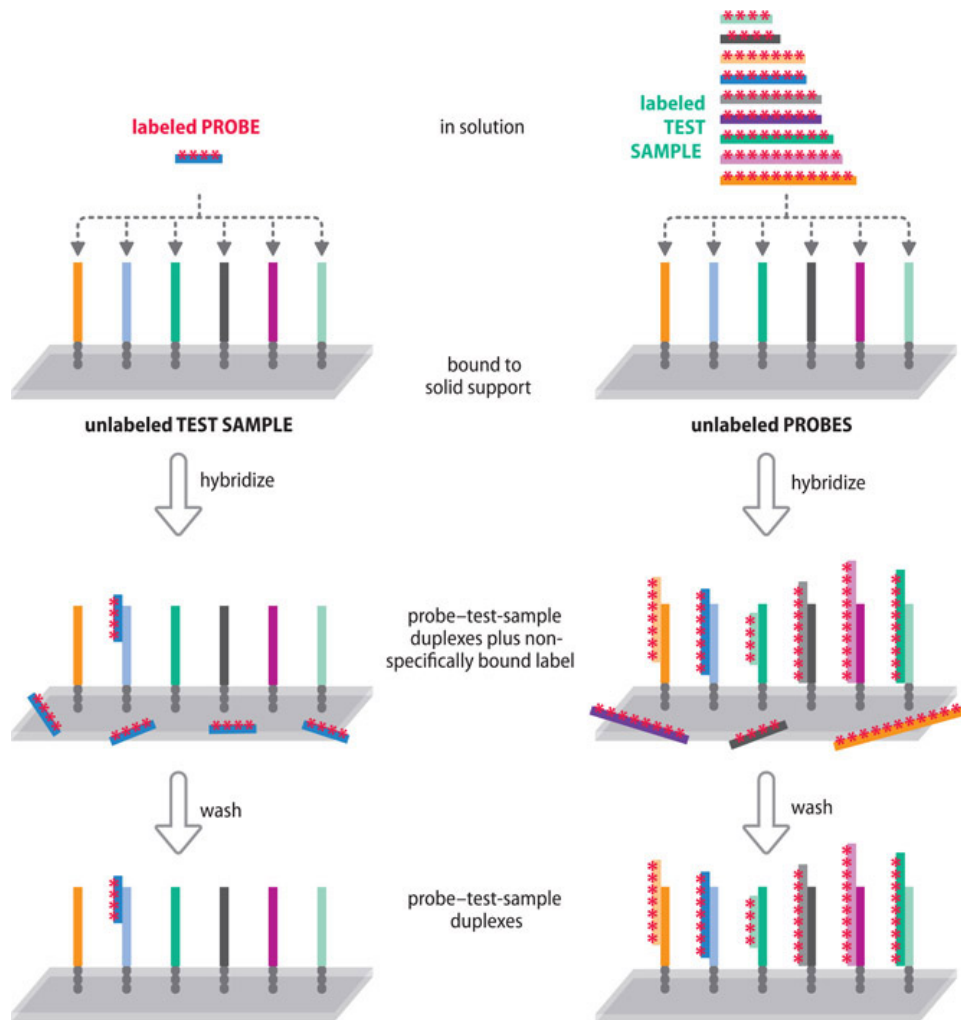


Figure 3.8 The two fundamental classes of hybridization assay and the use of solid supports to capture labeled probe-test sample duplexes. In both a standard hybridization assay, where the probe is labeled (left column), and a reverse hybridization assay, where the test sample is labeled (right column), the unlabeled nucleic acid/oligonucleotide population is bound to a solid support and denatured, before being exposed to an aqueous solution of the labeled nucleic acid/oligonucleotide population that has also, as required, been denatured. Single-stranded molecules in the labeled population can hybridize to complementary sequences in the unlabeled population, and so become bound to the solid support. Other labeled sequences that have not bound, or have bound nonspecifically at incorrect locations on the support, can be washed off. In the past, most hybridization assays were standard assays (see [Table 3.2](#) for examples), but reverse hybridization assays became popular after microarray hybridization was developed.

TABLE 3.2 EXAMPLES OF STANDARD HYBRIDIZATION ASSAYS

Probe and test sample labeling	Hybridization method	Applications	Examples
Labeled probe and unlabeled test sample (Figures 3.6 and 3.7A)	Southern blot	looking for medium-sized changes (hundreds of base pairs to several kilobases) in genes/DNA in test sample	Clinical Box 3 Figure 2 on page 171
	tissue <i>in situ</i>	tracking RNA transcripts in tissues and embryos	
	chromosome <i>in situ</i>	studying large-scale changes using fixed chromosomes on a slide as the test sample	Figures 10.7A and 11.4

Standard hybridization assays have been used for different purposes (**Table 3.2** gives some examples). For decades, almost all hybridization assays used a homogeneous labeled probe (often, a single type of DNA clone) to search for related sequences in an immobilized complex test nucleic acid sample (see [Figure 3.6](#) and the left part of [Figure 3.8](#)). As described in the next section, microarray-based hybridization assays use a reverse type of hybridization where unlabeled complex probe populations bound to a surface are used to interrogate a labeled test sample (the right column of [Figure 3.8](#) shows the principle).

Microarray hybridization: large-scale parallel hybridization to immobilized probes

Innovative and powerful hybridization technologies developed in the early 1990s permit numerous hybridization assays to be conducted simultaneously on a common sample under the same conditions. A DNA or oligonucleotide microarray consists of many thousands or millions of

different unlabeled DNA or oligonucleotide probe populations that have been fixed to a glass or other suitable surface within a high-density grid format. Within each grid square are millions of identical copies of just one probe (a grid square with its probe population is called a *feature*). For example, oligonucleotide microarrays often have a 1.28 cm × 1.28 cm surface that contain millions of different features, each occupying about 5 or 10 μm² ([Figure 3.9](#)).

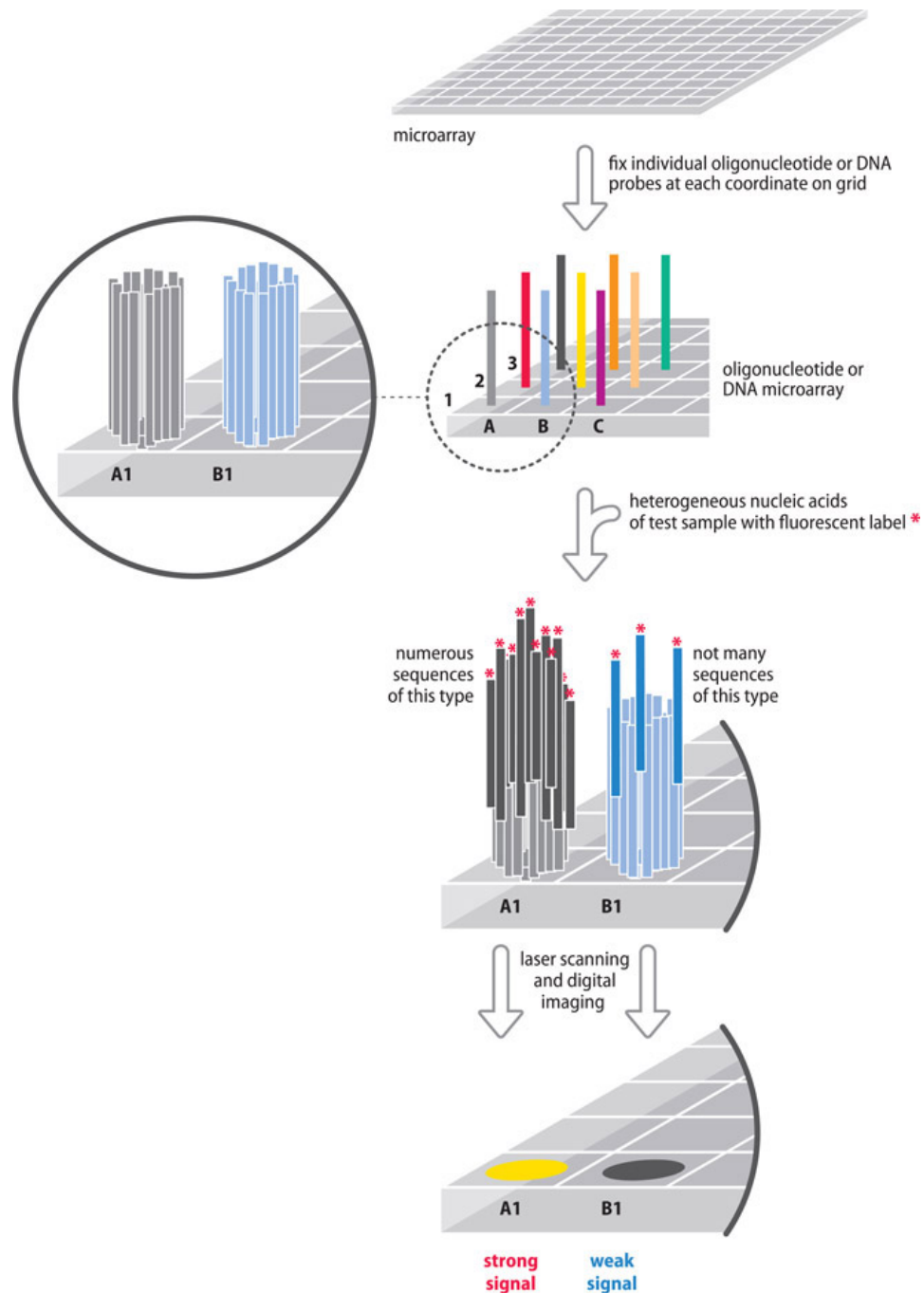


Figure 3.9 Principle of microarray hybridization. A microarray is a solid surface on which molecules can be fixed at specific coordinates in a high-density grid format. Oligonucleotide or DNA microarrays have thousands to millions of different synthetic single-stranded oligonucleotide or DNA probes fixed at specific predetermined positions in the grid. As shown by the expanded item enclosed within dashed lines, each grid square will have many thousands of identical copies of a single type of

oligonucleotide or DNA probe (a *feature*). An aqueous test sample containing a heterogeneous collection of labeled DNA fragments or RNA transcripts is denatured and allowed to hybridize with the probes on the array. Some probes (for example the A1 feature) may find numerous complementary sequences in the test population, resulting in a strong hybridization signal; for other probes (for example the B1 feature) there may be few complementary sequences in the test sample, resulting in a weak hybridization signal. After washing and drying of the grid, the hybridization signals for the numerous different probes are detected by laser scanning, giving huge amounts of data from a single experiment. (For ease of illustration, we show test-sample nucleic acids with end labels, but sometimes they contain labels on internal nucleotides.)

A test sample—an aqueous solution containing a complex population of fluorescently labeled denatured DNA or RNA—is hybridized to the different probe populations on the microarray. After a washing step to remove nonspecific binding of labeled test-sample molecules to the array, the remaining bound fluorescent label is detected with a high-resolution laser scanner. The signal emitted from each feature on the array is analyzed with digital imaging software that converts the fluorescent hybridization signal into one of a palette of colors according to its intensity ([Figure 3.9](#)).

Because the intensity of each hybridization signal reflects the number of labeled molecules that have bound to a feature, microarray hybridization is used to *quantitate* different sequences in complex test-sample populations such as different samples of genomic DNA or total cellular RNA (or cDNA). Frequent applications include quantifying different transcripts (expression profiling) and also scanning genomes to look for large-scale deletions and duplications, as described in [Chapter 11](#).

3.4 PRINCIPLES OF DNA SEQUENCING

DNA sequencing is the ultimate DNA test. Until quite recently, Sanger dideoxy DNA sequencing was the predominant method. It relies on amplifying individual DNA sequences. For each amplified DNA, nested

sets of labeled DNA copies are made and then separated according to size by gel electrophoresis.

In the last few years completely different technologies have allowed massively parallel DNA sequencing. No attempt is made to obtain the sequence of just a purified DNA component; instead, millions of DNA fragments present in a complex DNA sample are simultaneously sequenced without the need for gel electrophoresis.

Dideoxy DNA sequencing remains widely used for investigating specific DNA sequences, for example testing whether individuals have mutations in a particular gene. What the newer DNA sequencing technologies offer is a marked increase in sequencing capacity and the ability to sequence complex DNA populations, such as genomic DNA sequences, very rapidly. As a result of fast-developing technology, the running costs of DNA sequencing are plummeting, and very rapid sequencing of whole genomes is quickly becoming routine.

Dideoxy DNA sequencing

Like PCR, dideoxy DNA sequencing (also called Sanger sequencing) uses primers and a DNA polymerase to make DNA copies of specific DNA sequences of interest. To obtain enough DNA for sequencing, the DNA sequences are amplified by PCR (or sometimes by cloning in cells). The resulting purified DNAs are then sequenced, one after another, in individual reactions. Each reaction begins by denaturing a selected purified DNA. A single oligonucleotide primer is then allowed to bind and is used to make labeled DNA copies of the desired sequence (using a provided DNA polymerase and the four dNTPs).

Instead of making full-length copies of the sequence, the DNA synthesis reactions are designed to produce a population of DNA fragments sharing a common 5' end sequence (defined by the primer sequence) but with variable 3' ends. This is achieved by simultaneously having the standard dNTP precursors of DNA plus low concentrations of ddNTPs, dideoxynucleotide analogs that differ from a standard deoxynucleotide only

convenient to use labeled ddNTPs that have different fluorescent groups according to the type of base, as shown here. The DNA copies will have a common 5' end (defined by the sequencing primer) but variable 3' ends, depending on where a labeled dideoxynucleotide has been inserted, producing a nested set of DNA fragments that differ by a single nucleotide in length. A series of nested fragments that differ incrementally by one nucleotide from their common 5' end are fractionated according to size by gel electrophoresis; the fluorescent signals are recorded and interpreted to produce a linear base sequence. (C) Example of DNA sequence output, showing a succession of dye-specific (and therefore base-specific) intensity profiles. This example shows a cDNA sequence from the *PHC3* polyhomeotic gene, provided by E. Tonkin, Newcastle University.

DNA synthesis continues smoothly when dNTPs are used, but once a dideoxynucleotide has been incorporated into a growing DNA molecule, chain extension is immediately terminated (the dideoxynucleotide lacks a 3'-OH group to form a phosphodiester bond). To keep the balance tilted toward chain elongation, the ratio of each ddNTP to the corresponding dNTP is set to be about 1:100, so that a dideoxynucleotide is incorporated at only about 1 % of the available nucleotide positions.

If we consider competition between ddATP and dATP in the example in [Figure 3.10B](#), there are four available positions for nucleotide insertion: opposite the T at nucleotide positions 2, 5, 13, and 16 in the starting DNA. Because the DNA synthesis reaction results in numerous DNA copies, then by chance some copies will have a dideoxyA incorporated opposite the T at position 2, some will have a dideoxyA opposite the T at position 5, and so on. Effectively, chain elongation is *randomly* inhibited, producing sets of DNA strands that have a common 5' end but variable 3' ends.

Fluorescent dyes are used to label the DNA. One convenient way of doing this, as shown in [Figure 3.10B](#), is to arrange matters so that the four different ddNTPs are labeled with different fluorescent dyes. The reaction products will therefore consist of DNA strands that have a labeled dideoxynucleotide at the 3' end carrying a distinctive fluorophore according to the type of base incorporated.

All that remains is to separate the DNA fragments according to size by electrophoresis ([Box 3.3](#)) and to detect the fluorescence signals. In modern dideoxy sequencing, as the DNA fragments migrate in the gel, they pass a laser that excites the fluorophores, causing them to emit fluorescence at distinct wavelengths. The fluorescence signals are recorded and an output is provided in the form of intensity profiles for the differently coloured fluorophores, as shown in [Figure 3.10c](#).

BOX 3.3 SLAB GEL ELECTROPHORESIS AND CAPILLARY GEL ELECTROPHORESIS FOR SEPARATING NUCLEIC ACIDS ACCORDING TO SIZE

Nucleic acids carry numerous negatively charged phosphate groups and will migrate toward the positive electrode when placed in an electric field. By arranging for them to migrate through a porous gel during electrophoresis, nucleic acid molecules can be fractionated according to size. The porous gel acts as a sieve: small molecules pass easily through the pores of the gel, but larger fragments are impeded by frictional forces.

Standard gel electrophoresis with agarose gels allows the fractionation of moderately large DNA fragments (usually from about 0.1 kb to 20 kb). Pulsed-field gel electrophoresis can be used to separate much larger DNA fragments (up to megabases long). It uses specialized equipment in which the electrical polarity is regularly changed, forcing the DNA molecules to alter their conformation periodically in preparation for migrating in a different direction. Polyacrylamide gel electrophoresis allows the superior resolution of smaller nucleic acids (it is usually used to separate fragments in size ranges up to 1 kb) and is used in dideoxy DNA sequencing to separate fragments that differ in length by just a single nucleotide.

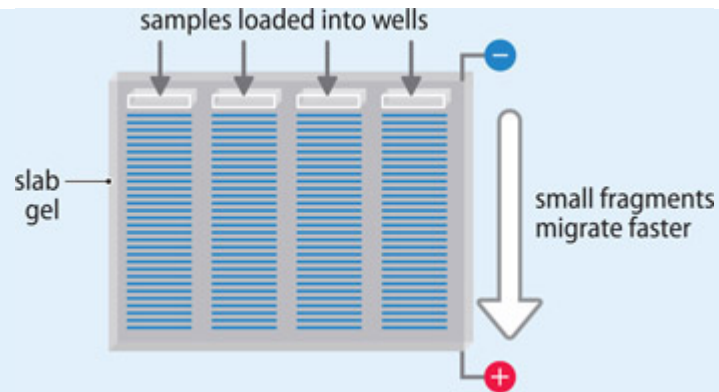


Figure 1 Slab gel electrophoresis.

In slab gel electrophoresis, individual samples are loaded into cut-out wells at one end of a solid slab of agarose or polyacrylamide gel. They migrate in parallel lanes toward the positive electrode ([Figure 1](#)). The separated nucleic acids can be detected in different ways. For example, after the end of an electrophoresis run, the gels can be stained with chemicals such as ethidium bromide or SYBR green that bind to nucleic acids and fluoresce when exposed to ultraviolet radiation. Sometimes the nucleic acids are labeled with fluorophores before electrophoresis, and during electrophoresis a recorder detects the fluorescence of individual labeled nucleic acid fragments as they sequentially pass a recorder placed opposite a fixed position in the gel.

The disadvantage of slab gel electrophoresis is that it is labor-intensive. The modern trend is to use capillary gel electrophoresis, which is largely automated. Fluorescently labeled DNA samples migrate through individual long and very thin tubes containing polyacrylamide gel, and a recorder detects fluorescence emissions as samples pass a fixed point ([Figure 2](#)). Modern dideoxy DNA sequencing uses capillary electrophoresis, as do many different types of diagnostic DNA screening methods that we outline in [Chapter 11](#).

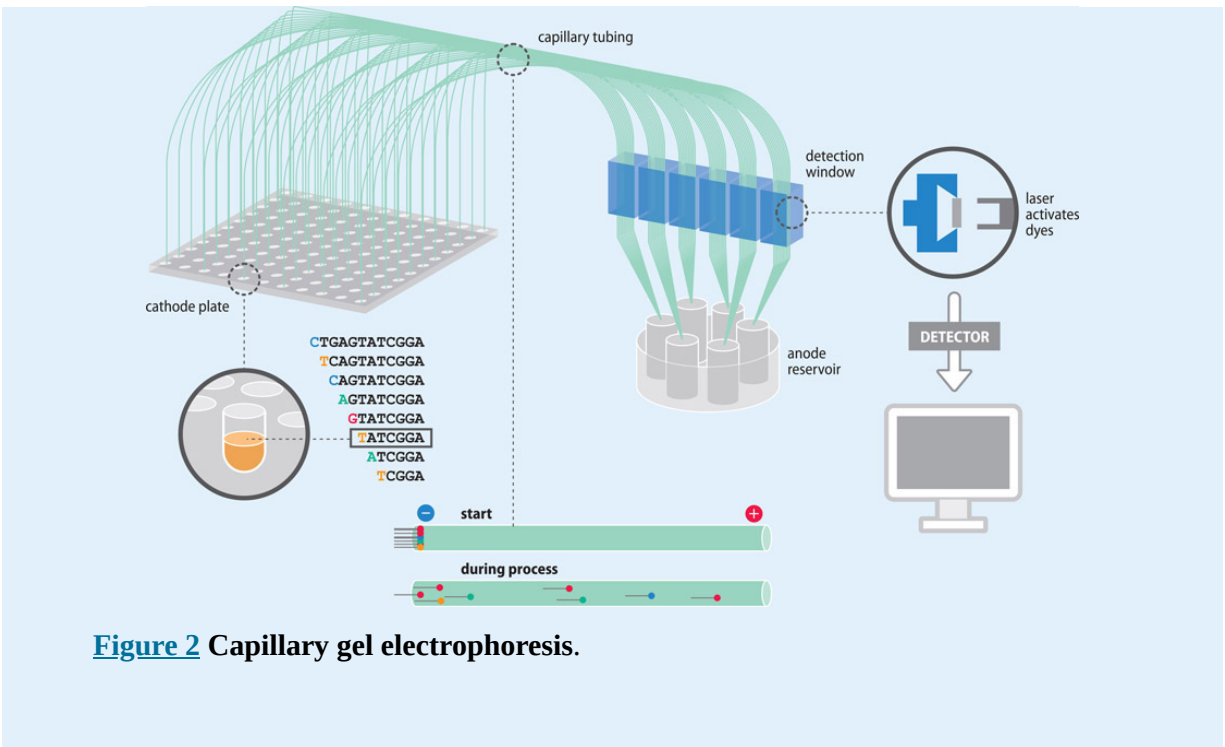


Figure 2 Capillary gel electrophoresis.

Dideoxy DNA sequencing is disadvantaged by relying on gel electrophoresis (slab polyacrylamide gels were used initially; more modern machines use capillary electrophoresis [Box 3.3]). Because gel electrophoresis is not suitable for handling large numbers of samples at a time, dideoxy sequencing has a limited sequence capacity. It is therefore not well suited to genome sequencing (although it has been used in the past to obtain the first human genome sequences). In modern times it is often used for analyzing variation over small DNA regions, such as regions encompassing individual exons.

Massively parallel DNA sequencing (next-generation sequencing)

In the early to mid-2000s new sequencing-by-synthesis methods were developed that could record the DNA sequence while the DNA strand is being synthesized. That is, the sequencing method was able to monitor the incorporation of each nucleotide in the growing DNA chain and to identify which nucleotide was being incorporated at each step.

The new sequencing technologies, often called next-generation sequencing (NGS), represent a radical step-change in sequencing technology. Standard dideoxy sequencing is a highly targeted method requiring the purification of specific sequences of interest that are then selected to be sequenced, one after another. By contrast, massively parallel DNA sequencing is indiscriminate: all of the different DNA fragments in a complex starting DNA sample can be *simultaneously* sequenced without any need for gel electrophoresis. The difference in sequencing output is therefore vast. As listed in [Table 3.3](#), various NGS technologies are commercially well established. Some of them require amplification of the starting DNA; others rely on unamplified starting DNA (“single-molecule sequencing”).

TABLE 3.3 MAJOR CHARACTERISTICS OF SOME COMMERCIALY AVAILABLE DNA SEQUENCING TECHNOLOGIES

Technology class	Sequencing platform	Read length (nucleotides)	Throughput (DNA sequence per run)
Conventional (dideoxy chain termination sequencing)	ABI prism 3730 Sanger dideoxy sequencing	~700	65 kb
Massively parallel sequencing of PCR-amplified DNAs	Illumina/Solexa NextSeq 2000	300	300 Gb
	Life Technologies Ion Torrent	200	50Gb
Massively parallel sequencing of unamplified (single-	Pacific Biosciences Sequel II	~25000	~2Gb

molecule) DNAs

Technology class	Sequencing platform	Read length (nucleotides)	Throughput (DNA sequence per run)
	Oxford Nanopore MinION	> 2 000 000	30 Gb*

* Per flow cell

They vary in different parameters, such as read lengths (the length of DNA sequence generated per starting DNA), run lengths, and the number of different DNA sequences that can be conducted in parallel.

By comparison with the standard Sanger dideoxy sequencing, the NGS methods generally have high intrinsic error rates in base calling but the final reported sequences are much more accurate than the initial reads (after quality filtering and comparison of multiple sequence reads). And, importantly, they have significantly cheaper running costs per base (but are not suited to low-capacity sequencing).

A variety of additional single-molecule sequencing technologies are currently also being piloted, and sequencing capacity is likely to be increased in the near future, with yet further decreases in sequencing costs. We will describe two widely used massively parallel DNA sequencing technologies in [Chapter 11](#).

SUMMARY

- In complex genomes, an individual gene, exon, or other sequence of interest is often a tiny fraction of the genome. To study a specific short DNA sequence like this either we must first *purify* it by selectively amplifying its copy number using some a DNA polymerase) or use some method to specifically *track* the sequence.

- Making multiple copies of a DNA sequence can be done within cells (DNA cloning), or in a cell-free system (notably by using PCR).
- In DNA cloning, the DNA sequence of interest is first attached to a vector DNA molecule that can self-replicate in a suitable host cell (often a bacterial cell). Vector molecules are modified DNAs that can readily replicate in the host cell, such as small circular plasmids or different types of bacteriophage.
- Restriction nucleases are used to cut large DNA molecules, such as chromosomal DNAs, into small pieces of discrete sizes that can easily be joined to similarly cut vector molecules, producing recombinant DNA molecules.
- Recombinant DNA molecules can be induced to enter a suitable host cell (transformation). Transformation is selective: each transformed cell has normally taken up a *single* DNA molecule. A transformed bacterial cell can multiply many times, and large numbers of identical copies of the recombinant DNA are produced.
- A DNA library is a bank of DNA clones that collectively include many different DNA sequences representing a complex starting population of genomic DNA (or cDNA copies of a complex RNA population).
- In PCR, a DNA sequence of interest can be copied many times from a complex source of DNA by *in vitro* DNA synthesis. Specific oligonucleotide primers are designed to bind to the starting DNA at positions flanking the sequence of interest and then used to make DNA copies that can themselves serve as templates for making further copies, rapidly increasing the copy number of the sequence of interest.
- Nucleic acid hybridization is the key method used to track a DNA or RNA sequence of interest. The method relies on the specificity of base pairing—if two different nucleic acids are

related in sequence, they may be able to form an artificial heteroduplex that is stable under selected experimental conditions.

- To perform nucleic acid hybridization, a test nucleic acid population with some sequence or sequences of interest is made single-stranded (denatured) and mixed with a probe population of known denatured nucleic acids. The object is to identify heteroduplexes in which a single-stranded sequence of interest in the test sample has formed a stable hybrid with a known sequence within the probe population.
- In many types of nucleic acid hybridization, a homo geneous labeled probe population is used to identify related sequences in an unlabeled test population that is typically bound to a solid surface.
- In microarray hybridization, many thousands of unlabeled oligonucleotide probe populations are attached to a solid surface in a regular grid formation and hybridized in parallel with a labeled test nucleic acid population provided in solution. According to the amount of labeled DNA bound to each type of oligonucleotide, it is possible to quantitate specific sequences that are complementary to each of the different probes.
- In DNA sequencing, DNA samples are made single-stranded and a DNA polymerase is used to synthesize a complementary DNA in a way that provides a readout of the base sequence.
- In standard dideoxy DNA sequencing, selected individual DNA samples are sequenced. The DNA synthesis step uses a mixture of normal and chain-terminating nucleotides, producing a nested set of fragments that differ incrementally by one nucleotide and that can be separated by gel electrophoresis.

- In massively parallel DNA sequencing (next-generation sequencing), a complex population of very many (often millions of) DNA templates are sequenced simultaneously and indiscriminately. There is no gel electrophoresis. Instead, the methods rely on being able to monitor which of the four nucleotides is being incorporated at each step in synthesizing the cDNA.

QUESTIONS

Questions can be downloaded by visiting the following link, under Support Materials: www.routledge.com/9780367490812.

FURTHER READING

Brown TA (2016) *Gene cloning and DNA analyses. An Introduction*, 7th ed. Wiley-Blackwell.

Geschwind DH (2003) DNA microarrays: translation of the genome from laboratory to clinic. *Lancet Neurol* 2:275–282; PMID 12849181.

Goodwin S (2016) Coming of age: ten years of next-generation sequencing technologies. *Nature Rev Genet* 17:333–351; PMID 27184599.

4

Principles of genetic variation

DOI: [10.1201/9781003044406-4](https://doi.org/10.1201/9781003044406-4)

CONTENTS

[4.1 DNA SEQUENCE VARIATION ORIGINS AND DNA REPAIR](#)

[4.2 POPULATION GENOMICS AND THE SCALE OF HUMAN
GENETIC VARIATION](#)

[4.3 FUNCTIONAL GENETIC VARIATION AND PROTEIN
POLYMORPHISM](#)

[4.4 EXTRAORDINARY GENETIC VARIATION IN THE IMMUNE
SYSTEM](#)

[SUMMARY](#)

[QUESTIONS](#)

[FURTHER READING](#)

The human genome reference sequence, described in [Section 2.3](#), is just a single representative snapshot of our genome, and an artificial one (different parts of the reference sequence originated from different people). We may speak about *the* human genome, but, in reality, there are many billions of different human genomes that owe their differences to genetic variation.

Genetic variation is very largely inherited, transmitted between generations in gametes. During life, every fertile man makes billions of sperm cells, but each sperm cell—and each egg cell—is genetically unique (pre-existing genetic variation is shuffled at meiosis by recombination and independent chromosome assortment). As a result, every one of us arose from a single fertilized egg cell that contained a unique diploid genome. (Occasionally, however, splitting of the very early embryo generates genetically identical embryos that can give rise to twins or, rarely, triplets.)

The vast majority of our genetic information is stored in the nuclear DNA molecules contained within our chromosomes. Each of us inherits two different haploid nuclear genomes, a paternal genome and a maternal genome, and so inherited genetic variation occurs within, as well as between, individuals. At any genetic **locus** (DNA region having a unique chromosomal location) the maternal and paternal DNA sequences (**alleles**) normally have identical or slightly different DNA sequences (we are said to be **homozygotes** if the alleles are identical, or **heterozygotes** if they differ by even a single nucleotide).

Two regions of the human genome are always inherited from a single parent. The nonrecombining portion of the Y chromosome has no sequence counterparts on the X chromosome and is transmitted exclusively by fathers to sons. Men are said to be **hemizygous** for sequences in this region, having inherited just a single allele at each locus. And all of us inherit the tiny mitochondrial DNA (mtDNA) exclusively from our mothers. (The transmitted maternal egg, however, contains about 100 000 mtDNA molecules that may show some differences in sequence; this type of mitochondrial DNA sequence variation is described as *heteroplasmy*.)

The genetic variation inherited in the fertilized egg, and present in all our nucleated cells is known as *constitutional variation*. Additional changes occur in the DNA of each of our nucleated cells throughout life, constituting *post-zygotic* or *somatic* genetic variation. (Note, however, that the qualifying term *somatic* can be used to describe genetic variation in two ways [see [Table 4.1](#)]).

TABLE 4.1

SOME COMMON WAYS OF CLASSIFYING GENETIC VARIATION

Classification according to:	Type	Description
Timing during development	Constitutional	Present in the zygote and transmitted to descendent cells
	Post-zygotic [*]	Not in the zygote, but occurring in some post-zygotic cell and transmitted to descendants of that cell
Possibility of transfer	Germline	Occurring in gametes or in any direct precursor cell.
to next generation	Somatic* (=non-germline)	Occurring in cells other than gametes or their direct precursors (such as lymphocytes, neurons, and so on)
Changes in different copies of same DNA	Allelic	Variation between maternal and paternal copies of the same chromosomal DNA sequence in a person
sequence	Heteroplasmic	Variation between different copies of the mtDNA sequence in a

^{*}

A post-zygotic mutation is often loosely described as somatic but some occur in precursor cells in the germ line.

Classification according to:	Type	Description
		person

*

A post-zygotic mutation is often loosely described as somatic but some occur in precursor cells in the germ line.

Most post-zygotic DNA changes occur in a rather random fashion, causing small differences in the DNA within different body cells. However, *programmed* DNA changes are also programmed to occur in certain genes in some cells, notably in maturing B and T cells (to ensure that each of us can make huge numbers of different antibodies and different T-cell receptors).

Individuals differ from each other mostly because our DNA sequences differ, but genetic variation is not the only explanation for differences in **phenotype** (our observable characteristics). A fertilized egg cell can split in two in early development and give rise to genetically identical twins (monozygotic twins) that nevertheless grow up to be different: although hugely important, genetic variation is not the only influence on phenotype. During development, additional effects on the phenotype occur by a combination of stochastic (random) factors, differential gene-environment interactions and additional *epigenetic* variation that is not attributable to changes in base sequence. We consider epi-genetic effects and environmental factors in later chapters when we examine how our genes are regulated.

In this chapter we look at general principles of human genetic variation and how variation in DNA relates to variation in the sequences of proteins and non-coding RNAs. We are not concerned here with the very small fraction of genetic variation that causes disease. That will be covered in later chapters, especially in [Chapter 7](#) (where we look primarily at genetic variation in relation to monogenic disorders), [Chapter 8](#) (genetic variation

in relation to complex inherited diseases), and [Chapter 10](#) (genetic variation and cancer).

In [Section 4.1](#) we consider the origins of DNA sequence variation. We take a broad look at the extent of human genetic variation in [Section 4.2](#) and at the different forms in which this variation manifests. In [Section 4.3](#) we deal with functional genetic variation. Here, we examine in a general way how variation in the sequences of protein products is determined both by genetic variation and by post-transcriptional modification. In this section we also deal with aspects of population genetics that relate to the spread of advantageous DNA variants through human populations (but the population genetics of harmful disease-associated DNA variants is examined in later chapters).

Genetic variation is most highly developed in genes that work in recognizing foreign, potentially harmful, molecules that have been introduced into the body. These molecules are often under independent genetic control, as in the case of infection by microbial pathogens. When that happens, two types of Darwinian natural selection may oppose each other. Thus, natural selection works on us to maximize genetic variation in the frontline immune system genes needed to recognize antigens on the invaders. Some genetic variants in these genes will be more effective than others; accordingly, some individuals in the population will be more resistant than others to the potential harmful effects of specific microbial pathogens. But natural selection also works on the microbial pathogens to maximize genetic diversity of external molecules in an effort to escape detection by our immune defense systems.

As described in [Section 4.4](#), the frontline genes in our immune system defenses need to recognize a potentially huge number of foreign antigens. Here, we describe the basis for the quite exceptional variability of some human leukocyte antigen (HLA) proteins and the medical significance of this variability. We also consider how exceptional post-zygotic genetic variation is created at our immunoglobulin and T-cell receptor loci so that an individual person can make a huge number of different antibodies and T-cell receptors.

4.1 DNA SEQUENCE VARIATION ORIGINS AND DNA REPAIR

Underlying genetic variation are changes in DNA sequences. **Mutation** describes both a process that produces altered DNA sequences (either a change in the base sequence or in the number of copies of a specific DNA sequence) and the outcome of that change (the altered DNA sequence). As events, mutations can occur at a wide variety of levels, and can have different consequences. They may contribute to a normal phenotype (such as height) or to a disease phenotype, and very rarely, they may have some beneficial effect. However, as explained below, the great majority of mutations have no obvious effect on the phenotype.

Mutations originate as a result of changes in our DNA that are not corrected by cellular DNA repair systems. The DNA changes are occasionally induced by exposure to certain environmental mutagens that include certain types of radiation (notably ionizing radiation and excessive ultraviolet irradiation), and also certain chemicals that we come into contact with. However, the great majority of mutations arise from endogenous sources: both spontaneous errors in normal chromosome and DNA function and also spontaneous chemical damage to DNA.

Mutations are inevitable. They may have adverse effects on individual organisms, causing aging and contributing to many human diseases. But they also provide the raw fuel for natural selection of beneficial adaptations that allow evolutionary innovation and, ultimately, the origin of new species.

Genetic variation arising from errors in chromosome and DNA function

Natural errors in various processes that affect chromosome and DNA function—chromosome segregation, recombination, and DNA replication—are important contributors to genetic variation. No cellular function can occur with 100 % efficiency and occasional mistakes are inevitable. Errors

in the above processes may often not have harmful consequences, but some make important contributions to disease. We examine in detail how they can cause disease in [Chapter 7](#); in this section we take a broad look into how they affect genetic variation in general.

DNA replication errors

General errors in DNA replication are unavoidable. Each time the DNA of a human diploid cell replicates, six billion nucleotides need to be inserted in the correct order to make new DNA molecules. Not surprisingly, DNA polymerases very occasionally insert the wrong nucleotide, resulting in mispaired bases (a base mismatch; the likelihood of such an error simply reflects the relative binding energies of correctly paired bases and mispaired bases).

In the great majority of cases, the errors are quickly corrected by the DNA polymerase itself. The major DNA polymerases engaged in replicating our DNA have an intrinsic 3' → 5' exonuclease activity with a *proofreading function*. If, by error, the wrong base is inserted, the polymerase's 3' → 5' exonuclease is activated and degrades the newly synthesized DNA strand from its 3' end, removing the wrongly inserted nucleotide and a short stretch before it. Then the DNA polymerase resumes synthesis again. If mispaired bases are not eliminated by the DNA polymerase, a DNA mismatch repair system is activated.

Another type of DNA replication error commonly occurs within regions of DNA where there are short tandem oligonucleotide repeats. If, for example, the DNA polymerase encounters a 30-nucleotide sequence with 15 sequential repeats of the AT dinucleotide, or 10 sequential repeats of the CAA trinucleotide, there will be an increased chance that during DNA replication a mistake is made in aligning the growing DNA strand with its template strand. A frequent result is that the template strand and newly synthesized strand pair up out of register by one (or sometimes more) repeat units, causing **replication slippage**, as detailed below. Errors like this are

also often repaired successfully by the DNA *mismatch repair* system. We detail this repair system in [Chapter 10](#) because of its importance in understanding cancer.

Although the vast majority of DNA changes caused by DNA replication errors are identified and corrected, some persist. We have many very effective DNA repair pathways, but DNA repair is also not 100 % effective: unrepaired changes in DNA sequence are an important source of mutations.

Chromosome segregation and recombination errors

Errors in chromosome segregation result in abnormal gametes, embryos, and somatic cells that have fewer or more chromosomes than normal and so have altered numbers of whole DNA molecules. Changes in chromosomal DNA copy number are not uncommon. If they occur in the germ line they often cause embryonic lethality or a congenital disorder (such as Down syndrome, which is commonly caused by an extra copy of chromosome 21), but changes in copy number of sex chromosomes are more readily tolerated. In somatic cells, changes in chromosomal DNA copy number are a common feature of many cancers.

Various natural errors can also give rise to altered copy number of a specific sequence within a DNA strand that may range up to megabases in length. That can occur by different recombination (and recombination-like) mechanisms in which nonallelic (but often related) sequences align so that chromatids are paired with their DNA sequences locally out of register. Subsequent crossover (or sister chromatid exchange) produces chromatids with fewer or more copies of the sequences. The ensuing duplication or deletion of sequences may, or may not, have functional consequences—we cover the mechanisms and how they can result in disease in [Chapter 7](#).

Various endogenous and exogenous sources can cause damage to DNA by altering its chemical structure

DNA is a comparatively stable molecule. Nevertheless, there are constant threats to its integrity, causing breakage of covalent bonds within DNA or inappropriate bonding of chemicals to DNA. Most of the damage originates spontaneously within cells (normal cellular metabolism generates some chemicals that are harmful to cells). A minority of the damage is induced by external sources.

Chemical damage to DNA can involve the cleavage of covalent bonds in the sugar-phosphate backbone of DNA, causing single-strand or double-strand breaks. Alternatively, bases are deleted (by cleavage of the N-glycosidic bond connecting a base to a sugar) or they are chemically modified in some way: certain chemical groups on bases may be replaced, or chemical groups may be added to bases, or unusual covalent bonds may form between two bases on the same strand or on complementary strands (*cross-linking*)—see [Figure 4.1](#) for examples. The chemically modified bases may block DNA or RNA polymerases, and cause base mispairing; if not repaired, they may induce mutations.

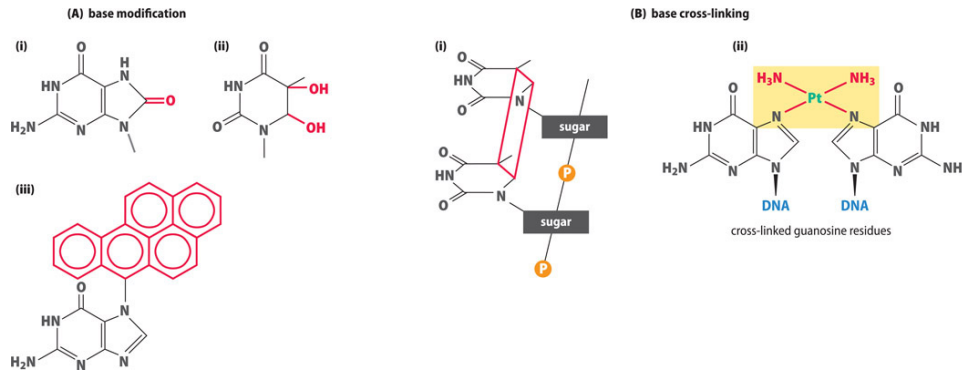


Figure 4.1 Examples of base modification and cross-linking. (A) Base modification.

Altered bonding or added chemical groups are shown in red. Examples are: 8-oxoguanine (i), which base pairs to adenine and so induces mutations; thymidine glycol (ii), which is not a mutagen but blocks DNA polymerase; and a DNA adduct (iii) formed by covalent bonding, in this case of an aromatic hydrocarbon such as benzo(*a*)pyrene to N7 of a guanine residue. (B) Base cross-linking. Cyclobutane pyrimidine dimers, the most prevalent form of damage induced by solar UV light, arise by bonding of carbon atoms 4 and 5 on adjacent pyrimidines on a DNA strand (i). The anticancer agent cisplatin, (NH₃)₂PtCl₂, causes interstrand cross-links by covalently bonding the

N7 nitrogen atoms of guanines on opposite strands (ii).

Spontaneous and environmentally induced DNA damage

Most of the chemical damage to our DNA arises spontaneously and is unavoidable. Every day, under normal conditions, around 20 000–100 000 lesions are generated in the DNA of each of our nucleated cells. Hydrolytic damage can disrupt bonds that hold bases to sugars to produce an abasic site (depurination is particularly common), and also strips amino groups from some bases (deamination). Oxidative damage is also very common because normal cellular metabolism generates some strongly electrophilic (and therefore highly reactive) molecules or ions. The most significant are **reactive oxygen species (ROS)** formed by the incomplete one-electron reduction of oxygen, including superoxide anions (O_2^-), hydrogen peroxide (H_2O_2), and hydroxyl radicals ($OH\cdot$). ROS are generated in different cellular locations and have important natural roles in certain inter-cellular and intracellular signaling pathways, but they mostly originate in mitochondria (where electrons can prematurely reduce oxygen).

A minority of the chemical damage to our DNA is caused by external **mutagens**, agents that can induce mutation, including radiation and harmful chemicals. Ionizing radiation (X-rays, gamma rays, and so on) interacts with cellular molecules to generate ROS that break DNA strands (see below). Non-ionizing ultraviolet radiation causes covalent bonding between adjacent pyrimidines on a DNA strand (see [Figure 4.1Bi](#)).

Our bodies are also exposed to harmful environmental chemicals in our food and drink, in the air that we breathe, and so on. Some chemicals interact with cellular molecules to generate ROS. Other chemicals covalently bond to DNA, often forming bulky DNA adducts that distort the double helix. Large aromatic hydrocarbons found in cigarette smoke and automobile fumes can bond to DNA (see the example in [Figure 4.1Aiii](#)). Electrophilic alkylating agents can result in base cross-linking.

The wide range of DNA repair mechanisms

Cells have different systems for detecting and repairing DNA damage, according to the type of DNA damage. Some types of DNA damage may be minor—the net effect might simply be an altered base. Others, such as DNA cross-linking, are more problematic: they may block DNA replication (the replication fork stalls) or block transcription (the RNA polymerase stalls).

Different molecular sensors identify different types of DNA damage, triggering an appropriate DNA repair pathway. If the DNA lesion is substantial and initial repair ineffective, cell cycle arrest may be triggered that may be temporary (we consider this in the context of cancer in [Chapter 10](#)), or be more permanent. In other cases, as often happens in lymphocytes, apoptosis may be triggered.

The DNA repair process very occasionally involves a simple reversal of the molecular steps causing DNA damage. Usually, however, DNA repair pathways do not directly reverse the damage process. Instead, according to the type of lesion, one of several alternative DNA repair pathways is used. Most of the time, the repair needs to be made to one DNA strand only; sometimes both strands need to be repaired, as in interstrand cross-linking and double-strand DNA breaks.

Errors in DNA replication and chemical damage to DNA are a constant throughout life. Inevitably, however, some mistakes are made in repairing DNA, and there are also inherent weaknesses in detecting some base changes, as described below. Inefficiency in detecting and repairing DNA damage is an important contributor to generating mutation. We consider the health consequences of defective DNA repair in [Clinical Box 1](#), at the end of this section. Before this, we consider the major DNA repair mechanisms in the next two subsections.

Repair of DNA damage or altered sequence on a single DNA strand

DNA damage or an error in DNA replication usually results in one strand having a DNA lesion or a wrongly inserted base but leaves the complementary DNA strand unaffected at that location. In that case, the undamaged complementary strand may be used as a template to direct accurate repair.

- *Base excision repair (BER)*. This pathway is specifically aimed at lesions where a single base has either been modified or excised by hydrolysis to leave an abasic site).
- *Single-strand break repair*. Simple single-strand breaks—also called *DNA nicks*—are caused by breakage of a single phosphodiester bond and are common. They are easily reversed by DNA ligase. More complex breaks occur when oxidative attack causes deoxyribose residues to disintegrate. A type of base excision repair is then employed, whereby strand breaks are rapidly detected and briefly bound by a sensor molecule, poly(ADP-ribose), that initiates repair by attracting suitable repair proteins to the site.
- *Nucleotide excision repair (NER)*. This mechanism allows the repair of bulky, helix-distorting DNA lesions. After the lesion is detected, the damaged site is opened out and the DNA is cleaved some distance away on either side of the lesion, generating an oligonucleotide of about 30 nucleotides containing the damaged site, which is discarded. Resynthesis of DNA is performed with the opposite strand as a template. The priority is to rapidly repair bulky lesions that block actively transcribed regions of DNA. A specialized subpathway, *transcription-coupled repair*, initiates this type of repair after detection of RNA polymerases that have stalled at the damaged site. Otherwise, an alternative global genome NER pathway is used.
- *Base mismatch repair*. This mechanism corrects errors in DNA replication. Errors in base mismatch repair are important in cancer

and we describe this mechanism in [Chapter 10](#).

Repair of DNA lesions that affect both DNA strands

Double-strand DNA breaks (DSBs) are normally rare in cells. They can occur by accident, as a result of chemical attack on DNA by endogenous or externally induced reactive oxygen species (but at much lower frequencies than SSBs). DNA repair is required but can sometimes be difficult to perform: when the two complementary DNA strands are broken simultaneously at sites sufficiently close to each other, neither base pairing nor chromatin structure may be sufficient to hold the two broken ends opposite each other. The DNA termini will often have sustained base damage and the two broken ends are liable to become physically dissociated from each other, making alignment difficult.

Unrepaired DSBs are highly dangerous to cells. The break can lead to the inactivation of a critically important gene, and the broken ends are liable to recombine with other DNA molecules, causing chromosome rearrangements that may be harmful or lethal to the cell. Cells respond to DSBs in different ways. Two major DNA repair mechanisms can be deployed to repair a DSB, as listed below; if repair is incomplete, however, apoptosis is likely to be triggered.

- *Homologous recombination (HR)-mediated DNA repair*. This highly accurate repair mechanism requires a homologous intact DNA strand to be available to act as a template strand. Normally, therefore, it operates after DNA replication (and before mitosis), using a DNA strand from the undamaged sister chromatid as a template to guide repair ([Figure 4.2](#)). It is important in early embryogenesis (when many cells are proliferating rapidly), and in the repair of proliferating cells after the DNA has replicated.

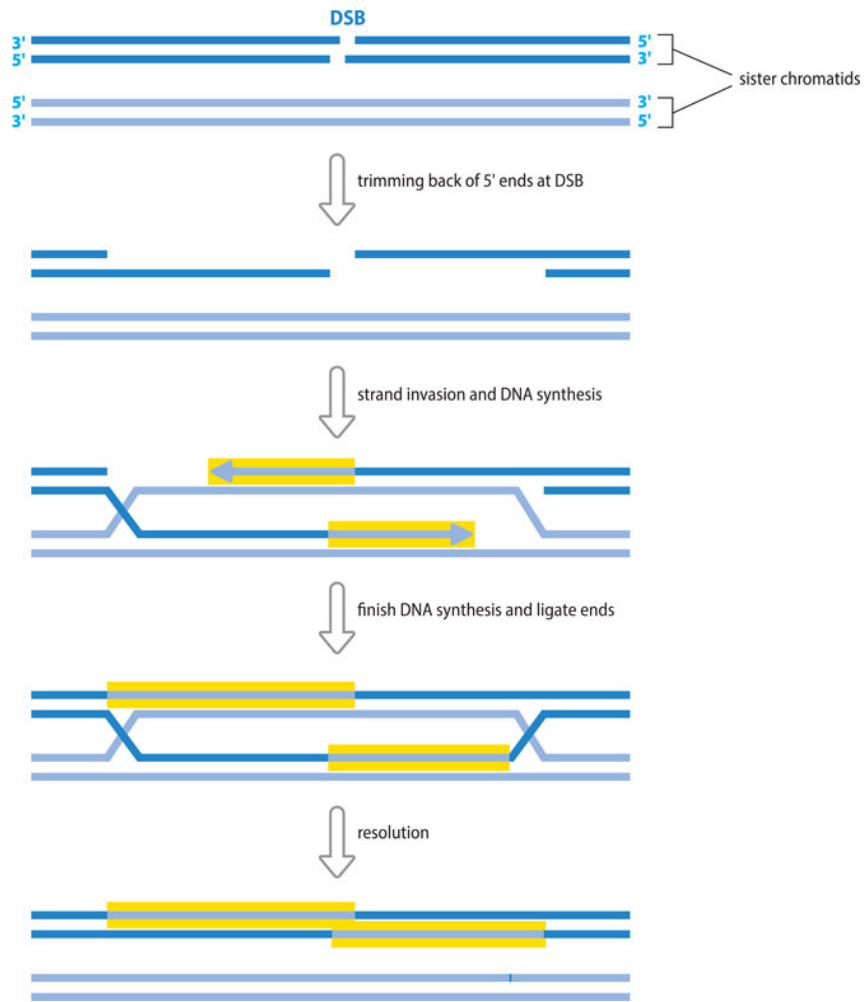


Figure 4.2 Homologous recombination-mediated repair of double-strand DNA breaks. The double-strand break (DSB) in the chromatid at the top is repaired using as a template the undamaged DNA strands in the sister chromatid (*note: to make the mechanism easier to represent, the upper chromatid has, unconventionally, the 3' → 5' strand placed above the 5' → 3' strand*). The 5' ends at the DSB are cut back to leave protruding single-strand regions with 3' ends. After strand invasion, each single-strand region forms a duplex with an undamaged complementary DNA strand from the sister chromatid, which acts as a template for new DNA synthesis (shown by the arrows highlighted in yellow). After DNA synthesis, the ends are sealed by DNA ligase (newly synthesized DNA copied from the sister chromatid DNA is highlighted in yellow). The repair is highly accurate because for both broken DNA strands the undamaged sister chromatid DNA strands act as

templates to direct the incorporation of the correct nucleotides during DNA synthesis.

- **Nonhomologous end joining (NHEJ).** No template strand is needed here because the broken ends are fused together. Specific proteins bind to the exposed DNA ends and recruit a special DNA ligase, DNA ligase IV, to rejoin the broken ends. Unlike HR-mediated DNA repair, NHEJ is, in principle, always available to cells. However, it is most important for the repair of differentiated cells and of proliferating cells in G1 phase, before the DNA has replicated.

Undetected DNA damage, DNA damage tolerance, and translesion synthesis

DNA damage may sometimes go undetected. In vertebrates cytosines occurring within the dinucleotide CG are highly mutable due to inefficient DNA repair. The CG dinucleotide is a frequent target for DNA methylation, converting cytosine to 5-methylcytosine (5-meC). Deamination of cytosine residues normally produces uracil, which is efficiently recognized as a foreign base in DNA and eliminated by uracil DNA glycosylase. Deamination of 5-meC, however, produces a normal DNA base, thymine, that may go undetected as an altered base ([Figure 4.3](#)). As a result, C ↔ T substitutions are the most frequent type of single-nucleotide change in our DNA.

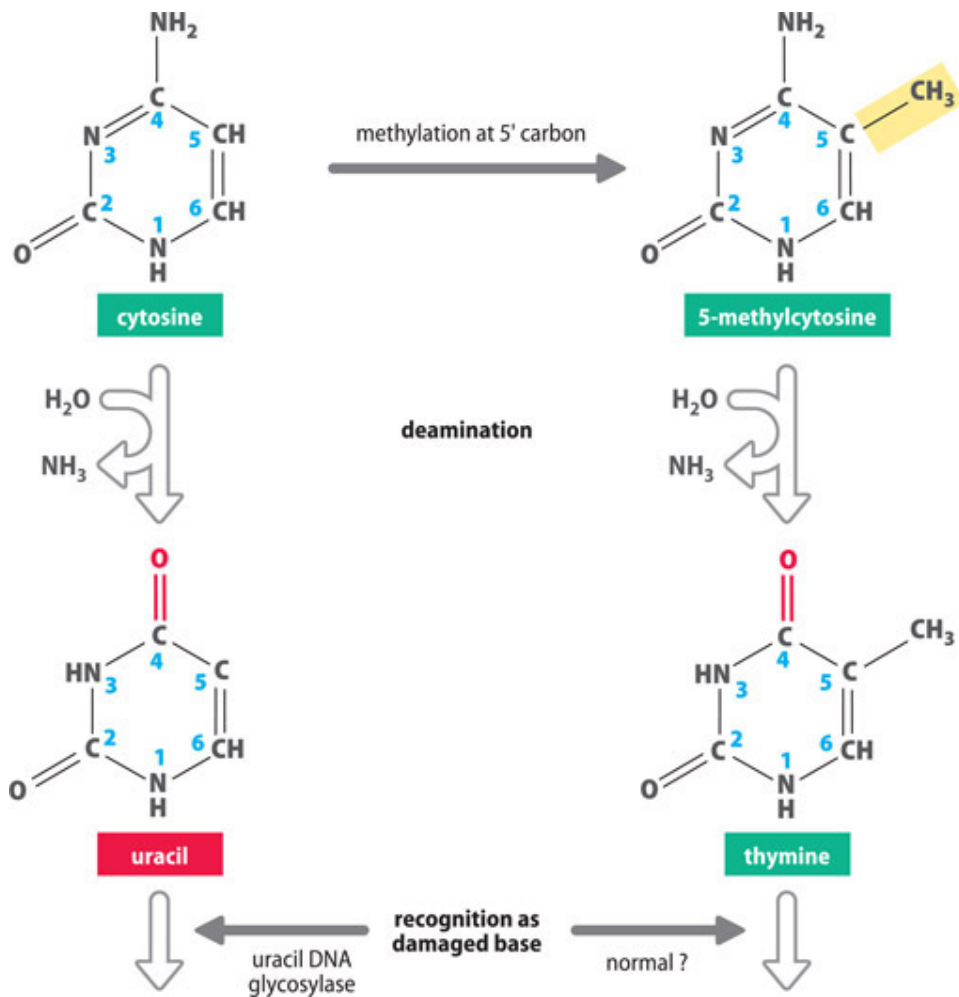


Figure 4.3 Why C → T mutations are so common in human and vertebrate DNA.

Deamination of cytosine is a very common reaction in our cells and normally produces uracil, a base usually found in RNA, not DNA. In our cells a specialized enzyme, uracil DNA glycosylase, recognizes uracil residues in DNA and removes them. However, as in the DNA of other vertebrates, many of our cytosines are methylated at carbon atom 5. Deamination of 5-methylcytosine produces thymine, a base normally found in DNA. Although a stable CG base pair has been replaced by a TG base mismatch, the base mismatch may often escape detection by the base mismatch repair system (which focuses on DNA replication events). At the subsequent round of DNA replication the thymine will form a TA base pair, effectively producing a C → T mutation.

Sometimes, DNA lesions may be identified but are not repaired before DNA replication (damage tolerance). For example, DNA lesions that block replication may be bypassed rather than repaired, and non-classical DNA

polymerases are required to resume DNA synthesis past the damaged site (*translesion synthesis*). Subsequently, the gap in the daughter strand opposite the lesion is filled in; the lesion can be repaired later on, by using the daughter strand as a template in nucleotide excision repair. The nonclassical DNA polymerases used in translesion synthesis exhibit a low fidelity in DNA replication. They have a higher success in incorporating bases opposite a damaged site, but they are prone to error by occasionally inserting the wrong base. As a result, replication forks are preserved, but at the cost of mutagenesis.

CLINICAL BOX 1 DISEASES ARISING FROM DEFECTIVE DNA DAMAGE RESPONSE/DNA REPAIR

DNA damage accumulates in all of us throughout our lives. Inevitably, as we grow older, the incidence of somatic mutations increases, with consequences for increased risk of developing cancer and of declining efficiency in a variety of cellular processes, contributing to the aging process. More than 170 human genes are known to be involved in DNA damage responses and DNA repair (see Further Reading), and a wide variety of single-gene disorders are known to result from germline mutations in genes that work in these pathways ([Table 1](#) gives some examples).

TABLE 1

EXAMPLES OF INHERITED DISORDERS OF DNA REPAIR/DNA DAMAGE RESPONSES

DNA repair/DNA damage response system	Associated single-gene disorders	Disease features*			
		C	P	N	I

DNA repair/DNA damage response system	Associated single-gene disorders	Disease features [*]			
		C	P	N	I
Mismatch repair (described in Chapter 10)	HNPC (Lynch syndrome)	+	-	-	-
Nucleotide excision repair (NER)	xeroderma pigmentosum	+	-	+	-
NER (transcription-coupled repair)	Cockayne syndrome	-	+	+	-
	trichothiodystrophy	-	+	+	-
Single-strand break (SSB) repair	ataxia oculomotor apraxia 1	-	-	+	-
	spinocerebellar ataxia with axonal neuropathy 1	-	-	+	-
Interstrand cross-link repair	Fanconi anemia	+	+	+	+
Double-strand break (DSB) repair (NHEJ)	Lig4 syndrome	+	-	+	+
	severe combined immunodeficiency	-	-	-	+
DNA damage signaling/DSB repair	ataxia telangiectasia	+	-	+	+
	Seckel syndrome	-	-	+	+
	primary microcephaly 1	-	-	+	-

DNA repair/DNA damage response system	Associated single-gene disorders	Disease features [*]			
		C	P	N	I
Homologous recombination (HR)	Bloom syndrome	+	-	+	+
Base excision repair (BER) in mtDNA	spinocerebellar ataxia-epilepsy	-	-	+	-
	progressive external ophthalmoplegia	-	-	-	-
Telomere maintenance (TM)	Dyskeratosis congenita	+	+	+	+
HR, BER, TM	Werner syndrome	+	+	-	-

^{*}

C, cancer susceptibility; P, progeria; N, neurological features; I, immunodeficiency. HNPC, hereditary nonpolyposis cancer. SCA, spinocerebellar ataxia

As expected, increased susceptibility to cancer and accelerated aging are often found in these disorders, and developmental abnormalities and neurological features are also common. Although many cell types are regularly replaced, nondividing neurons are especially vulnerable. They have high oxygen and energy needs (with a resulting high frequency of oxidative damage), and they accumulate DNA damage over very long periods. Cellular abnormalities are frequently seen, with respect to chromosome and genome instability as listed below.

DISEASE FEATURES

Cancer (C) susceptibility. This is apparent in many inherited DNA repair deficiencies. Genome instability in mismatch repair deficiencies can induce cancer in highly proliferating tissues, notably intestinal epithelium. Individuals with xeroderma pigmentosum have little protection against UV radiation, and exposure to sunlight induces skin tumors ([Figure 1A](#)).

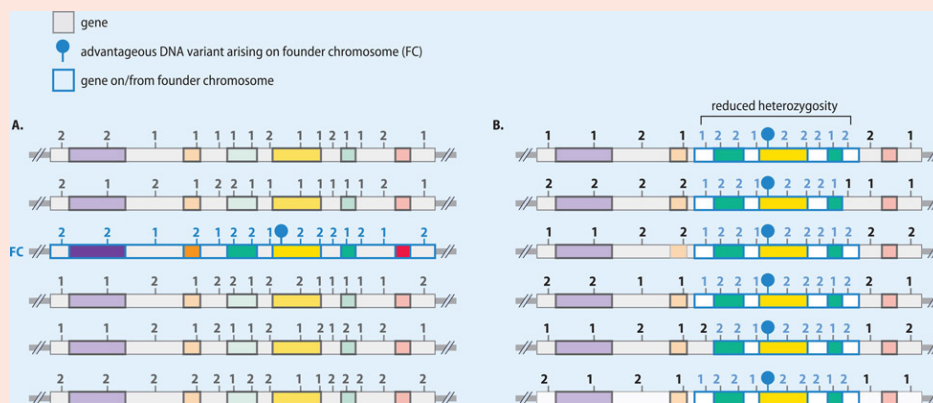


Figure 1 General effect of a selective sweep for an advantageous DNA variant.

(A) Heterozygosity profile before selection. Imagine that an advantageous DNA variant has just occurred in the gene shown in yellow on a founder chromosome 22 (FC) (with genes outlined in blue). Now imagine assaying genetic variation by using intronic and extragenic microsatellite markers, each with four common alleles (1 to 4), over each copy of chromosome 22 in the population. We might expect significant heterozygosity, as shown by the six representative chromosome 22s. (B)

Heterozygosity profile after positive selection over many generations. Vertical transmission of the founder chromosome 22, recombination, and continued positive selection for the advantageous variant will result in an increased frequency of the advantageous DNA variant plus closely linked DNA variants, causing reduced heterozygosity for that chromosome segment. Some tightly linked neighboring genes will also increase in frequency in the population because of selection for the variant. They are often described as *hitchhiking alleles* (shown here in blue and green).

- *Progeria (P)*. Some disorders have clinical features that mimic accelerated aging, notably individuals with Werner syndrome ([Figure 1B](#)), who prematurely develop gray hair, cataracts, osteoporosis, type 2 diabetes, and atherosclerosis, and generally

die before the age of 50 as a result of cancer or atherosclerosis.

- *Neurological (N) features.* Neuronal death and neuro-degeneration are common features. Individuals with ataxia telangiectasia experience cerebellar degeneration leading to profound ataxia and become confined to a wheelchair before the age of 10. Microcephaly is common, sometimes accompanied by neurodegeneration and learning difficulties.
- *Immunodeficiency (I).* Some DNA repair proteins also work in specialized genetic mechanisms in B and T lymphocytes. For example, components of the NHEJ repair pathway are needed to make immunoglobulin and T-cell receptors, and when they are lacking, hypogammaglobulinemia and lymphopenia or severe combined immunodeficiency result.

CELL ANALYSES REVEALING GENOME AND CHROMOSOMAL INSTABILITY

The DNA of individuals with disorders of mismatch repair (described in [Section 10.3](#)) shows striking evidence of genome instability when short tandem repeat DNA polymorphisms known as microsatellite DNA are assayed. Cells from individuals with a DNA repair disorder quite often also show an increased frequency of spontaneous chromosome aberrations characteristic of the disorder, as in the case of ataxia telangiectasia, Fanconi anemia, and Bloom syndrome (which shows very high levels of sister chromatid exchange).

Chromosome analyses can also provide a simple route to laboratory-based diagnosis. Fanconi anemia (where there are variably assorted developmental abnormalities, plus progressive bone marrow failure and an increased risk of malignancy) can be caused by mutations in any one of at least 13 different genes that repair interstrand cross-links, making DNA-based diagnosis difficult. Chromosome-based diagnosis is more straightforward: lymphocyte cultures are treated with diepoxybutane or

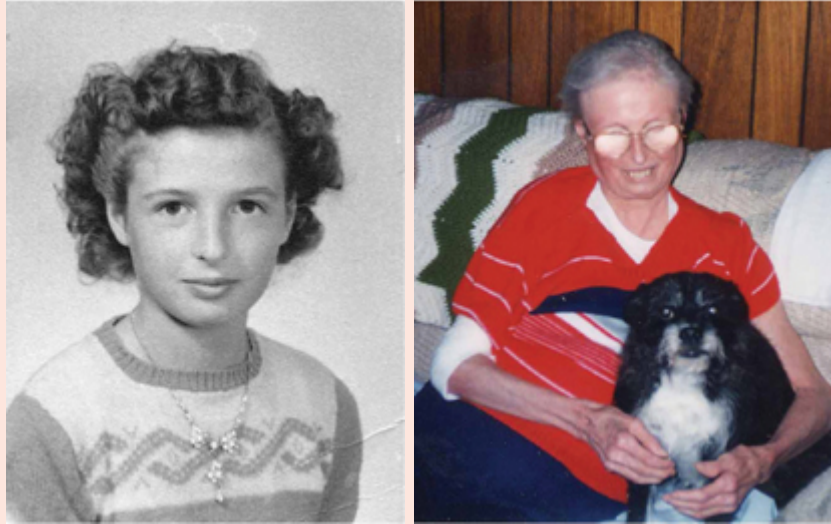
mitomycin C—chemicals that induce DNA interstrand cross-links—and chromosomes are analyzed for evidence of chromatid breakage, which can produce characteristic abnormal chromosome formations ([Figure 1C](#)).

Figure 1 Examples of abnormal phenotypes in DNA-repair disorders.

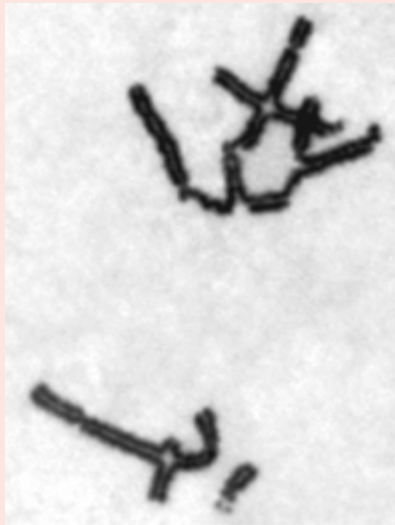
(A) Extensive skin cancer in xeroderma pigmentosum. (B) Accelerated aging in Werner syndrome—portraits of the same woman at age 13 (left) and age 56 (right). (C) Characteristic quadriradial and triradial chromosome formations in Fanconi anemia cells after treatment with mitomycin C. (A, courtesy of Himynameislax (CC BY-SA 3.0). B, from Hisama FM, Bohr VA, and Oshima J [2006] *Sci Aging Knowl Environ* 10:pe18. With permission from the AAAS (left) and the International Registry of Werner Syndrome (right). C, courtesy of Niall Howlett from Harney JA, Shimamura A and Howlett NG [2008] *Pediatr Health* 2:175–187. With permission from Future Medicine Ltd.)



(A)



(B)



(C)

4.2 POPULATION GENOMICS AND THE SCALE OF HUMAN GENETIC VARIATION

Genetic variation is caused by DNA sequence changes that can be sorted into different classes according to the underlying mechanisms and the scale. But all DNA changes can be classified into two broad categories:

1. **Changes that do not affect the DNA content.** Here the number of nucleotides is unchanged. Quite often, for example, a single nucleotide is replaced by a different nucleotide. More rarely, multiple nucleotides at a time may be sent to a new location (by chromosome breakage and rejoining) without net loss or gain of DNA content; the great majority are balanced translocations and inversions resulting in chromosome breakage without net loss or gain of DNA.
2. **Changes in copy number.** Here there is a net loss or gain of DNA sequence. At one extreme, abnormal chromosome segregation produces fewer or more chromosomes than normal, and therefore a change in the copy number of whole nuclear DNA molecules. They are almost always harmful. At the other extreme are deletions or insertions of a single nucleotide. In between are copy number changes that range from altered numbers of sequences that may be short (specific oligonucleotide sequences, for example) or long (sequences extending over multiple Mb of DNA).

Overall, the most common DNA changes are those that change only a single nucleotide or a very small number of nucleotides. Small-scale changes like this (often called **point mutations**) may often have no obvious effect on the phenotype; in that case they would be considered to be neutral mutations. That happens mostly because more than 90 % of our DNA has been poorly conserved during evolution and may have no or very little functional value to the cell. Small, and sometimes quite large, changes in this fraction of the genome seem to be without obvious effect.

DNA variants, polymorphisms, and human population genomics

Alternative forms of DNA produced by mutation are generally described as DNA **variants**. Until quite recently, it was usual to describe a common DNA variant, with a frequency > 0.01 , as a **polymorphism**; DNA variants

with frequencies < 0.01 were traditionally described as *rare variants*. (The 0.01 cut-off might seem arbitrary; it was initially proposed so as to exclude recurrent mutation.)

The general use of the term *polymorphism* has been declining, partly because of the arbitrary nature of the 0.01 cut-off, and partly because of ambiguity in how the term is used. In medical disciplines, for example, *polymorphism* is often used to denote any sequence variation that does not cause disease, whereas *mutation* is used to describe a disease-causing sequence variant.

In modern times, the term *polymorphism* is largely avoided in population genomics projects; instead, it is customary to use DNA *variants*; they are often classified as: common (frequency $> 5\%$); low frequency (from 0.5% to 5%); and rare ($< 0.5\%$). And it is also now customary to describe a change to a single nucleotide as a *single nucleotide variant (SNV)*.

The Human Genome Project delivered an artificial *reference* sequence for the human genome, a patchwork of partial genome sequences from multiple individual anonymized donors that were combined into a single sequence. To obtain detailed knowledge of human genetic variation, however, genome-wide sequences from multiple individuals (each with two nuclear genomes, a maternal and a paternal genome) need to be analyzed. Clearly, the greater the number of individual samples analyzed and the greater the fraction of the genome sequenced the more information is obtained. Analysis of very large numbers of genome sequences is important because rare genetic variants can be medically important.

Personal genome sequencing first became a reality in 2007–2008 (interested readers can find descriptions of the first two individual human genomes to be sequenced, one by laborious Sanger sequencing (PMID 17803354) and the other by rapid massively parallel DNA sequencing (PMID 18421352). The ability to sequence whole genomes rapidly ushered in the era of human *population genomics* (population-based genome sequencing). Detailed information on human genetic variation has rapidly become available and large-scale projects have been launched to correlate genotypes with phenotypes (see [Box 4.1](#)).

Structural variation as opposed to small-scale variation

The vast majority of DNA changes are errors of DNA replication and repair. They typically affect one or a very small number of nucleotides. Because of the predominance of small-scale mutations, the study of human genetic variation was very largely focused on this type of variation until quite recently.

Whole genome sequencing has shown that additional moderate to large-scale DNA changes (> 50 bp), which include the outcomes of specific types of DNA breakage and rejoining mechanisms, are also highly significant. Such **structural variation** can involve very large changes, and although structural variants are rather infrequent, significantly more nucleotides across the genome are altered as a result of structural variation than as a result of small-scale mutations.

The borderline between small-scale genetic variation and structural variation is, of course, an arbitrary one. (In the past, structural variation used to be applied to sequences that were one kilobase or longer, but the modern tendency has been to include smaller variants as long as the sequence change involves more than 50 bp).

BOX 4.1 LARGE-SCALE HUMAN POPULATION GENOMICS AND GENOTYPE–PHENOTYPE CORRELATION PROJECTS

Once rapid personal genome sequencing became possible in 2008, the first human population genomics project was launched, the 1000 Genomes Project, the initial aim of which was to sequence 1000 individual genomes from 26 different human populations across the world. As well as getting more information generally on human genetic variants, a major goal was to compare the genetic diversity of different ethnic populations (substantial differences in genetic variants exist between different human populations and are important in explaining differential population-based susceptibility to many disorders).

Since then, there has been a plethora of human population genomics projects, some focused on sequencing whole **exomes** (concentrating on exons from protein-coding genes), and others on whole genome sequencing. The latter had the notable advantage of vastly increasing our knowledge of formerly neglected structural variation, as well as offering a wide range of single nucleotide variants outside the more intensively studied protein-coding regions.

Initially, different human population genomics projects often used diverse ways of analyzing the data. To improve efficiency, consortia were formed to aggregate the sequencing data from different projects, and re-analyze the data in a common pipeline. In 2016, the Exome Aggregation Consortium (ExAC) catalogued genetic variation in the protein-coding parts of the genome from 60 000 people. And in 2020, the genome aggregation database consortium (gnomAD) reported on analysis of sequences from 125 748 human exomes and 15 708 whole genomes.

The large scale of the gnomAD study has been important for identifying rare variants, and the substantial number of whole genomes analyzed has provided important information on human noncoding DNA variation. For a short overview, see PMID 32461645; for seminal gnomAD research publications see under Further Reading.

GENOTYPE-PHENOTYPE CORRELATION PROJECTS

To maximize the value to medicine and health of the burgeoning data on human genetic variation, large projects have recently been developed with the aim of correlating genotypes with phenotypes. The UK Biobank Project has been a pioneering project in this respect, collecting deep genetic and phenotypic data on 500 000 individuals from across the UK. For the UK's 100 000 Genomes Project, genome sequences from 85 000 UK National Health Service (NHS) patients affected by a rare disease or cancer were sequenced by the end of 2018. That project is being expanded to 1 000 000 genomes, including those of the 500 000 UK Biobank volunteers, and a further expansion has been planned towards genome

sequencing of 5 million UK individuals. The All of Us project organized by the US National Institutes of Health seeks to correlate geno-types in 1 000 000 volunteers with their health data. We consider genotype-phenotype correlations more fully in later chapters.

Small-scale variation: single nucleotide variants and small insertions and deletions

Base substitution is the most common type of point mutation. Two major classes of base substitution are recognized, as listed below:

1. a *transition* (a purine is replaced by another purine, or a pyrimidine by another pyrimidine)
2. a *transversion* (a purine is replaced by a pyrimidine, or a pyrimidine by a purine).

If, for example, an A were substituted, there are three possibilities: A → C (transversion); A → G (transition) or A → T (transversion).

Base substitution can generate ([Figure 4.4](#)) a **single nucleotide variant (SNV)**. For example, at a defined position on a DNA molecule most sampled sequences might happen to have a G, but a minority might have a C. In many cases the minority variant occurs at a low frequency in the population and may be described as a rare variant, or even a *private variant* (as a result of very recent mutation). In some cases, however, a minority variant is present at a population frequency of 0.01 or more, a frequency that is too high to explain by recurrent mutation. In that case, the DNA variation has traditionally been described as a **single nucleotide polymorphism** or **SNP** [pronounced “snip”].)

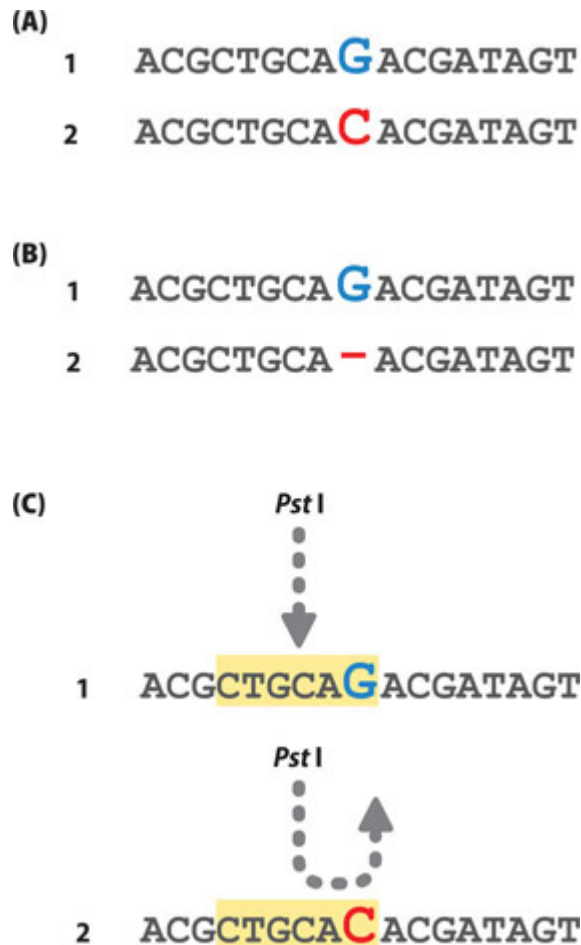


Figure 4.4 Classes of DNA variation affecting a single nucleotide position. (A) Single nucleotide variant (SNV) in which two variants differ by having a G or a C. (B) Insertion/deletion (*indel*) variation in which variant 1 has a G not present in variant 2. (C) Sometimes differences at a single nucleotide position can lead to the variable presence of a restriction site. Here variant 1 shown in (A) can be seen to have a recognition sequence (CTGCAG) for the restriction enzyme *Pst*I; in variant 2 the equivalent sequence (CTGCAC) will not be cleaved by *Pst*I. If the variants are common this would be an example of a *restriction fragment length polymorphism (RFLP)* that can conveniently be assayed by designing PCR primers to amplify the sequence containing it and then cutting the PCR product with *Pst*I.

The pattern of single nucleotide variation in the human genome is nonrandom. Different DNA regions and different DNA sequences can undergo different mutation rates, and there is a large excess of C ↔ T substitutions in the human genome (see [Table 4.2](#)).

TABLE 4.2**SOME NONRANDOM FEATURES OF SINGLE NUCLEOTIDE VARIATION IN THE HUMAN GENOME**

Feature	Description
Excess of transitions	The expected ratio of transitions to transversions is 1:2. In reality there is an excess of transitions because C → T transitions are unexpectedly frequent (see Figure 4.3)
Mutational bias	A general bias towards A-T base pairs (also observed in a wide range of other species)
SNP inheritance in germline DNA	Alternative nucleotides at SNP sites mark ancestral chromosomal segments that are common in the present-day population (see text)
Local suppression	Some regions of the genome, notably coding sequences, are subject to purifying selection to minimize harmful substitutions
Local enhancement	A higher rate in condensed chromatin (which is late replicating; possibly the condensed structure impedes access to the mismatch repair machinery)

Another reason for nonrandom variation comes from our evolutionary ancestry. Readers might reasonably wonder why only certain nucleotides should be polymorphic and be surrounded by stretches of nucleotides that only rarely show variants. In general, the nucleotides found at SNP sites are not particularly susceptible to mutation, and SNPs are stable over evolutionary time. Instead, the alternative nucleotides at SNP sites mark alternative ancestral chromosome segments that just happen to be common

in the present-day population. As described in [Chapter 8](#), using SNPs to define ancestral chromosome segments is important in mapping genetic determinants of disease. We cover methods for assaying specific single nucleotide changes in [Chapter 11](#).

Small insertions and deletions

Some point mutations create DNA variants that differ by the presence or absence of a single nucleotide, or by a small number of nucleotides at a specific position. This is described as **insertion/deletion** variation or *indel* for short—see Figure 4.4B. A subset of SNVs or indels leads to the gain or loss of a restriction site, in which case cutting the DNA with the relevant restriction nuclease can generate restriction fragment length polymorphism (RFLP; see Figure 4.4C).

Although indels could be considered to be copy number variants, the modern convention is to reserve the term **indel** to describe deletions or insertions of from one nucleotide up to an arbitrary 50 or so nucleotides (chosen because many massively parallel DNA sequencing methods produce quite short sequences and are often not suited to detecting deletions or insertion of greater than 50 nucleotides). The term **copy number variation (CNV)** is mostly used for changes in copy number of sequences that result in larger deletions and insertions, usually more than 100 nucleotides up to megabases.

The frequency of insertion/deletion polymorphism in the human genome is about one-tenth the frequency of single nucleotide substitutions. Short insertions and deletions are much more common than long ones. Thus, 90 % of all insertions and deletions are of sequences 1–10 nucleotides long, 9 % involve sequences from 11 to 100 nucleotides, and only 1 % involve sequences greater than 100 nucleotides. Nevertheless, because many of the last category involve huge numbers of nucleotides, CNV affects more nucleotides than single nucleotide substitutions.

Microsatellites and other variable number of tandem repeat (VNTR) polymorphisms

As described in [Section 2.5](#), repetitive DNA accounts for a large fraction of the human genome. Tandem copies of quite short DNA repeats (1 bp to fewer than 200 bp) are common, and those with multiple tandem repeats are especially prone to DNA variation. A continuous sequence of multiple tandem repeats is known as an array. Different organizations are evident and the repeated sequences are classified as belonging to three classes, according to the total length of the array and genomic location:

1. *microsatellite DNA* (array length: fewer than 100 bp long; widely distributed in euchromatin)
2. *minisatellite DNA* (array length: 100 bp to 20 kb; found primarily at telomeres and subtelomeric locations)
3. *satellite DNA* (array length is often from 20 kb to many hundreds of kilo-bases; located at centromeres and some other heterochromatic regions).

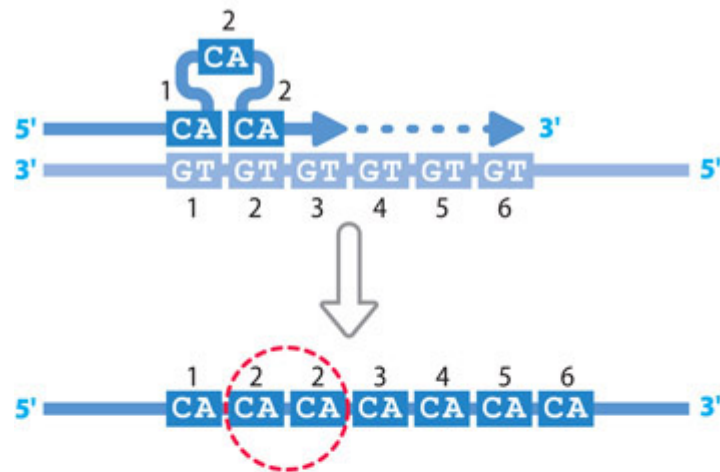
The instability of tandemly repeated DNA sequences results in DNA variants that differ in the numbers of tandem repeats, that is, variable number of tandem repeats (VNTR) polymorphism. Because microsatellite DNA arrays (usually called **microsatellites**) are frequently distributed within human euchromatin (roughly once every 30 kb) and are often highly polymorphic, they have been widely used in genetic mapping.

Because microsatellites have very short repeats (one to four base pairs long), microsatellite polymorphisms are sometimes known as *short tandem repeat polymorphisms (STRPs)*. They usually result in short insertions and deletions. But unlike SNPs (which almost always have just two alleles), microsatellite polymorphisms usually have multiple alleles ([Figure 4.5](#)).



Figure 4.5 Length polymorphism in a microsatellite. Here, a microsatellite locus is imagined to have three common alleles that differ in length as a result of having variable numbers of tandem CA repeats. See [Figure 4.6](#) for the mechanism that gives rise to the variation in copy number.

(A) backward slippage of nascent strand causes insertion



(B) forward slippage of nascent strand causes deletion

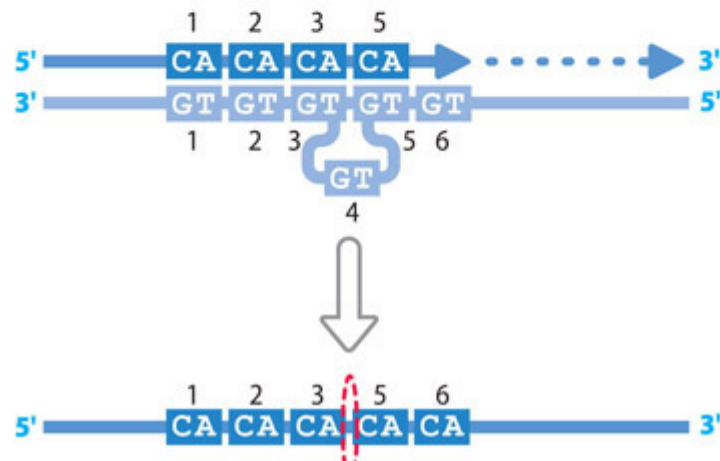


Figure 4.6 Microsatellite polymorphism results from strand slippage during DNA replication. The dark blue strand represents the synthesis of a new (nascent) DNA strand from the pale blue template DNA strand. During normal DNA replication, the nascent strand often partly dissociates from the template and then reassociates. When there is a tandemly repeated sequence, the nascent strand may mispair with the template strand when it reassociates, so that the newly synthesized strand has more repeat units (A) or fewer repeat units (B) than the template strand, as illustrated within the dashed red circles.

The variation in copy number arises as a result of **replication slippage**: during DNA replication, the nascent (newly synthesized) DNA strand slips

relative to the template strand so that the two strands are slightly out of alignment—see [Figure 4.6](#).

Individual microsatellite polymorphisms can be assayed using PCR to amplify a short sequence containing the array, and then separating PCR products according to size by gel electrophoresis. Although extensively used in family studies and DNA profiling (to establish identity), it is not easy to automate assays of microsatellite polymorphisms, unlike for SNP assays.

Structural variation and low copy number variation

Until quite recently, the study of human genetic variation was largely focused on small-scale variation such as changes affecting single nucleotides and micro-satellite polymorphisms. We now know that variation due to moderately large-scale changes in DNA sequence is very common. Such **structural variation** can be of two types: balanced and unbalanced.

In balanced structural variation, the DNA variants have the same DNA content but differ in that some DNA sequences are located in different positions within the genome. They originate when chromosomes break and the fragments are incorrectly rejoined, but without loss or gain of DNA. That can involve inversions and translocations that do not involve change in DNA content ([Figure 4.7](#)).

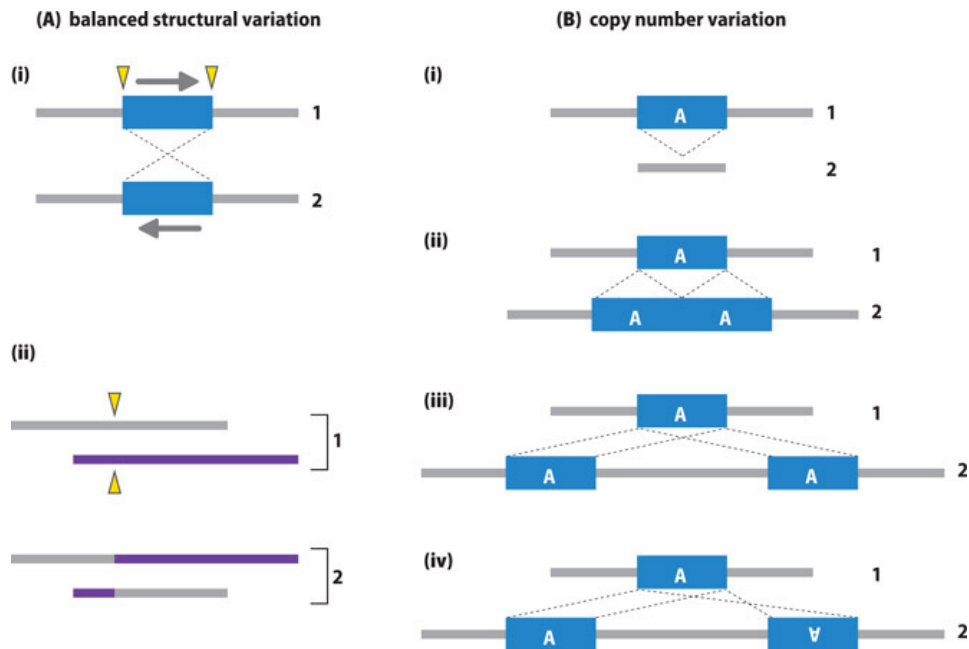


Figure 4.7 Structural variation and low copy number variation. The numbers 1 and 2 refer to alternative variants throughout. (A) Balanced structural variation involves large-scale changes that produce variants with the same number of nucleotides, including many inversions (i) and balanced translocations (ii). (B) Unbalanced structural variation includes unbalanced inversions and unbalanced translocations (not shown here) plus different types of low copy number variation (CNV). Copy number variants have different numbers of copies of a moderately long sequence (represented here by the box marked A). (i) CNV in which variants possess (1) or lack (2) a sequence, effectively a large-scale indel. This can result from an insertion (for example of a mobile element) or a deletion. (ii) CNV due to tandem duplication, effectively a large-scale VNTR that can sometimes have several copies, rather than just the one or two copies shown here. Sometimes additional insertion and inversion events can result in interspersed duplication with normal orientation of copies (iii) and interspersed duplication with inversion of a copy (iv).

In unbalanced structural variation, the DNA variants differ in DNA content. In rare cases in which a person has gained or lost certain chromosomal regions (as when a parent with a balanced reciprocal translocation passes one of the translocation chromosomes, but not the other, to a child), the gain or loss of substantial chromosomal segments

often results in disease. Unbalanced structural variation also includes commonly occurring CNV in which the variants differ in the number of copies of a moderately long to very long DNA sequence. Some CNVs such as this contribute to disease, but very many CNVs are commonly found in the normal population.

Copy number variation can take different forms. One form is effectively simple insertion/deletion variation on a large scale in which DNA variants either lack or possess a specific sequence (see Figure 4.7Bi). Other forms result from tandem duplication that may be complicated by subsequent insertion and inversion events (see Figure 4.7Bii–iv). In some CNVs, the DNA sequence that varies in copy number can include part of a gene sequence or regulatory sequence and sometimes multiple genes. As a result, some CNVs are important contributors to disease.

Taking stock of human genetic variation

The data from population-based genome sequencing projects indicate that single nucleotide changes are the most common type of genetic variation, accounting for close to 75 % of DNA changes. A study of 1092 individuals in 14 human populations by the 1000 Genomes Project Consortium reported a total of 38 million human SNPs (more than 1 per 100 nucleotides). However, the vast majority of these SNPs are rare in any population. Any one individual has just two haploid genomes and will be homozygous at many SNP loci. Personal genome sequencing shows that single nucleotide differences between the maternal and paternal genomes in one individual occur about once every 1000 nucleotides. Much of that variation falls outside coding sequences and mostly represents neutral mutations.

Structural variation is less common, accounting for close to one-quarter of mutational events and is dominated by CNV. Because CNV often involves very long stretches of DNA, however, the number of nucleotides involved in CNVs significantly exceeds those involved in SNVs. Various

databases have been established to curate basic data on genetic variation in humans (and other species)—see [Table 4.3](#).

TABLE 4.3

GENERAL HUMAN GENETIC VARIATION DATABASES

Database	Description	Website
dbSNP	SNPs and other short genetic variations	http://www.ncbi.nlm.nih.gov/SNP/index.html
dbVar	genomic structural variation	http://www.ncbi.nlm.nih.gov/dbvar/
DGV		http://dgy.tcag.ca/
ALFRED	allele frequencies in human populations	http://alfred.med.yale.edu/alfred/index.asp

Databases focusing on mutations that relate to phenotypes and disease are described in [Chapters 7](#), [8](#), [10](#) and [11](#).

4.3 FUNCTIONAL GENETIC VARIATION AND PROTEIN POLYMORPHISM

Until now we have focused on the different types of human genetic variation at the DNA level, and their origins. Most DNA variation is neutral, having no detectable effect on the phenotype, but a small fraction of DNA variation alters how gene products are made, causing disease. Surprisingly, however, a normal person carries an average of about 120 gene-inactivating

variants, with about 20 genes being predicted to be inactivated in both alleles.

Such a level of gene inactivation might seem alarming. However, some predicted loss-of-function variants occur in exons that are variably used in transcripts. And quite a few of our genes do not carry out vital functions. For example, people with blood group O are homozygotes for an inactivating mutation in the *ABO* gene. They fail to make an enzyme that transfers a monosaccharide group onto the H antigen (an oligosaccharide attached to certain lipids or proteins on the surface of some cell types, notably red blood cells). People with blood groups A, B, and AB do produce this enzyme. Effectively, therefore, people with blood group O make the H-antigen, and people with other blood groups make a modified H-antigen carrying an extra monosaccharide. Inactivating mutations are also common in some large gene families, such as the olfactory receptor super-family that we describe near the end of this section.

In this section we primarily consider how genetic variation is expressed at the level of gene products, notably proteins because they are overwhelmingly the major functional endpoint of our genes. (We have numerous RNA genes but either they assist protein synthesis directly or they are involved in gene regulation pathways ultimately affecting patterns of protein production. The great majority of molecular pathogenesis is ultimately due to abnormal protein expression).

Variation in protein sequences can occasionally result from recent gene duplication (as described below) but is usually due to variation at a single gene locus. In the latter case, the variation may result from changes at the level of DNA sequence or RNA sequence. Protein variants produced by changes in RNA sequence are usually described as **isoforms**. See [Table 4.4](#) for a summary.

[TABLE 4.4](#)

DIFFERENTWAYS OF PRODUCING VARIATION IN PROTEINS

Level	Mechanism	Examples
DNA (multilocus)	Gene duplication	Olfactory receptors (this section)
DNA (single locus)	Allelic variation	Numerous
	Post-zygotic DNA changes	Immunoglobulins, Tcell receptors (Section 4.4)
RNA	Alternative splicing initiated from alternative promoters	p14 and p16 from <i>CDKN2A</i> gene (Figure 6.8B)
	Alternative splicing causing variable possession of internal exons or sequence motifs	The +KTS and -KTS isoforms of the Wilms Tumor <i>WT1</i> gene
	RNA editing*	See Section 6.1
	Alternative polyA-addition*	Dystrophin Dp40 isoform

*

We explain the underlying mechanisms in [Chapter 6](#) when we consider gene regulation.

The vast majority of genetic variation has a neutral effect on the phenotype, but a small fraction is harmful

Functional DNA variants are primarily those affecting the function of genes, changing the structure of a gene product or altering the rate at which it is produced. Only a very small fraction of nucleotides in our DNA is important for gene function, however. Coding DNA sequences make up

close to 1.3% of the human genome. Some additional DNA sequences make functional noncoding RNAs; others regulate gene expression in some way—at the transcriptional, post-transcriptional, or translational level.

Estimating how much of the genome is functionally important is not straightforward. The traditional way is to carry out cross-species comparisons to identify how much of the genome is subject to purifying selection to conserve functionally important sequences. Population-based human genome sequencing also offers insights into evolutionarily recent functional constraint. Current estimates suggest that perhaps a maximum of about 10 % of the genome is under functional constraint. Mutation at 90 % or more of our nucleotides may have essentially no effect.

Even within the small target of sequences that are important for gene function, many small DNA changes may still have no effect. For example, many coding DNA mutations are silent: they do not change the protein sequence and would usually have no effect (unless they cause altered splicing—we show examples in [Chapter 7](#)). Single nucleotide changes in regulatory sequences or in sequences that specify functional noncoding RNA (ncRNA) may often also have no, or very little, effect. We know, however, little about the functional significance of changes in noncoding RNA; most studies have focused on coding DNA. Of course, harmful mutations also occur in functional DNA, and very occasionally mutations have a beneficial effect.

Different types of Darwinian natural selection operate in human lineages

A small fraction of genetic variation is harmful, and we consider the detail in other chapters, notably [Chapters 7](#) and [10](#). Harmful mutations are subject to a type of negative selection called **purifying selection**: people who possess them will tend to have lower reproductive success rates and the mutant allele will gradually be eliminated from populations over several generations. Harmful DNA changes include many different types of small-scale mutations, both in coding DNA (resulting in changes in amino acid

sequence) and noncoding DNA (causing altered splicing, altered gene regulation, or altered function).

In addition, structural variation can often have negative effects on gene function. Genes may be inactivated by balanced structural variation if breakpoints affect how they are expressed. Copy number variation can lead to a loss or gain of gene sequences that can be harmful because the levels of some of our gene products need to be tightly controlled, as explained in later chapters.

Occasionally, a DNA variant has a beneficial effect on the phenotype that can be transmitted to offspring. DNA variants like this become prevalent through **positive selection**. Here, individuals who possess the advantageous DNA variant may have increased survival and reproductive success rates; the DNA variant then increases in frequency and spreads throughout a population.

Positive selection has occurred at different times in human lineages. It has been responsible for fostering different features that distinguish us from the great apes, notably human innovations in brain development and increased cognitive function. The great majority of the selected variants seem to occur in noncoding regulatory DNA and result in altered gene expression. As described below, positive selection is also important in response to microbial pathogens and to various alterations in our environment.

Positive selection in response to microbial pathogens

Human populations are subject to survival pressure from infectious diseases, notably ones that can develop through efficient transmission between humans to become pandemics. The Black Death plague in the mid-fourteenth century had mortality rates of 80–100 % and killed 25–50 % of European and Chinese populations; and 500 million people, one quarter of people on the planet at that time, were infected in the “Spanish” influenza pandemic of 1918 with a mortality rate of perhaps 10 %.

In the front line to protect us from diseases like this are HLA genes. They make proteins involved in recognizing viral and other foreign antigens in host cells and in activating T cells to recognize infected cells and counter the pathogen. The selection here is thought to favor heterozygosity at the key HLA loci. An individual then may often produce slightly different proteins at several HLA loci, increasing the chances of protective immune system responses. As a result, HLA proteins are the most polymorphic of all our proteins—we provide details in [Section 4.4](#).

Adaptations to altered environments

Positive selection has also been responsible for various instances of adaptive evolution in human populations. After out-of-Africa migrations 50 000–100 000 years ago, modern humans settled in different geographic regions and were exposed to different environments, including living in latitudes with low levels of sunlight, or living at high altitudes. As shown in [Table 4.5](#) different adaptations developed in the migratory human populations.

TABLE 4.5

EXAMPLES OF GENETIC VARIANTS IN ADAPTIVE EVOLUTION IN HUMAN POPULATIONS

Altered environment	Adaptation and its effects	Associated genetic variants
Reduced sunlight (low UV exposure)	decreased pigmentation (decreased melanin in skin allows more efficient transmission of the depleted UV to a deep layer of the dermis—see text)	an <i>SLC24A5</i> variant (replacing the ancestral ALA at position 111 by THR) is prevalent in European populations (see Box 4.2)

Altered environment	Adaptation and its effects	Associated genetic variants
High-altitude settlements (low O ₂ tension)	in Tibetan* populations lowered hemoglobin levels and a high density of blood capillaries provide protection against hypoxia	variants in <i>EPAS1</i> , a key gene in the hypoxia response
Malaria-infested environments	alterations in red blood cell physiology, affecting transmission of the mosquito-borne parasites <i>P. falciparum</i> or <i>P. vivax</i> and conferring increased resistance to malaria	pathogenic mutations** in <i>HBB</i> or <i>G6PD</i> for <i>P. falciparum</i> malaria; inactivating <i>DARC</i> variants that do not express the Duffy antigen* in <i>P. vivax</i> malaria
Lifelong intake of fresh milk	persistence of lactase production in adults, allowing efficient digestion of lactose	the -13910T allele about 14 kb upstream of the lactase gene, <i>LCT</i>
High levels of dietary starch	increased production of enzyme needed to digest starch efficiently	high <i>AMY1A</i> copy number (Figure 4.8)

Altered environment	Adaptation and its effects	Associated genetic variants
---------------------	----------------------------	-----------------------------

Gene symbols are as follows: *SLC24A5*, solute carrier family 24, member 5; *EPAS1*, endothelial PAS domain protein 1; *HBB*, (β -globin gene; *G6PD*, glucose-6-phosphate dehydrogenase; *LCT*, lactase gene (converts lactose to galactose plus glucose); *AMY1A*, salivary α -amylase gene (converts starch into a mixture of constituent monosaccharides).

*

Andean populations show different anti-hypoxia adaptations.

**

Includes sickle-cell, thalassemia, and glucose-6-phosphate dehydrogenase deficiency mutations.

The Duffy antigen is a ubiquitously expressed cell surface protein required for infection of erythrocytes by *Plasmodium vivax*.

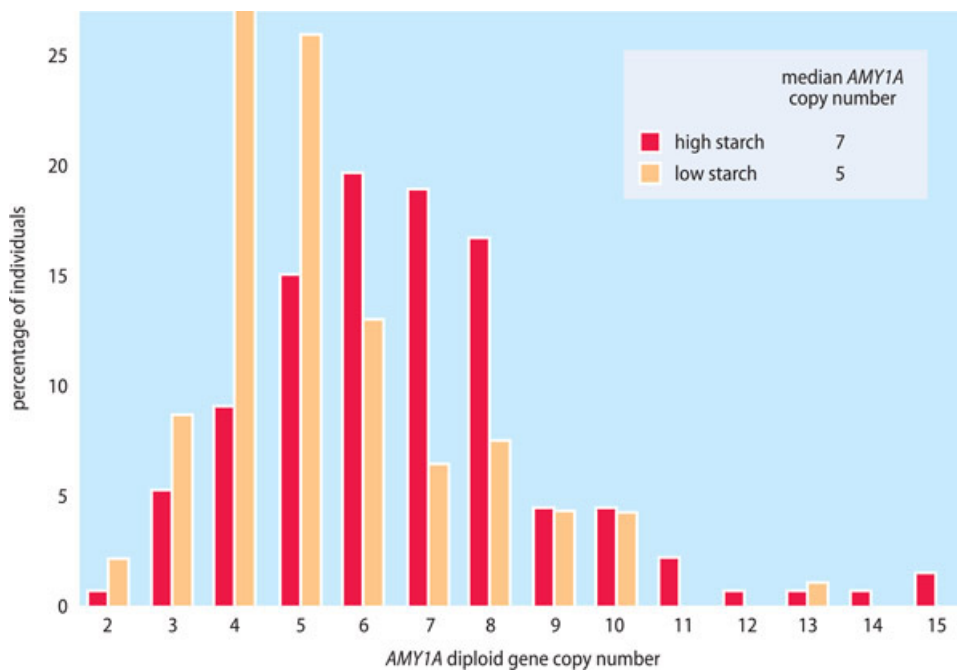


Figure 4.8 Recent acquisition of multiple genes encoding salivary α -amylase as an adaptation to high starch diets in some human populations. The graph illustrates

two points. First, the diploid copy number of the human salivary α -amylase gene *AMY1A* is quite variable (between 2 and 15 from this data set) but generally high (chimpanzees have a single copy of this gene). Secondly, individuals in populations that have high-starch diets have significantly more *AMY1A* gene copies than those in populations with low-starch diets. (From Perry GH et al. [2007] *Nat Genet* 39:1256–1260; PMID 17828263. With permission from Macmillan Publishers Ltd.)

Some adaptations also provide some protection against infectious diseases that are endemic in certain areas of the planet, notably mosquito-borne malaria. And as agriculture developed, DNA variants were selected in response to changes of diet (see [Table 4.5](#)).

Adaptations to local environments can involve downregulating a physiological function, as in reduced skin pigmentation in Europeans. Ultraviolet (UV) radiation in sunlight is needed for a photolytic reaction that occurs in a deep layer of the dermis, and this is the principal source of vitamin D3. Dark skins in equatorial populations protect skin cells from DNA damage caused by intense exposure to UV. Populations that migrated to northern latitudes were exposed to less UV, but the potentially reduced ability to make vitamin D3 was offset by an adaptation that reduced the amount of melanin, maximizing UV transmission through skin. The most significant contributor was a nonsynonymous change in the *SLC24A5* gene, resulting in the replacement of alanine at position 111 by threonine (A111T). The *SLC24A5* protein is a type of calcium transporter that regulates melanin production, and the A111T change results in defective melanogenesis. The A111T variant became fixed in European populations as a result of what is called a *selective sweep* (**Box 4.2**).

BOX 4.2 RECENT STRONG POSITIVE SELECTION CAN LEAD TO A SELECTIVE SWEEP WITH LOCAL SUPPRESSION OF GENETIC VARIATION

Positive selection for an advantageous DNA variant can leave distinctive signatures of genetic variation in the DNA sequence. Imagine a large

population of individuals before positive selection occurs for some advantageous DNA variant on a region of, say, chromosome 22. If we were able to scan each chromosome 22 in the population before selection we might expect to find hundreds of thousands of different combinations of genetic variants ([Figure 1](#)).

Now imagine that an advantageous DNA variant arises by mutation on one chromosome 22 copy and then gets transmitted through successive generations. If the advantageous variant is subject to strong positive selection, people who carry it will have significantly higher survival and reproductive success rates. As descendants of the original chromosome 22 copy carrying the variant become more and more common, the selected DNA variant will increase in frequency to become a common allele ([Figure 1B](#)).

The entire chromosome 22 copy is not passed down as a unit: recombination will result in the replacement of some original segments by equivalent regions from other chromosome 22s. A short segment from the original chromosome 22 copy containing the favorable DNA variant and nearby “hitchhiking alleles” will increase in prevalence in a **selective sweep** ([Figure 1B](#)), but the segment will be slowly reduced in size by recombination.

A genomic region that has been subject to a recent selective sweep will demonstrate extremely low heterozygosity levels. The genomic region on chromosome 15 that contains the *SLC24A5* locus in Europeans provides a good practical example—see [Figure 2](#).

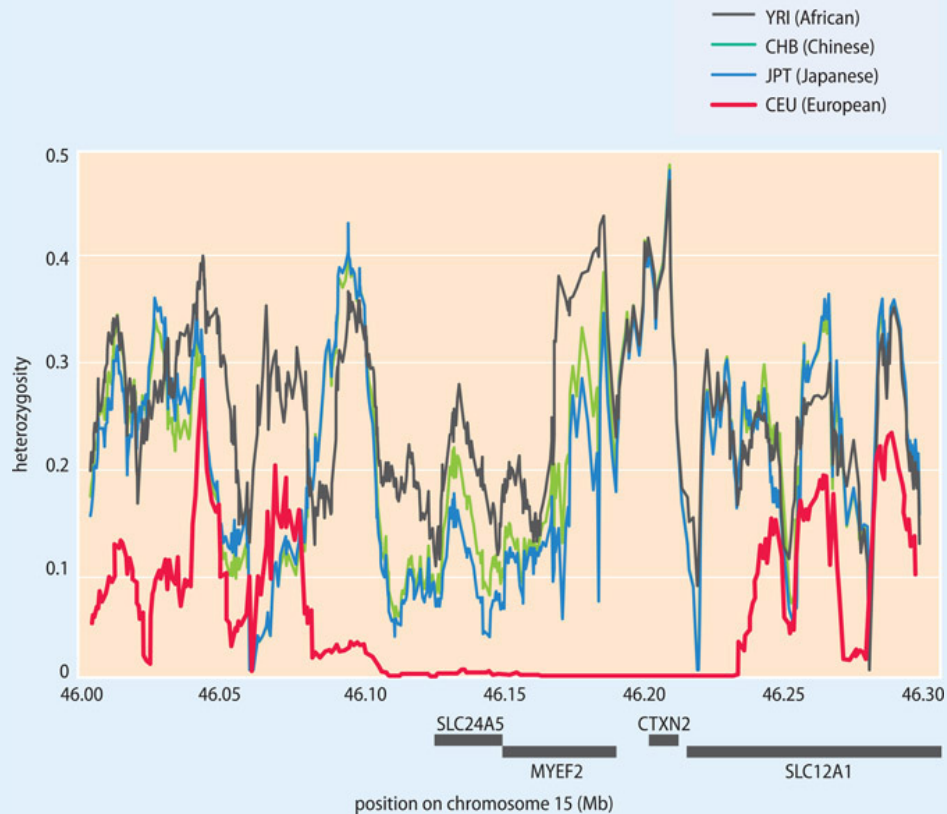


Figure 2 A strong selective sweep acting on an advantageous DNA variant in the *SLC24A5* gene in European populations. Heterozygosity levels in the region containing the *SLC24A5* gene on chromosome 15 were determined for a high density of common SNPs and averaged over 10 kb windows. The observed heterozygosity profiles for this chromosome region are unremarkable in African, Chinese, and Japanese populations. However, in the European population a strong selective sweep for a specific *SLC24A5* variant associated with reduced skin pigmentation has meant that almost all European chromosome 15s share a segment containing the favorable *SLC24A5* variant and hitchhiker alleles at the neighboring *MYEF2* and *CTXN2* loci. The result is a sharp decrease in heterozygosity for this chromosome region. (Adapted from Lamason RL, Mohideen MA, Mest JR et al. [2005] *Science* 310:1782–1786; PMID 16357253. With permission from the AAAS.)

Adaptations to living in malaria-infested regions have often involved increased frequencies of harmful alleles associated with certain blood

disorders, notably sickle-cell disease, thalassemia, and glucose-6-phosphate dehydrogenase deficiency in the case of *Plasmodium falciparum* malaria. Heterozygotes with mutant alleles associated with these diseases exhibit small changes in phenotype that make them comparatively resistant to malaria. **Balancing selection** is involved: the mutant heterozygotes have higher rates of reproductive success than both mutant and normal homozygotes (**heterozygote advantage**)—we consider the details in [Chapter 5](#).

The development of agriculture brought significant changes to the human diet. The domestication of wheat and rice led to high-starch diets, and the domestication of cows and goats led to lifelong consumption of fresh milk in some human populations. Adaptive responses to high-starch diets and extended milk consumption both involved the increased production of enzymes required to metabolize starch or lactose (the major sugar in milk). But the adaptive genetic changes that permitted increased enzyme production were quite different as described below.

- *Gene amplification to permit enhanced starch metabolism.* Salivary α -amylase, the major enzyme needed to break down starch, is produced by the *AMY1A* gene. Our closest animal relative, the chimpanzee, has a single gene copy in the haploid genome but humans normally have multiple *AMY1A* genes. Individuals who take in a large amount of starch in their diets have a significantly higher *AMY1A* copy number and increased capacity to make salivary α -amylase than those used to low-starch diets (**Figure 4.8**).
- *Gene upregulation to permit enhanced lactose metabolism.* Like other mammals, most of the world's human population are lactose intolerant: the ability to digest lactose declines rapidly after weaning as levels of the enzyme lactase fall in the small intestine. In populations who had domesticated cows and goats, however, a cultural tradition developed of lifelong drinking

of animal milk. Strong vertical transmission of this cultural practice led to selection for regulatory DNA variants allowing lifelong expression of the lactase gene, *LCT* (lactase persistence). In each case mutations occur in a regulatory DNA region located about 14 kb upstream of the start codon.

Generating protein diversity by gene duplication: the example of olfactory receptor genes

Diverse forms of a protein – protein isoforms – can be generated by alternative mechanisms. Some occur at the level of post-transcriptional processing of an individual gene, including alternative splicing, alternative use of promoters, RNA editing and alternative polyA-addition (summarized in [Table 4.4](#) above). Here we turn our attention to a different way of generating protein isoforms: gene duplication. Recall that gene duplication has been important in evolution in generating families of genes that can develop somewhat different functions. But in some cases gene duplication has been exploited to provide large numbers of protein isoforms that carry out the same basic function. The outstanding example is how gene duplication provides an extraordinary number of olfactory receptors.

The interaction of olfactory receptors on sensory neurons with odorants in the lining of the nose allows us to detect a huge number, possibly even one trillion, different smells, but there is pronounced variation between individuals in the ability to detect specific odors. As a single odorant may be recognized by multiple ORs, and one OR may recognize multiple odorants, there seems to be a combinatorial code of binding of different receptor so that ultimately, different odorants are represented as different combinations of activated ORs.

Perhaps not surprisingly, the olfactory receptor (OR) gene family is our largest protein-coding gene family (with ~400 OR genes plus ~450 OR pseudogenes) and this family demonstrates the greatest variation in gene content of any human gene family. In addition to the pseudogenes, alleles

for deleterious mutations at functional OR gene loci are both common in the population and highly variable between individuals ([Figure 4.9](#)).

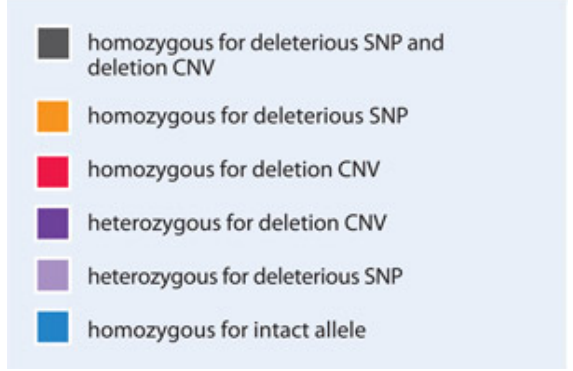
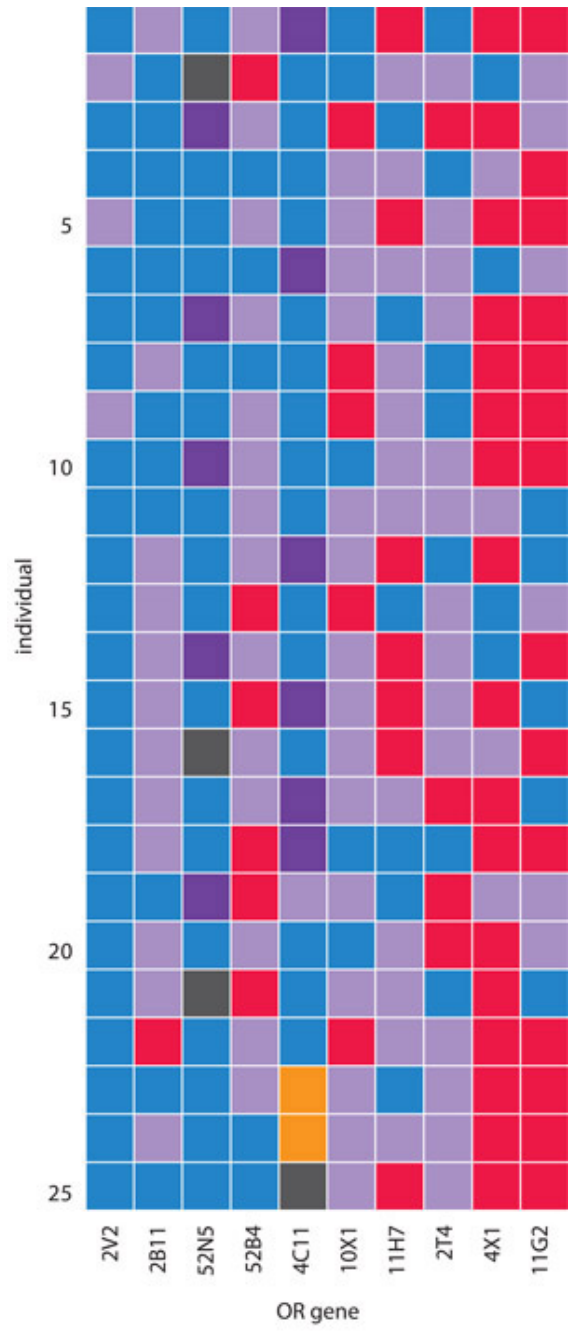


Figure 4.9 Common deleterious variants in human olfactory receptor (OR) genes.

Colored squares represent recorded genotypes for different deleterious allele combinations at ten functional OR gene loci, for which both an intact and inactive allele are common in the population. Each column represents one of the ten OR genes studied; each row represents an individual person. (Data courtesy of Doron Lancet and Tsviya Olender, Department of Molecular Genetics, Weizmann Institute of Science, Rehovot.)

4.4 EXTRAORDINARY GENETIC VARIATION IN THE IMMUNE SYSTEM

Protein variants originating from a single gene locus through sequence changes at the DNA level are often rare. However, some proteins need to be able to recognize harmful foreign molecules introduced into the body that are subject to independent genetic control, and they can show very significant variation. The most extreme variation occurs in the case of immune system proteins working to recognize foreign molecules from microbial pathogens, ultimately leading to killing of microbes or virus-infected cells.

Pronounced genetic variation in four classes of immune system proteins

Our immune systems have a tough task. They are engaged in a relentless battle to protect us from potentially harmful microbial and viral pathogens. Not only must we be protected against a bewildering array of pathogens but, in addition, new forms of a pathogen can rapidly develop by mutation (in an effort to escape detection); that provides new challenges to which we must continuously adapt.

Four types of proteins are primarily involved, belonging to two broad classes as listed below and illustrated in [Figure 4.10](#).

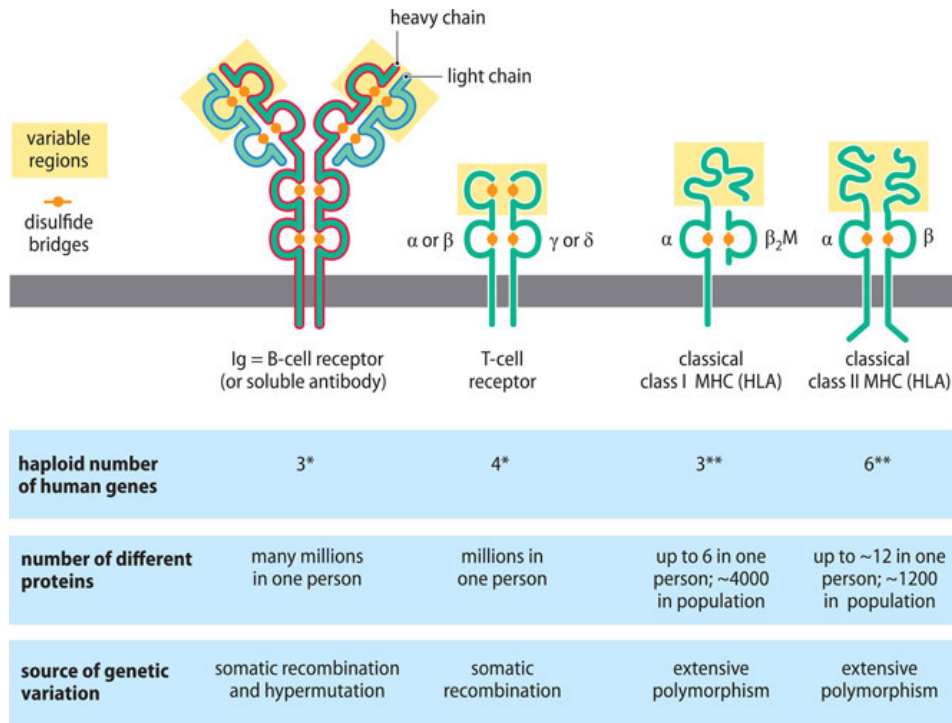


Figure 4.10 Extreme variation in four types of proteins needed to recognize foreign antigens. Immunoglobulins (Igs), T-cell receptors and MHC (major histocompatibility complex) proteins are heterodimers with similar structures: globular domains (maintained by intrachain disulfide bridges) and N-terminal *variable regions* that bind foreign antigens (but otherwise have a conserved sequence, known as *constant regions* in Igs). They are cell surface receptors (except that Igs in activated B cells become secreted antibodies). Only a few human genes encode an Ig or T-cell receptor, but nevertheless many different proteins are made because of special genetic mechanisms in B and T lymphocytes and because of selection for heterozygosity of HLA antigens. B2M is β_2 -microglobulin, the non-polymorphic light chain of class I HLA antigens. *It is estimated that we can make 10^7 – 10^8 different antibodies and close to the same number of different T-cell receptors. **Figure 1 of [Box 4.3](#) on page 105 shows genes encoding the classical (= highly polymorphic) HLA proteins.

- **Immunoglobulins.** Expressed on the surface of B cells in the bone marrow or secreted as soluble immunoglobulins (antibodies) by activated B cells, their main task is to recognize and bind specific foreign antigens. They can bind and neutralize toxins released by microbes, inhibit viruses from infecting host cells, and activate both

complement-mediated lysis of bacteria and phagocytosis.

- *T-cell receptors*. Displayed on the surface of T cells, they work in cell-mediated immunity, along with proteins encoded by the major histocompatibility complex (MHC; known in humans as the HLA complex).
- *Classical class I MHC (HLA) proteins*. After synthesis they can recognize and bind foreign antigens within cells that have been infected by a virus or other cell surface pathogen. They transport the foreign antigens to the cell surface to be recognized by cytotoxic T lymphocytes (*antigen presentation*), after which the infected cells are killed by the cytotoxic T lymphocytes.
- *Classical class II MHC (HLA) proteins*. They are displayed on the surface of very few types of cells, notably immune system cells, and transport foreign intracellular antigens to the cell surface to be recognized by helper T lymphocytes.

As shown in [Figure 4.10](#), there is an extraordinary variety of each of the above types of proteins, but their diversity arises by two different fundamental mechanisms. For MHC proteins, natural selection works to promote heterozygosity (if you have two protein variants at multiple HLA genes, you have a higher chance of recognizing and dealing with harmful antigens). But the extraordinary variation of immunoglobulins (Igs) and T-cell receptors (TCRs) comes from specialized mechanisms that are *programmed* to rearrange Ig genes in maturing B cells and TCR genes in maturing T cells.

A key point is that any *one* individual makes a huge variety of immunoglobulins and T-cell receptors because the rearrangement of Ig or TCR genes is *cell-specific*. Different B cells in a single individual can produce different immunoglobulins, and different T cells can produce different T-cell receptors. By contrast, the extensive variety of classical MHC proteins is apparent at a *population* level. Classical MHC proteins are

highly polymorphic but a single person has a limited number of them, having at most two alleles at each of a small number of polymorphic HLA loci.

Programmed and random post-zygotic genetic variation

As well as the genetic variation that we inherit from our parents, our DNA undergoes some changes as we develop from the single-celled zygote and throughout life. Post-zygotic genetic variation can involve mutations that occur randomly in all our cells. Therefore, although all our cells originate from the zygote, we are genetic **mosaics** who carry genetically different cells.

Much of the somatic genetic variation between the cells of an individual is due to copy number variants; mosaic patterns of copy number variations are a feature of human neurons, for example. Small-scale mutations also arise post-zygotically that often have no functional consequences. Whereas an inherited mutation will appear in all of our nucleated cells, a somatic mutation will only be present in the cell in which it arose plus any cell lineages that arise by cell division from the progenitor cell. Some somatic mutations give rise to disease if they occur at an early stage in development or result in abnormal tumor cell populations (we consider mosaicism for pathogenic mutations in [Box 5.3](#)).

In addition to random somatic mutations, programmed DNA changes are targeted to occur at immunoglobulin genes in the DNA of maturing B cells and at T-cell receptor genes in maturing T cells. We inherit from each parent just three immunoglobulin genes (*IGH*, *IGK*, and *IGL*) and four T-cell receptor genes (*TRA*, *TRB*, *TRD*, and *TRG*). However, the immunoglobulin genes in maturing B cells and the T-cell receptor genes in maturing T cells are programmed to undergo DNA changes *in a cell-specific way*: specific types of somatic DNA changes occur at these genes, but there is also a high degree of randomness so that the precise DNA changes vary from one B cell to the next B cell in the same individual, or from one T cell to the next. The net effect of these post-zygotic changes is to endow a single individual

with huge numbers of different immunoglobulin gene variants and of different T-cell receptor gene variants that can be pressed into service. Four mechanisms are responsible, as described in the next section.

Somatic mechanisms allow cell-specific production of immunoglobulins and T-cell receptors

Although a human zygote has a total of six immunoglobulin genes and eight T-cell receptor genes, somatic DNA changes in maturing B cells and T cells allow us to develop millions of different immunoglobulin (Ig) gene variants and millions of different T-cell receptor gene variants. Up to four mechanisms are involved, as described below.

Combinatorial diversity via somatic recombination

Each Ig and T-cell receptor gene is made up of a series of repeated gene segments that specify discrete segments of the protein, and different combinations of gene segments are used in protein production in different B cells or in different T cells of each individual. The different combinations of gene segments are made possible by somatic recombinations that occur in Ig genes in mature B cells and in T-cell receptor genes in mature T cells.

As an example, consider the gene segments that specify an Ig heavy chain. The variable region, which is involved in antigen recognition, is encoded by three types of repeated gene segment: V (encoding the first part of the variable region), D (diversity region), and J (joining region). The constant region defines the functional class of immunoglobulin (IgA, IgD, IgE, IgG, or IgM) and is encoded by repeated C gene segments (that have coding sequences split by introns). For each type of segment, the repeats are similar in sequence but nevertheless show some differences.

The first step in making an Ig heavy chain requires two sequential recombination events within the *IGH* gene of a maturing B cell. The end result is that one V gene segment, one D gene segment, and one J gene

segment are fused together to form a continuous VDJ coding sequence that will specify the variable region ([Figure 4.11](#)).

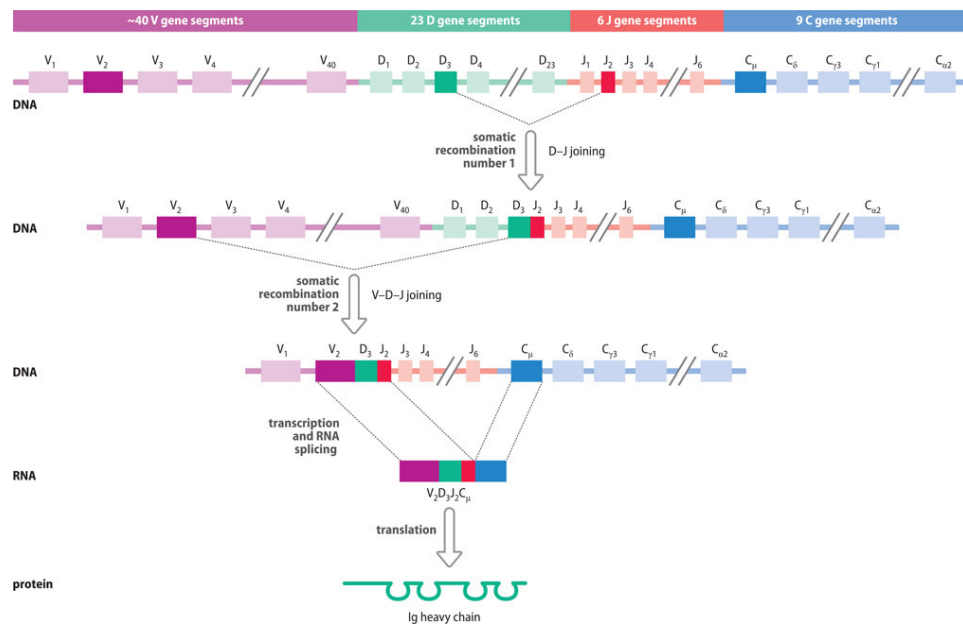


Figure 4.11 Somatic recombination in the *IGH* gene of B cells is used to make cell-specific immunoglobulin heavy chains. The human *IGH* gene has multiple but slightly different repeats for each of four types of gene segments: V (first part of variable region), D (diversity region), J (joining region), and C (constant region; although not shown here, each of the C gene segments has a coding sequence split by introns). An immunoglobulin heavy chain is made by bringing together coding sequences from one each of these four types of segments (shown here as filled boxes). Two sequential somatic recombinations produce first D–J joining, then a mature, functional VDJ coding sequence unit, which is effectively a large novel exon. In this example, the successful combination is V₂D₃J₂, but the choice of combinations is *cell-specific*. Once a functional VDJ exon has been assembled, transcription is initiated starting with this exon and RNA splicing joins the VDJ coding sequence to coding sequences in the *closest* constant (C) region gene segment, in this case C_μ. Another type of somatic recombination (known as *class switching*, but not shown here) can change the position of C gene segments so that other C gene segments can be used instead of C_μ to give alternative classes of immunoglobulin.

Once assembled, a VDJ coding unit activates transcription. RNA splicing fuses the VDJ transcript to transcribed coding sequences within the nearest C gene segments, initially C μ (see [Figure 4.11](#)) and then, through alternative splicing, either C μ or C γ . The first immunoglobulins to be made by a B cell are membrane-bound IgM and then IgD. Subsequently, as B cells are stimulated by foreign antigen and helper T lymphocytes, they secrete IgM antibodies.

Later in the immune response, B cells undergo class-switching (also called isotype switching) to produce different antibody classes. Here, another type of somatic recombination positions an alternative C gene segment to be nearest to the J gene segments: either a C, C ϵ , or C α gene segment to produce respectively an IgG, IgE, or IgA antibody.

The key point about the somatic recombination events that diversify the variable regions is that they occur randomly in each maturing B or T cell, respectively (with the proviso that one each of the different repeated gene segments are brought together). That is, the genetic variation is produced by *cell-specific* recombinations. The V2D3J2 combination in [Figure 4.11](#) might occur in one maturing B cell, but a V4D2J9 unit or a V38D5J4 unit, for example, might be generated in neighboring B cells in the same person. The variable region of the T-cell receptor β chain is also formed by the same kind of VDJ recombination, but for both Ig light chains and T-cell receptor α chains a single VJ recombination is involved because their genes lack diversity gene segments.

Additional diversity generation

Two or three additional mechanisms are responsible for generating diversity in Igs and T-cell receptors as listed below. These mechanisms, together with V(D)J recombination, endow each of us with the potential of making many trillions of different antigen-binding sites, both for immunoglobulins and T-cell receptors. As required, individual B and T cells that enable successful

recognition of foreign antigen are induced to proliferate to make identical clones with the same antigen specificity as the original cell.

- *Junctional diversity.* The somatic recombination mechanisms that bring together different gene segments in Ig or T-cell receptor genes variably add or subtract nucleotides at the junctions of the selected gene segments.
- *Protein chain combinatorial diversity.* Igs and T-cell receptors are heterodimers, and diversity is compounded by unique combinations of two unique protein chains. Note, however, that a B cell, for example, makes just one type of Ig. Although each diploid B cell has six Ig genes, in each B cell only one of the two *IGH* alleles is (randomly) selected to make a heavy chain (*allelic exclusion*) and only one of the four light chain genes in each B cell is ever used (a combination of *light chain exclusion*—to select either a k or l light chain—plus allelic exclusion).
- *Somatic hypermutation.* This mechanism applies only to Igs and is used to further increase variability in the variable region after somatic recombinations have produced functional VDJ or VJ units. When B cells are stimulated by a foreign antigen, an activation-induced cytidine deaminase is produced by the activated B cell that deaminates cytidine to uridine. The uridines are variably repaired by base excision repair (see above), and the end result is that multiple nucleotides in the variable region are mutated.

MHC (HLA) proteins: functions and polymorphism

The HLA complex is the human major histocompatibility complex (MHC). The latter name came from the observation that certain MHC genes are the primary determinants in transplant rejection. That, of course, is an artificial situation: the normal function of MHC genes is to assist certain immune

system cells, notably helping T cells to identify host cells that harbor an intracellular pathogen such as a virus.

Some MHC genes—called classical MHC genes—are extremely polymorphic. They are subject to positive selection to maximize genetic variation (people who are heterozygous for multiple MHC loci will be better protected against microbial pathogens and have higher reproductive success rates). The classical MHC proteins are deployed on the cell surface as heterodimers ([Figure 4.10](#)). They serve to bind peptide fragments derived from the intracellular degradation of pathogen proteins and display them on the surface of host cells (**antigen presentation**) so that they can be recognized by T cells. Appropriate immune reactions are then initiated to destroy infected host cells. There are two major classes of classical MHC proteins, as detailed below.

Class I MHC proteins

Class I MHC proteins are expressed on almost all nucleated host cells. Their job is to help cytotoxic T lymphocytes (CTLs) to recognize and kill host cells that have been infected by a virus or other intracellular pathogen. When intracellular pathogens synthesize protein within host cells, a proportion of the protein molecules get degraded by proteasomes in the cytosol. The resulting peptide fragments are transported into the endoplasmic reticulum. Here, a newly formed class I MHC protein binds a peptide and is exported to the cell surface, where it is recognized by a CTL with a suitable receptor.

Because of cell-specific somatic recombinations (similar to those in [Figure 4.11](#)), individual CTLs make unique T-cell receptors that recognize *specific* class I MHC–peptide combinations. If the bound peptide is derived from a pathogen, the CTL induces killing of the host cell. Note that a proportion of normal host cell proteins also undergo degradation in the cytosol and the resulting self-peptides are bound by class I MHC proteins and displayed on the cell surface. But there is normally no immune

response (starting in early fetal life, CTLs that recognize MHC-self peptide are programmed to be deleted, to minimize autoimmune responses).

Class II MHC proteins

Class II MHC proteins are expressed in professional antigen-presenting cells: dendritic cells, macrophages, and B cells. These cells also express class I MHC proteins but, unlike most cells, they make co-stimulatory molecules needed to initiate lymphocyte immune responses.

Whereas class I MHC proteins bind peptides from *endogenous* proteins (those made within the cytosol, such as a viral protein made after infection of that cell), class II MHC proteins bind peptides derived from *exogenous* proteins that have been transported into the cell (by endocytosis of a microbe or its products) and delivered to an endosome, where limited proteolysis occurs. The resulting peptide fragments are bound by previously assembled class II MHC proteins and transported to the cell surface so that a helper T lymphocyte with an appropriate receptor recognizes a specific class II MHC–peptide combination. (Helper T cells have critical roles in coordinating immune responses by sending chemical signals to other immune system cells.)

MHC restriction

T cells recognize a foreign antigen only after it has been degraded and become associated with MHC molecules (***MHC restriction***). A proportion of all normal proteins in a cell are also degraded, and the resulting peptides are displayed on the cell surface, complexed to MHC molecules. MHC proteins cannot distinguish self from nonself, and even on the surface of a virus-infected cell the vast majority of the many thousands of MHC proteins on the cell surface bind peptides derived from host cell proteins rather than from virus proteins.

The rationale for MHC restriction is that it provides a simple and elegant solution to the problem of how to detect intracellular pathogens—it allows T cells to survey a peptide library derived from the entire set of proteins in a cell *but only after the peptides have been displayed on the cell surface*.

MHC polymorphism

MHC polymorphism is pathogen-driven: strong selection pressure favors the emergence of mutant pathogens that seek to evade MHC-mediated detection. The MHC has evolved two counterstrategies to maximize the chance of detecting a pathogen. First, gene duplication has provided multiple MHC genes that make different MHC proteins with different peptide-binding specificities. Secondly, many of the MHC genes are extraordinarily polymorphic, producing the most polymorphic of all our proteins ([Table 4.6](#)).

TABLE 4.6

STATISTICS FOR THE SIX MOST POLYMORPHIC HLA LOCI

HLA gene	–A	–B	–C	–DPB1	–DQB1	–DRB
Number of alleles or DNA variants	7354	8756	7307	1909	2193	3094
Number of protein variants	4302	5287	4042	1198	1386	2107

Data were derived from the European Bioinformatics Institute’s IPD-IMGT/HLA database (release 3.47, January 2022). The statistics for these and additional loci are available at <http://www.ebi.ac.uk/ipd/imgt/hla/about/statistics>

The polymorphism of classical MHC proteins is focused on amino acids that form the antigen-binding pockets: different alleles exhibit different

peptide-binding specificities. A form of long-standing balancing selection (also called overdominant selection) seems to promote MHC polymorphism. Heterozygosity is favored (presumably the ability to produce many different HLA proteins affords us greater protection against pathogens), and certain heterozygote geno-types seem to display greater fitness than others.

The balancing selection seems to have originated before the speciation event leading to evolutionary divergence from the great apes. HLA polymorphism is therefore exceptional in showing trans-species polymorphism: a human HLA allele may be more closely related in sequence to a chimpanzee HLA allele than it is to another human HLA allele. For example, human HLA-DRB1*0701 and HLA-DRB1*0302 show 31 amino acid differences out of 270 amino acid positions, but human HLA-DRB1*0701 and its chimpanzee equivalent, Patr-DRB1*0702, show only 2 differences out of 270.

The medical importance of the HLA system

The HLA system is medically important for two principal reasons. First, the high degree of HLA polymorphism poses problems in organ and cell transplantation. Secondly, certain HLA alleles are risk factors for individual diseases, notably many autoimmune diseases and certain infectious diseases; other HLA alleles are protective factors, being negatively correlated with individual diseases.

Transplantation and histocompatibility testing

After organ and cell transplantation, the recipient's immune system will often mount an immune response against the transplanted donor cells (the graft), which carry different HLA antigens from those of the host cells. The immune reaction may be sufficient to cause rejection of the transplant (but corneal transplants produce minimal immune responses—the cornea is one of a few immune privileged sites that actively protect against immune

responses in several ways, including having a much reduced expression of class I HLA antigens).

Bone marrow transplants and certain stem cell transplants can also result in graft-versus-host disease (GVHD) when the graft contains competent T cells that attack the recipient's cells. GVHD can even occur when donor and recipient are HLA-identical because of differences in minor (non-HLA) histocompatibility antigens.

Immunosuppressive drugs are used to suppress immune responses after transplantation, but transplant success depends largely on the degree of HLA matching between the cells of the donor and the recipient. Histocompatibility testing (also called tissue typing) involves assaying HLA alleles in donor tissues so that the best match can be found for prospective recipients. The key HLA loci are the most polymorphic ones: *HLA-A*, *-B*, *-C*, *-DRB1*, *-DQB1*, and *-DPB1* ([Table 4.6](#) and **Box 4.3**).

BOX 4.3 HLA GENES, ALLELES, AND HAPLOTYPES

HLA GENES

The HLA complex spans 3.6 Mb on the short arm of chromosome 6. The 253 genes in the complex include the 18 protein-coding HLA genes shown in [Figure 1](#), ranging from *HLA-DPB1* (closest to the centromere) to *HLA-F*. Genes in the class I region make the heavy chain of class I HLA antigens (the non-polymorphic class I HLA light chain, b2-microglobulin, is encoded by a gene on chromosome 15); the class II region has genes encoding both chains of class II HLA antigens. The intervening region does not contain any HLA genes, but it does contain multiple genes with an immune system function and is sometimes referred to as the class III region.

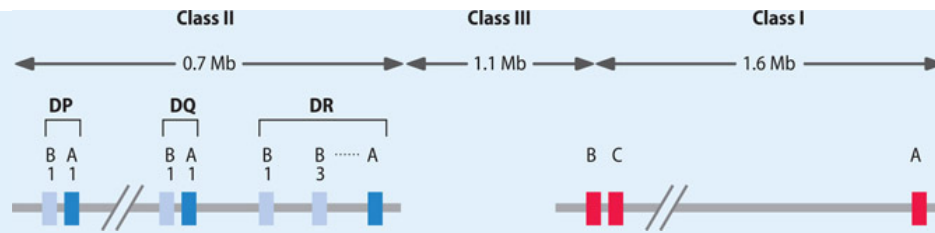


Figure 1 Classical (polymorphic) HLA genes within the HLA complex at 6p21.3.

Genes in the class II HLA region encode a chains (dark shading) and b chains (pale shading) that pair up to form heterodimers within specific classes as indicated by horizontal bars above (DP, DQ, DR). Classical class I HLA genes encode a polymorphic class I a chain that forms a heterodimeric protein with the non-polymorphic b-microglobulin chain encoded by a gene on chromosome 15. Within the class I and class II HLA regions are several other non-polymorphic HLA genes and many HLA-related pseudogenes not shown here. The class III region includes certain complement genes. Some additional genes with an immune system function are found within the HLA complex plus some functionally unrelated genes such as the steroid 21-hydroxylase gene.

HLA ALLELES

Because of their extraordinary polymorphism, alleles of the classical, highly polymorphic HLA genes have been typed for many decades at the protein level (using serological techniques with panels of suitably discriminating antisera). The number of alleles that can be distinguished in this way is very high, for example 28 HLA-A alleles, 50 HLA-B alleles, and 10 HLA-C alleles (called Cw for historical reasons; the “w” signifies workshop because nomenclature was updated at regular HLA workshops).

Serological HLA typing is still used when rapid typing is required, as in the case of solid organ transplants (in which the time between the chilling of an organ and the time it is warmed by having the blood supply restored needs to be minimized). However, much of modern HLA typing is performed at the DNA level, where very large numbers of alleles can be identified (see [Table 4.6](#)). The complexity means a rather cumbersome nomenclature for HLA alleles identified at the DNA level—see [Table 1](#) for examples.

TABLE 1**HLA ALLELE NOMENCLATURE**

NOMENCLATURE	MEANING
<i>HLA-DRB1</i>	an HLA gene (encoding the β chain of the HLA-DR antigen)
<i>HLA-DRB1*13</i>	alleles that encode the serologically defined HLA-DR13 antigen
<i>HLA-DRB1*13:01</i>	one specific HLA allele that encodes the HLA-DR13 antigen
<i>HLA-DRB1*13:01:02</i>	an allele that differs from <i>DRB1*13:01:01</i> by a synonymous mutation
<i>HLA-DRB1*13:01:01:02</i>	an allele that differs from <i>DRB1*13:01:01</i> by having a mutation outside the coding region
<i>HLA-A*24:09N</i>	a null allele related by sequence to alleles encoding the HLA-A24 antigen

For more details see <http://hla.alleles.org/>.

HLA HAPLOTYPES

The genes in the HLA complex are highly clustered, being confined to an area that represents only about 2 % of chromosome 6. Genes that are close to each other on a chromosome are usually inherited together because there is only a small chance that they will be separated by a recombination event occurring in the short interval separating the genes. Such genes are said to be *tightly linked* (we consider genetic linkage in detail in [Section 8.1](#)).

A **haplotype** is a series of alleles at linked loci on an *individual* chromosome; haplotypes were first used widely in human genetics with reference to the HLA complex. See [Figure 2](#) for how haplotypes are established by tracking the inheritance of alleles in family studies. Note that because the HLA genes are very closely linked, recombination within the HLA complex is rare.

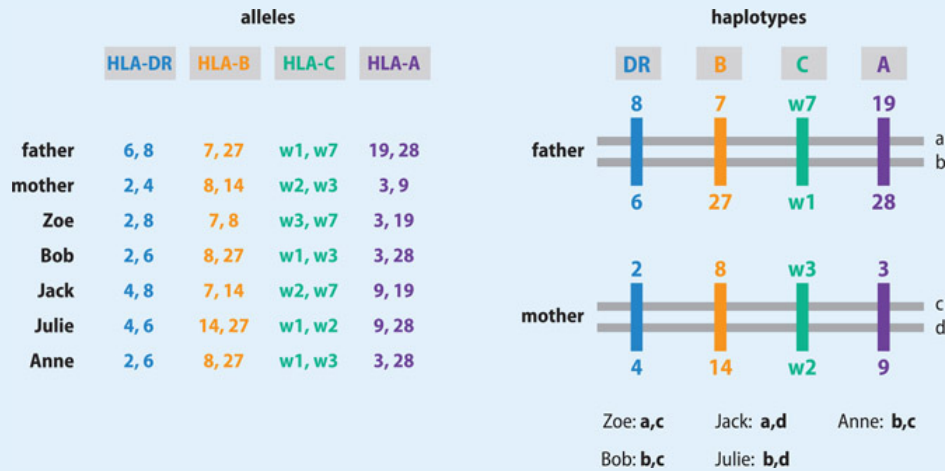


Figure 2 Deriving HLA haplotypes from family studies. Father, mother, and their three daughters, Zoe, Julie, and Anne, and two sons, Bob and Jack, have been tissue typed using serological reagents for four HLA antigens as shown at the left. By tracking which parental alleles have been passed on to individual children it is possible to deduce the parental HLA haplotypes. Father has one chromosome 6 with the HLA haplotype DR8, B7, Cw7, A19 (haplotype a) and another chromosome 6 with the HLA haplotype DR6, B27, Cw1, A28 (haplotype b). Similarly, mother has haplotypes c (DR2, B8, Cw3, A3) and d (DR4, B14, Cw2, A9). Father has transmitted haplotype a to Zoe and Jack, and haplotype b to Bob, Julie, and Anne. Mother has transmitted haplotype c to Zoe, Bob, and Anne, and haplotype d to Jack and Julie.

HLA disease associations

By displaying peptide fragments on host cell surfaces, HLA proteins direct T cells to recognize foreign antigens and initiate an immune response

against cells containing viruses or other intracellular pathogens. Because HLA proteins differ in their ability to recognize specific foreign antigens, people with different HLA profiles might be expected to show different susceptibilities to some infectious diseases.

In autoimmune diseases, the normal ability to discriminate self-antigens from foreign antigens breaks down, and autoreactive T cells launch attacks against certain types of host cells. Certain HLA antigens are very strongly associated with individual diseases, such as type 1 diabetes and rheumatoid arthritis; in general, genetic variants in the HLA complex are the most significant genetic risk factors that determine susceptibility to autoimmune diseases. Determining to what extent HLA variants are directly involved in the pathogenesis and how much is contributed by other variants that lie in the immediate vicinity of the HLA genes (and outside the HLA complex) is a major area of research—we consider HLA associations with individual diseases in some detail in [Chapter 8](#).

SUMMARY

- The DNA in our cells accumulates changes over time (mutations) that usually have no significant effect on the phenotype.
- Some mutations adversely affect how genes work or are expressed; they can be associated with disease, and because at least some people carrying them have a lower reproductive fitness they tend to be removed from populations (purifying selection).
- Very occasionally, a mutation may result in some benefit and may accumulate in frequency if it endows individuals with increased reproductive fitness (positive selection).

- Large-scale changes to DNA can result from abnormalities in chromosome segregation and recombination. Smaller-scale changes typically result from unrepaired errors in DNA replication or unrepaired chemical attacks on DNA.
- DNA is damaged within living cells and organisms by various types of chemical attack that break covalent bonds in DNA or form inappropriate covalent bonds with bases. One or both DNA strands may be broken, bases or nucleotides may be deleted, or inappropriate chemical groups may be covalently bonded to the DNA.
- Much of the chemical damage to DNA is caused by highly reactive chemicals produced naturally inside our cells.
- According to the type of chemical damage to DNA, different cellular pathways are used to repair a DNA lesion. Direct reversal of the damage-causing chemical steps is rare, and individual pathways often involve many molecular components.
- DNA variants often have low frequencies. More common variants, with a frequency of more than 0.01, are sometimes described as DNA polymorphisms.
- A single nucleotide variant (or polymorphism) involves the substitution of one nucleotide for another at a specific location. Nucleotide substitutions are nonrandom—for example, C → T substitutions are particularly common in vertebrate DNA.
- An indel is a site where variants differ by lacking or possessing one or a few nucleotides.

- Some DNA variants differ by having different numbers of copies of a tandemly repeated DNA sequence, producing length variation. Microsatellite variants are DNA sequences that show small length differences as a result of having fewer or more tandem copies of a simple repeat sequence with between one and four nucleotides.
- Structural variation results from large-scale changes in DNA. In balanced structural variation, the variants do not differ in DNA content. In unbalanced structural variation, there is substantial length variation between variants that often occurs as a result of copy number variation for a long nucleotide sequence.
- In population-based genome sequencing, whole diploid genomes from multiple individuals are sequenced, providing comprehensive data on human genetic variation.
- Recent positive selection for genetic variants in different human populations has allowed adaptation to different local environments and to major dietary changes.
- Gene duplication is the basis of our diverse repertoire of olfactory receptors.
- To identify foreign antigens efficiently, each of us makes a huge variety of immunoglobulins and T-cell receptors. We inherit only three immunoglobulin and four T-cell receptor genes from each parent, but cell-specific somatic rearrangements in maturing B and T cells endow us with huge numbers of different immunoglobulin and T-cell receptor gene variants.

- Our most polymorphic proteins are produced by genes in the HLA complex (the human major histo-compatibility complex). HLA proteins recognize and bind peptides from processed foreign proteins and present them on cell surfaces so that they can be recognized by specific T-cell receptors.
- The extreme polymorphism of HLA proteins means that recipients of tissue or organ transplants often mount strong immune responses to the foreign tissue. Tissue typing seeks to find reasonable matches between HLA antigens expressed by donor tissue and prospective recipients.

QUESTIONS

Questions can be downloaded by visiting the following link, under Support Materials: www.routledge.com/9780367490812.

FURTHER READING

DNA damage and DNA repair

Barnes DE & Lindahl T (2004) Repair and genetic consequences of endogenous DNA base damage in mammalian cells. *Annu Rev Genet* 38:445–476; PMID 15568983.

Ciccia A & Elledge SJ (2010) The DNA damage response: making it to safe to play with knives. *Mol Cell* 40:179–204; PMID 20965415. [An authoritative review on both DNA damage and repair, including detailed tabulation of frequencies of different types of DNA damage and of inherited disorders of DNA damage responses/DNA repair.]

Rass U (2007) Defective DNA repair and neurodegenerative disease. *Cell* 130:991–1004; PMID 17889645.

Large-scale human population genomics and genotype-phenotype correlation projects

Bycroft C (2018) The UK Biobank resource with deep phenotyping and genomic data. *Nature* 562:203–209; PMID 30305743.

Karczewski KJ (2020) The mutational constraint spectrum quantified from variation in 141 456 humans. *Nature* 581:434–443; PMID 32461654.

Nature's Genome Aggregation Database (gnomAD) Literature Collection available at <https://www.nature.com/immersive/d42859-020-00002-x/index.html>

The All of Us Research Program at www.allofus.nih.gov

The gnomAD genome aggregation database at <https://gnomad.broadinstitute.org/>

Structural variation, copy number variation, and indels

Alkan C (2011) Genome structural variation and genotyping. *Nature Rev Genet* 12:363–376; PMID 21358748.

Collins RL (2020) A structural variation reference for medical and population genetics. *Nature* 581:385–386; PMID 32461652.

Conrad DF (2010) Origins and functional impact of copy number variation in the human genome. *Nature* 464:704–712; PMID 19812545.

Mullaney JM (2010) Small insertions and deletions (INDELs) in human genomes. *Hum Mol Genet* 19:R131–R136; PMID 20858594.

Human mutation distribution and mutation rates

Campbell CD & Eichler EE (2013) Properties and rates of germline mutations in humans. *Trends Genet* 29:575–584; PMID 23684843.

Gonzalez-Perez A (2019) Local determinants of the mutational landscape of the human genome. *Cell* 177:101–114; PMID 30901533.

5

Single-gene disorders: inheritance patterns, phenotype variability, and allele frequencies

DOI: [10.1201/9781003044406-5](https://doi.org/10.1201/9781003044406-5)

CONTENTS

[5.1 INTRODUCTION: TERMINOLOGY, ELECTRONIC RESOURCES, AND PEDIGREES](#)

[5.2 THE BASICS OF MENDELIAN AND MITOCHONDRIAL DNA INHERITANCE PATTERNS](#)

[5.3 UNCERTAINTY, HETEROGENEITY, AND VARIABLE EXPRESSION OF MENDELIAN PHENOTYPES](#)

[5.4 ALLELE FREQUENCIES IN POPULATIONS](#)

[SUMMARY](#)

[QUESTIONS](#)

[FURTHER READING](#)

Genes are functional units of DNA that make some product needed by cells, ultimately either the polypeptide chain of a protein or a functional noncoding RNA. In this chapter, however, we will view genes very largely

as abstract entities and we consider them within the context of single-gene disorders—diseases in which the genetic contribution is determined primarily by one gene locus. Although individually rare, single-gene disorders are important contributors to disease. Knowledge of single-gene disorders also provides a framework for understanding the more complex genetic susceptibility to common disease described in later chapters.

We look first at the patterns of inheritance of single-gene disorders and provide an introductory basis for estimating disease risk according to the inheritance pattern (we provide more advanced disease risk calculations within the context of genetic counseling in [Chapter 11](#)).

We also consider how genes affect our observable characteristics. The term phenotype may be used broadly to describe the observable characteristics of a person, an organ, or a cell. But geneticists also use the word phenotype in a narrower sense to describe only those specific manifestations that arise in response to the differential expression of just one or a small number of genes. These manifestations may be harmful, and we can talk of a disease phenotype.

When the observable manifestations are not disease-associated we normally refer to a character or **trait**, for example blue eyes or blood group O. We can measure and record aspects of the phenotype, such as anatomical and morphological features, behavior, or cognitive functions. Sophisticated laboratory procedures can be used to perform more extensive investigations at the physiological, cellular, and molecular levels.

Genetic variation—changes in the base sequence of our DNA—is the primary influence on the phenotype (identical twins are remarkably similar, after all). But it is not the only determinant of the phenotype: environmental factors also make a contribution. Gene expression can be regulated by various *epigenetic* mechanisms (which, unlike genetic mechanisms, are independent of the base sequence of DNA), and stochastic factors also make a contribution.

As we will see, there can be considerable complexity in the link between genetic variation and phenotype: even in single-gene disorders the phenotype is often variable in affected members of one family. Note that we

do not deal with the molecular basis of single-gene disorders here; that will be covered in later chapters, notably [Chapter 7](#).

We end the chapter by generally looking at the factors that affect allele frequencies in populations, and then focusing on frequencies of disease alleles (which are important practically for calculating disease risk for some types of single-gene disorder). And we explain why some single-gene disorders are common but others are rare.

5.1 INTRODUCTION: TERMINOLOGY, ELECTRONIC RESOURCES, AND PEDIGREES

Background terminology and electronic resources with information on single-gene disorders

An individual gene or DNA sequence in our nuclear DNA has a unique chromosomal location that defines its position, its **locus** (plural **loci**). We can refer to the ABO blood group locus, for example, or the *D3S1563* locus (a polymorphic DNA marker sequence located on chromosome 3).

In human genetics an **allele** means an individual copy of a gene or other DNA sequence that is carried at a locus on a *single* chromosome. Because we are diploid, we normally have two alleles at any one chromosomal locus, one inherited from each parent: a maternal allele and a paternal allele). The term **genotype** describes the combination of alleles that a person possesses at a single locus (or at a number of loci). If both alleles are the same at an individual locus, a person is said to be **homozygous** at that locus and may be referred to as a homozygote. If the alleles are different, even by a single nucleotide, the person is said to be **heterozygous** at that locus, a heterozygote.

Although we are essentially diploid, men have two types of sex chromosomes, X and Y, which are very different in both structure and gene content. As a result, most DNA sequences on the X chromosome do not

have a direct equivalent (allele) on the Y chromosome, and vice versa. Men are therefore **hemizygous** for such loci (because they normally only have one allele). Women normally have two alleles at each locus on the X chromosome.

In humans any genetic character is likely to depend on the expression of a large number of genes and environmental factors. For some, however, a particular genotype at a *single* locus is the primary determinant, being both necessary and sufficient for the character to be expressed under normal circumstances. Such characters are often said to be **Mendelian**, but that implies that a chromosomal locus is involved; a more accurate term is monogenic (which takes into account both chromosomal loci and loci on mitochondrial DNA). Although collectively important, individual single-gene disorders are rare, and common genetic disorders depend on multiple genetic loci.

When a human monogenic disorder (or trait) is determined by a nuclear gene, the disorder (or trait) is said to be **dominant** if it is manifested in the heterozygote (who carries a normal allele and a mutant allele), or **recessive** if it is not. Sometimes two different phenotypes that result from mutations at a single gene locus can be simultaneously displayed by the heterozygote and are said to be co-dominant. For example, the AB blood group is the result of **co-dominant** expression of the A and B blood group phenotypes that are determined by different alleles at the *ABO* blood group locus. As described below, the inheritance of mitochondrial DNA is rather different, with important implications for associated phenotypes.

Various electronic resources provide extensive information on human single-gene disorders and characters ([Box 5.1](#)). GeneReviews[®] provides excellent summaries for many of the more common single-gene disorders that are accessible through the widely used PubMed system for electronic searching of biomedical research literature. We therefore often provide the eight-digit PubMed identifier (**PMID**) for relevant GeneReviews articles on single-gene disorders. The Online Mendelian Inheritance in Man (**OMIM**) database is comprehensive, and we provide six-digit OMIM database numbers for some disorders.

BOX 5.1 ELECTRONIC RESOURCES WITH INFORMATION ON HUMAN SINGLE-GENE DISORDERS AND UNDERLYING GENES

Some of the more comprehensive and stable resources are listed below. There are also many disease-specific databases; we describe some of these in [Section 7.2](#).

GeneReviews (<http://www.ncbi.nlm.nih.gov/books/NBK1116/>; see PMID 20301295 for an alphabetic listing). This series of clinically and genetically orientated reviews of single-gene disorders is made available through NCBI's Bookshelf program. Individual reviews are assigned a PubMed identifier (PMID), an eight-digit number that in this case normally begins with 2030—for example, Huntington disease is at PMID 20301482. The series covers the most common single-gene disorders, and for listed disorders there is more clinical information than in OMIM (see below).

OMIM (<http://www.ncbi.nlm.nih.gov/omim>). The Online Mendelian Inheritance in Man database is the most comprehensive single source of information on human Mendelian phenotypes and the underlying genes. Entries have accumulated text over many years, and the early part of an entry may often reflect historical developments rather than current understanding.

Each OMIM entry has an identifying six-digit number in which the first digit indicates the mode of inheritance. The initial convention for the first digit was: 1, autosomal dominant; 2, autosomal recessive; 3, X-linked; 4, Y-linked; and 5, mitochondrial. However, the distinction between autosomal dominant and autosomal recessive was discontinued for new entries after May 1994. After that date all new entries for auto-somal traits and genes were assigned a six-digit number beginning with the number 6. See the review by [McKusick \(2007\)](#) under Further Reading for further details.

GeneCards[®] (<http://www.genecards.org>). A gene-centered database, this contains a large amount of automatically generated entries, mostly relating to specific human genes. It provides substantial biological information about each gene.

Investigating family history of disease and recording pedigrees

The extent to which a human disorder has a genetic basis can often be established by taking a family history. Medical records may be available to health service professionals for some family members; details of deceased family members and others who may be difficult to contact may be obtained by consulting more accessible family members.

A **pedigree** is a graphical representation of a family tree that uses the standard symbols depicted in [Figure 5.1](#). Generations are often labeled with Roman numerals that increase from top to bottom of the page (toward the youngest generation). Individuals within each generation are given Arabic numerals that increase from left to right. An extended family covering many generations may be described as a kindred. A family member through whom the family is first ascertained (brought to the attention of health care professionals) is known as the *proband* (also called *propositus*—feminine *proposita*) and may be marked with an arrow.

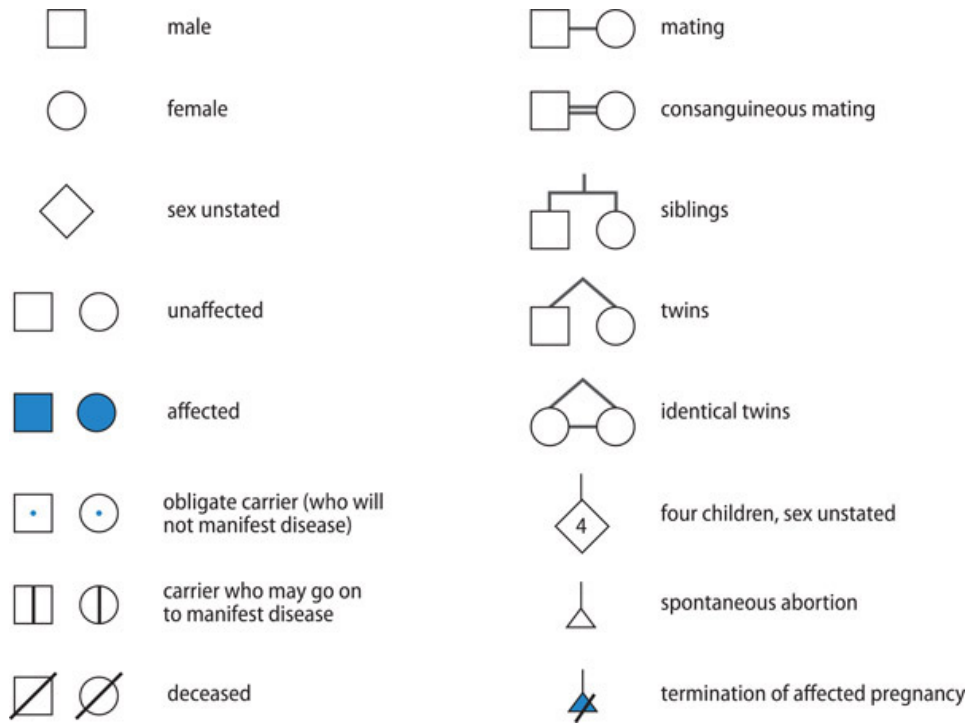


Figure 5.1 Pedigree symbols.

The term **sib** (sibling) is used to indicate a brother or sister, and a series of brothers and sisters is known as a sibship. According to the number of steps in the pedigree that links two family members, they may be classified relatives of the first degree (parent and child; sibs); second degree (grandparent and grandchild; uncle/aunt and nephew/niece; half-sibs); third degree (first cousins), and so on. Couples who have one or more recent ancestors in common are said to be **consanguineous**.

5.2 THE BASICS OF MENDELIAN AND MITOCHONDRIAL DNA INHERITANCE PATTERNS

Mendelian characters are determined by chromosomal loci, either on an auto-some (human chromosomes 1 to 22) or on a sex chromosome (X or Y). Females are diploid for all loci (they have 23 pairs of homologous chromosomes). Males are different. Like females they have two copies of

each autosomal locus and of **pseudoautosomal** sequences found at the tips of the sex chromosomes (see below). However, they are hemizygous for the great majority of loci on the X and the Y (males have only one copy of the great majority of loci that are located on the X and the Y but outside the pseudoautosomal regions).

As a result of the above, there are five basic Mendelian inheritance patterns: autosomal dominant, autosomal recessive, X-linked dominant, X-linked recessive, and Y-linked (not Y-linked dominant or Y-linked recessive because males are never heterozygous for Y-linked sequences; the two Y chromosomes in rare XYY males are duplicates). In addition there is the unique pattern of inheritance of mitochondrial DNA mutations, which are substantial contributors to human genetic disease.

Autosomal dominant inheritance

A dominantly inherited disorder is one that is manifested in heterozygotes: affected persons usually carry one mutant allele and one normal allele at the disease locus. In autosomal dominant inheritance, the disease locus is present on an autosome (any chromosome other than the X or the Y), and so an affected person can be of either sex.

When an affected person has children with an unaffected person, each child would normally have a 50 % chance of developing the disease (the affected parent can transmit either the mutant allele or the normal allele). Affected persons often have an affected parent (see a typical example of autosomal dominant inheritance in [Figure 5.2](#)).

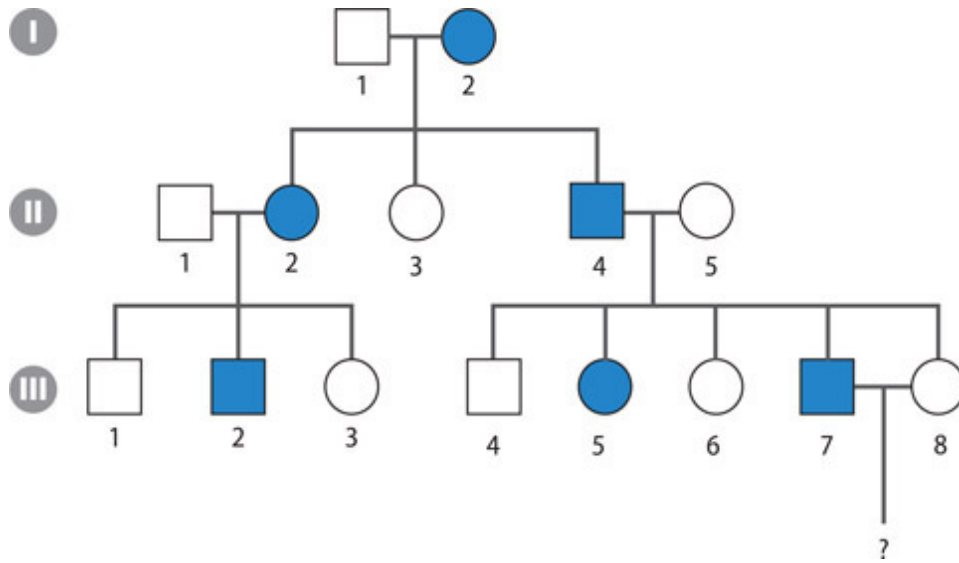


Figure 5.2 Pedigree showing autosomal dominant inheritance. Both sexes are affected and are equally likely to transmit the disorder. Affected individuals are typically heterozygotes (with one mutant allele and one normal allele) and usually have at least one affected parent. The question mark indicates the chance of having an affected child, which is 1 in 2 because one parent, III-7, is affected (there is a 50 % chance of transmitting the mutant allele and a 50 % chance of transmitting the normal allele).

Because the disorders are rare, affected individuals are almost always heterozygotes. Very occasionally, however, affected homozygotes are born to parents who are both affected heterozygotes. According to the effect of the mutation on the gene product, the affected homozygotes may show the same phenotype as the affected heterozygote. More commonly, affected homozygotes have a more severe phenotype than affected heterozygotes, as reported for conditions such as achondroplasia (PMID 20301331) and Waardenburg syndrome type I (PMID 20301703), or they have a much earlier age at onset of the disease, as in familial hypercholesterolemia (OMIM 143890).

In model organisms, a distinction is often made between different phenotypes seen in affected homozygotes and in affected heterozygotes (which are respectively called dominant and semidominant phenotypes in mice, for example). In human genetics, however, we refer to dominant phenotypes in

affected heterozygotes simply because affected homozygotes are so rarely encountered.

Autosomal recessive inheritance

A person affected by an autosomal recessive disorder can be of either sex and is usually born to unaffected parents who are heterozygotes (the parents would be described as asymptomatic **carriers** because they carry one mutant allele without being affected). Affected individuals carry two mutant alleles at the disease locus, one inherited from each parent. Assuming that both parents of an affected child are phenotypically normal carriers, the chance that each future child born to these parents is also affected is normally 25 % (the risk that one parent transmits the mutant allele is 1/2, so the risk that they both transmit the mutant allele to a child is $1/2 \times 1/2 = 1/4$).

Every one of us carries a single harmful allele at multiple loci associated with recessive phenotypes (carrying two such alleles can lead to disease, or even lethality in the prenatal period). When an autosomal recessive disorder is quite frequent, carriers will be common. In that case an affected child may often be born to two parents who carry different mutant alleles. The affected individual with two different mutant alleles would be described as a **compound heterozygote** ([Figure 5.3A](#)).

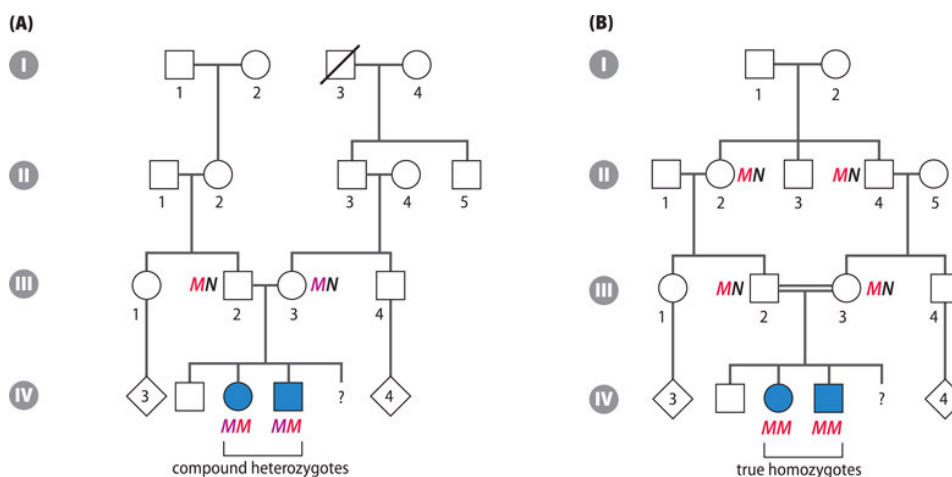


Figure 5.3 Pedigree showing autosomal recessive inheritance. (A) A pedigree for a common autosomal recessive disorder. The parents of the affected children in generation IV are carriers, with one normal allele (N) and one mutant allele (M). If they are not known to be related, they might well have different mutant alleles (shown by pink or red M) and the affected children would be compound heterozygotes. From the pedigree alone, we would not know who carried mutant alleles in generations I and II. For each subsequent child of III-2 and III-3, the risk of being affected is 1 in 4, irrespective of sex (each parent has a 50 % chance of transmitting the mutant allele, and the chance of inheriting both alleles is $1/2 \times 1/2 = 1/4$). (B) Involvement of consanguinity. Here we know that III-2 and III-3 are first cousins. They can be expected to be carriers, with one mutant allele (M) and one normal allele (N). We could infer that II-2 and II-4 inherited the same mutant allele (red M) from one parent (either I-1 or I-2). That means that III-2 and III-3 have the same mutant allele and their affected children will have inherited two identical mutant alleles and be true homozygotes. The chance that their fourth child will be affected (question mark) remains 1 in 4, irrespective of its sex.

Consanguinity

A feature of many recessive disorders, especially rare conditions, is that affected individuals often have two identical mutant alleles because the parents are close relatives; such couples are said to be *consanguineous*. In the example in [Figure 5.3B](#) the parents of the affected child in generation IV are first cousins, and they will have 1/8 of their genes in common by genetic descent ([Box 5.2](#) shows how these calculations are made). The two parents, III-2 and III-3, have each inherited the same mutant allele ultimately from the same common ancestor (in this case, a common grandparent, either I-1 or I-2).

BOX 5.2 CONSANGUINITY AND THE DEGREE TO WHICH CLOSE RELATIVES ARE GENETICALLY RELATED

Ultimately, all humans are related to one another, but we share the highest proportion of our genes with close family relatives. Mating between the most closely related family members (with 50 % of their genes in common, such as parent/child, and sibs) is very likely to result in homozygotes for recessive disease and is legally prohibited and/or socially discouraged in just about all societies. Cousin marriages can, however, be quite frequent in some communities from the Middle East, parts of the Indian subcontinent and other parts of Asia. Because cousins share a significant proportion of their genes, the offspring of cousin marriages can have a high degree of homozygosity with increased chance of being affected by a recessive disorder.

Because a child's risk of being homozygous for a rare recessive allele is proportional to how related the parents are, it is important to measure consanguinity. When one person is a direct descendant of another, the proportion of genes they have in common is $(1/2)^n$, where n is the number of generational steps separating the two. This gives: parent-child, $1/2$ of genes in common; grandparent-grandchild, $(1/2)^2 = 1/4$ of genes in common; greatgrandparent-greatgrandchild, $(1/2)^3 = 1/8$ of genes in common.

CALCULATING THE COEFFICIENT OF RELATIONSHIP

The **coefficient of relationship** is the proportion of alleles shared by two persons as a result of common genetic descent from one or more recent (definable) common ancestors (or, more loosely, the proportion of genes in common as a result of common genetic descent). To calculate this, one considers paths of genetic descent linking the two individuals through *each* common ancestor in a family. A single generational step in such a path reduces the shared genetic component from the common ancestor by $1/2$.

Consider the example in [Figure 1](#). I-2 has had three children, a brother and sister who are sibs because they also have a common father, I-1, and their half brother, II-5. Half-sibs, such as II-3 and II-5, have a single

ancestor in common and so there is a single path connecting them to their common parent. So, the orange path connecting II-3 to II-5 via their common mother has two steps, making a contribution of $1/2 \times 1/2 = 1/4$ of genes in common.

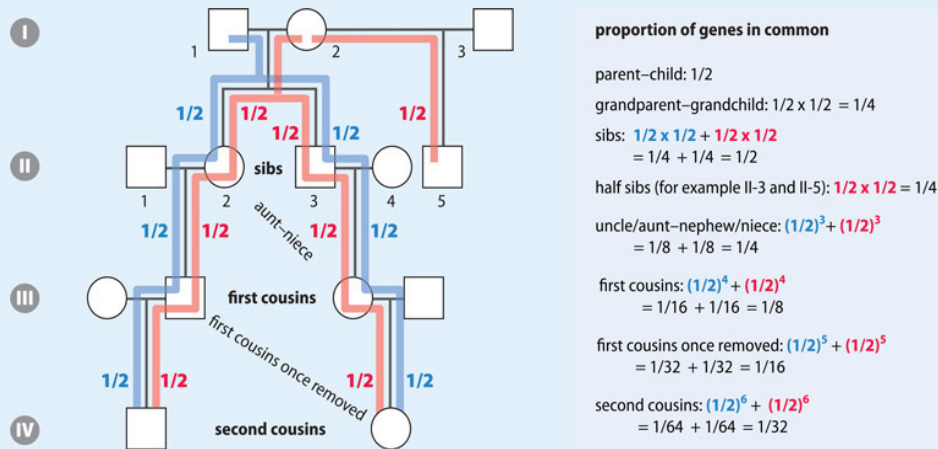


Figure 1 The proportion of genes in common between family members.

I-1 and I-2 are common ancestors for the sibs in generation II, for the first cousins in generation III and for the second cousins in generation IV. To calculate the coefficient of relationship for relatives linked by two or more common ancestors, we need to calculate the contributions made by each path and then sum them. Thus, for the first cousins in generation III, the green path that links them through their common grandfather, I-1, has four steps, making a contribution of $(1/2)^4 = 1/16$, and the orange path that links them through the common grandmother, I-2, also has four steps, making a contribution of $(1/2)^4 = 1/16$. Adding the two paths gives $2/16$ or $1/8$ of genes in common. More complicated inbreeding may mean that individuals have four or more recent common ancestors, but the principle is always the same: work out paths for each common ancestor and sum the contributions.

The **coefficient of inbreeding** is the probability that a homozygote has identical alleles at a locus as a result of common genetic descent from a recent ancestor. It is also the proportion of loci at which a person is expected to be homozygous because of parental consanguinity and is one-

half of the coefficient of relationship of the parents. So, if the parents are first cousins, the coefficient of inbreeding is $1/16$. Note that even quite highly inbred pedigrees result in relatively moderate coefficients of inbreeding.

For rare disorders when there is doubt concerning the mode of inheritance, known parental consanguinity will strongly indicate autosomal recessive inheritance in a pedigree in which affected individuals have unaffected parents. But if consanguinity is not apparent, as in the pedigree in [Figure 5.3A](#), alternative explanations are possible, as described below.

Disease-related phenotypes in carriers

Although carriers of an autosomal recessive disorder are considered asymptomatic, they may nevertheless express some disease-related trait that can distinguish them from the normal population. Take sickle-cell disease (OMIM 603903), for example. Affected individuals are homozygous for a b-globin mutation and produce an abnormal hemoglobin, HbS, that causes red blood cells to adopt a rigid, crescent (or sickle) shape. The sickle cells have a shorter life span that leads to anemia, and they can block small blood vessels, causing hypoxic tissue damage.

Carriers of the sickle-cell mutation are not quite asymptomatic. The sickle-cell allele produces a mutant b-globin that is co-dominantly expressed with the normal b-globin, and heterozygotes can have mild anemia (sickle-cell trait). However, under intense, stressful conditions such as exhaustion, hypoxia (at high altitudes), and/or severe infection, sickling may occur in heterozygotes and result in some of the complications associated with sickle-cell disease. Note that whereas sickle-cell disease is recessively inherited, the sickle-cell trait is expressed in the heterozygote and is therefore a dominant trait.

Sex-linked inheritance

In sex-linked inheritance, the inheritance patterns are controlled by genes that reside on the X and/or Y chromosomes. Before we go on to consider sex-linked inheritance, we need to take account of mechanisms that compensate for the variable number of sex chromosomes in humans (and other mammals): females have two X chromosomes but males have one X and one Y.

Having different numbers of chromosomes usually has severe, often lethal consequences—the loss of just one of our 46 chromosomes is lethal except for 45,X (Turner syndrome), and having an extra chromosome is usually lethal or results in a developmental syndrome such as trisomy 21 (Down syndrome). This happens because of problems with *gene dosage*: for some of our genes, the amount of gene product made must be tightly controlled (having one or three copies of these genes can be harmful because too little or too much product is made).

The sex difference regarding the Y chromosome is minimized by the conspicuous lack of genes on the Y. Most of the very few genes on the Y chromosome have male-specific functions, or they have an equivalent gene copy on the X (these X–Y gene pairs are mostly concentrated at the tips of the sex chromosomes in the pseudoautosomal regions).

X-chromosome inactivation

Unlike the Y chromosome, the human X chromosome has many hundreds of important genes. To compensate for having different numbers of X chromosomes in males and females, a special mechanism is needed: genes on one of the two X chromosomes in each female cell are silenced so that they do not produce any gene products (**X-inactivation**). Whereas males are *constitutionally* hemizygous for most genes on the X chromosome, X-inactivation means that at the *functional* level, females behave as if they were hemizygous for most genes on the X.

The X-inactivation mechanism is initiated after a cellular mechanism counts the number of X chromosomes in each cell of the early embryo. If

the number of X chromosomes is two (or more), all except one of the multiple X chromosomes is inactivated. Each such X chromosome is induced to form a highly condensed chromosome that is mostly transcriptionally inactive, known as a Barr body ([Figure 5.4](#)). Note that some genes, including genes in the pseudoautosomal regions, escape inactivation (we consider the mechanism of X-inactivation in [Chapter 6](#)).

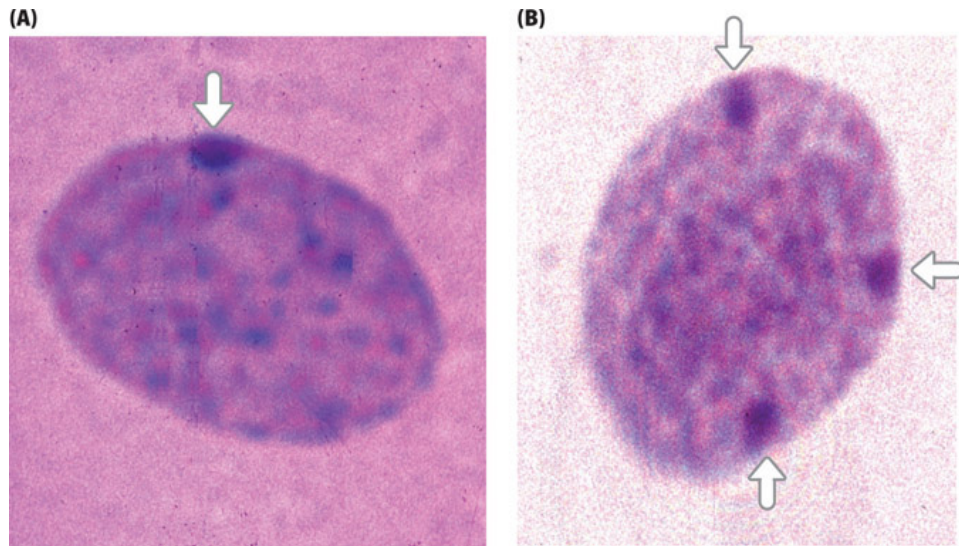


Figure 5.4 Barr bodies. (A) A cell from an XX female has a single inactivated X chromosome that forms a Barr body (arrow). (B) A cell from a 49,XXXXY male has one active X chromosome plus three inactivated X chromosomes that form Barr bodies (arrows). (Images courtesy of Malcolm Ferguson-Smith.)

In humans the initial decision to inactivate one of the two X chromosomes is randomly made in the preimplantation embryo, beginning at around the eight-cell stage; some cells inactivate the paternal X and others inactivate the maternal X. Once a cell has chosen which X to inactivate in the early embryo, however, that pattern of X-inactivation is continued in all descendant cells. Thus, a female who is heterozygous at a disease locus will be a genetic **mosaic**, containing cell clones in which the normal allele is expressed and clones in which the mutant allele is expressed. As described below, this has implications for the female phenotype in X-linked disorders.

X-linked recessive inheritance

In X-linked recessive disorders, affected individuals are mostly male, and affected males are usually born to unaffected parents. The mother of an affected male is quite often a carrier (and clearly so if she has affected male relatives). A distinguishing feature is that there is no male-to-male transmission because males pass a Y chromosome to sons ([Figure 5.5A](#)). However, a pedigree may *appear* to show male-to-male transmission when an affected man (with a condition such as hemophilia, for example) and a carrier woman produce an affected son ([Figure 5.5B](#)). The same parents could each potentially transmit a mutant X to produce an affected daughter.

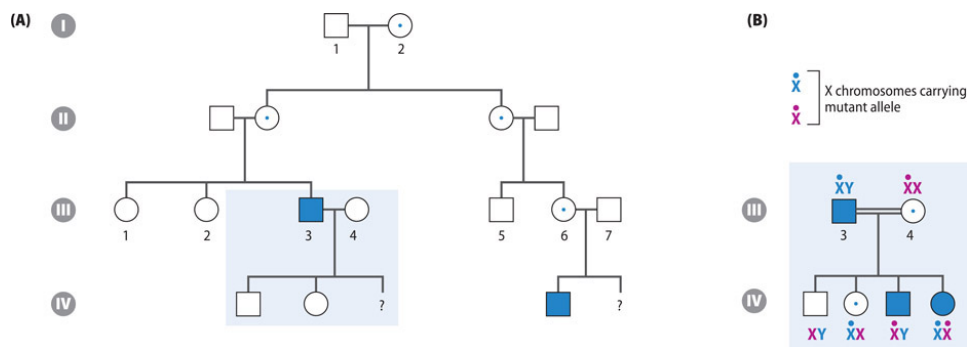


Figure 5.5 Pedigree showing X-linked recessive inheritance. (A) Affected males in generations III and IV have inherited (via female carriers) a common mutant allele from I-2. For each child of a carrier mother (such as III-6), the overall chance of being affected is 1 in 4, but this is sex-dependent: a son will have a 1 in 2 risk but a daughter will not be at risk (though she has a 1 in 2 risk of being a carrier). In the highlighted box, III-3 and III-4 have had two normal children, and the risk of having affected children would normally be very low (the father cannot transmit the mutant X allele to sons—he must transmit a Y—and any daughter will inherit a normal X from her mother). (B) The complication of inbreeding. Imagine that III-3 and III-4 were consanguineous and had the same mutant allele. We now have mating between an affected individual and a carrier, and there is a 1 in 2 chance that a child would be affected, irrespective of whether it was a boy or girl. The apparent male-to-male transmission is an illusion (the affected son has inherited a mutant allele from his mother, not from his father). The affected daughter is homozygous and although one

mutant allele will be silenced in each of her cells by X-inactivation, she does not have a normal allele.

In X-linked recessive disorders, female carriers with a single mutant allele can occasionally be quite severely affected and are known as *manifesting heterozygotes*. Because of X-inactivation, female carriers of an X-linked mutation are mosaics: some of their cells have the normal X chromosome inactivated and other cells have the mutant X inactivated, as seen most readily in skin disorders. Manifesting heterozygotes can occur by chance when most cells of a tissue critically important in disease development happen to have an inactivated X carrying the normal allele.

Manifesting heterozygotes can occasionally occur because of nonrandom X-inactivation. That can happen when there is some advantage in inactivating the normal X chromosome instead of the mutant X chromosome. For example, an X-linked disorder may manifest in a woman who has an X-autosome translocation in which the breakpoint on the X is the cause of the disorder. If the X-autosome translocation chromosome were to be inactivated, neighboring autosomal genes would also be silenced, causing gene dosage problems, and so the normal X is preferentially inactivated. Skewing of X-inactivation can often work in the other direction: some female carriers are asymptomatic because of nonrandom inactivation of the mutant X chromosome. We consider the mechanisms in [Chapter 6](#).

X-linked dominant inheritance

As in autosomal dominant disorders, affected individuals with an X-linked dominant disorder can be of either sex and usually at least one parent is affected. However, there are significantly more affected females than affected males, and affected females typically have milder (but more variable) expression than affected males.

The excess of affected females arises because there is no male-to-male transmission of the disorder. All children born to an affected mother (and an unaffected father) have a 50 % chance of being affected, but an affected father with a single X chromosome will consistently have unaffected sons (they do not inherit his X chromosome), but his daughters will always be at risk because they will always inherit his affected X ([Figure 5.6A](#) gives an example pedigree).

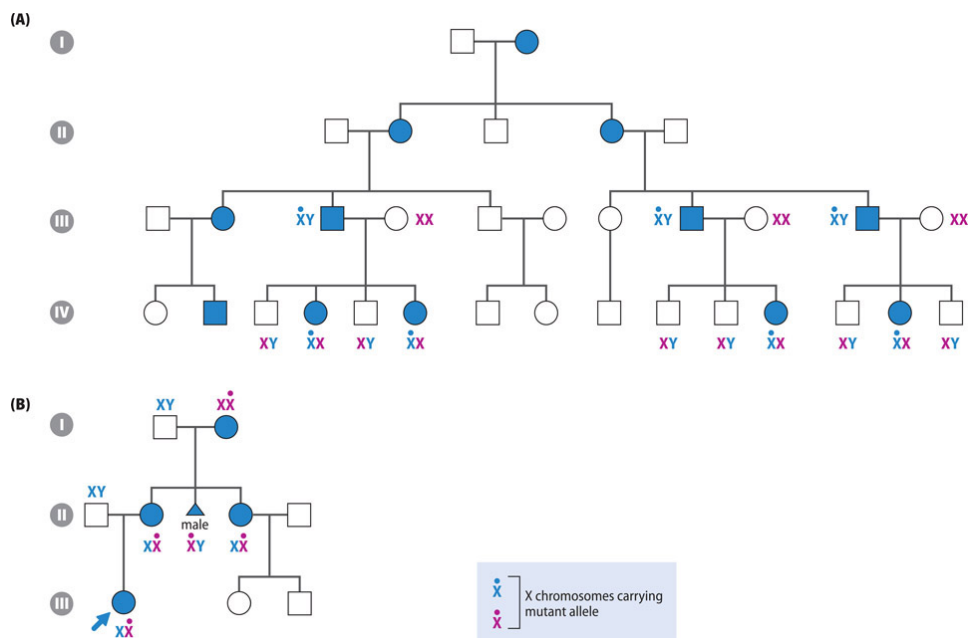


Figure 5.6 Pedigrees showing X-linked dominant inheritance. (A) Each child of an affected parent has a 1 in 2 chance of being affected. There is a 50 % chance that an affected female can transmit a mutant X allele to sons and to daughters; the risk is the same, irrespective of the sex of the child. However, the risk to the children of an affected father depends crucially on the sex of the child: every son would be expected to be unaffected; instead, the risk of being affected is focused on daughters (as shown for the three affected males in generation III who each have had children—the father must pass on a Y chromosome to each son and so does not transmit the mutant X, but must transmit the mutant X chromosome to each daughter). (B) X-linked dominant inheritance with early male lethality. This example shows four affected females in a three-generation family with incontinentia pigmenti that was followed up after the birth of the affected granddaughter (arrowed). An affected male in generation II had

spontaneously aborted. (Adapted from Minić S et al. [2010] *J Clin Pathol* 63:657–659; PMID 20591917. With permission from BMJ Publishing Group Ltd.)

The milder phenotype seen in affected females is a result of X-inactivation—the mutant allele is located on an inactivated X in a proportion of their cells. For certain X-linked dominant disorders, virtually all affected individuals are female: the phenotype is so severe in males that they die in the prenatal period, but the milder phenotype of affected females allows them to survive and reproduce. We illustrate this with the example of incontinentia pigmenti in [Figure 5.6B](#); another disorder like this, Rett syndrome, is profiled in [Section 6.3](#).

X-Y recombination and X-Y homology

In female meiosis, the two X chromosomes recombine like any pair of homologous chromosomes; in male meiosis, however, recombination between the X and Y chromosomes is very limited. The X and Y are very different in size, and pairing between the X and Y at meiosis is very limited.

Despite their considerable differences in size and gene content, the X and Y nevertheless have some short gene-containing regions in common, notably the pseudoautosomal regions located just before the telomere-associated repeats at the ends of both short and long chromosome arms ([Figure 5.7](#)). The pseudoautosomal regions are distinctive: they are the only regions of the X and Y that can pair up during male meiosis and undergo recombination like paired sequences do on homologous autosomal chromosomes (at each meiosis, there is an obligate X–Y crossover in the major pseudoautosomal region; recombination is less frequent in the minor pseudoautosomal region).

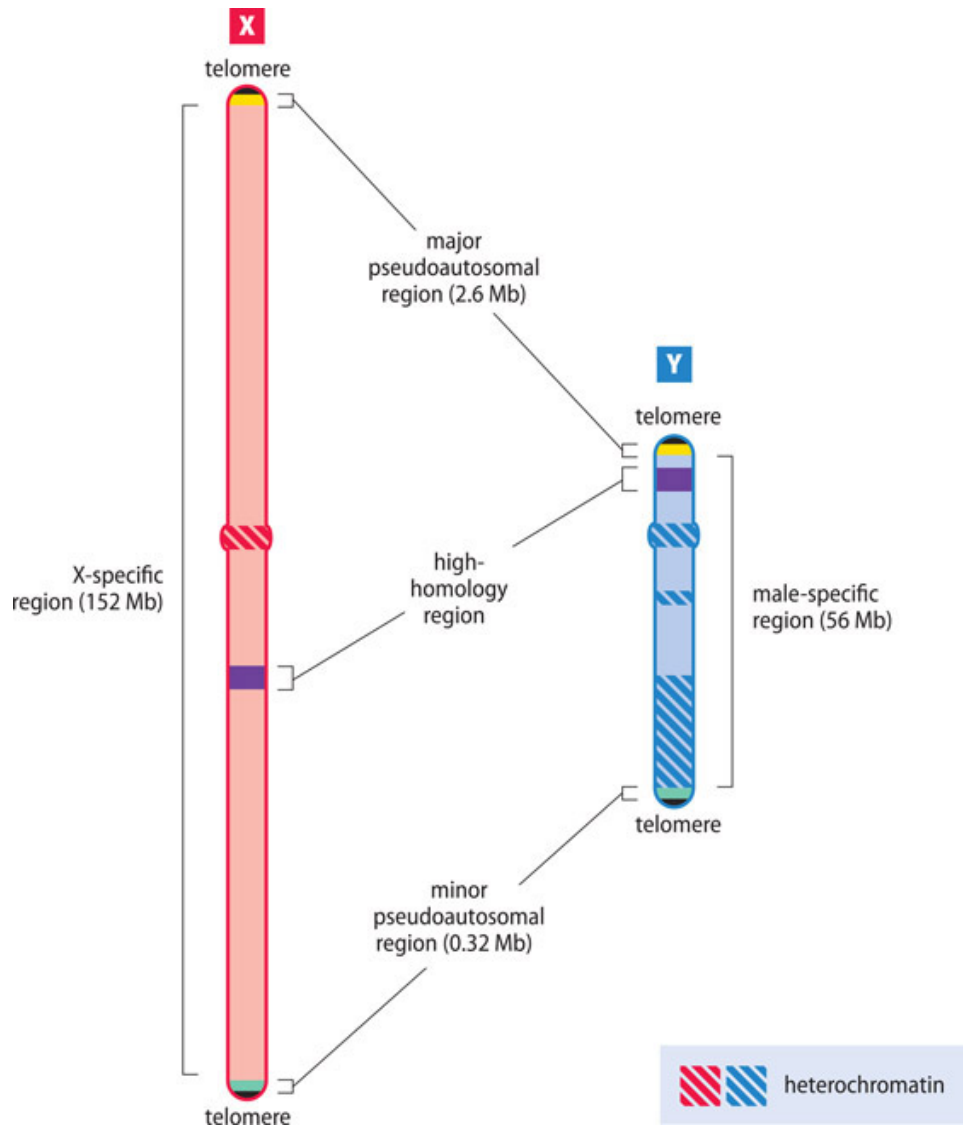


Figure 5.7 The human X and Y chromosomes: differences and major homology regions. The X and Y chromosomes differ greatly in size, heterochromatin content (much of the long arm of the Y is composed of heterochromatin), and DNA sequence. Colors indicate sequences that are X-specific (pink), Y-specific (blue), heterochromatin (hatched), or the major sequences shared by the X and Y chromosomes only (other colors). The major pseudoautosomal regions on the X and Y (yellow) are essentially identical, as are the minor pseudoautosomal regions on the long arms (green). The pseudoautosomal regions are involved in X–Y pairing and recombination in male meiosis. Note that the large central regions of the X and Y do not engage in recombination and are X-specific or Y-specific; the Y-specific region is also a *male-specific region* because it is not normally transmitted to females. As a result of an

evolutionarily recent X–Y transposition event there is also roughly 99 % sequence homology between certain sequences on Yp, lying close to the major pseudoautosomal region, and sequences at Xq21 (shown by purple boxes).

Outside the pseudoautosomal regions there is no recombination between the X and Y, and the remaining large central regions are X-specific and Y-specific regions. The X-specific region can engage in recombination in female meiosis, and sequences in this region can be transmitted to males or females; the Y-specific region is never involved in recombination and so is also called the *male-specific region*. The sequences in the X-specific and male-specific regions are very different, with just a few exceptions (see [Figure 5.7](#) for an example).

Pseudoautosomal inheritance

As a result of recombination in male meiosis, the individual X–Y gene pairs in the pseudoautosomal regions are effectively alleles. An individual allele in these regions can move locations between the X and Y chromosomes and so is neither X-linked nor Y-linked; instead, the pattern of inheritance resembles autosomal inheritance ([Figure 5.8](#)).

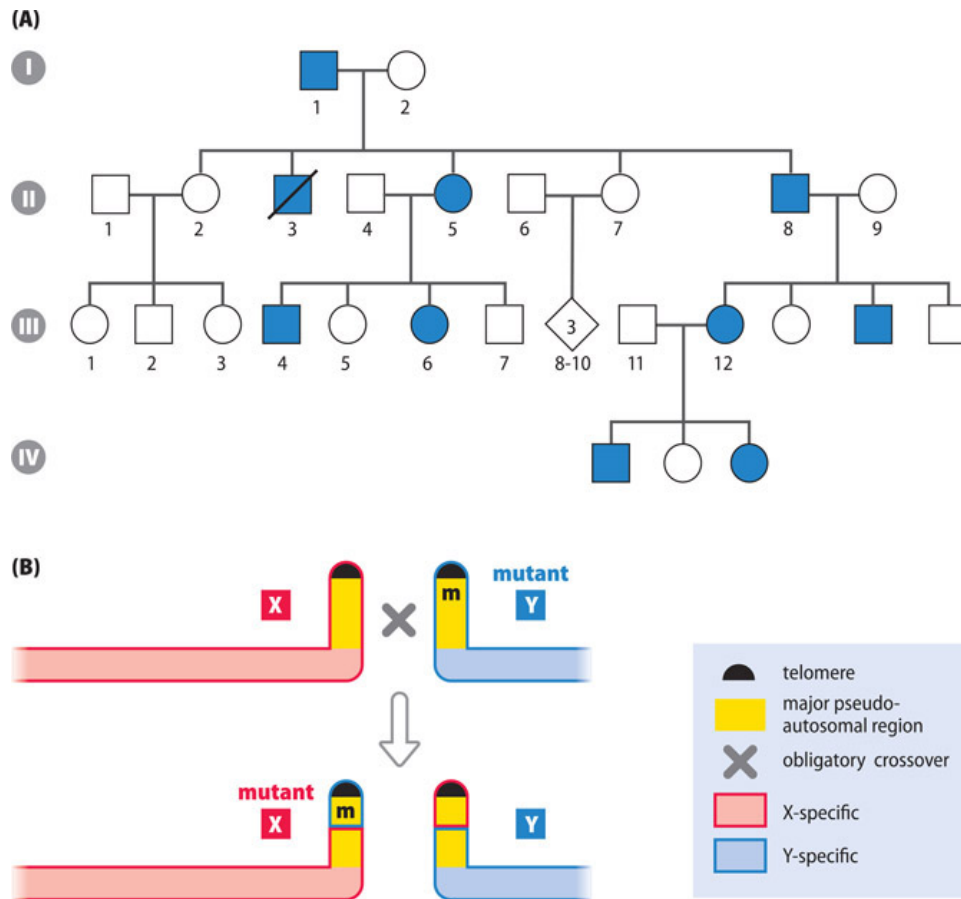


Figure 5.8 Pseudoautosomal inheritance and X–Y recombination. (A) In this pedigree, affected individuals are heterozygous for a mutation in the major pseudoautosomal region, and the disorder shows dominant inheritance. Affected females carry the mutation on an X chromosome that can be passed to both sons and daughters. Affected males pass a Y chromosome containing the mutant allele to affected sons but can also pass an X chromosome containing the mutant allele to affected daughters. This happens as a result of the obligatory X–Y recombination, which occurs within the major pseudoautosomal region. (B) When the crossover point occurs proximal to the mutant allele, the allele is transposed between the X and Y chromosomes.

There are few genes in the pseudoautosomal regions, so that few pseudoautosomal conditions have been described. However, mutations in *KAL* can cause Kallmann syndrome OMIM 308700, and the *SHOX* homeobox gene is a locus for two disorders. If one *SHOX* gene copy is damaged by mutation, the resulting heterozygotes have Leri-Weill

dyschondreostosis (OMIM 127300). Homozygotes with mutations in both *SHOX* genes have a more severe condition, Langer mesomelic dysplasia (OMIM 249700).

Y-linked inheritance

The Y-specific region of the Y chromosome is a nonrecombining male-specific region. Population genetics dictates that nonrecombining regions must gradually lose DNA sequences (as a way of deleting acquired harmful mutations because they cannot be removed by recombination). Over many millions of years of evolution, the Y chromosome has undergone a series of contractions and now has only 38 % of the DNA present in the X (the X and Y are thought to have originated as a homologous pair of autosomes that began to diverge in sequence after one of them acquired a sex-determining region). As a result of DNA losses, the male-specific region of the Y chromosome has few genes and makes a total of only 31 different proteins, most of which are involved in male-specific functions.

In Y-linked inheritance, males only should be affected and there should be exclusive male-to-male transmission. However, because of the lack of genes, Y-linked disorders are rare. Claims for some Y-linked traits, such as hairy ears (OMIM 425500), are now known to be dubious, but maleness is indisputably Y-linked. Interstitial deletions on the long arm of the Y chromosome are an important cause of male infertility (but infertile males are not normally able to transmit chromosomes unless conception is assisted by procedures such as intracytoplasmic sperm injection).

Matrilineal inheritance for mitochondrial DNA disorders

The mitochondrial genome is a small (16.5 kb) circular genome that has 37 genes (see [Figure 2.11](#)). It is much more prone to mutation than nuclear DNA, partly because of its proximity to reactive oxygen species (the mitochondrion is a major source of reactive oxygen species in the cell). As a result, mutations in mitochondrial DNA (mtDNA) are a significant cause

of human genetic disease. Tissues that have a high energy requirement—such as muscle and brain—are primarily affected in mtDNA disorders.

Individuals with a mitochondrial DNA disorder can be of either sex, but affected males do not transmit the condition to any of their children. The sperm does contribute mtDNA to the zygote, but the paternal mtDNA is destroyed in the very early embryo (after being tagged by ubiquitin), and a father's mtDNA sequence variants are not observed in his children. That is, inheritance occurs exclusively through the mother (*matrilineal* inheritance). An additional, common feature of mitochondrial DNA disorders is that the phenotype is highly variable within families ([Figure 5.9](#)).

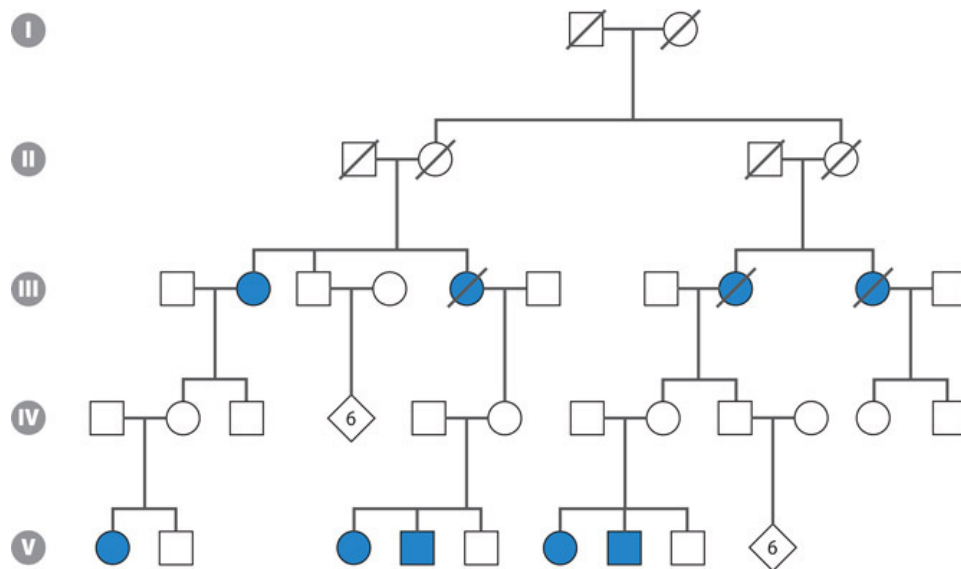


Figure 5.9 A pedigree illustrating matrilineal inheritance for a mitochondrial DNA disorder. Mitochondrial DNA disorders are transmitted by females only (because any mtDNA originating from the sperm is quickly degraded in the early embryo). However, an affected female can pass on the condition to both sons and daughters. A common feature of mtDNA disorders is incomplete penetrance, as shown here by the absence of clinical phenotypes in several individuals who must be gene carriers, including three clear carrier females in generation IV, each of whom were born to an affected mother and went on to produce affected children of their own (the females in generations I and II might also have been expected to be carriers of the mutant mtDNA). *One* cause of this intrafamilial variability is variable heteroplasmy. The mutation here was shown to be a nucleotide substitution in the mitochondrial 12S rRNA gene that was associated

with variable hearing loss. (From Prezant TR et al. [1993] *Nat Genet* 4:289–294; PMID 7689389. With permission from Macmillan Publishers Ltd.)

Variable heteroplasmy and clinical variability

Each cell contains multiple mitochondria, and there are often several hundred to thousands of mtDNA copies per cell. In some affected persons, every mtDNA molecule carries the causative mutation (homoplasmy), but affected individuals frequently have cells with a mixed population of normal and mutant mtDNAs (**heteroplasmy**). The clinical features depend mostly on the proportion of mutant to normal mtDNA molecules in the cells of tissues with high energy requirements.

Although a human egg cell is haploid for nuclear DNA, it contains more than 100 000 mtDNA molecules. A heteroplasmic mother can give rise to children who differ widely from her and from each other in the ratio of mutant to normal mtDNA molecules in their tissues (variable heteroplasmy). As a result, there can be very significant clinical variability between affected members of the same family.

To explain rapid shifts in heteroplasmy that occur over only one generation, the mitochondrial genetic bottleneck hypothesis envisages that, during early development, germline cells pass through a bottleneck stage in which they contain very few mtDNA molecules. By chance, germline cells at this stage may have a much higher or much lower proportion of mutant mtDNA molecules than the somatic cells. As a result, a heteroplasmic mother could give rise ultimately to eggs with a much higher or much lower proportion of mutant mtDNA molecules than are present in her affected tissues. We consider this in more detail in [Section 7.6](#).

Another contributor to variable heteroplasmy is the rapid evolution of a mtDNA variant within an individual. Mutant mtDNAs that have a large deletion or a large duplication can evolve rapidly, so that different tissues or even the same tissue at different times may show different distributions of the mtDNA variant.

5.3 UNCERTAINTY, HETEROGENEITY, AND VARIABLE EXPRESSION OF MENDELIAN PHENOTYPES

[Section 5.2](#) dealt with the different modes of inheritance for phenotypes that are determined principally by single genes. Some complications were covered, including the effects of X-inactivation in females and hemizygoty in males, occasional differences between homozygotes and heterozygotes for autosomal dominant disorders, the occasional expression of disease symptoms in carriers of an autosomal recessive disorder, the mimicking of autosomal inheritance by genes in the pseudoautosomal regions of the X and Y chromosomes, and the unique features of mitochondrial DNA inheritance.

In this section we discuss broader complications that relate to uncertainty of mode of inheritance, and difficulties posed by heterogeneity in the links between DNA variation and phenotypes. In addition, we consider how affected individuals within a single family can show variable phenotypes for Mendelian disorders.

Difficulties in defining the mode of inheritance in small pedigrees

Many families are small and may have only a single affected person. If the disorder is rare and we do not know the underlying disease gene, how can we work out the mode of inheritance? Knowing the mode of inheritance is important in genetic counseling (calculating the risk of having a subsequent affected child is made on that basis). Unless a disease gene has been identified and screened for mutations, however, the mode of inheritance inferred from examining the pedigree should be regarded simply as a working hypothesis.

Having a single affected child in a family with no previous history of a presumed genetic disorder might suggest the possibility of a recessive disorder, with a 1 in 4 risk that each subsequent child would be affected.

Alternatively, it could be a dominant disorder and the affected individual could be a heterozygote. In that case, one parent carries the disease gene but does not display the phenotype, or the disorder is due to a *de novo* mutation (see below).

One possible way to work out the mode of inheritance is to study multiple families with the same disorder and calculate the overall proportion of affected children (called the *segregation ratio*). But there are many difficulties with this approach. First, the disorder may be heterogeneous and be due to different genes in different families. Secondly, the total numbers of children who can be studied are often too small to get reliable estimates.

There are also problems in how the families are ascertained (that is, in finding the people and families who will be studied). In the pre-genomics era trying to establish that a disorder was autosomal recessive was difficult. Then the priority would have been to collect a set of families and try to show a segregation ratio of 1 in 4. However, there was the complication of *ascertainment bias*: if there is no independent way of recognizing carriers, the families will be identified only through an affected child (families with two carrier parents and only unaffected children would seem perfectly normal and not be included).

Happily, in the genomics era underlying disease genes can quickly be found even for rare single-gene disorders. Rapid next-generation DNA sequencing is now being widely used to screen *exomes* (in practice, all exons of protein-coding genes) of affected individuals with the same condition. As is described in later chapters, genes underlying some rare single-gene disorders have been successfully identified after sequencing exomes from only a very few unrelated individuals with the disorder.

For a single-gene disorder, the observed incidence of mutant alleles in a defined population can be quite stable over time. A proportion of mutant alleles are transmitted from one generation to the next, and a proportion are lost because some individuals possessing mutant alleles do not transmit them. To keep the frequency of mutant alleles constant, new mutations

make up for the loss of mutant alleles that are not transmitted to the next generation.

Persons who have a severe disorder usually do not reproduce or have a much-reduced reproductive capacity (unless the disorder is not manifested until later in life). In severe autosomal recessive conditions, however, for each affected individual there are very many asymptomatic carriers who can transmit mutant alleles to the next generation. Because only a very small proportion of mutant alleles go untransmitted, the incidence of new mutation is low.

For severe dominant disorders, the mutant alleles are concentrated in affected individuals. If most individuals who carry the disease allele do not reproduce (because the disorder is congenital, say), the incidence of new mutation will be very high. If, however, there is a relatively late age at onset of symptoms, as with Huntington disease, individuals with the mutant allele may reproduce effectively, and the rate of new mutation may be very low.

For severe X-linked recessive disorders, the incidence of new mutation will also be quite high to balance the loss of mutant alleles when affected males do not reproduce. However, female carriers will usually be able to transmit mutant alleles to the next generation.

As a result of a new mutation, an affected person may be born in a family with no previous history of the disorder and would present as an isolated (*sporadic*) case. In rare disorders that have not been well studied, a sporadic case poses difficulty for calculating the risk that subsequent children could also be affected. The affected individual could be a heterozygote (as a result of *de novo* mutation, or the failure of the disorder to be expressed in one parent), but alternatively could be a homozygote born to carrier parents, or a hemizygous boy whose mother is a carrier of an X-linked recessive condition.

Post-zygotic mutations and mosaicism

Most mutations arise as a result of endogenous errors in DNA replication and repair. Mutations can occur during gametogenesis and produce sperm and eggs with a new mutant allele. In addition, *de novo* pathogenic mutations can also occur at any time in post-zygotic life. As a result of post-zygotic mutations, each individual person is a genetic **mosaic** with genetically distinct populations of cells that have different mutational spectra.

Post-zygotic mutations may result in somatic mosaicism that will have consequences only for that individual ([Box 5.3](#)). But certain post-zygotic mutations, often occurring comparatively early in development, may also result in *germline mosaicism*. A person who has a substantial proportion of mutant germline cells (a germline mosaic or gonadal mosaic) may not show any symptoms but will produce some normal gametes and some mutant gametes. The risks of having a subsequently affected child are much higher than if an affected child carries a mutation that originated in a meiotic division.

BOX 5.3 POST-ZYGOTIC MUTATIONS AND WHY WE ARE ALL GENETIC MOSAICS

A pathogenic new mutation can be imagined to occur during gamete formation in an entirely normal person. Most mutations arise as a result of endogenous errors in DNA replication and repair, and although mutations do occur during gametogenesis and produce sperm and eggs with a new heritable mutant allele, they can also occur at any time in post-zygotic life. As a result of post-zygotic mutations, each individual person is a genetic *mosaic* with genetically distinct populations of cells that have different mutational spectra.

Human mutation rates are around 10^{-6} per gene per generation, and so a person with a wild-type allele at conception has a roughly one in a million chance of transmitting it to a child as an altered (mutant) allele. In this case we are considering the chance of a mutation occurring in a lineage of germline cells from zygote to gamete, involving a series of about 30 cell

divisions in females and several hundred divisions in males (about 400 by age 30 and increasing by about 23 per year because spermatogenesis continues through adult life—see [Figure 7.5](#) on page 190).

Now consider post-zygotic mutations in somatic cell lineages. The journey from single-celled zygote to an adult human being involves a total of about 10^{14} mitotic cell divisions. With so many cell divisions, post-zygotic mutation is unavoidable—we must all be mosaics for many, many mutations. Having so many potentially harmful somatic mutations is usually not a concern because the number of cells that will fail to function correctly is normally very small. A cell will usually function normally after sustaining a harmful mutation in a gene that is not normally expressed in that cell type, and even if the cell does function abnormally as a result of mutation it might not give rise to many mutant descendants.

A person may be at risk of disease, however, if a mutated cell is able to give rise to substantial numbers of descendant cells that act abnormally ([Figure 1](#)). The biggest disease risk posed by post-zygotic mutations is that they set off or accelerate a process that leads to cancer. As we describe in [Chapter 10](#), cancers are unusual in that although they can be inherited, the biggest contribution to disease comes from somatic mutations.

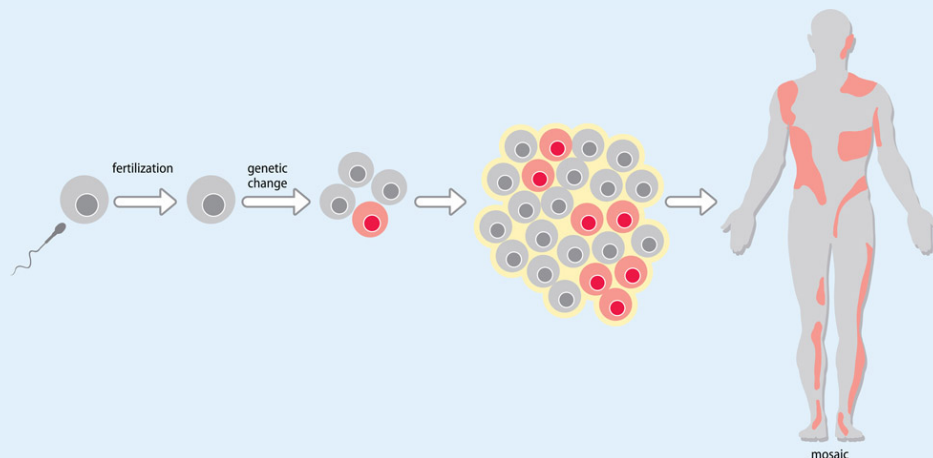


Figure 1 Genetic mosaicism. As illustrated here, post-zygotic mosaicism may often have consequences just for the individual who possesses the mutant cells; that is, the mutation affects somatic cells only. Sometimes, however, post-zygotic mutations can

occur in germ-cell precursors (*germline mosaicism*), and that has important implications concerning the possibility of transmitting a disorder.

Heterogeneity in the correspondence between phenotypes and the underlying genes and mutations

There is no one-to-one correspondence between phenotypes and genes. Three levels of heterogeneity are listed below. As we will see below and in later chapters, both nongenetic factors (environmental and epigenetic) and additional genetic factors can also influence the phenotypes of single-gene disorders.

Locus heterogeneity

The same clinical phenotype can often be produced by mutations in genes at two or more loci. The different genes often make related products that work together as a complex or in a common pathway; sometimes one gene is the primary regulator of another gene.

Locus heterogeneity explains how parents who are both affected with the same common recessive disorder produce multiple unaffected children. Recessively inherited deafness is the classic example (sensorineural hearing impairment mostly shows autosomal recessive inheritance, and deaf people often choose to have children with another deaf person). If two deaf parents are homozygous for mutations at the same gene locus, one would expect that all their children would also have impaired hearing. If, instead, the parents are homozygous for mutations at two different recessive deafness loci, all their children would be expected to be double heterozygotes and have normal hearing ([Figure 5.10](#)).

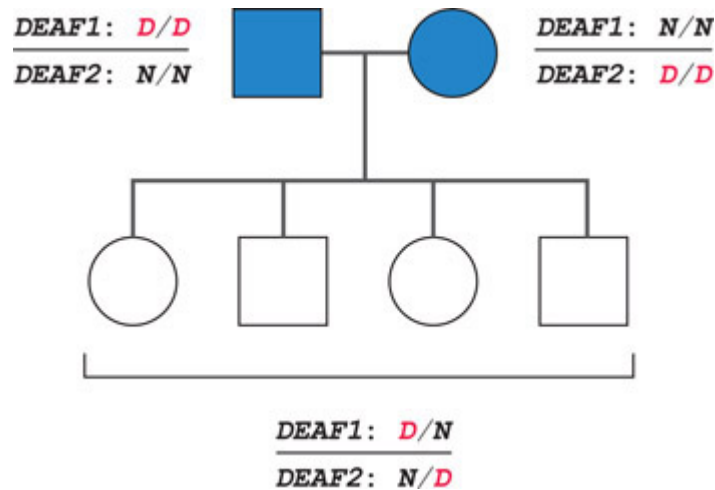


Figure 5.10 Locus heterogeneity explains why two parents with autosomal recessive deafness can consistently produce unaffected children. Imagine that the two parents are deaf because they have two mutant alleles at different autosomal recessive deafness loci, which we represent here as *DEAF1* and *DEAF2*. We represent normal alleles as *N* and deafness-associated alleles as *D*. In this case, sperm produced by the father would carry the *DEAF1***D* allele and the *DEAF2***N* allele, and eggs produced by the mother would carry the *DEAF1***N* allele and the *DEAF2***D* allele. All children would therefore be unaffected because they would be heterozygous at both loci. The normal phenotypes of each child result from complementation between normal alleles at the two loci. If, instead, both parents had autosomal recessive deafness caused by different mutations in the *same* gene, all their children would be expected to be deaf as a result of inheriting two mutant alleles at that locus.

As the underlying genes for single-gene disorders become known, it has become clear that very many conditions show locus heterogeneity. One might anticipate that many different genes contribute at different steps to broad general pathways (responsible for hearing or vision, for example). It is therefore unsurprising that autosomal recessive deafness or retinitis pigmentosa (hereditary retinal diseases with degeneration of rod and cone photoreceptors) can result from mutations in different genes.

More specific phenotypes can also be caused by mutations at any one of many different gene loci. Usher syndrome, for example, involves profound sensorineural hearing loss, vestibular dysfunction, and retinitis pigmentosa;

autosomal recessive forms can be caused by mutations at any one of at least 11 different gene loci.

Bardet-Biedl syndrome (PMID 20301537) provides another illustrative example. It is a *pleiotropic* disorder (many different body systems and functions are impaired) and the primary features are: degeneration of light-sensitive cells in the outer regions of the retina (causing night blindness, tunnel vision, reduced visual acuity), learning disabilities, kidney disease, extra toes and/or fingers, obesity, and abnormalities of the gonads. Autosomal recessive inheritance is the typical inheritance pattern, and the disorder is caused by mutations in any of at least 21 genes, all involved in regulating how cilia function ([Figure 5.11](#)).

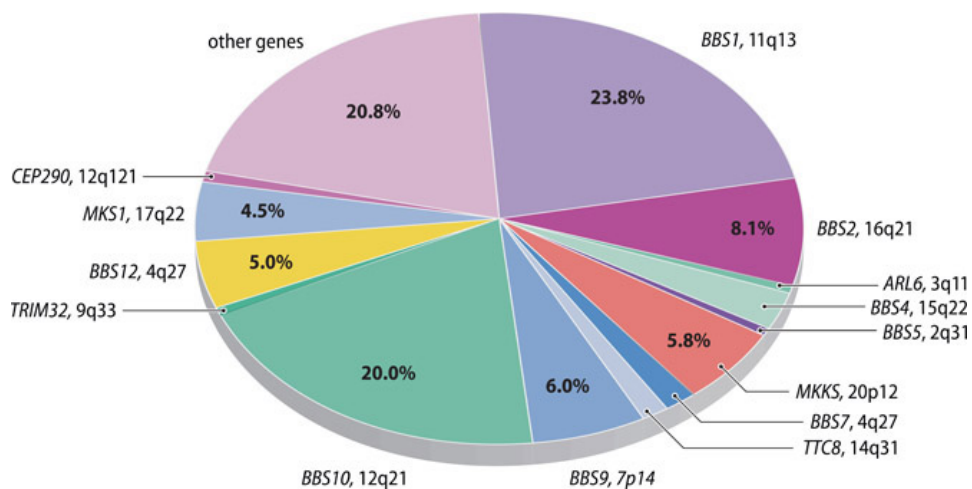


Figure 5.11 Extraordinary locus heterogeneity for Bardet-Biedl syndrome (BBS).

Segments represent the proportion of total mutant alleles attributable to the first 14 genes known to be mutated in BBS, with 20.8 % initially unidentified genes represented here by the segment labeled “Other genes”. Seven of the genes – *BBS1*, *BBS2*, *MKKS*, *BBS9*, *BBS10*, *BBS12*, and *MKS1* – account for ~70 % of the identified pathogenic mutations in BBS. Note: some of the genes are also mutated in other disorders, such as *MKS1* (in Meckel syndrome and Joubert syndrome) and *MKKS* (in McKusick-Kaufman syndrome). Since this figure was first published, seven more identified BBS genes have been described (see PMID 29487844). (Adapted from Zaghoul NA & Katsanis N [2009] *J Clin Invest* 119:428–437; PMID 19252258. With permission from the American Society for Clinical Investigation.)

Allelic and phenotypic heterogeneity

Many different mutations in one gene can have the same effect and produce similar phenotypes. For example, b-thalassemia results from a deficiency of b-globin and can arise by any number of different inactivating mutations in the hemoglobin b chain (*HBB*) gene. Different mutations in a single gene can also often result in different phenotypes. That can arise in two ways: either different types of mutation somehow have different effects on how the underlying gene works—which we consider here—or some factors outside the disease locus have varying effects on the phenotype (described later).

Phenotype variation due to different mutations at a single gene locus may differ in degree (severe or mild versions of the same basic phenotype) or be extensive and result in rather different disorders. For example, Duchenne and Becker muscular dystrophies (OMIM 310200 and 300376, respectively) represent severe and mild forms of the same type of muscular dystrophy and are both examples of dystrophinopathies (PMID 20301298). More extreme phenotype heterogeneity can result from mutations at some genes (see the example of the lamin A/C gene in [Table 5.1](#)).

TABLE 5.1 REMARKABLE HETEROGENEITY OF CLINICAL PHENOTYPES
RESULTING FROM MUTATION IN THE LAMIN A (*LMNA*) GENE

Class of disorder	Disorder	Inheritance pattern	OMIM No.
Lipodystrophy	lipodystrophy, familial partial, type 2	AD	151660
	mandibulosacral dysplasia type A with lipodystrophy	AR	248370
	Emery-Dreifuss muscular dystrophy type 2	AD	181350
	Emery-Dreifuss muscular dystrophy type 3	AR	181350

Class of disorder	Disorder	Inheritance pattern	OMIM No.
Neuropathy	congenital muscular dystrophy	AD	613205
	cardiomyopathy, dilated type IA	AD	115200
	Malouf syndrome (cardiomyopathy, dilated, with hypertrophic hypogonadism)	AR	212112
	heart-hand syndrome, Slovenian type	AD	610140
	Charcot-Marie-Tooth disease, type 2B1	AR	605588
Progeria	Hutchinson-Gilford progeria syndrome	AD, AR	176670

Clinical phenotypes can also vary between affected members of the same family even although they have identical mutations. As we saw in [Section 5.2](#), heteroplasmy can explain divergent phenotypes in family members affected by a mitochondrial DNA disorder. But single-gene disorders can also show intrafamilial variation in phenotype that may be due to genetic and nongenetic factors as described below.

Nonpenetrance and age-related penetrance

The **penetrance** of a single-gene disorder is the probability that a person who has a mutant allele will express the disease phenotype. Dominantly inherited disorders, by definition, are manifested in heterozygotes and might be expected to show 100 % penetrance. That might be true for certain dominant disorders. For many others, however, penetrance is more variable and the disorder can sometimes appear to skip a generation so that a person who must have inherited the disease allele is unaffected (**nonpenetrance**—see [Figure 5.12](#)).

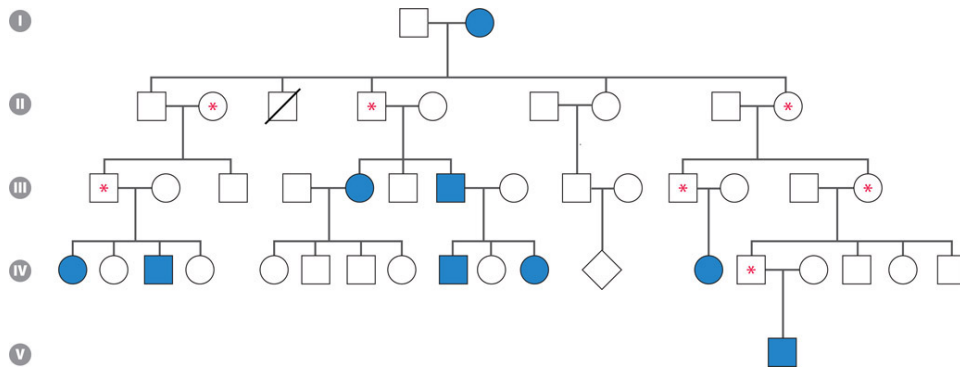


Figure 5.12 Nonpenetrance in an autosomal dominant disorder. Individuals with a red asterisk are asymptomatic disease gene carriers: they have inherited a mutant allele ultimately from the affected great-great-grandmother in generation I, but none of them expresses the disease phenotype. In this example, the disorder is evident only in individuals who have inherited a mutant allele from their father (in each case the unaffected individuals with a red asterisk inherited a mutant allele from their mother). As described in the text, an epigenetic mechanism known as *imprinting* can result in this type of parent-of-origin effect on the phenotype.

Nonpenetrance should not be viewed as surprising. Even in single-gene disorders—in which, by definition, the phenotype is largely dictated by the genotype at just one locus—other genes can play a part, as can epigenetic and environmental factors.

Variable age at onset in late-onset disorders

A disease phenotype may take time to manifest itself. If a disorder is present at birth, it is said to be congenital. In some disorders, however, there is a late age at onset so that the penetrance is initially very low but then increases with age. Age-related penetrance means a late onset of symptoms, and quite often the disease first manifests in adults.

The slow development of disease in adult-onset disorders may occur in different ways. Harmful products may be produced slowly but build up over time, for example. If pathogenesis involves a gradual process of cell death, it may take some time before the number of surviving cells drops to

critically low levels that produce clinical symptoms. In hereditary cancers, a mutation is inherited at a tumor-suppressing gene locus and a second, somatic mutation is required to initiate tumor formation. The second mutation occurs randomly, but the probability of a second mutation increases with time and therefore with age.

Huntington disease is a classic example of a late-onset single-gene disorder. In this case, mutant alleles produce an abnormal protein that is harmful to cells and especially toxic to neurons. The loss of neurons is gradual but eventually results in a devastating neurodegenerative condition. Huntington disease is highly penetrant. The onset of symptoms typically occurs in middle to late adult life, but juvenile forms are also known ([Figure 5.13](#)).

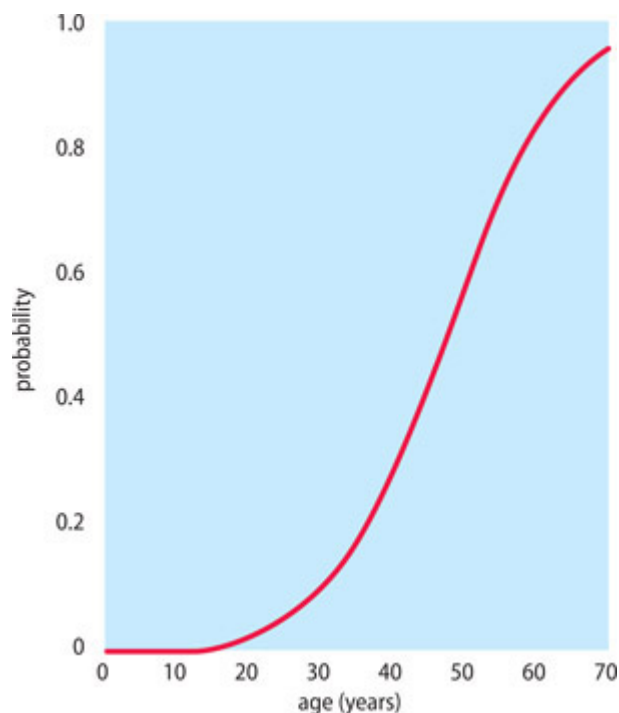


Figure 5.13 Age-related onset of Huntington disease. The curve shows the probability that an individual carrying a Huntington disease allele will have developed symptoms by a given age. (From Harper PS [2010] *Practical Genetic Counselling*, 7th ed. With permission from Taylor & Francis Group LLC.)

Age-at-onset curves for late-onset disorders are used in genetic counseling to calculate the chance that an asymptomatic person at risk of

developing the disease carries the mutation. In Huntington disease an unaffected person who has an affected parent will have a 50 % *a priori* risk that decreases with age (see [Figure 5.13](#)); if one is still free of symptoms by age 60, for example, the chance of developing the disease falls to less than 20 %.

Phenotypes resulting from mutation in mitochondrial DNA are highly variable because of the special mitochondrial property of heteroplasmy (see [Section 5.2](#)). Some types of Mendelian disorders, notably dominant phenotypes, are also prone to variable expression, and different family members show different features of disease (sometimes called *variable expressivity*—see [Figure 5.14](#) for an example pedigree). But, like nonpenetrance, variable phenotype expression is occasionally seen in recessive pedigrees.

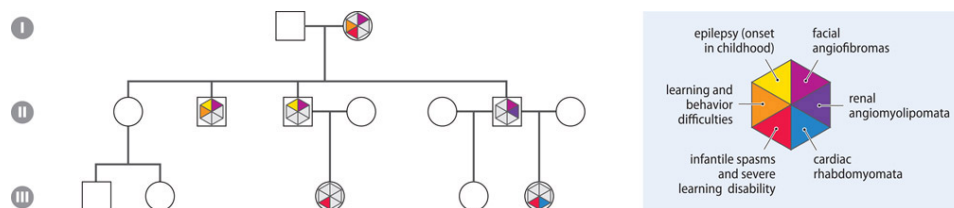


Figure 5.14 Variable phenotypes in a tuberous sclerosis family. Tuberous sclerosis is an autosomal dominant disorder caused by mutations in either the *TSC1* or *TSC2* gene. These two genes make two subunits of a tumor suppressor protein complex that regulates cell growth and proliferation. The disorder affects multiple body systems with characteristic tumor-like lesions in the brain, skin, and other organs, and is often associated with seizures and learning difficulties. However, as is evident in this family from the northeast of England, there can be considerable differences in *expressivity* of the disorder. (Pedigree information provided by Dr Miranda Splitt, Northern Region Genetics Service UK.)

Nonpenetrance can be regarded as an extreme endpoint of variable expression, and the factors that produce variable expression of phenotypes within families are the same as those that result in nonpenetrance. They include nongenetic factors—epigenetic regulation and environmental factors ([Figure 5.15B](#)) and also stochastic factors. Additional genetic

factors are also involved, notably *modifier genes* that regulate or interact with a Mendelian locus, affecting how it is expressed. Different alleles at a modifier gene locus may have rather different influences on the expression of the Mendelian locus ([Figure 5.15B](#)).

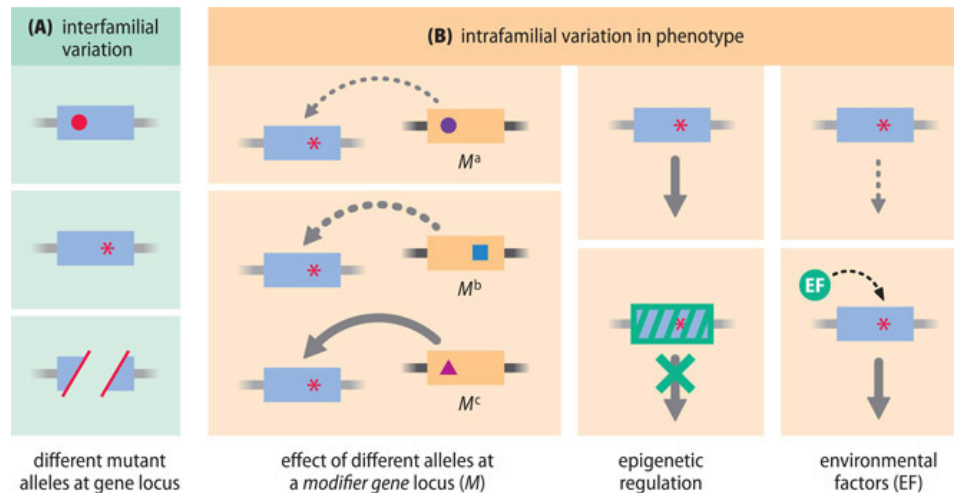


Figure 5.15 Main explanations for phenotype variation in Mendelian disorders. (A) *Interfamilial variation in phenotype*. Unrelated individuals with the same Mendelian disorder may often have different mutations (red symbols) at the disease gene locus with different consequences for gene expression and disease. (B) *Intrafamilial variation in phenotype*. Affected members of a single family can be expected to have the same mutation at the disease gene locus but can nevertheless show differences in phenotype because of genetic or nongenetic factors. In the former case, the affected individuals may have different alleles at one or more modifier gene loci. Modifier genes make products that interact with the primary gene locus so as to modulate the phenotype, and different alleles of a modifier gene can have different effects. Alternatively, nongenetic factors can explain phenotype variation; an example is epigenetic regulation, in which the disease allele can be differently regulated in some individuals by an altered chromatin conformation (green hatched box) or by variable exposure to an environmental factor (green circle) such as a specific virus or chemical during development *in utero*.

Imprinting

Certain phenotypes show autosomal dominant inheritance with parent-of-origin effects. Both sexes are affected, and the mutant allele can be transmitted by either sex but is expressed only when inherited from a parent of one particular sex. For some conditions, a mutant allele must be inherited from the father for the disease to be expressed (see [Figure 5.12](#) for an example). For other conditions, such as Beckwith-Wiedemann syndrome, the disease phenotype is expressed only if the disease allele is inherited from the mother ([Figure 5.16](#)).

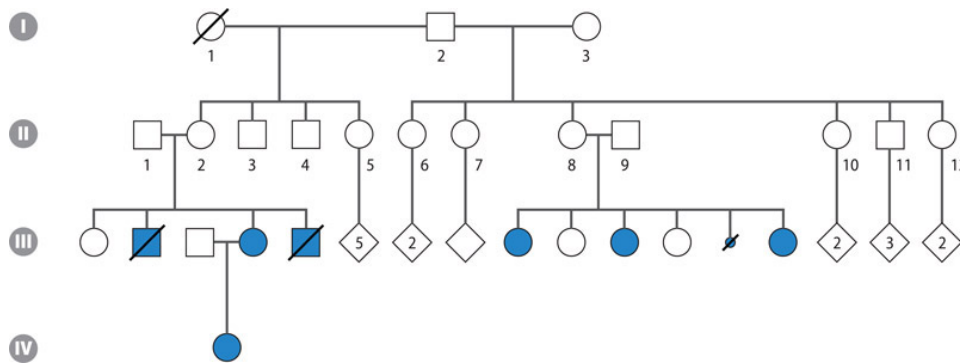


Figure 5.16 Parent-of-origin effect on the expression of an inherited disorder. This pedigree shows autosomal dominant Beckwith-Wiedemann syndrome (PMID 20301568), which manifests only when the underlying mutant allele is maternally inherited. The affected individuals in generation III must have inherited the mutant allele from their common grandfather I-2 but none of his 10 children in generation II have symptoms of disease, including two daughters, II-2 and II-8, who have gone on to have multiple affected children. (From Viljoen D & Ramesar R [1992] *J Med Genet* 29:221–225; PMID 1583639. With permission from BMJ Publishing Group Ltd.)

The parent-of-origin effects are due to an epigenetic mechanism known as imprinting, which we describe in detail in [Chapter 6](#). The mutant allele that is not expressed is often described as the imprinted allele. Accordingly, Beckwith-Wiedemann syndrome is said to be paternally imprinted, because paternally inherited alleles are not expressed.

Anticipation

Some disorders show consistent generational differences in phenotype. Disorders such as fragile X mental retardation syndrome, myotonic dystrophy, and Huntington disease are caused by unstable mutations (often called *dynamic mutations*) whose characteristics can change after they undergo DNA replication. As a result, the phenotype can vary between affected individuals in families but in a directional way; that is, it can be expressed at an earlier age and become increasingly severe with each new generation of affected individuals. This phenomenon is known as **anticipation** ([Figure 5.17](#)). We consider the molecular mechanisms in detail in [Chapter 7](#).



Figure 5.17 A three-generation family affected with myotonic dystrophy.

The degree of severity increases in each generation. The grandmother (right) is only slightly affected, but the mother (left) has a characteristic narrow face and somewhat limited facial expression. The baby is more severely affected and has the facial features of children with neonatal-onset myotonic dystrophy, including an open, triangular mouth. The infant has more than 1000 copies of the trinucleotide repeat, whereas the

mother and grandmother each have about 100 repeats. (From Jorde LB, Carey JC & Bamshad MJ [2009] *Medical Genetics*, 4th ed. With permission from Elsevier.)

5.4 ALLELE FREQUENCIES IN POPULATIONS

Genetic disorders that are comparatively common and serious have somehow avoided being eliminated by natural selection. This raises two questions. First, are high mutation rates enough to explain why harmful disease alleles persist? And if so, why should some single-gene disorders be comparatively common but others very rare? In this section we are concerned primarily with allele frequencies and the factors that affect them.

The frequency of a single-gene disorder in a population relates to the frequency in the population of pathogenic alleles at the relevant disease locus (or loci). A high disease allele frequency might result if a gene were to be particularly susceptible to mutation. Large genes may contain many repetitive sequences that confer structural instability, such as the very large dystrophin gene that is very prone to intragenic deletions and duplications.

Some of the most common single-gene disorders, such as sickle-cell anemia and the thalassemias, result from mutation in tiny genes—as we will see below, autosomal recessive disorders do not require high mutation rates to be common. Even in some autosomal dominant disorders, a high incidence of the disorder may not necessarily mean that the underlying gene loci have high mutation rates, as described below.

Some disorders may be caused by a *selfish mutation*. Achondroplasia (PMID 20301331) is a common single-gene disorder but is caused exclusively by mutation at just a single nucleotide, producing a highly specific change (glycineto-arginine substitution at residue 380) in the FGFR3 (fibroblast growth factor receptor type 3) protein. The nucleotide that is altered is not thought to be highly mutable. Instead, the mutation may promote its own transmission: male germ-line cells that contain it may have a proliferative advantage and make a disproportionate contribution to

sperm. As a result, there is a high allele frequency even although the mutation rate is not so exceptional. We consider selfish mutations in detail in [Section 7.2](#).

We also need to explain why some single-gene disorders are common in some human populations but very rare in others. Cystic fibrosis is particularly common in northern European populations, for example, and sickle-cell anemia is especially frequent in tropical Africa but virtually absent from many other human populations.

In all of these considerations, what do we mean by a human *population*? We could mean anything from a small tribe to the whole of humanity. An idealized population would be large with no barriers to random mating; as we will see below, some important principles in population genetics are based on this kind of population.

In practice, mating is often far from random because of different types of barriers. Geographic barriers can mean that people who live in locations that are remote (or otherwise difficult to access) form populations with limited genetic diversity and with distinctive allele frequencies. But even within single cities there are also many ethnic populations with distinctive allele frequencies. And, as we will see below, even within these populations, mating is not random.

Allele frequencies and the Hardy-Weinberg law

The frequency of an allele in a population can vary widely from one population to another. The concept of the **gene pool** (all of the alleles at a specific gene locus within the population) provides the reference point for calculating allele frequencies (which are often inaccurately represented in the literature as *gene frequencies*).

For a specific allele, say allele A^*1 at locus A , the **allele frequency** is the proportion of all the alleles in the population at locus A that are A^*1 and is given as a number between 0 and 1. Effectively, the allele frequency for A^*1 is the *probability* that an allele, picked at random from the gene pool, would be A^*1 .

The Hardy-Weinberg law

The Hardy-Weinberg law (or equilibrium, principle, theorem) provides a mathematical relationship between allele frequencies and genotype frequencies in an *idealized* large population where matings are random and allele frequencies remain constant over time.

Imagine that locus A has only two alleles, A^*1 and A^*2 , and that their respective frequencies are p and q (so that $p + q = 1$). The respective genotypes are combinations of two alleles at a time. To calculate the frequency of a genotype, we therefore first need to estimate the probabilities of picking first one specified allele from the gene pool (as the paternal allele, say), and then picking a second allele to be the maternal allele.

Imagine we pick A^*1 first (with a probability of p) and then we pick A^*1 again (with a probability of p). If the population is large, the two probabilities are independent events and so the joint probability of picking A^*1 first and then A^*1 again is the product of the two probabilities, namely p^2 . This is the only way that we can arrive at the genotype $A^*1.A^*1$, whose frequency is therefore p^2 ([Figure 5.18](#)).

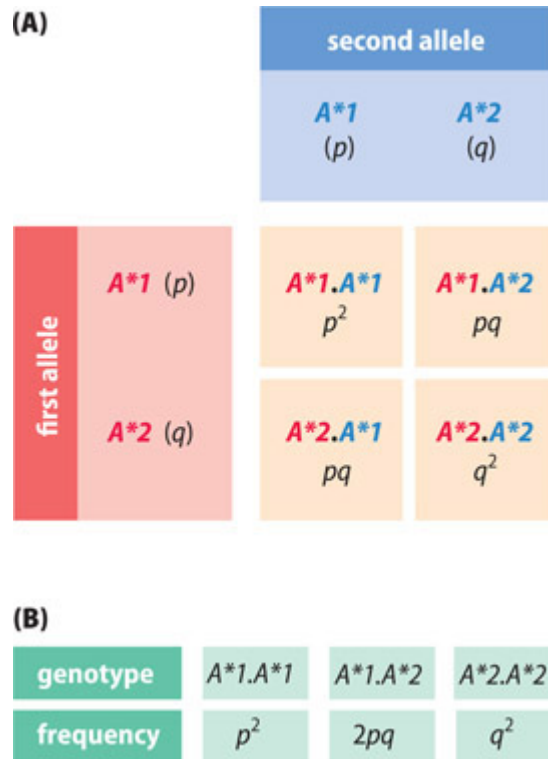


Figure 5.18 Visualizing how genotype frequencies are related to allele frequencies.

In this example we consider a locus A that has two alleles, A^*1 and A^*2 , with respective frequencies p and q . Genotypes are unique combinations of *two* alleles, one from a father and one from a mother. (A) We can first construct a matrix of all possible pairwise allele combinations, whose frequencies are simply the products of the frequencies of the two alleles. (B) We then integrate any pairwise combinations that have the same two alleles ($A^*1.A^*2$ is effectively the same as $A^*2.A^*1$) to get the frequencies of the three unique genotypes. Note that the Hardy-Weinberg law relates genotype frequencies to allele frequencies by a binomial expansion: $(p + q)^2 = p^2 + 2pq + q^2$ for two alleles (as shown here), $(p + q + r)^2$ for three alleles, $(p + q + r + s)^2$ for four alleles, and so on.

Now consider the genotype $A^*1.A^*2$. We can get this in two ways. One way is to first pick A^*1 (probability p) and then A^*2 (probability q), giving a joint probability of pq . But a second way is to pick A^*2 first (probability q) and then pick A^*1 (probability p), again giving a combined probability of pq . As a result, the frequency of the genotype $A^*1.A^*2$ is $2pq$.

In summary, in a suitably ideal population, the Hardy-Weinberg law gives the frequencies of homozygous genotypes as the square of the allele frequency, and the frequencies of heterozygous genotypes as twice the product of the two allele frequencies. An important consequence is that if allele frequencies in a population remain constant from generation to generation, the genotype frequencies will also not change.

Applications and limitations of the Hardy-Weinberg law

The major clinical application of the Hardy-Weinberg law is as a tool for calculating genetic risk. In a family with a single-gene disorder, only one or two mutant alleles are normally found in the causative gene, but within a population there may be many different mutant alleles at the disease locus. To apply the Hardy-Weinberg law to single-gene disorders, all the different mutant alleles are typically lumped together to make one disease allele. That is, we envisage just two alleles according to their effect on the disease phenotype: a normal allele (N), with no effect on the phenotype, and a disease allele (D), which can be *any* mutant allele. If we assign frequencies of p for allele N and q for allele D , the genotype frequencies would be as follows: p^2 for NN (normal homozygotes), $2pq$ for ND (heterozygotes), and q^2 for DD (disease homozygotes).

Practical application of the Hardy-Weinberg law to single-gene disorders is largely focused on autosomal recessive disorders, where it allows the frequency of carriers to be calculated without having to perform relevant DNA tests on a large number of people ([Box 5.4](#)). Its utility depends on certain assumptions—notably random mating and constant allele frequencies—that may not be strictly upheld. As described below, allele frequencies can change in populations, but the changes are often slow and in small increments, and often have minor effects in disturbing the Hardy-Weinberg distribution of genotypes. However, certain types of nonrandom mating can substantially upset the relative frequency of genotypes predicted by the Hardy-Weinberg law.

BOX 5.4 USING THE HARDY-WEINBERG LAW TO CALCULATE CARRIER RISKS FOR AUTOSOMAL RECESSIVE DISORDERS

Genetic counseling for autosomal recessive conditions often requires calculations to assess the risk of being a carrier. The proband who seeks genetic counseling is typically a prospective parent with a close relative who is affected. He/she is worried about the high risk of being a carrier and then about the risk that his/her spouse could also be a carrier.

The proband's chance of being a carrier can be calculated by using the principles of Mendelian inheritance, but the Hardy-Weinberg law is used to calculate the risk that his/her spouse could also be a carrier. If both parents were to be carriers, each child would have a 1 in 4 risk of being affected.

Take the specific example in [Figure 1](#). The healthy proband (arrowed) has a sister with cystic fibrosis and is worried about the prospect that he and his wife might have a child with cystic fibrosis. His wife is Irish, and the Irish population has the highest incidence of cystic fibrosis in the world, affecting one birth in 1350.

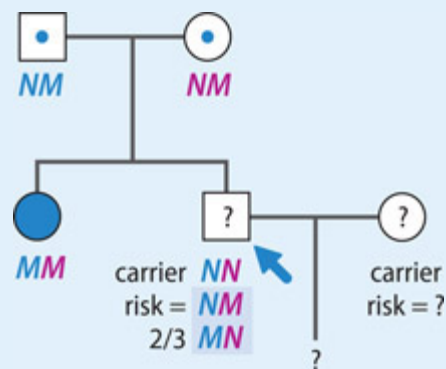


Figure 1 Using a combination of Mendelian principles and the Hardy-Weinberg law to estimate disease risk. The arrow indicates the proband. N , normal allele. M , mutant allele.

The proband's parents can be presumed to be carriers, each with one normal allele N and one mutant allele M . Because the proband is healthy,

he must have inherited one of three possible combinations of parental alleles: N from both parents (homozygous normal); N from father and M from mother (carrier); and M from father and N from mother (carrier). So from Mendelian principles, he has a risk of $2/3$ of being a carrier (see [Figure 1](#)).

The risk that his wife is a carrier is the same as the probability that a person, picked at random from the Irish population, is a carrier. If we assign a frequency of p for the normal allele and q for the cystic fibrosis allele, the Hardy-Weinberg law states that the frequency of affected individuals will be q^2 and the frequency of carriers will be $2pq$. Because population surveys show that cystic fibrosis affects 1 in 1350 births in the Irish population, $q^2 = 1/1350$ and so $q = 1/\sqrt{1350}$, or $1/36.74 = 0.027$.

Since $p + q = 1$, the value of $p = 0.973$. The risk of the wife being a carrier ($2pq$) is therefore $2 \times 0.973 \times 0.027 = 0.0525$, or 5.25 %. The combined risk that both the proband and his wife are carriers is $2/3 \times 0.0525 = 0.035$, or 3.5 %.

For rare autosomal recessive disorders, the value of p very closely approximates 1, so the carrier frequency can be taken to be $2q$. However, if the disorder is especially rare, the chances that the prospective parents are consanguineous is much higher, making the application of the Hardy-Weinberg law much less secure.

Nonrandom mating

In addition to geographical barriers to random mating, people also preferentially select mates who are similar to themselves in different ways. They may be members of the same ethnic group and/or sect, for example. Because breeding is less frequent between members from different communities, allele frequencies can vary significantly in the different communities. Geneticists therefore need to define populations carefully and calculate genetic risk by using the most appropriate allele frequencies.

Additional types of assortative mating occur. We also tend to choose a mate of similar relative stature and intelligence to us, for example. Positive assortment mating of this type leads to an increased frequency of homozygous genotypes and a decreased frequency of heterozygous genotypes. It extends to medical conditions. People who were born deaf or blind have a tendency to choose a mate who is similarly affected.

Inbreeding is a powerful expression of assortative mating that is quite frequent in certain societies and can result in genotype frequencies that differ significantly from Hardy-Weinberg predictions. Consanguineous mating results in an increased frequency of mating between carriers and a correspondingly increased frequency of autosomal recessive disease.

Ways in which allele frequencies change in populations

Allele frequencies can change from one generation to the next in different ways. Often changes in allele frequency are quite slow, but occasionally the composition of populations can change quickly, producing major shifts in allele frequency. Principal ways in which alleles change in the frequency of a population are listed below.

- *Purifying selection.* If a person affected by genetic disease is unlikely to reproduce, disease alleles are lost from the population (a form of negative natural selection). This effect is much more pronounced in early-onset dominant conditions, in which—with the exception of nonpenetrance—anyone with a mutant allele is affected by the time of puberty.
- *New mutations.* New alleles are constantly being created by the mutation of existing alleles. Some mutations produce new disease alleles by causing genes to lose their function or to function abnormally. There are numerous different ways in which a “forward” mutation can cause a gene to lose its function, but a “back mutation” (*revertant* mutation) that can restore the function of a nonfunctioning allele has to be very specific and so is comparatively very rare.

- *Influx of migrants.* If a population absorbs a large influx of migrants with rather different allele frequencies, then the overall gene pool will change.
- *Random sampling of gametes.* Only a certain proportion of individuals within a population reproduce. Out of all the alleles within the population, therefore, only those present in people who reproduce can be transmitted to the next generation. That is, a *sample* of the total alleles in the population is passed on and that sample is never exactly representative of the total population for purely statistical reasons. The smaller the size of a population, the larger will be the random fluctuations in allele frequency. This effect is known as genetic drift and in small populations it can cause comparatively rapid changes in allele frequencies between generations ([Figure 5.19](#)).

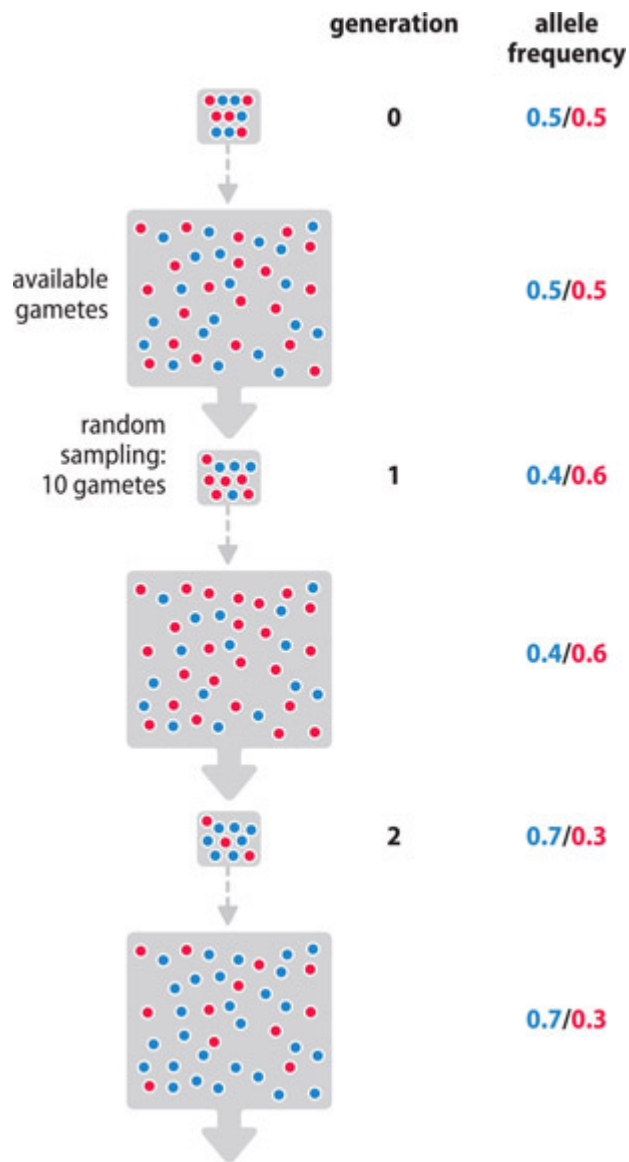


Figure 5.19 Random sampling of gametes in small populations can lead to **considerable changes in allele frequencies**. Small boxes represent gametes transmitted by reproducers to the next generation; large boxes represent all available gametes in the population. The comparative frequencies of the red and blue alleles can change significantly between generations when by chance the samples of transmitted gametes have allele frequencies that are rather different from the allele frequencies in the population. Such *genetic drift* is significant in small populations. (Adapted from Bodmer WF and Cavalli-Sforza LL [1976] *Genetics, Evolution and Man*. With permission from WH Freeman & Company.)

Population bottlenecks and founder effects

Genetic drift is most significant when population sizes are small. There have been several occasions during our evolution when the human population underwent a *population bottleneck*, a severe reduction in size before the reduced population (now with much less genetic variation) expanded again ([Figure 5.20A](#)). As a result, genetic variation in humans is very much less than in our nearest relative, the chimpanzee.

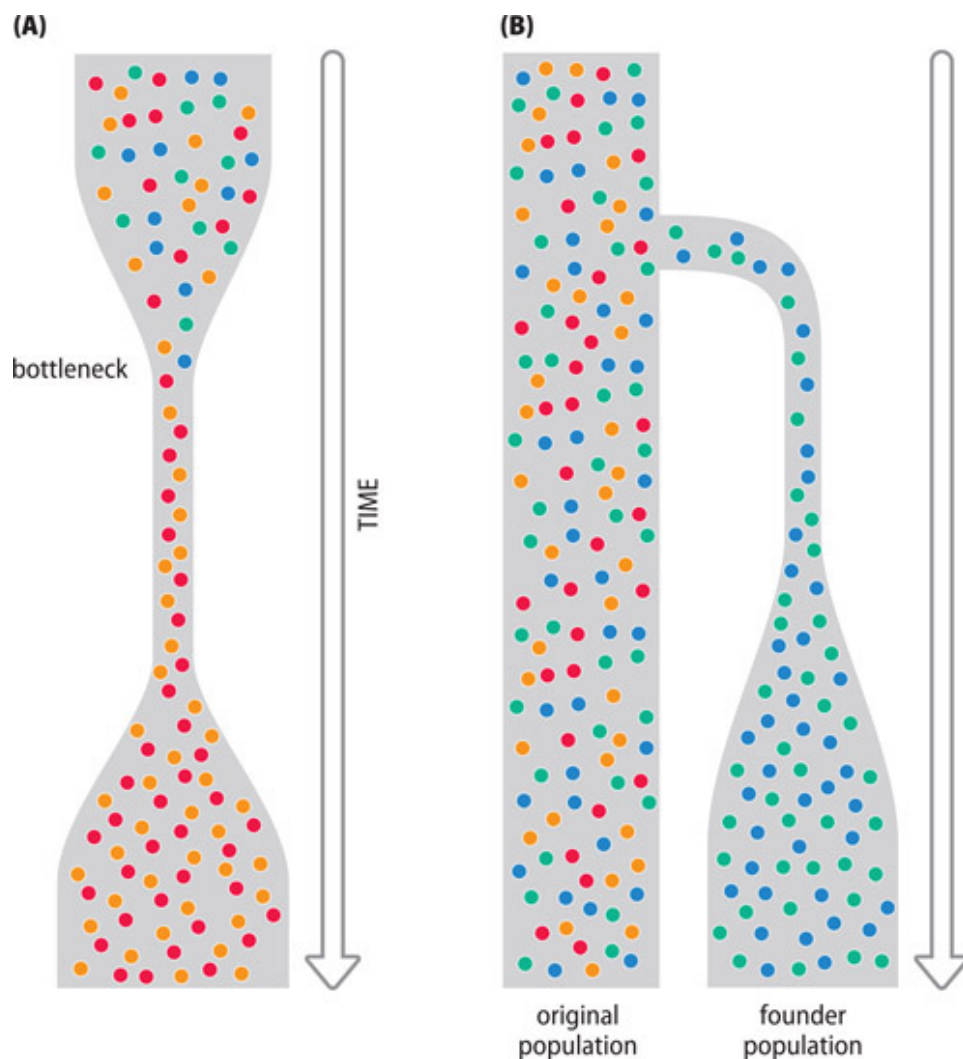


Figure 5.20 Altered allele frequencies after a population bottleneck and formation of a founder population. (A) In a population bottleneck, a severe reduction in the size of the population can lead to altered allele frequencies and much less genetic variation in the subset of the surviving population. Subsequent expansion will reestablish a large

population but with reduced genetic variation compared with the time before the bottleneck. (B) A small group of individuals, with a subset of the genetic variation of the larger population, migrate to establish a separate colony (founder population) that can expand but continues to show different allele frequencies from the original foundation. In both images the vertical arrows indicate the passage of time.

Another type of population reduction has periodically happened during human migrations, when a small group of individuals emigrated to form a separate colony. Again, the small population would represent a subset of the genetic variation in the original population and have different allele frequencies. Subsequent expansion of the founding colony would lead to a new population that continued to have limited genetic variation and distinctive allele frequencies reflecting those of the original settlers (a **founder effect**—see [Figure 5.20B](#)).

If a founder colony happens to have an increased frequency of a disease allele, the new population that will descend from it can be expected to have an increased frequency of the disease. Various populations throughout the world have elevated frequencies of certain single-gene disorders as a result of a founder effect. In autosomal recessive disorders, the great majority of mutant alleles are found in asymptomatic carriers who transmit the mutant alleles to the next generation.

The Finnish and Ashkenazi Jewish populations have been particularly amenable to investigations of founder effects because of rapid recent population expansion, high education levels, and very well-developed medical services. After the introduction of agriculture from the Middle East in prehistoric times, Finland was one of the last regions of Europe to be populated, and the major expansion that led to the present population began only 2000–2500 years ago as migrants entered southern Finland. Thereafter, in the seventeenth century a second large population expansion began with the occupation of the uninhabited north of Finland.

Ashkenazi Jews (descended from a population that migrated to the Rhineland in the ninth century and from there to different countries in eastern Europe) and Sephardic Jews (primarily from Spain, Portugal, and

north Africa) have been distinct populations for more than a thousand years. Until just a few hundred years ago, Ashkenazi Jews used to represent a minority of Jews, but they have undergone a rapid population expansion and now account for 80 % of the global Jewish population. Founder effects have been documented in many other populations; see [Table 5.2](#) for examples.

TABLE 5.2 EXAMPLES OF SINGLE-GENE DISORDERS THAT ARE COMMON IN CERTAIN POPULATIONS BECAUSE OF A FOUNDER EFFECT

Disorder and inheritance (OMIM)	Population	Comments
Aspartylglucosaminuria; AR (208400)	Finnish	carrier frequency = 1 in 30
Ellis-van Creveld syndrome; AR (225500)	Amish, Pennsylvania	carrier frequency \approx 1 in 8. Traced to a single couple who immigrated to Pennsylvania in 1774
Familial dysautonomia; AR (223900)	Ashkenazi Jews	carrier frequency = 1 in 30
Hermansky-Pudlak syndrome; AR (203300)	Puerto Ricans	thought to have been introduced by migrants from southern Spain
Alzheimer disease type 3, early onset; AR (607822)	in remote villages in the Andes	all descended from a couple of Basque origin who settled in Colombia in the early 1700s
Huntington disease (HD); AD (143100)	in fishing villages around Lake Maracaibo, Venezuela	more people with HD here than in rest of world. About 200 years ago, a single woman with the HD allele bore 10 children. Many current residents of Lake

Disorder and inheritance (OMIM)	Population	Comments
		Maracaibo can trace their ancestry and disease-causing allele back to this lineage
Myotonic dystrophy, type I, AD (160900)	in Saguenay-Lac-Saint-Jean, Quebec	prevalence of 1 in 500 (30–60 times more frequent than in most other populations). Introduced by French settlers

A distinguishing feature of a founder effect is that affected individuals will usually have mutant alleles with the same ancestral mutation. For example, affected individuals in nine Amish families with Ellis-van Creveld syndrome were shown to be homozygous for the same pathogenic mutation in the *EVC* gene and for a neighboring nonpathogenic sequence change that is absent from normal chromosomes. In this case, genealogy studies were able to confirm a founder effect: all affected individuals could trace their ancestry to the same couple, a Mr Samuel King and his wife, who immigrated in 1774.

Mutation versus selection in determining allele frequencies

If we consider stable, large populations (so that migrant influx and genetic drift are not significant factors), the frequencies of mutant alleles (and genetic diseases) in a population are determined by the balance between two opposing forces: mutation and selection.

Purifying selection removes disease alleles from the population when a disorder causes affected individuals to reproduce less effectively than the normal population. The genetic term **fitness** (*f*) is applied here and is really a measure of reproductive success: it uses a scale from 0 to 1 to rank the capacity of individuals to reproduce and have children who survive to a reproductive age. Thus, a fitness of 0 (genetic lethal) means consistent failure to reproduce, and so mutant alleles are not transmitted vertically to

descendants. Loss of mutant alleles from the population by purifying selection is balanced by the creation of new mutant alleles by fresh mutation, keeping constant the disease allele frequency in the population.

For autosomal dominant disorders, all people who have a disease allele might be expected to be affected (if we discount nonpenetrance). Yet, according to the disorder, the fitness of individuals varies enormously. In many cases, affected individuals have severely or significantly reduced fitness. However, individuals affected by a late-onset disorder can have fitness scores that approach those of normal individuals—they are healthy in their youth and can reproduce normally ([Figure 5.21](#) gives some examples).

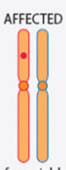
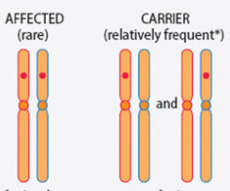
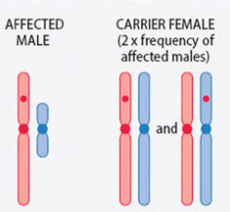
	individuals carrying mutant alleles	fitness of individuals		mutant alleles transmitted and created	
		affected individuals	carriers	vertical transmission and longevity	proportion created by new mutation
autosomal dominant	 <p>AFFECTED</p> <p>$f = \text{variable}$</p>	<p>$f = 0$ (genetic lethal)</p> <p>$f = \text{intermediate}$ (for example 0.5 for NF1; 0.2 for achondroplasia)</p> <p>f approaches 1 (late-onset disease for example Huntington disease)</p>	n/a	<p>none transmitted</p> <p>intermediate proportion transmitted and survive for several generations</p> <p>transmitted effectively and very long-lived in the population</p>	<p>100%</p> <p>intermediate (50% for NF1; 80% for achondroplasia)</p> <p>very small percentage</p>
autosomal recessive	 <p>AFFECTED (rare)</p> <p>$f = 0$ or low</p> <p>CARRIER (relatively frequent*)</p> <p>and</p> <p>$f = 1$</p>	$f = 0$ or low	$f = 1$ or >1	<p>not transmitted or rarely transmitted by affecteds; but transmitted very effectively by numerous carriers</p> <p>mutant alleles are therefore very long-lived in the population</p>	<p>very small percentage</p>
X-linked recessive	 <p>AFFECTED MALE</p> <p>$f = \text{variable}$</p> <p>CARRIER FEMALE (2 x frequency of affected males)</p> <p>and</p> <p>f often = 1</p>	<p>$f = 0$ (genetic lethal) (for example Duchenne muscular dystrophy)</p> <p>$f = \text{intermediate}$ (for example 0.7 for hemophilia)</p>	$f = 1$	<p>2/3 mutant alleles transmitted (by carriers)</p> <p>$>2/3$ mutant alleles transmitted</p>	<p>1/3</p> <p>$<1/3$</p>

Figure 5.21 Fitness of individuals and mutant allele transmission/creation in single-gene disorders. Note that carriers of certain autosomal recessive disorders may have a higher fitness than normal individuals (*heterozygote advantage*—see following page).

For recessive disorders, mutant alleles are also found in carriers who have a single mutant allele. In autosomal recessive disorders, carriers vastly outnumber affected individuals. Recall the Hardy-Weinberg law that gives a ratio of $2pq$ (carriers) to q^2 (affecteds) = $2p/q \approx 2/q$ (p is very close to 1 for almost all recessive disorders). To take one example, cystic fibrosis occurs in roughly 1 in 2000 births in northern European populations, so $q^2 = 1/2000$. This gives $q \approx 1/45$ and $2/q \approx 2/(1/45) = 90$. That is, there would be about 90 carriers of cystic fibrosis for each affected individual in this population. Because carriers of autosomal recessive disease are normally asymptomatic, they are very effective at transmitting mutant alleles and so new mutations are rare in autosomal recessive disease.

For X-linked recessive disorders, there are two female carriers per affected male. This happens because mutant alleles that reside on an X chromosome get transferred by recombination between three types of X chromosome: a single X chromosome in males and two X chromosomes in females. If we discount manifesting heterozygotes and take an approximation that the fitness of carriers is close to 1, then for conditions in which affected males do not reproduce, natural selection removes 1 out of 3 mutant alleles from the population. Because the lost alleles are replaced by new mutant alleles, 1 out of 3 mutations are new mutations.

Heterozygote advantage: when natural selection favors carriers of recessive disease

We saw above how some populations have a particularly high incidence of a genetic disorder as a result of a founder effect. Another reason why a recessive condition may be especially common in one population is that under certain conditions a type of natural selection can favor a particularly high frequency of carriers.

Recall that natural selection works to eliminate disadvantageous alleles within the population (purifying selection) and also to promote an increase in frequency of advantageous alleles (positive selection). That occurs because natural selection works through the genetic *fitness* of individuals

(their ability to reproduce and have children who survive to a reproductive age): disadvantageous alleles are alleles that reduce fitness; advantageous alleles increase fitness. But sometimes a disadvantageous allele can also simultaneously be an advantageous allele. A form of natural selection called **balancing selection** can cause a harmful disease allele to increase in frequency in a population because carriers of the mutant allele have a higher fitness than normal individuals (**heterozygote advantage**).

Sickle-cell anemia provides a classic example of heterozygote advantage. It is very common in populations in which malaria caused by the *Plasmodium falciparum* parasite is endemic (or was endemic in the recent past) but is absent from populations in which malaria has not been frequent. In some malaria-infested areas of west Africa, the sickle-cell anemia allele has reached a frequency of 0.15—far too high to be explained by recurrent mutation.

Sickle-cell heterozygotes have red blood cells that are inhospitable to the malarial parasite (which spends part of its life cycle in red blood cells). As a result, they are comparatively resistant to falciparum malaria. Normal homozygotes, however, frequently succumb to malaria and are often severely, sometimes fatally, affected. Heterozygotes therefore have a higher fitness than both normal homozygotes and disease homozygotes (who have a fitness close to zero because of their hematological disease).

Heterozygote advantage through comparative resistance to malaria has also been invoked for certain other autosomal recessive disorders that feature hemolytic anemia, such as the thalassemias and glucose-6-phosphate dehydrogenase deficiency. The high incidence of cystic fibrosis in northern European populations and Tay–Sachs disease in Ashkenazi Jews is also likely to have originated from heterozygote advantage, possibly through a greater resistance of carriers to infectious disease.

If continued over many generations, even a small degree of heterozygote advantage can be enough to change allele frequencies significantly (invalidating Hardy–Weinberg predictions that assume constant allele frequencies).

Distinguishing heterozygote advantage from founder effects

Diseases that are common in a population because of a founder effect typically originate from one (or occasionally two) mutant alleles. Most people in the population who carry mutant alleles can be expected to have the same ancestral mutation. Heterozygote advantage, by contrast, could be conferred by multiple different mutations of similar effect in the same gene.

If genealogical evidence is not available, strong support for a founder effect can still be obtained. DNA analyses may show that multiple individuals from the population have alleles with the same pathogenic mutation located within a common haplotype of nonpathogenic alleles at neighboring marker DNA loci. If, by contrast, it can be shown that there are multiple different disease alleles in the population or that the disease alleles are embedded in different haplotypes (suggesting different mutational events), heterozygote advantage is likely to apply. But sometimes it is difficult to distinguish between different possible contributions made by founder effects, heterozygote advantage, and even genetic drift when population sizes are very small.

SUMMARY

- Some human disorders and traits are very largely determined by genetic variation at a single gene locus.
- Multiple members of a human kinship (extended family) may be affected by the same single-gene disorder as a result of genetic transmission of mutant alleles (individual versions of a gene at one locus) from one generation to the next.
- In dominantly inherited disorders, an affected person is usually a heterozygote—one allele at the disease locus is defective or harmful, but the other allele is normal.
- In recessive disorders, an affected person has defective alleles only at the disease locus. A person with one disease allele and

one normal allele is usually an unaffected carrier who can transmit the harmful (mutant) allele to the next generation.

- A person with an autosomal recessive disorder may have two identical mutant alleles (a true homozygote) or two different mutant alleles (a compound heterozygote).
- Because the X and Y chromosomes have very different genes, men are hemizygous by having a single functional allele for most genes on these chromosomes.
- In X-linked recessive disorders, men are disproportionately affected (they have a single allele, but women with one mutant allele are usually asymptomatic carriers).
- One X chromosome is randomly inactivated in each cell of the early female embryo; descendant clonal cell populations have an inactivated maternal X or an inactivated paternal X. A female carrier of a mutant X-linked allele may be affected if the normal X has been inactivated in a disproportionately large number of cells.
- A genetic mosaic with a mixture of normal and mutant cells may be mildly affected but transmit the mutant allele to descendants who would have harmful mutations in each cell and be more severely affected.
- Some types of mutation are dynamically unstable and become more severe from one generation to the next (anticipation).
- Affected individuals in the same family can also show differences in phenotype because they have different alleles at some other genetic locus (modifier) that interacts with the disease gene locus.
- Our cells each contain many copies of mitochondrial DNA (mtDNA) and affected individuals in a family with a mtDNA disorder may be variably affected because of heteroplasmy (variable ratios of mutant to normal mtDNA copies per cell).

- In disorders of imprinting, individuals who have inherited a mutant gene may or may not be affected, depending on whether the mutant allele was inherited from the paternal or maternal line.
- There is no one-to-one correspondence between genes and phenotypes. Different mutations in the same gene can sometimes cause different disorders, and yet the same disorder is quite often caused by mutations in different genes.
- Some single-gene disorders are notably common in certain ethnic populations. For recessive disorders, a high carrier frequency may arise because asymptomatic carriers of the mutant allele have been reproductively more successful than individuals with two normal alleles (the single mutant allele may have given heterozygotes an advantage by providing greater protection against certain infectious diseases).
- Mutant alleles lost from the population (when individuals fail to reproduce) are balanced by new mutant alleles (created by fresh mutation). For recessive disorders and late-onset dominant disorders, comparatively few alleles are lost from the population and so fresh mutation rates are low. For a severe dominant disorder that manifests before puberty, the rate of fresh mutation may be very high.
- Allele frequencies can be calculated in populations by using the Hardy-Weinberg law, which gives the frequency of a homozygous genotype as the square of the allele frequency, and the frequency of a heterozygous genotype as twice the product of the allele frequencies.

QUESTIONS

Questions can be downloaded by visiting the following link, under Support Materials: www.routledge.com/9780367490812.

FURTHER READING

Single-gene disorders

[McKusick VA](#) (2007) Mendelian inheritance in Man and its online version, OMIM. *Am J Hum Genet* 80:588–604; PMID 17357067.

Pagon RA (eds) *GeneReviews*TM.

<http://www.ncbi.nlm.nih.gov/books/NBK1116/>; PMID 20301295 (see also **Box 5.1**).

General Mendelian inheritance

Bennett RL (2008) Standardized human pedigree nomenclature: update and assessment of the recommendations of the National Society of Genetic Counselors. *J Genet Counsel* 17:424–433; PMID 18792771.

Wilkie AOM (1994) The molecular basis of dominance. *J Med Genet* 31:89–98; PMID 8182727.

Zschocke J (2008) Dominant versus recessive: molecular mechanisms in metabolic disease. *J Inherit Metab Dis* 31:599–618; PMID 18932014.

X-linked inheritance and X-inactivation

Franco B & Ballabio A (2006) X-inactivation and human disease: X-linked dominant male-lethal disorders. *Curr Opin Genet Dev* 16:254–259; PMID 16650755.

Mangs AH & Morris BJ (2007) The human pseudoautosomal region (PAR): origin, function and future. *Curr Genomics* 8:129–136; PMID 18660847.

Migeon BR (2007) Females are Mosaics. *X Inactivation and Sex Differences in Disease*. Oxford University Press.

Orstavik KH (2009) X chromosome inactivation in clinical practice. *Hum Genet* 126:363–373; PMID 19396465.

Allele frequencies, mosaicism, and calculating genetic risk

Aidoo M (2002) Protective effects of the sickle cell gene against malaria morbidity and mortality. *Lancet* 359:1311–1312; PMID 11965279.

Clarke A (2019) *Harper's Practical Genetic Counselling*, 8th ed. CRC Press.

Hartl D & Clark AG (2007) *Principles of Population Genetics*, 4th ed. Sinauer Associates.

Hurst LD (2009) Fundamental concepts in genetics: genetics and the understanding of selection. *Nature Rev Genet* 10:83–93; PMID 19119264.

McCabe LL & McCabe ER (1997) Population studies of allele frequencies in single gene disorders: methodological and policy considerations. *Epidemiol Rev* 19:52–60; PMID 9360902.

Van der Meulen MA (1995) Recurrence risks for germinal mosaics revisited. *J Med Genet* 32:102–104; PMID 7760316.

6

Principles of gene regulation and epigenetics

DOI: [10.1201/9781003044406-6](https://doi.org/10.1201/9781003044406-6)

CONTENTS

[6.1 GENETIC REGULATION OF GENE EXPRESSION](#)

[6.2 CHROMATIN MODIFICATION AND EPIGENETIC FACTORS IN GENE REGULATION](#)

[6.3 ABNORMAL EPIGENETIC REGULATION IN MENDELIAN DISORDERS AND UNIPARENTAL DISOMY](#)

[SUMMARY](#)

[QUESTIONS](#)

[FURTHER READING](#)

All our cells develop ultimately from the zygote. Each nucleated cell in a person contains the same set of genes. However, only a subset of the genes in a cell are expressed to make functional end products, and that subset varies according to the type of cell. The global gene expression pattern of a cell dictates the form of a cell, how it behaves, and ultimately its identity—whether it will be a hepatocyte, for example, or a macrophage, or a sperm cell.

In [Chapter 2](#) we outlined the basic details of gene expression. Here, we are concerned with how the expression of genes is regulated. Different levels of gene regulation affect the production or stability of gene products: transcription, post-transcriptional processing (to make final mRNA or noncoding RNA products), translation of mRNA, post-translational modification, folding of protein products, incorporation into a multisubunit functional molecule, and degradation of gene products.

We explore aspects of mRNA degradation and protein folding in [Chapter 7](#). Here we mostly deal with gene regulation at the levels of transcription, post-transcriptional processing, and translation. Complex networks of interacting regulatory nucleotide sequences and proteins are involved.

Two fundamental types of gene regulation

All the cells in our body originate by cell division ultimately from one cell, the fertilized egg cell. Given that in each person the nucleated cells all contain the same DNA molecules, readers might reasonably wonder how we could ever come to have different cell types with distinct gene expression patterns. However, it is not just the sequence of nucleotides in DNA that determines gene expression. Chromatin structure is also crucially important, and gene expression is regulated at two fundamental levels listed below, one of which is not genetic.

- *Genetic regulation.* Here control of gene expression is *dependent* on the nucleotide sequence. If a promoter sequence is deleted, for example, the expected transcript is not produced.
- *Epigenetic regulation.* Here control of gene expression is *independent* of the nucleotide sequence. As detailed in [Section 6.2](#), various non-genetic control mechanisms can affect chromatin structure, causing it to be tightly compacted (preventing expression of genes) or more open (facilitating gene expression).

Major control mechanisms involve certain chemical modifications of DNA and histones, changes in the positioning of nucleosomes, and interactions of certain regulatory noncoding RNAs with chromatin.

We explain in [Section 6.2](#) how epigenetic controls are required in very early development to initiate programs of cell differentiation that progressively leads to different cell lineages and ultimately different cell types. We will also explain how epigenetic controls can be heritable (but can also be reset), and how they can be influenced by environmental and stochastic factors.

Cis-acting and trans-acting effects in gene regulation

Genetic control of gene expression largely depends on collections of short regulatory nucleotide sequences in both DNA and RNA; they act as target sequences that can be bound by certain regulatory RNA molecules and proteins.

A regulatory sequence is said to be **cis-acting** when its function is limited to the *single* DNA or RNA molecule it resides on. Take gene promoters. The promoter upstream of the insulin gene on a paternally inherited chromosome 11 regulates the *paternal* insulin gene only, not the allelic insulin gene on maternal chromosome 11. In addition, an allele may be regulated by more distantly related *cis-acting* sequences on the same chromosomal DNA molecule ([Figure 6.1A,B](#)).

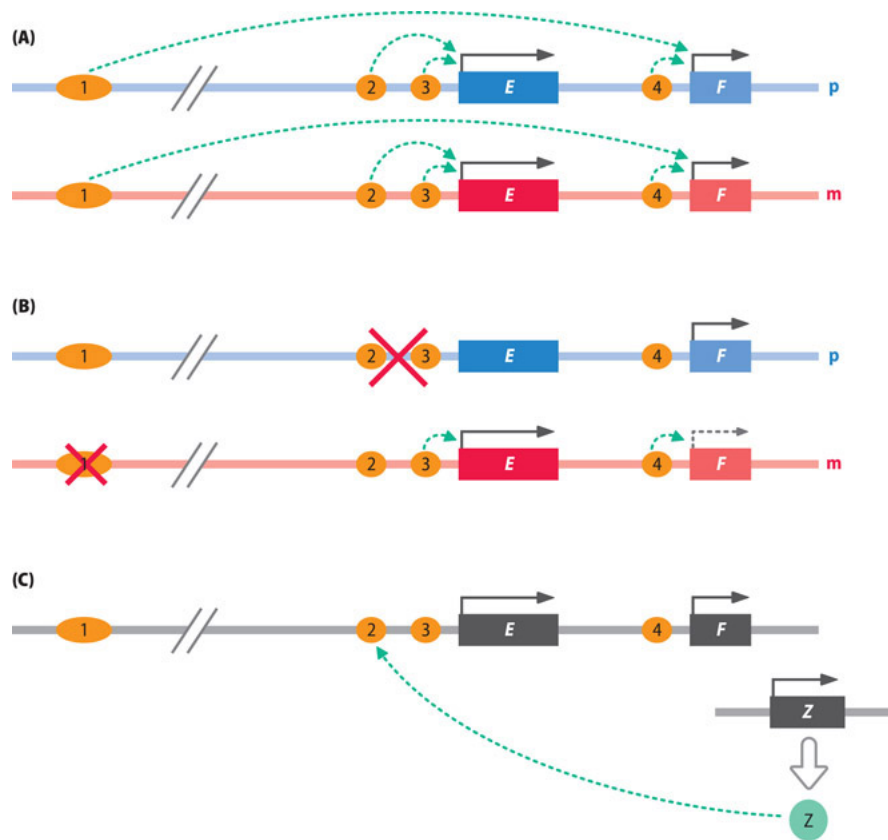


Figure 6.1 Examples of *cis-acting* and *trans-acting* effects at the DNA level. (A) Normal regulation of neighboring genes *E* and *F* on homologous chromosomes by positive *cis-acting* regulatory elements 1 to 4 (orange ovals). Paternal DNA (p) is shown in blue; maternal DNA (m) is in pink. Gene *E* is controlled by elements 2 and 3; gene *F* is controlled by proximal element 4 and remote element 1. (B) Effects of mutation (large red X) abolishing regulatory elements. Deletion of paternal elements 2 and 3 inactivates the paternal *A* allele only; deletion of maternal element 1 selectively reduces expression of the maternal *B* allele. (C) *Trans-acting* gene regulation. A remote gene *Z* on another chromosome (as shown here), or on the same chromosome as genes *E* and *F*, but distantly located (not shown) produces a *trans-acting* regulatory protein *Z* that binds to regulatory element 2 on *both* paternal and maternal chromosomes (represented here by a single, generic black chromosome).

A **trans-acting** gene regulator is a regulatory protein or regulatory RNA molecule that can migrate by diffusion to recognize and bind *specific* short regulatory nucleotide sequences in DNA or RNA. Unlike a *cis-acting* gene regulator, a *trans-acting* gene regulator can regulate the expression of *both* alleles on distantly located genes (Figure 6.1C). And some individual *trans-acting* regulators regulate a set of genes at multiple loci (all of which possess a target nucleotide sequence it can bind to).

Many RNA transcripts also contain *cis-acting* regulatory elements (whose effect is limited to regulating the expression of the RNA transcript on which they reside). Untranslated sequences in mRNA molecules, for example, generally contain *cis-acting* sequences that regulate the expression of the transcript. They may be recognized and bound by *trans-acting* regulatory proteins or *trans-acting* regulatory noncoding RNAs, notably microRNAs (see Figure 6.2).

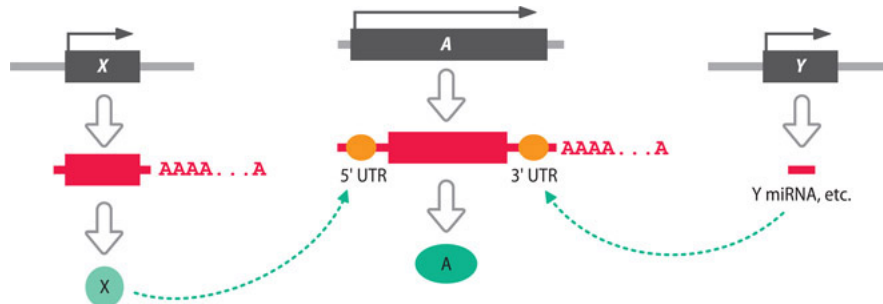


Figure 6.2 *Cis-acting and trans-acting regulation at the RNA level.* *Cis-acting* regulatory elements in mRNA are often located in untranslated regions (UTRs). Expression of protein-coding gene A is imagined to be regulated at the transcript level by *trans-acting* regulatory protein X and by microRNA Y that bind to *cis-acting* elements in the 5' and 3' UTRs, respectively, of the mRNA. Note that microRNAs often bind to target nucleotide sequences in RNA transcripts from multiple different genes and thereby regulate the expression of specific sets of genes. Figure 6.9 shows some examples of *transacting* regulatory proteins.

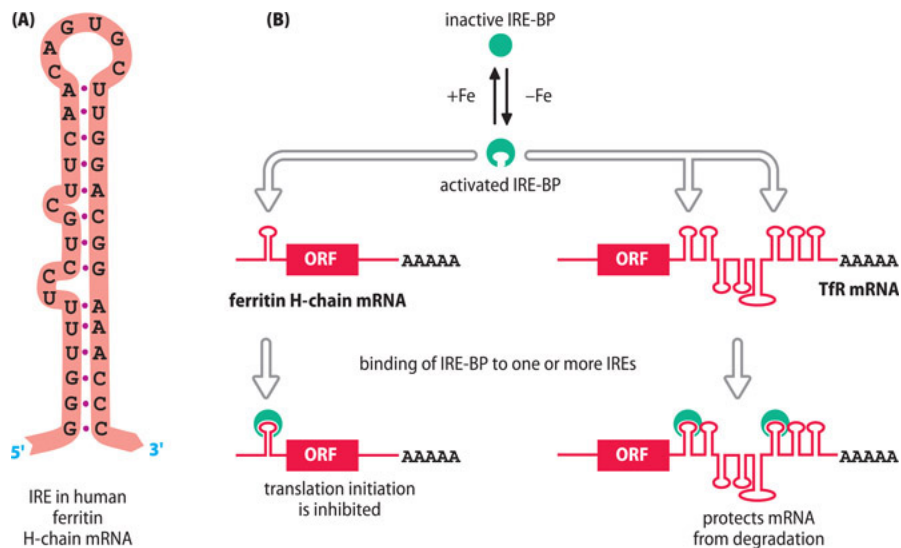


Figure 6.9 *Iron-response elements in the ferritin and transferrin mRNAs.* (A) Stem-loop structure of an iron-response element (IRE) in the 5' untranslated region of the ferritin heavy (H)-chain mRNA. (B) When iron levels are low, a specific IRE-binding protein (IRE-BP) is activated and binds the IRE in the ferritin heavy-chain gene and also IREs in the 3' untranslated region of the transferrin receptor (TfR) mRNA. Binding inhibits the translation of ferritin but protects the transferrin receptor mRNA from degradation, maximizing the production of transferrin receptor. When iron levels are high, the IRE-binding protein is inactivated, maximizing the production of ferritin and decreasing the production of transferrin receptor. ORF (open reading frame) designates the central coding DNA of the mRNAs.

Not shown in [Figure 6.2](#) are additional types of *cis*- and *trans*-regulation in which regulatory long noncoding RNA acts on DNA. Unlike *trans*-acting RNA regulators, *cis*-acting RNA regulators are not free to move away by diffusion; instead, they remain within chromatin, attached to the individual DNA strand they were transcribed from. We show in [Section 6.2](#) specific examples of how *cis*-acting regulatory RNA molecules work in epigenetic gene regulation in processes such as X-chromosome inactivation and genome imprinting.

We begin this chapter by looking at how genetic regulation governs how our genes are expressed. We then consider principles of epigenetic regulation. We end with a section on abnormal epigenetic regulation that results from abnormal chromosome inheritance, or from single-gene disorders in which mutations affect a gene involved in epigenetic regulation. We describe a very different type of epigenetic dysregulation that relates to protein folding in [Chapter 7](#), and we examine epigenetic contributions to complex disease in [chapters 8](#) and [10](#).

6.1 GENETIC REGULATION OF GENE EXPRESSION

As described in [Figure 2.12](#), the mitochondrial genome is transcribed from fixed start points, generating large multigenic transcripts from each DNA strand, which are subsequently cleaved. By contrast, it is usual for nuclear genes to be transcribed individually, and transcription is regulated by genetic factors, as described in this section, and also by epigenetic factors, which we consider in [Sections 6.2](#) and [6.3](#). Recently, geneticists have also become increasingly aware of the importance of post-transcriptional controls, notably at the level of RNA processing and translation.

Promoters: the major on–off switches in genes

Along the lengths of each DNA strand of our very long chromosomal DNA molecules are **promoters**, *cis*-acting regulatory DNA sequences that are important in establishing which segments of a DNA strand will be transcribed. Each promoter is a collection of very short sequence elements that are usually clustered within a few hundred nucleotides from the transcription start site. For each DNA strand, transcription begins at fixed points on the DNA where the chromatin has been induced to adopt a relaxed, “open” structure (see below).

Nuclear genes are transcribed by three different types of RNA polymerases. A nucleolar RNA polymerase, RNA polymerase I, is dedicated to making three of the four different ribosomal RNAs (rRNAs) in our cytoplasmic ribosomes (the 28S, 18S, and 5.8S rRNAs). It transcribes clusters of about 50 tandem DNA repeats (each containing sequences for the 28S, 18S, and 5.8S rRNAs) on each of the short arms of chromosomes 13, 14, 15, 21, and 22. RNA polymerase II transcribes all protein-coding genes, genes making long noncoding RNAs and some short RNA genes (including many miRNA genes). RNA polymerase III transcribes tRNA genes, the 5S rRNA gene, and some other genes that make short RNAs.

None of the RNA polymerases acts alone; each is assisted by dedicated protein complexes. In the case of RNA polymerase II, for example, a core transcription initiation complex is formed by the sequential assembly of five multisubunit proteins (*general* transcription factors—see below) at specific sites on the DNA.

Some of the protein subunits of the transcription initiation complexes recognize and bind specific short DNA sequence elements of a promoter; others are recruited by binding to previously bound proteins. For a protein-coding gene, most core promoter elements are upstream of the start site, and the spacing of the elements is important.

[Figure 6.3](#) illustrates some important core promoter elements, but note that the composition of core promoter elements is highly variable—some promoters lack all the elements shown in this figure. We describe additional *cis*-acting elements in the next section.

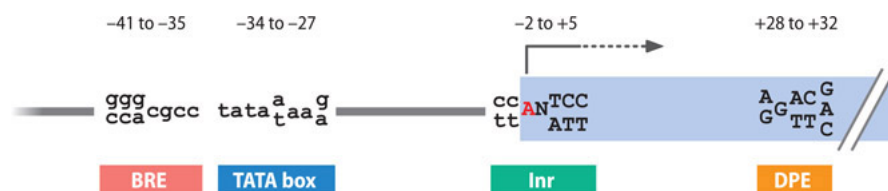


Figure 6.3 Consensus sequences for some core promoter elements often found in genes transcribed by RNA polymerase II. The TATA box is bound by the TATA-binding protein subunit of transcription factor IID. The initiator (Inr) element defines the transcription start site (the highlighted A) when located 25–30 bp from a TATA box. The downstream core promoter (DPE) element is only functional when placed precisely at +28 to +32 bp relative to the highlighted A of an Inr element. TFIIB binds to the BRE (TFIIB recognition element) and accurately positions RNA polymerase at the transcription start site. However, none of these elements is either necessary or sufficient for promoter activity, and many active polymerase II promoters lack all of them. N represents any nucleotide. (Adapted from Smale ST & Kadonga JT [2003] *Annu Rev Biochem* 72:449–479; PMID 12651739. With permission from Annual Reviews.)

Once the basal transcription apparatus has fully assembled, a component with DNA helicase activity is responsible for locally unwinding the DNA helix, and the activated RNA polymerase accesses the template strand.

Modulating transcription and tissue-specific regulation

As a metaphor for gene expression, imagine the output from a radio. The basal transcription apparatus described above would be the radio's ON switch that is needed to get started. It is required in all cells and uses ubiquitous transcription factors that bind to *cis*-acting elements in the core promoter. But there is also the need for a VOLUME control to amplify or reduce the signal, as required in different cell types or at different stages in a cell's life or development.

The role of the volume control is performed by additional circuits. First, additional, often non-ubiquitous, transcription factors bind to *cis*-acting regulatory elements other than those of the core promoter. These elements are sometimes distantly located from the gene they regulate, as described immediately below. Then there are co-activator or co-repressor proteins that are recruited by bound transcription factors. In addition, diverse types of long noncoding RNAs regulate transcription. Because they often work in the epigenetic regulation of transcription, we consider them in [Section 6.2](#).

Cis-acting regulators as modifiers of basal gene expression

As seen from [Figure 6.3](#) a promoter is made up of sequence elements whose orientation and spacing are important. Two types of additional *cis*-acting regulatory DNA elements modify the transcriptional output in a way that is independent of their orientation:

- **enhancers** amplify transcription;
- **silencers** repress transcription.

Enhancers and silencers may be located close to a transcriptional start site, from shortly upstream of the promoter (of the gene they regulate) to the beginning of its first intron. But quite often, too, they may be rather distantly located from the promoter ([Figure 6.4A](#)). To allow remote elements such as these to work, the intervening DNA needs to be looped out so that regulatory proteins bound to the enhancer can now physically interact with proteins bound to the promoter of the target gene (see [Figure 6.4B](#)).

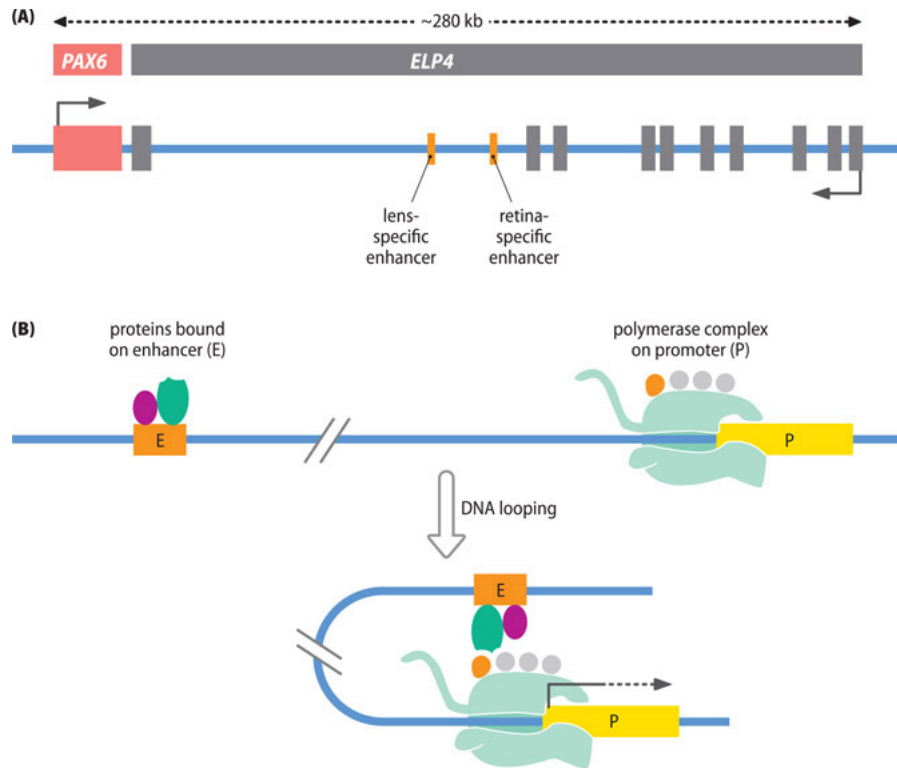


Figure 6.4 DNA looping is required to bring a distant enhancer in close proximity to the promoter of the gene it regulates. (A) Example of remote enhancers. The 33 kb *PAX6* gene (upper left) is mutated in aniridia type I (OMIM 106210) and is known to be regulated by two distantly located enhancers: a lens-specific and a retina-specific enhancer residing within a long intron of the large neighboring *ELP4* gene. Vertical gray boxes represent *ELP4* exons (for clarity, *PAX6* is represented here by a single box representing both exons and introns). Exons and enhancer elements are not to scale. (B) General enhancer-promoter interaction. Bending of the intervening DNA allows direct physical interactions between proteins bound to an enhancer (or other remote *cis*-acting element) with some of the many proteins bound to the promoter of the gene that it regulates. For clarity, only the RNA polymerase is shown at the promoter.

Cis-acting regulators as boundary elements

The long-distance action of elements such as enhancers needs to be targeted to the correct genes. To ensure that signals from regulatory elements do not affect genes other than the intended targets, **boundary elements** are needed to establish physical separation between euchromatin and constitutive heterochromatin, and also between different regions of euchromatin. Two important classes are:

- **barrier elements.** They maintain differences in chromatin structure between neighboring euchromatic and heterochromatic regions.
- **insulators.** They block inappropriate interactions between enhancers and promoters in a region of euchromatin.

We give examples in [Section 6.3](#) to illustrate the use of both types of boundary element.

Transcription factor binding and specificity

A protein **transcription factor** is a sequence-specific DNA-binding protein that binds to specific short target DNA sequences (often four to nine nucleotides long) within or close to genes that it regulates. In addition to ubiquitous general transcription factors (which bind to core promoter elements), many other transcription factors bind to target sequences within additional, often remote, *cis*-acting sequences, such as enhancers, as described below.

Some genes need to be expressed in all cells (“housekeeping” genes), but many are expressed only in specific tissues and/or at specific developmental stages (the activity of the promoters relies on tissue-specific or developmentally regulated transcription factors that bind to noncore elements). Like other DNA-binding proteins, a transcription factor typically recognizes its target sequence using a DNA-binding domain that contains some motifs that physically bind DNA, such as zinc fingers or leucine zippers ([Figure 6.5](#)).

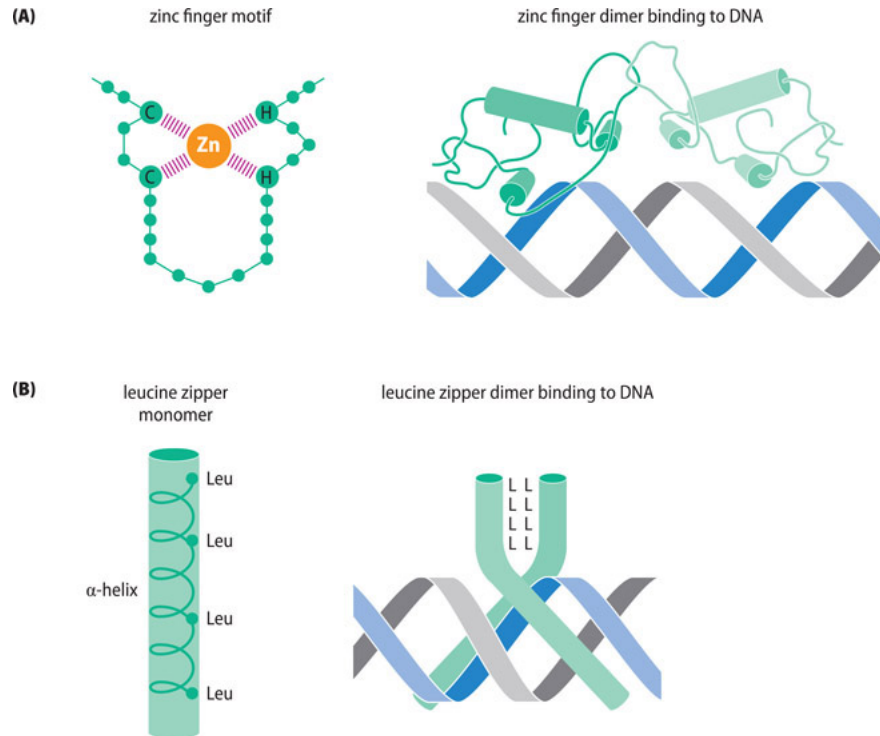


Figure 6.5 Examples of common DNA-binding motifs found in transcription factors and other DNA-binding proteins. (A) In the zinc finger motif a Zn^{2+} ion is bound by four conserved amino acids (normally either histidine or cysteine) to form a loop (finger). Clusters of sequential zinc fingers are common. The so-called C2H2 (Cys2/His2) zinc finger typically comprises ~23 amino acids and neighboring fingers are separated by a stretch of about seven or eight amino acids. The structure of a zinc finger may consist of an α -helix and a β -sheet (held together by coordination with the Zn^{2+} ion), or of two α -helices, as shown here. In either case, the primary contact with the DNA is made by an α -helix binding to the major groove. (B) The leucine zipper is a helical stretch of amino acids rich in hydrophobic leucine residues, aligned on one side of the helix. These hydrophobic patches allow two individual α -helical monomers to join together over a short distance to form a coiled coil. Beyond this region, the two α -helices separate, so that the overall dimer is a Y-shaped structure. The dimer is thought to grip the double helix much like a clothes peg grips a clothesline. Leucine zipper proteins normally form homodimers (but can occasionally form heterodimers).

Transcription factor specificity

Transcription factors recognize short target sequences, and so for each transcription factor there are from tens of thousands to hundreds of thousands of potential binding sites across the human genome. Only a tiny fraction of potential target sequences are used, however, for two reasons. First, the binding site must be accessible (in an open chromatin conformation and away from direct contact with nucleosomes). Secondly, transcription factor binding is *combinatorial*—different transcription factors work in concert by binding to adjacent recognition sequences.

A transcriptional activation domain is present in transcription factors that stimulate transcription; transcriptional repressors often recruit specialized protein complexes to silence gene expression, as described in [Section 6.2](#). Other proteins modulate transcription without binding to DNA. Instead, they work by protein–protein interactions

that support other regulatory proteins (which bind DNA directly). There are two types: transcriptional *co-activators* (which enhance transcription) and *co-repressors* (which downregulate transcription).

Genetic regulation during RNA processing: RNA splicing and RNA editing

Understanding the genetic control of splicing is important for understanding pathogenesis because mutations causing abnormal RNA splicing are a relatively common cause of disease. RNA editing is a less well understood form of RNA processing.

Regulation of RNA splicing

Like transcription, RNA splicing is subjected to different controls, and some splicing patterns are ubiquitous; others are tissue-specific. As illustrated in [Figure 6.6A](#), three fundamental *cis*-acting regulatory RNA sequences are required for the basic splicing mechanism, which is performed by large ribonucleoprotein complexes known as spliceosomes. The **splice donor site** contains an invariant GU dinucleotide that defines the 5' end of an intron at the RNA level. The **splice acceptor site** contains an invariant AG dinucleotide that defines the 3' end of an intron at the RNA level and is embedded within a larger sequence that includes a preceding polypyrimidine tract. An additional control element, the branch site, is located very close to the splice acceptor; it contains an invariant A nucleotide and is responsible for initiating the splicing reaction. Note that the sequence surrounding the invariant GU and AG signals is variable—some splice sites are strong and readily used, whereas others are weak and used only occasionally.

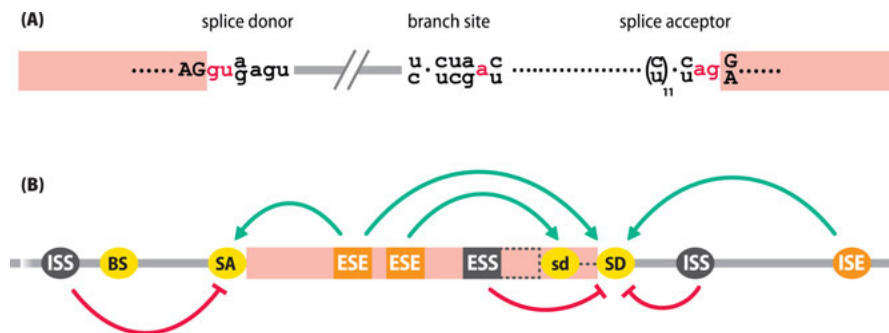


Figure 6.6 *Cis*-acting sequences that regulate RNA splicing. Pink boxes represent transcribed exon sequences. (A) The three fundamental RNA target sequences in the splicing mechanism. Bases in red are essentially invariant. The gap shown by the // symbol can vary in length from tens of nucleotides up to several hundred kilobases long in extreme cases. Spliceosomes contain several types of small nuclear RNA (snRNA), including U1 snRNA, which base pairs with the splice donor sequence, and U2 snRNA, which base pairs with the branch site sequence. (B) In addition to the splice donor (SD), splice acceptor (SA), and branch site (BS) sequences, other regulatory RNA sequences stimulate splicing (orange) or inhibit splicing (black). In this example, an exon has two exonic splice enhancers (ESE), and an exonic splice suppressor (ESS) and is flanked by introns that have intronic splice suppressors (ISS). The dotted black lines indicate an alternative 3' end to the exon, due to the use of an alternative splice donor (sd) sequence instead of the usual splice donor sequence (SD).

Splicing is also regulated by two additional classes of short (often hexanucleotide) *cis*-acting regulatory RNA elements:

- *splice enhancer* sequences (which stimulate splicing);
- *splice suppressor* sequences (which inhibit splicing).

These sequences are located close to splice junctions and can lie within exons or introns ([Figure 6.6B](#)). To help keep the spliceosome in place, splicing enhancers bind SR proteins (so called because they have a domain based on repeats of the serine-arginine dipeptide). Splicing suppressors bind hnRNP proteins that are active in removing

bound spliceosomes. Because different tissues and cell types can express different SR proteins and different hnRNP proteins, splicing patterns can vary between tissues.

Alternative splicing

More than 90 % of human protein-coding genes undergo some kind of alternative splicing, when primary transcripts of a single gene are spliced in different ways (Figure 6.7 gives some variations, and Figure 6.6B shows how they can be generated). Sometimes, some transcripts retain transcribed intronic sequence. **Exon skipping** occurs when one or more full-length exons are not represented in some transcripts. In other cases, there is some variability in the precise locations of exon–intron junctions, so that transcripts from one gene may have short or long versions of an exon.

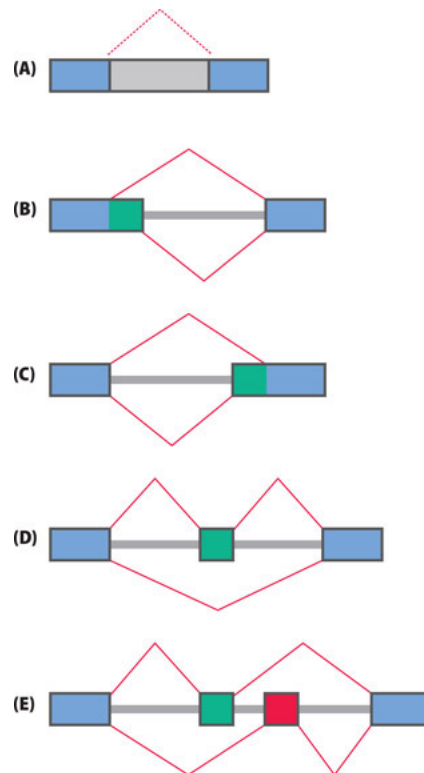


Figure 6.7 Types of alternative splicing event. (A) An intronic sequence (gray) is either excluded from a transcript or retained. (B,C) The use of alternative splice donor sites (B) or of alternative splice acceptor sites (C) results in the inclusion or exclusion of the sequences in green. (D) The exon in green may be either included or skipped (a *cassette exon*). (E) Alternative exons: the mature mRNA includes either the exon in green or the exon in red, but not both or neither. Blue boxes represent exons that are always included in the mature mRNA.

Some of the variable transcripts may be functionally unimportant (splicing accidents must happen occasionally). Often, however, alternative splicing patterns show tissue specificity (so that, for example, one splice pattern is consistently produced in brain but another pattern is normally found in liver), or there may be consistent differences in the use of specific splice patterns at different stages in development. By producing alternative products (**isoforms**) from individual genes, alternative splicing can increase functional variation.

Alternative isoforms may be retained in cells, or secreted, or sent to different cellular compartments (to interact with different molecules and perform different roles). For example, the –KTS isoforms of the WT1 Wilms tumor protein (Figure 6.8A) function as DNA-binding transcription factors, but the +KTS isoforms associate with pre-mRNA and may have a general role in RNA splicing. This pattern of alternative splicing has been conserved over hundreds of millions of years. The ERBB4 protein, a tyrosine protein kinase that is a member of the epidermal

growth factor receptor family, has CYT1 and CYT2 isoforms that respectively possess or lack a binding site for the phosphatidylinositol-3-kinase signaling molecule ([Figure 6.8A](#)).

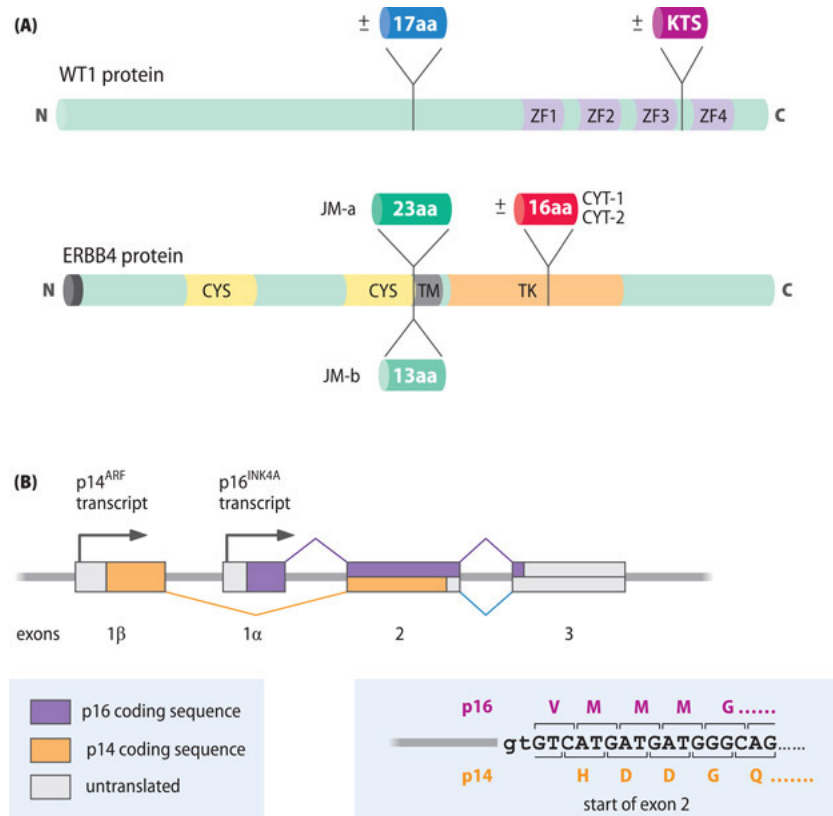


Figure 6.8 Examples of alternative splicing in human genes. (A) Alternative splicing results in the variable presence of a 17 amino acid (17aa) peptide sequence near the middle of the WT1 Wilms tumor protein, and of a Lys-Thr-Ser tripeptide sequence (KTS) between the third and fourth zinc finger (ZF) domains. (B) Four different isoforms also exist for the human ERBB4 protein. Just before the transmembrane (TM) domain there is the alternative presence of a 23aa or a 13aa peptide sequence (JM-a and JM-b isoforms, respectively). Within the tyrosine kinase (TK) domain a 16-amino-acid peptide sequence with a binding site for phosphatidylinositol-3-kinase is present in the CYT-1 isoform but absent in the CYT-2 isoform. (C) Alternative splicing of the *CDKN2A* gene produces two tumor suppressor proteins, p16-INK4A and p14-ARF, that both work in cell cycle control but with entirely different amino acid sequences. Exon 2, the only exon that has a coding sequence for both proteins, is translated in different reading frames (bottom right).

Very occasionally, quite different proteins are created from a common gene by alternative splicing. The *CDKN2A* gene provides a prime example by producing two entirely different proteins that, nevertheless, have similar functions (see [Figure 6.8B](#)).

RNA editing

In some RNA transcripts, certain nucleotides naturally undergo deamination or transamination. When this happens in the coding sequences of mRNAs, the amino acid sequence of the protein will differ from that predicted by the genomic DNA sequence. For example, certain adenines in some RNA transcripts are naturally deaminated to give the base inosine (I), which behaves like guanine (by base pairing with cytosine). In coding sequences, A \rightarrow I editing is most commonly directed at CAG codons, which specify glutamine (Q). The resulting CIG codons behave like CGG and code for arginine (R), and so this type of RNA editing is therefore also called Q/R editing.

Q/R editing is quite commonly found during the maturation of mRNAs that make neurotransmitter receptors or ion channels.

Some other types of RNA editing are known, including C → U editing (used in making apolipoprotein B mRNA, for example) and U → C editing (used in making mRNA from the *WT1* Wilms tumor gene). The extent of RNA editing is still controversial, and its significance is unclear.

Translational regulation by *trans*-acting regulatory proteins

Regulation at the level of translation allows cells to respond more rapidly to altered environmental stimuli than is possible by altering transcription. According to need, stores of inactive mRNA may be held in reserve so that they can be translated at the optimal time. Controls are also exerted over where an mRNA is translated: some mRNAs are transported as ribonucleo-protein particles to specific locations within a cell; for example, alternative splicing allows tau mRNA, to be selectively localized to the proximal regions of axons rather than to dendrites.

To control gene expression at the level of mRNA, *trans-acting* regulatory factors bind to specific *cis*-acting RNA elements in the untranslated regions of the mRNA. Single-stranded RNA is quite flexible (unlike DNA, which has a rather rigid structure), but typically it has a very high degree of secondary structure as a result of intra-chain hydrogen bonding (shown in Figures 2.4 and 2.6). RNA elements that bind protein are often structured as hairpins, as in the example of the iron-response element shown in [Figure 6.9A](#).

As an example of translational regulation, consider how cells control the availability of two proteins involved in iron metabolism: ferritin (an iron-binding protein used to store iron in cells) and the transferrin receptor (which helps us absorb iron from the diet). When iron levels are low, the priority is to maximize the amount of iron that can be absorbed from the diet: transferrin receptor mRNA is protected from degradation so that it can make a protein product. Conversely, when iron levels are high, ferritin production is activated to store iron in cells. This happens without any change in the production of ferritin or transferrin receptor mRNAs. Instead, both these mRNAs have iron-response elements (IREs), which can be bound by a specific IRE-binding protein that regulates the production of protein from these mRNAs; the availability of this binding protein is also regulated by iron concentrations (see [Figure 6.9B](#)).

Post-transcriptional gene silencing by microRNAs

Trans-acting regulators such as the IRE-binding protein that work by binding to mRNA used to be viewed as quirky exceptions. The discovery of tiny RNA regulators, notably microRNA, changed all that. MicroRNAs (miRNAs) are single-stranded regulatory RNAs that downregulate the expression of target genes by base pairing to complementary sequences in their transcripts. Typically about 20–22 nucleotides long, they are formed by multiple processing events, including cleavages in the cytoplasm that are performed by the same endoribonucleases used in RNA interference, a natural cell defense mechanism that we detail in [Section 9.4](#).

A miRNA binds to any transcript that has a suitably long complementary sequence to form a stable heteroduplex (correct base pairing is important for the “seed” sequence covering the first eight or so nucleotides from the 5′ end of the miRNA; some mismatches are tolerated when the remaining part of the miRNA pairs up). Because miRNAs are short and some base mismatches are tolerated, a single miRNA can regulate transcripts from many different genes ([Figure 6.10](#)).

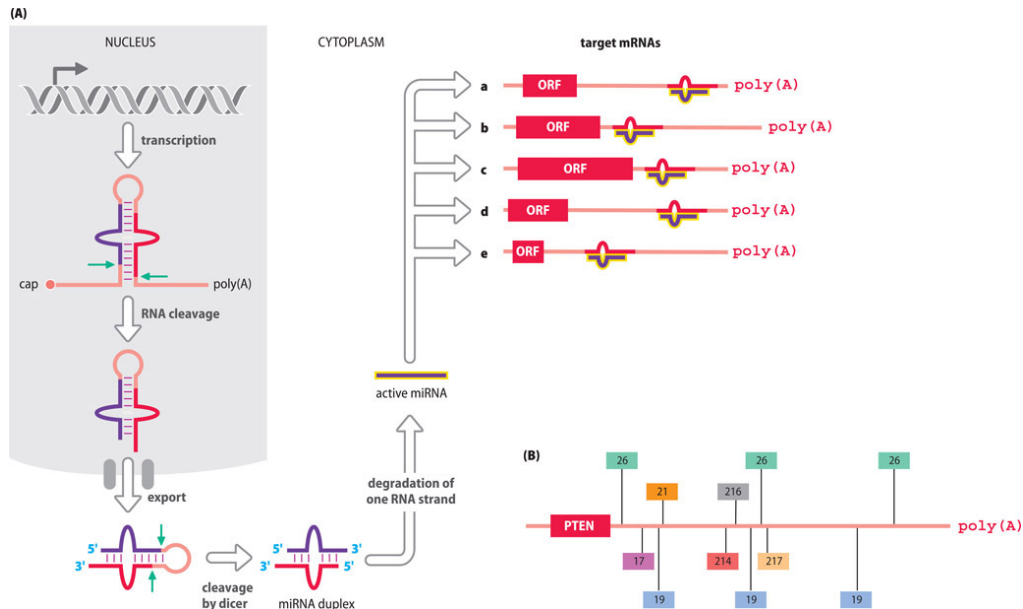


Figure 6.10 How microRNAs are produced and work in cells. (A) miRNA genes are transcribed and cleaved in the nucleus to generate a stem-loop RNA that is exported to the cytoplasm and further cleaved asymmetrically by the endoribonuclease dicer to generate a miRNA duplex with overhanging 3' ends. One strand of the duplex (the *passenger strand*, shown in red) is then cleaved and degraded, leaving the other strand (the *guide strand*, shown in purple) as a mature single-stranded miRNA. A typical human miRNA binds to and regulates transcripts produced by hundreds of different genes, and the vast majority of miRNA–target RNA heteroduplexes have imperfect base pairing. Shown here for illustration are five mRNAs produced from five different genes (a–e); the complementary sequence to which the miRNA binds is shown in red. ORF, open reading frame (= coding DNA). (B) An individual mRNA often has multiple miRNA-binding sites. The example here is the mRNA from the human *PTEN* tumor suppressor gene that has binding sites in the 3' untranslated region for miRNAs belonging to seven miRNA families: three binding sites each for miR-19 and miR-26, and one each for miR-17, miR-21, miR-214, miR-216, and miR-217.

We have several hundred miRNA genes, and they frequently show tissue-specific expression; many are important in early development, but miRNAs have been found to be important regulators in a whole range of different cellular and tissue functions. At least 50 % of our protein-coded genes are thought to be regulated by miRNAs, and individual types of mRNA often have recognition sequences for multiple miRNA regulators. Just like protein transcription factors, miRNAs seem to be involved in complex regulatory networks, and they are subject to negative regulation by a wide range of RNA classes as described in the next section.

Repressing the repressors: competing endogenous RNAs sequester miRNA

Many of our pseudogenes are known to be transcribed. Some of them seem to have undergone purifying selection, indicating that they are functionally important. A landmark study published in 2010 provided the first real insights into how functional pseudogenes work: it showed that the human *PTEN* gene at 10q23 is regulated by a highly related processed pseudogene, *PTENP1*, located at 9p21.

PTEN makes a protein tyrosine phosphatase that is very tightly controlled (cells are very sensitive to even subtle decreases in abundance of this protein, and aberrant *PTEN* expression is common in cancers). *PTENP1* does not make a protein (one of the changes from the *PTEN* sequence disrupts the initiator methionine codon). It does, however, make a noncoding RNA that retains many of the miRNA-binding sites in the 3' UTR of *PTEN* mRNA. The *PTENP1* RNA seems to regulate *PTEN* expression by binding to and sequestering miRNAs that would normally bind to the *PTEN* mRNA (Figure 6.11 gives the principle).

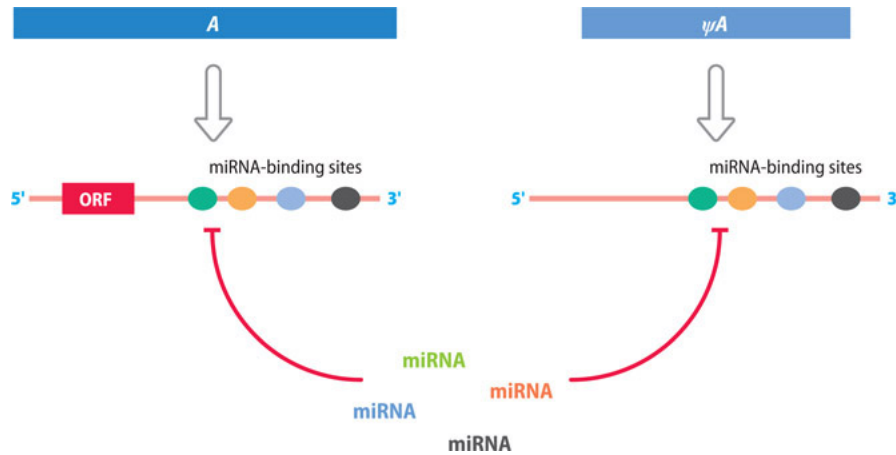


Figure 6.11 Different types of competing endogenous RNAs can act as miRNA sponges. In this example, a protein-coding gene A and a closely related pseudogene ψA produce RNA transcripts that have in common binding sites for certain miRNA classes. The pseudogene RNA can compete with the mRNA for binding by the same miRNA classes (by soaking up miRNA the ψA RNA acts as a “miRNA sponge”). Various other RNA classes can also act as miRNA sponges, including certain long noncoding RNAs and also *circular RNAs*. The latter are surprisingly abundant in our cells and form by head-to-tail splicing reactions (“back-splicing”) that join first and last exon sequences; since they largely overlap protein-coding sequences they can contain sequences corresponding to the untranslated sequences of mRNAs, including miRNA-binding sites.

6.2 CHROMATIN MODIFICATION AND EPIGENETIC FACTORS IN GENE REGULATION

During development, cell differentiation (and the production of different cell lineages) is dictated by programmes of altered gene expression that are independent of the DNA sequence. Instead, they depend on altered epigenetic settings (often called **epigenetic marks**) that affect chromatin structure (and thereby gene expression). In its broadest sense, **epigenetics** covers all phenomena that can produce heritable changes in how genomes function without affecting the base pairing properties of the DNA sequence.

An overview of the molecular basis of epigenetic mechanisms

Later in this section we provide detail on some individual classes of epigenetic mechanisms. First, we briefly outline the characteristics of five important classes. Three comparatively well understood classes involve certain types of chemical modification of DNA or of histones bound to DNA, plus substitution of standard histones by histone sequence variants. Nucleosome positioning and *cis*-acting regulatory non-coding RNAs are also important (see [Table 6.1](#)).

TABLE 6.1 FIVE IMPORTANT CLASSES OF EPIGENETIC MECHANISMS AFFECTING CHROMATIN STRUCTURE

Epigenetic mechanism	Comments
DNA methylation	Specifically, methylation of cytosines within CpG dinucleotides to give 5-methylcytosine (which base pairs like cytosine). The palindromic nature of CpG provides a simple way of transferring this epigenetic mark to daughter DNA strands (Figure 6.15). Chromatin with highly methylated DNA is condensed, and transcriptionally inactive but hypomethylation is associated with an open chromatin structure (Figure 6.12).
Histone modification	Post-translation chemical modification of side chains occurs at multiple relatively accessible amino acids on the C-terminus tails of histones. Three common types of modification are:

* The term *chromosome remodeling* encompasses repositioning of nucleosomes and histone substitution.

Epigenetic mechanism	Comments
	acetylation (at certain lysines); methylation (at certain arginines and lysines); and phosphorylation (mostly directed at certain serines and threonines)—(Table 6.3).
Histone substitution*	The replacement of standard histones by certain other histone sequence variants (Table 6.4).
Nucleosome repositioning*	Chromatin modeling complexes are large ATP-powered multiproteins that physically drive nucleosomes along the DNA to create areas of low or high nucleosome density. They help set up active or repressed chromatin states (an active transcription site typically has ~ 150 bp of nucleosome-free DNA, and highly ordered flanking nucleosomes).
Cis-regulation by noncoding RNA	These regulatory RNAs remain attached to the DNA they were transcribed from. They act either as antisense RNAs or as scaffolds for binding regulatory protein complexes to change chromatin structure. See below for important examples in imprinting and X-inactivation.

* The term *chromosome remodeling* encompasses repositioning of nucleosomes and histone substitution.

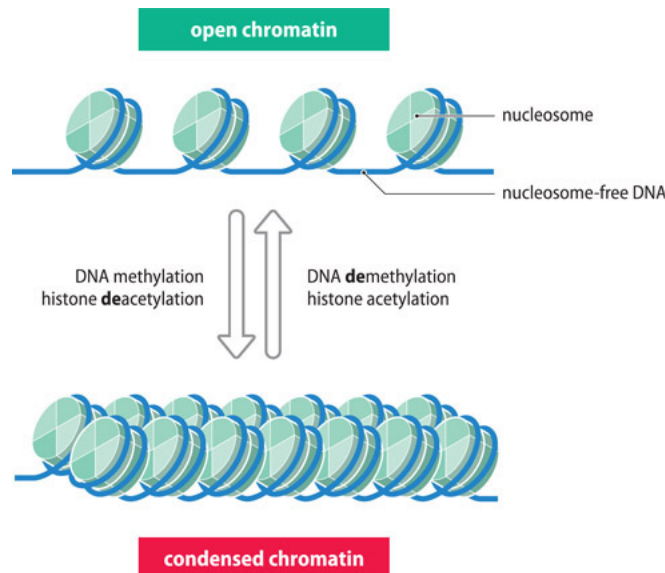


Figure 6.12 Altered chromatin states arise from DNA and chromatin modifications. In chromatin that has an open configuration the nucleosome-free stretches of DNA may include promoter elements and other regulatory sequences to which regulatory factors can bind, allowing gene expression. But in highly condensed (compacted) chromatin the transcription factors are denied access by the tight packing. Note that whereas DNA methylation (and histone deacetylation) is associated with condensed chromatin, different types of histone methylation are associated with open and condensed chromatin (see Table 6.3).

TABLE 6.3 EXAMPLES OF HISTONE MODIFICATIONS CHARACTERISTIC OF DIFFERENT CHROMATIN STATES*

AMINOACID	EUCHROMATIN				HETEROCHROMATIN		
	Promoters		Enhancers		Gene bodies	Facultative	Constitutive
	Active	Inactive	Active	Inactive	Inactive		
H3K4	H3K4me2/me3		H3K4me1/me2				
H3K9	H3K9ac	H3K9me3		H3K9me2/me3	H3K9me2/me3	H3K9me2	H3K9me3
H3K27	H3K27ac	H3K27me3	H3K27ac			H3K27me3	

* The nomenclature for histone modifications gives first the histone class, then the amino acid in one-letter code, followed by the position of the amino acid (counting from the N-terminus), and finally the chemical modification. So, for example, H3K9Ac represents acetylation of the lysine at the 9th amino acid counting from the N-terminus of histone H3.

AMINOACID	EUCHROMATIN				HETEROCHROM		
	Promoters		Enhancers		Gene bodies	Facultative	Consti
	Active	Inactive	Active	Inactive	Inactive		
H4K12	H4K12ac					H4K12ac	H4K12
H4K20							H4K20

* The nomenclature for histone modifications gives first the histone class, then the amino acid in one-letter code, followed by the position of the amino acid (counting from the N-terminus), and finally the chemical modification. So, for example, H3K9Ac represents acetylation of the lysine at the 9th amino acid counting from the N-terminus of histone H3.

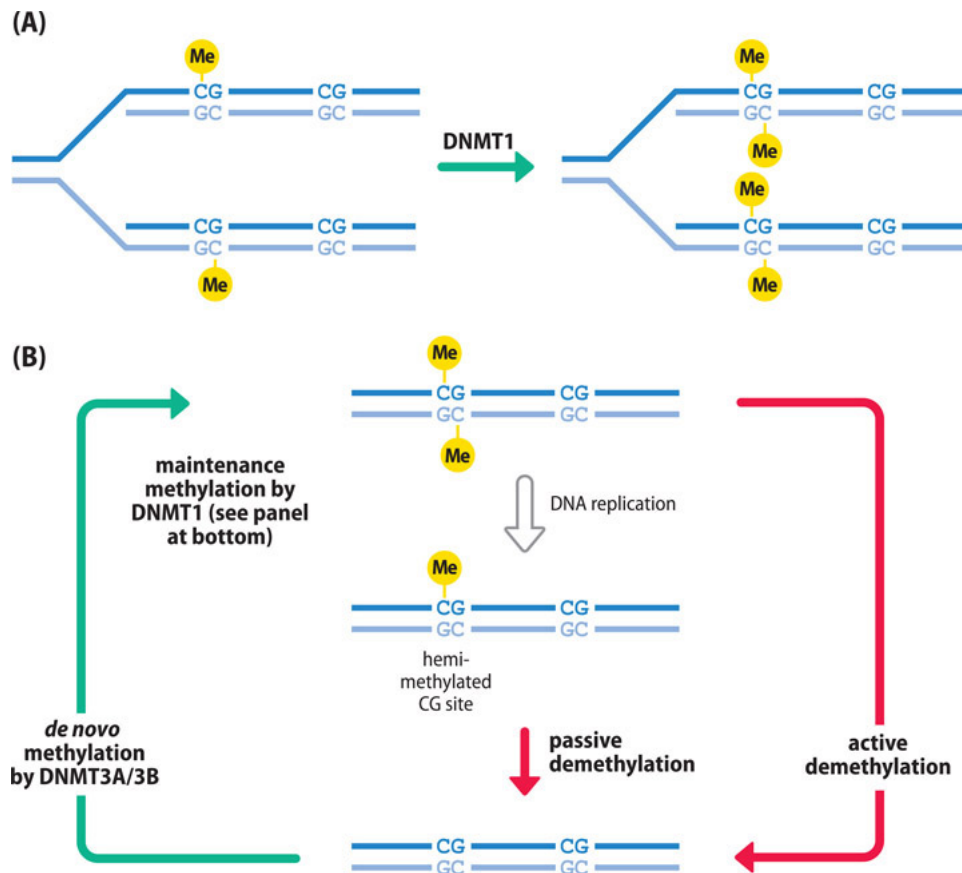


Figure 6.15 DNA methylation and demethylation mechanisms in mammalian cells. (A) Maintenance methylation. During replication of a DNA molecule containing methylated CG dinucleotides, the parental strand retains methylated cytosines, but the newly synthesized DNA incorporates unmethylated cytosines. DNMT1 is usually available, and it specifically methylates any CG dinucleotides on the newly synthesized strand that are paired with a methylated CG on the parental strand, regenerating the original methylation pattern. (B) Pathways towards DNA methylation (green arrows) and demethylation (red arrows). If DNMT1 is not available, the hemimethylated DNA can give rise in a subsequent DNA replication to unmethylated DNA (passive demethylation). Unmethylated DNA can also be generated by an active demethylation process at certain stages in development. DNMT3A and DNMT3B are used for *de novo* methylation at specific developmental stages (see [Figure 6.16](#)).

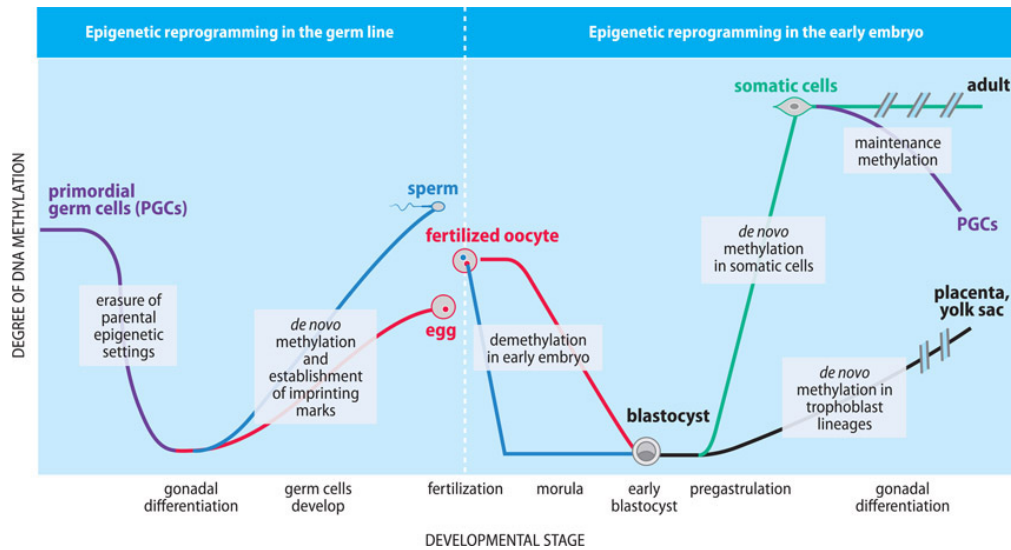


Figure 6.16 Changes in DNA methylation during mammalian development. Marked and often tissue-specific changes in overall methylation accompany gametogenesis and early embryonic development. Their causal role remains uncertain, although mice that are specifically unable to methylate sperm DNA are infertile. The horizontal time axis is necessarily abbreviated on the right-hand side of the figure leading toward birth and then adulthood (indicated by the use of double slashes) PGCs, primordial germ cells.

TABLE 6.4 EXAMPLES OF HISTONE H2A AND HISTONE H3 VARIANTS

Class	Variant	Description
H2A	H2AX	important in DNA repair and recombination (it is introduced at sites of double-strand breaks)
	H2A.Z	associated with the promoters of active genes. It also helps prevent the spread of silent heterochromatin and is important in maintaining genome stability
H3	H3.3	important in transcriptional activation
	CENP-A	a centromere-specific variant of H3. It is required for assembly of the kinetochore, to which spindle fibers attach

Heritability of epigenetic marks

Epigenetic marks can be stably transmitted from one cell generation to the next, providing a form of cellular memory. For example, once a cell has been epigenetically programmed to become an intestinal epithelial cell, daughter cells retain this programming so that they, too, are intestinal epithelial cells. And in addition to regulating how genes are expressed, epigenetic settings determine how some DNA sequences determine chromosome functions. DNA sequences at centromeres and telomeres, for example, have heritable epigenetic settings, and these sequences will continue to dictate centromere and telomere functions in the daughter cells. The patterns of constitutive heterochromatin in a parent cell are also reproduced in daughter cells.

There are some instances in nature, notably in plants, in which epigenetic effects can be transmitted through meiosis, from one organism to subsequent generations. In mammals, however, major waves of epigenetic reprogramming occur in gametogenesis and in the early embryo to remove parental DNA methylation marks and reset the global DNA methylation patterns (see [Table 6.2](#)). As a result, epigenetic marks are not normally transmitted across generations from parents to children (but we describe later some limited evidence in [Chapter 8](#)).

TABLE 6.2 EXAMPLES OF EPIGENETIC PHENOMENA INVOLVING DNA AND CHROMATIN MODIFICATION IN MAMMALIAN CELLS

Phenomenon	Mechanism/comments
------------	--------------------

Phenomenon	Mechanism/comments
Epigenetic reprogramming in gametogenesis	Readily detected as a wave of genome wide demethylation during germ cell development (erasing parental epigenetic marks) followed by comprehensive <i>de novo</i> DNA methylation to reset global patterns of DNA methylation and gene expression.
Epigenetic reprogramming in the early embryo	Eggs and sperm are differentiated cells, and their genomes have different epigenetic marks. They combine to give a zygote whose genome is gradually reprogrammed to erase the great majority of the inherited epigenetic marks. By the blastocyst stage, the cells of the inner cell mass are pluri potent and will give rise ultimately to all cells of the body. Epigenetic marks are reestablished in the descendants of the cells of the inner cell mass to establish different cell lineages and permit cell differentiation.
Establishment of centromeric heterochromatin	Centromere establishment relies on nucleosomes incorporating a specific histone H3 variant known as CENP-A.
X-chromosome inactivation	Initiated by the <i>XIST</i> long noncoding RNA that somehow coats most of one of the two X chromosomes in female cells, silencing most of its genes.
Genomic imprinting	Silencing of one allele, according to parent of origin, at diverse gene loci (often organized in gene clusters) on different chromosomes.
Position effects causing heterochromatinization	Large-scale changes in DNA, causing genes to be relocated to a region of constitutive heterochromatin where they are silenced.

Stability of epigenetic marks

Although epigenetic marks are often stable, they can be changed. A naturally occurring example is evident during germ cell development and in the very early embryo where epigenetic marks in mammalian genomes are programmed to be reset across the genome between generations. (And, of course, they can be reset artificially to clone animals, as in the case of the famous sheep Dolly, and in cultured cells to create induced pluripotent stem cells, for example.)

Some epigenetic marks can also be reset naturally in response to environmental conditions. Cells receive a wide range of extracellular chemical signals, notably from neighboring cells but also from chemicals in food that we ingest.

How changes in chromatin structure produce altered gene expression

Binding of DNA to a histone octamer and bending of the DNA on the surface of the histone octamer to form regular nucleosomes make it very difficult for regulatory factors such as transcription factors to bind to their target sequences. Depending on its chromatin environment, the properties of a DNA sequence can change. A functional gene when embedded in highly condensed chromatin may not be accessible to transcription factors and would be *silenced*. But if the chromatin structure is altered, adopting a more open, relaxed conformation, protein factors may be able to bind the promoter and related regulatory sequences to initiate transcription.

Sometimes a normally expressed gene is transposed (by a translocation or inversion) so that it takes up a new position within, or close to, a region of constitutive heterochromatin (permanently condensed heterochromatin). When that happens the gene would be silenced (an example of a **position effect**). For mammalian cells, the most striking evidence that gene expression is dependent not only on DNA sequence but also on chromatin structure comes from the X chromosome. In females one X chromosome is very highly condensed across nearly all its length and genes are silenced across most of the chromosome, unlike in the other X chromosome which has a comparatively open structure.

DNA methylation and chemical modification of histones are important regulators of chromatin structure. DNA methylation involves adding methyl groups to a small percentage of the cytosines and demethylation of DNA involves removing methyl groups from some of the methylated cytosines. (Note: because methylated cytosines behave like cytosine and base pair with guanines, the base sequence is not considered to be altered.) Extensive methylation of DNA sequences in vertebrates is generally characteristic of tightly packed chromatin; loosely structured chromatin (“open” chromatin) has low-level DNA methylation ([Figure 6.12](#)).

Histone modifications include different types of post-translational modification at specific amino acid positions on the different types of histone. Histone acetylation, for example, is associated with open chromatin, and histone deacetylation with condensed chromatin (see [Figure 6.12](#)).

The need for chromatin writers, erasers, and readers

Many different enzymes are responsible for creating or interpreting epigenetic marks, and belong to three classes as follows:

- “*writers*” add chemical groups to modify DNA or histones covalently; in the latter case, different enzymes are employed according to the chemical group deposited, and also according to the numerical position of the amino acid in the histone tail
- “*erasers*” work in the opposite direction to remove the chemical groups
- “*readers*” are involved in binding to specific chemical groups on DNA or histones to interpret defined epigenetic marks.

The readers may recruit additional factors to induce different changes in chromatin, such as chromosome compaction, or changes in nucleosome spacing and structure (**chromatin remodeling**). By adjusting the position of nucleosomes with respect to the DNA strand, promoters and other regulatory DNA sequences can become nucleosome-free, allowing access by transcription factors.

Histone modification and histone substitution in nucleosomes

Nucleosomes have 146 bp of DNA wrapped around a core of eight histone proteins, composed of two each of four different histone classes: H2A, H2B, H3, and H4. The histone proteins are positively charged (having multiple lysine and arginine residues) and have protruding N-terminal tails. Although the histone tails in [Figure 6.13A](#) are shown in isolation, they can make contact with adjacent nucleosomes.

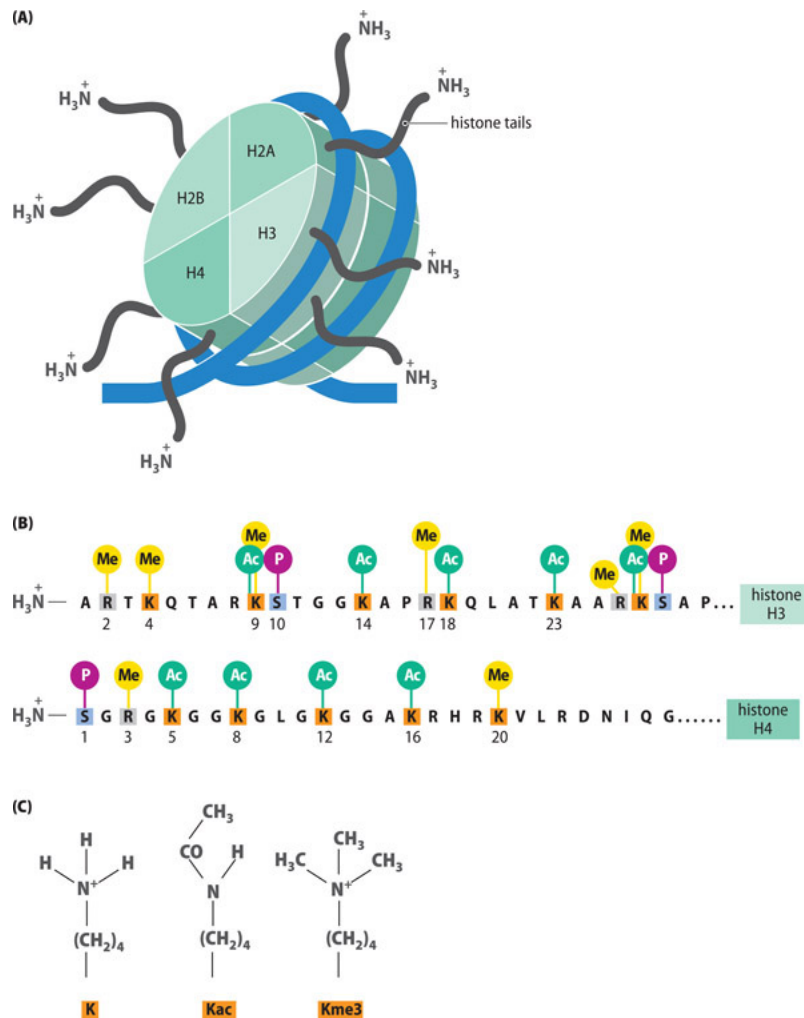


Figure 6.13 Nucleosome structure and histone modifications. (A) Positively charged N-terminal histone tails protrude from nucleosomes and can associate with other nucleosomes (not shown). (B) Map of histone H3 and H4 tail modifications. Note that lysine at position 9 in H3 can be methylated (H3K9me) or acetylated (H3K9ac), but never both. Some lysine residues may have two or three methyl groups (see Table 6.3). (C) Examples of lysine modification, showing the standard side chain of lysine (K), an acetylated lysine (Kac), and a trimethylated lysine (Kme3).

Each N-terminal histone tail shows a pattern of variable chemical modifications at specific amino acid positions. Individual amino acids in each tail may be methylated, acetylated, or phosphorylated (see Figure 6.13B), or subject to yet other types of modification, including tagging with ubiquitin. Particular types of amino acid are preferred targets for modifying the N-terminal histone tails: acetylation occurs only at lysine residues, phosphorylation mostly occurs at serines, but both lysine and arginine residues can be methylated.

Acetylation of lysines leads to loss of the positive charge (Figure 6.13C), and as a result acetylated histone tails interact less well with neighboring nucleosomes than do the unacetylated histone tails. Histone acetylation therefore results in a more relaxed chromatin conformation. According to the specific amino acid position, lysines may also be modified to contain one, two, or three methyl groups (but in these cases the positive charge is retained on the side chain—see Figure 6.13C).

Histone modifications are performed by a series of different enzymes that are devoted to adding or removing a chemical group at specific amino acid positions. Thus, for example, there are multiple histone acetyl transferases (HATs) and histone deacetylases (HDACs), and suites of histone lysine methyltransferases (KMTs) and histone lysine demethylases (KDMs).

Histone substitution

Nucleosome structure is also altered by histone substitution: standard histones in nucleosomes are substituted by minor *histone variants* that recruit regulatory factors. As described below, this can have different effects, such as activating transcription or defining centromeres. Although much of our knowledge of chromatin modification has come from studying patterns of DNA methylation and histone modification or substitution, numerous nonhistone proteins and noncoding RNAs also have important roles in modifying chromatin structure.

Modified histones and histone variants affect chromatin structure

Histone modifications are “read” by nonhistone proteins that recognize and bind the modified amino acids and then recruit other proteins to effect a change in chromatin structure. Proteins with a bromodomain recognize the acetylated lysines of nucleosomal histones, those with a chromodomain recognize methylated lysines, and different varieties of each domain can recognize specific lysine residues. Chromatin-binding proteins often have several domains that recognize histone modifications.

Certain individual types of histone modification are associated with open chromatin and transcriptional activation, or with condensed chromatin and transcriptional repression. For example, methylation of H3K4 (the lysine at position 4 on histone H3) is associated with open chromatin at the promoters of actively transcribed genes and at active enhancers (**Table 6.3**). By contrast, trimethylation of the lysine at position 9 on histone H3 (H3K9me3) is prominently associated with transcriptional repression, being widely found in constitutive heterochromatin and in inactive genes in euchromatin (see **Table 6.3**).

In addition to histone modification, core histone proteins can be replaced by minor variants, notably of histone classes 2A and 3 (the variants typically differ from the canonical histone by just a few amino acids). The minor histone variants are synthesized throughout interphase and are often inserted into previously formed chromatin by a histone exchange reaction powered by a chromatin remodeling complex. Once inserted, they recruit specific binding proteins to effect some change in the chromatin status for specific functions. A well-studied example is CENP-A, a centromere-specific histone H3 variant that is responsible for assembling kinetochores at centromeres. **Table 6.4** gives other examples.

Modified histones and histone variants typically work together with DNA methylation and demethylation in regulating gene expression (**Figure 6.14**). H3K9me3 can bind heterochromatin protein 1, which in turn recruits DNA methyltransferases that also serve to repress transcription. In turn, DNA methyltransferases and 5-meCG-binding proteins recruit histone deacetylases and appropriate histone methyltransferases to reinforce transcriptional repression.

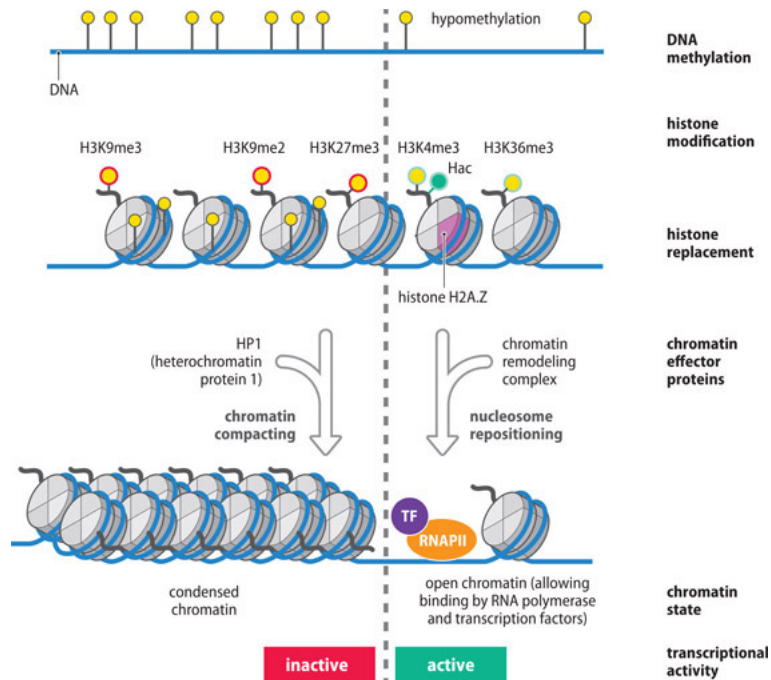


Figure 6.14 Contributions made by DNA and histone modification to different chromatin states in neighboring regions on a chromosome. For convenience, each nucleosome is shown with only one out of the eight N-terminal histone tails. Methylation modifications are shown by circles filled in yellow, comprising small circles (DNA methylation) and large circles (histone methylation). Note that some histone methylations (red outer lines) repress transcription, but other types (green outer lines) are associated with transcription. The filled green circle labeled Hac denotes histone acetylation (which applies to multiple lysines on H3 and H4). Chromatin effector proteins such as heterochromatin protein 1 (HP1)—or the repressive Polycomb group protein complexes PRC2 and PRC1 (not shown here)—are recruited to bind to specific histone modifications, often through the involvement of long noncoding RNAs. Chromatin remodeling can involve creating nucleosome-free regions of DNA that allows the binding of transcription factors. RNAPII, RNA polymerase II.

The function of DNA methylation in mammalian cells

Histones undergo very many different types of chemical modification, as do the nucleotide bases of RNA. However, only one major type of chemical modification is normally seen in DNA: methylation of certain cytosines, a chemical modification that is essential for mammalian development.

The principal function of DNA methylation is to regulate gene expression—it stabilizes, or locks in, patterns of gene silencing so that transcription is suppressed in highly methylated regions of chromatin. Highly repetitive DNA sequences, such as satellite repeats in pericentromeric heterochromatin and dispersed transposons, are extensively methylated. As detailed below, there is also significant—though more sporadic—methylation in the main body of genes (exons and introns) and in intergenic regions.

In the case of DNA methylation patterns, it is the extent of DNA methylation in the key *cis*-acting regulatory elements that distinguishes actively transcribing genes from silenced genes. Thus, the promoters and enhancers of actively transcribing genes are relatively free of DNA methylation. Along with characteristic histone modifications and histone variants such as H2A.Z and H3.3, such hypomethylated regions signal a locally open chromatin environment—transcription factors can gain access to and bind to their target sequences to stimulate transcription. Gene silencing is achieved when significant levels of DNA methylation in a gene’s promoter and enhancer elements cause the chromatin to be condensed, denying access to transcription factors.

Like other epigenetic marks, global DNA methylation patterns vary between different cell types and different stages in development. In addition to a general role in gene expression silencing, DNA methylation has important

roles in genomic imprinting and X-chromosome inactivation—we consider these epi-genetic phenomena in more detail below.

DNA methylation is also important for suppressing retrotransposon elements. Retrotransposons are evolutionarily advantageous to genomes because they can insert varied DNA sequences at different locations in the genome, potentially providing novel exon combinations in genes (see [Figure 2.16](#)) and giving birth to new regulatory sequences and new exons. About 43 % of the human genome is made up of retrotransposon repeats, but the number of actively transposing retrotransposons needs to be carefully regulated to prevent the genome from being overwhelmed. By suppressing retrotransposon transcription, DNA methylation acts as a necessary brake on excessive transposon proliferation.

DNA methylation: mechanisms, heritability, and global roles during early development and gametogenesis

In mammalian cells, DNA methylation involves adding a methyl group to certain cytosine residues, forming 5-methylcytosine (5-meC). The cytosines that are methylated occur within the context of a palindromic sequence, the CG dinucleotide (also called a CpG dinucleotide, where p represents phosphate).

The 5-meC base pairs normally with guanine (the methyl group is located on the outside of the DNA double helix, minimizing any effect on base pairing). It is recognized by specific 5-meCG-binding proteins that regulate chromatin structure and gene expression, as described below. In a somatic cell, about 70–80 % of CG dinucleotides will have a methylated cytosine, but the pattern of methylation is variable across the genome and across genes ([Box 6.1](#)).

BOX 6.1 CpG ISLANDS AND PATTERNS OF DNA METHYLATION ACROSS THE GENOME AND ACROSS GENES

DNA methylation is generally used to “lock in” transcriptional inactivity in regions of our cells that do not require expression. Accordingly, heterochromatin and intergenic regions are subject to high levels of DNA methylation. Hypermethylation of some regions, such as pericentromeric heterochromatin, is important for genome stability; a significant decrease in methylation levels in these regions can lead to mitotic recombination and genome instability. Our genes are also subject to DNA methylation; however, by comparison with heterochromatin and intergenic regions, the DNA methylation is generally reduced and more variable.

As described in the text, DNA methylation in our cells is limited to cytosines and occurs within the context of the dinucleotide CG (because CG is the target for cytosine methylation). The resulting 5-methylcytosine can undergo spontaneous deamination to give thymine ([Figure 4.3](#)), and during vertebrate evolution there has been a steady erosion of CG dinucleotides. As in other vertebrate genomes, therefore, the dinucleotide CG is notably under-represented in our DNA (41 % of our genome is made up of G–C base pairs, giving individual base frequencies of 20.5 % each for G and C; the expected frequency of the CG dinucleotide is therefore $20.5 \% \times 20.5 \% = 4.2 \%$, but the observed CG frequency is significantly less than 1 %).

Within the sea of our CG-deficient DNA are nearly 30 000 small islands of DNA in which the CG frequency is the expected value but cytosine methylation is suppressed. Such **CpG islands** (or CG islands; the p signifies the phosphate connecting C to G) are often 1 kb or less in length and are notably associated with genes. Approximately 50 % of CpG islands are located in the vicinity of known transcriptional start sites, as illustrated in [Figure 1](#). A further 25 % of CpG islands are found in the main gene body.

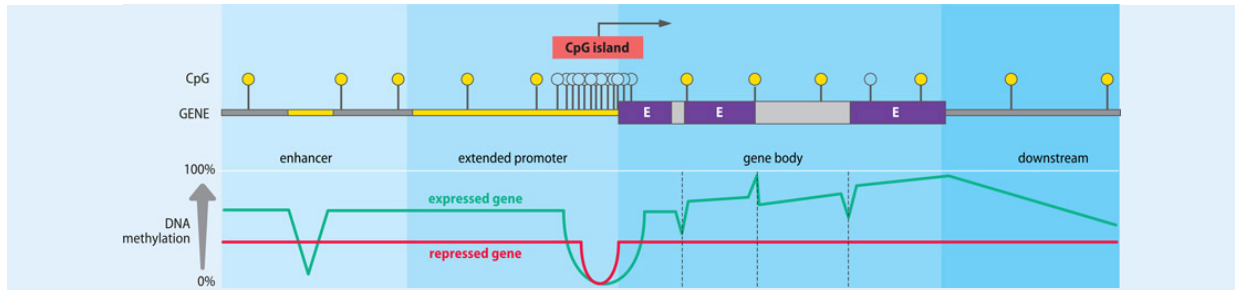


Figure 1 CpG density and DNA methylation levels across an idealized human gene. In this example we consider a gene that has single CpG island located in the vicinity of the transcriptional start site (marked by an arrow) and three exons (E). The CpG density across the gene is illustrated at the top, with open circles indicating unmethylated CpG dinucleotides and circles filled in yellow representing methylated CpG. If the gene is expressed, the gene body shows quite high DNA methylation levels, but the transcriptional start site and upstream enhancer are free of cytosine methylation, allowing access for *trans*-acting protein factors. Even when the gene is transcriptionally inactive, the cytosines remain unmethylated within the CpG island. (Adapted from Hassler MR & Egger G [2012] *Biochimie* 94:2219–2230. With permission from Elsevier Masson SAS.)

Note that CpG islands associated with transcriptional start sites remain unmethylated even when the gene is not being transcribed. Whether a gene is silenced or expressed seems instead to be related to the methylation status of other CpGs that are often located up to 2 kb from transcriptional start sites in CpG island “shores”.

DNA methylation mechanism

DNA methylation is performed by DNA methyltransferases (DNMTs). The DNMT1 enzyme serves to maintain an existing DNA methylation pattern. During replication of a methylated DNA molecule, each parental DNA strand retains its pattern of methylated cytosines. The newly synthesized complementary DNA strand is formed by incorporating unmethylated bases, and so in the absence of any further DNA methylation the result would be a hemimethylated DNA (**Figure 6.15A**).

To maintain the original DNA methylation pattern, DNMT1 is normally present at the replication fork and methylates the newly synthesized DNA strand. It methylates only those CG dinucleotides that are paired with a methylated CG on the opposing parental DNA strand. That is, the methylated parent DNA strands act as templates for copying the original methylation pattern (see **Figure 6.15A**). As a result, patterns of symmetric CG methylation can be faithfully transmitted from parent cell to daughter cells.

The enzymes DNMT3A and DNMT3B are *de novo* methyltransferases—they can methylate any suitable CG dinucleotide (see **Figure 6.15B**). They have important roles in epigenetic reprogramming when epigenetic marks are comprehensively reset across the genome, at two major stages. In each case, a wave of global DNA demethylation is followed by *de novo* DNA methylation that establishes a different methylation pattern.

DNA methylation in early development and gametogenesis

Major epigenetic reprogramming occurs in the early embryo. Egg cells and, notably, sperm cells have extensively methylated DNA (but rather different patterns of DNA methylation). Following fertilization, the introduced sperm genome (now within the male pronucleus) begins to undergo active DNA demethylation; after the male and female pronuclei fuse, global demethylation of the zygote begins and continues until the early blastula stage in the preimplantation embryo (**Figure 6.16**). Then a wave of genome re-methylation occurs, coincident with initial differentiation steps giving rise to different cell lineages. Genome methylation is extensive in somatic cell lineages but moderate in trophoblast-derived lineages (which will give rise to placenta, yolk sac, and so on).

Significant epigenetic reprogramming also occurs during gametogenesis. The **primordial germ cells** that will give rise ultimately to gametes are initially heavily methylated. As they enter the genital ridge, their genomes are progressively demethylated, erasing the vast majority of epigenetic settings (see [Figure 6.16](#)). Thereafter, *de novo* methylation allows epigenetic marks to be reset.

Long noncoding RNAs in mammalian epigenetic regulation

Diverse noncoding RNAs (ncRNAs) have important roles in gene regulation. Mammalian miRNAs are focused on post-transcriptional gene silencing (often by binding to specific target sequences in the untranslated regions of messenger RNAs). Although endogenous siRNAs are important in the epigenetic regulation of centromeres in some organisms, it is not clear that they have a similar role in mammals (if they do, it might be limited to gametogenesis or early embryo development).

The most recent GENCODE release (version 40, April 2022) gives a total of 18 805 human long noncoding RNA genes that have been identified, almost as many as the 19 988 human protein-coding genes. Although many of the long noncoding RNAs have not been well studied, a large proportion are retained in the nucleus and are associated with chromatin. Many of these genes are believed to play roles in chromosome architecture and gene regulation.

We describe later how specific long noncoding RNAs have critical roles in epi-genetic phenomena such as X-chromosome inactivation and imprinting. Here we describe some basic details about how long noncoding regulatory RNAs work.

Cis-acting and trans-acting long noncoding RNAs

Unlike microRNAs which regulate genes at the post-transcriptional level, many noncoding RNAs work within chromatin, and have roles in gene regulation. Two major classes are listed below.

- *Antisense RNAs*. These *cis*-acting RNAs can interfere with transcription of partially or fully overlapping genes on the opposite DNA strand, thereby downregulating them. The natural antisense RNAs may be unspliced or spliced, and may silence several genes in a cluster—see [Table 6.5](#) for examples.

TABLE 6.5 EXAMPLES OF DIFFERENT CLASSES OF REGULATORY LONG NONCODING RNAs

Mode of action	Example	Characteristics
trans-acting gene repression	HOTAIR	2.2 kb RNA encoded from within the <i>HOXC</i> cluster at 12q13 and represses <i>HOXD</i> genes at 2q31
cis-acting gene repression through antisense RNA (which remains tethered to the DNA strand from which it was transcribed)	KCNQ10T1 (LIT1)	92 kb unspliced antisense RNA that represses transcription of the <i>CDKN1C</i> gene on the opposite DNA strand at 11p15
	SNHG14 (SNRPN)	460 kb spliced and polyadenylated antisense RNA; part of the sequence is antisense to the <i>UBE3A</i> gene and represses it.
	CDKN2B-AS1 (ANRIL)	3.8 kb spliced and polyadenylated antisense RNA that represses transcription of <i>CDKN2B</i> gene at 9p31; also recruits PRC2* to silence co-located genes
cis-acting gene activation by recruiting a chromatin-activating protein complex	HOTTIP	3.8 kb RNA that works by inducing DNA looping to bring target genes into close proximity and then recruiting the WDR5-MLL1 protein complex to

* PRC2, Polycomb repressive protein complex 2 (initiates gene silencing and then recruits the PRC1 complex to maintain it).

Mode of action	Example	Characteristics
		deposit transcription-activating H3K4me epigenetic marks
cis-acting gene repression by recruiting a chromatin-repressing protein complex	XIST	19 kb RNA that initiates X-chromosome inactivation; represses transcription of most genes on the inactive X by recruiting the PRC2* complex to deposit transcription-repressing H3K27me3 epigenetic marks

* PRC2, Polycomb repressive protein complex 2 (initiates gene silencing and then recruits the PRC1 complex to maintain it).

- *Chromatin-modifying long noncoding RNAs.* After binding to their target genes (the genes they regulate), many long noncoding regulatory RNAs recruit chromatin-modifying protein complexes to the vicinity, allowing them to change the chromatin status of the target genes. Often long non-coding RNA bound to its target gene attracts a repressive protein complex to the chromatin to bring about local chromosome compaction to silence the target gene; some, however, work to activate transcription of a target gene—see [Table 6.5](#).

As an example, the Polycomb repressive complex 2 (PRC2) is often recruited by regulatory long noncoding RNAs to repress their target genes. It has a methyltransferase subunit that deposits the H3K27me3 epigenetic mark associated with facultative heterochromatin. Note that PRC2 cannot bind to chromatin directly; instead, it needs to be recruited to chromatin by a regulatory RNA that is able to bind both to chromatin and also to PRC2. After the repressive chromatin state has been initiated by PRC2, it is maintained by another Polycomb repressive complex 1 (PRC1).

[Figure 6.17](#) outlines the process for PRC2-induced repression of chromatin for *trans*-acting and *cis*-acting long noncoding RNAs. In the latter case, the long noncoding RNAs may work as newly synthesized RNA transcripts that are still physically associated with the chromatin where they act as scaffolds for assembling repressive protein complexes.

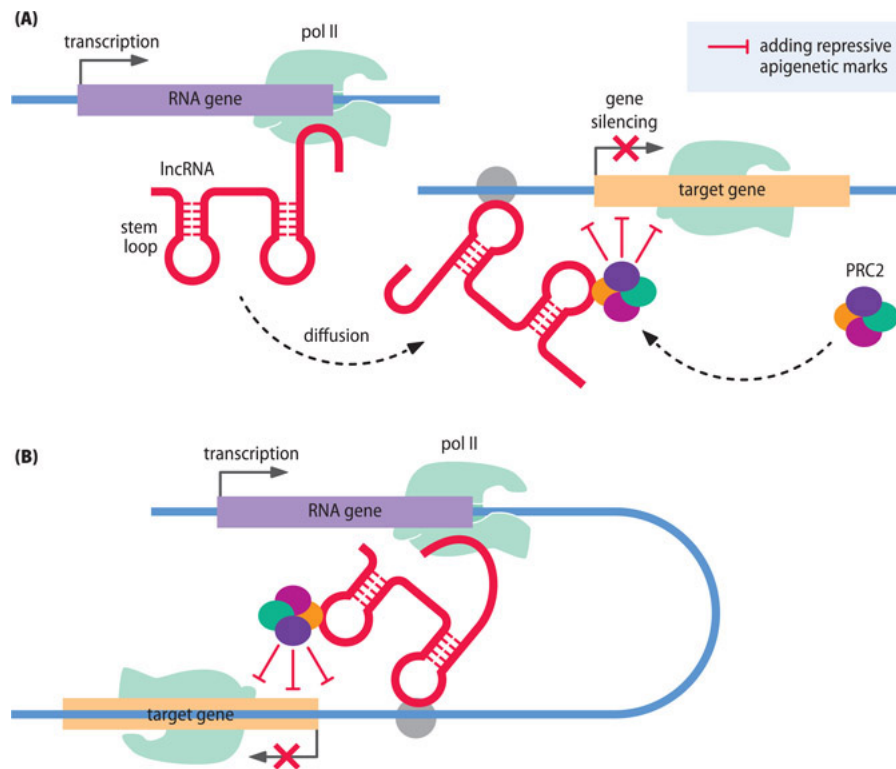


Figure 6.17 How *trans*-acting long noncoding RNA (A) and *cis*-acting long noncoding RNA (B) can silence target genes by recruiting a repressive protein complex such as the PRC2 Polycomb repressive complex. The long noncoding RNA (lncRNA) is envisaged to have two important binding sites (represented here, illustratively, as stem loops). One of these is for binding to the target gene (or to a sequence-specific binding protein bound to the gene, as shown here), and the other is for binding the PRC2 complex. The PRC2 complex cannot bind to target genes by itself. Instead, it relies on a long noncoding RNA with a PRC2-binding site that will bind it, and thereby position it next to a target gene. (A) For *trans*-acting lncRNAs, the RNA is synthesized, migrates to find and bind a target gene (often on a different chromosome), and then recruits PRC2 to deposit its epigenetic mark to silence the gene. (B) For *cis*-acting lncRNAs the newly synthesized RNA remains attached to the chromatin but can arrange (by DNA looping in this case) to bind to a nearby target gene on the same chromosome and recruit PRC2 to silence it.

Genomic imprinting: differential expression of maternally and paternally inherited alleles

We are accustomed to the idea that in mammalian diploid cells both the paternal and maternal alleles are expressed (biallelic expression). For a significant proportion of our genes, however, only one of the two alleles is normally expressed—the other allele is silenced (monoallelic expression).

Monoallelic expression at a locus can occur at random in an individual: in some cells the paternal allele of a gene is silenced, and in other cells the maternal allele is silenced ([Table 6.6](#)). However, for some genes a paternal allele is consistently silenced or a maternal allele is consistently silenced. That is, silencing of one allele occurs according to the parent of origin (**genomic imprinting**).

TABLE 6.6 MONOALLELIC EXPRESSION IN MAMMALS CAN OCCUR BY DIFFERENT MECHANISMS

Class	Mechanisms	Comments
Dependent on parent of origin	genomic imprinting	several genes are expressed only from paternally inherited chromosomes and several only from maternally inherited chromosomes
	X-inactivation in placenta	paternal X is always inactivated
Independent of parent of origin	X-inactivation in somatic cells*	inactivation of most genes on an X chromosome chosen at random, either the paternal or maternal X (see Figure 6.20A)
	production of <i>cell-specific</i> Ig and T-cell receptors	each mature B and T cell makes, respectively, Ig or T-cell receptor chains, using only one allele at a time. Once a functional chain is made by gene rearrangement at one randomly selected allele, a feedback mechanism inhibits further rearrangements (see Section 4.4)
	production of <i>cell-specific</i> olfactory receptors	each olfactory neuron expresses a single allele of just a single olfactory receptor (OR) gene (selected from several hundred OR genes) so that it fires in response to one specific odorant only. Depends on competition for a single monoallelic enhancer
	stochastic mechanisms	may be quite common

* At least in eutherian mammals (in marsupials the paternal X is consistently inactivated). Ig, immunoglobulin.

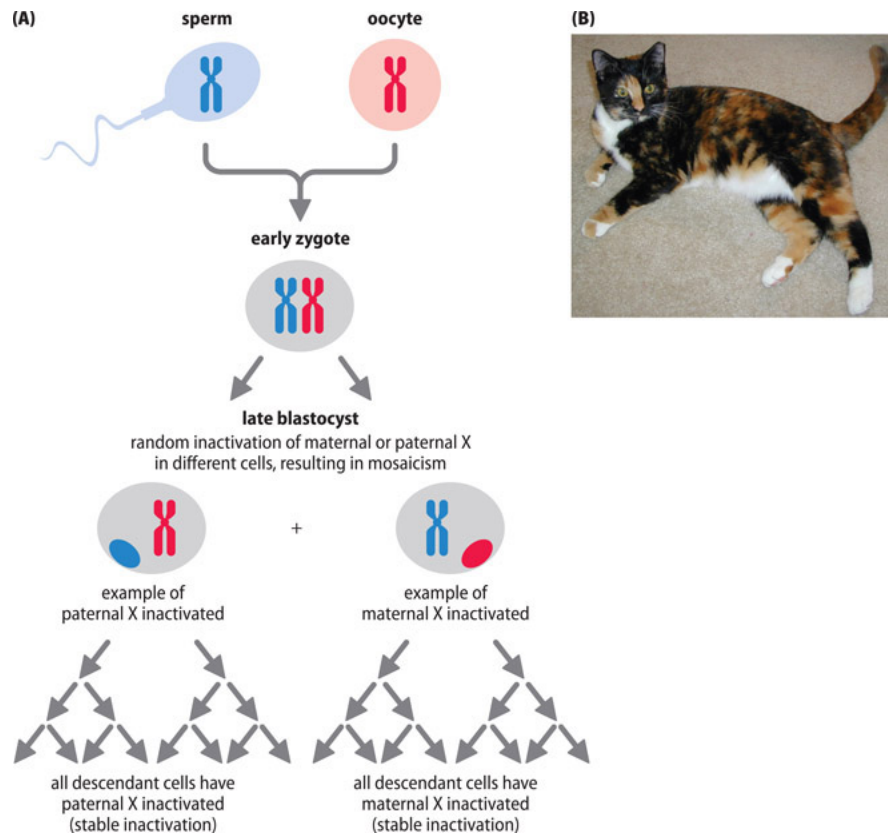


Figure 6.20 X-chromosome inactivation. (A) A randomly chosen X chromosome, either the maternal X or the paternal X, is inactivated in each cell of a 46,XX embryo. Once the choice is made, it is faithfully transmitted through all subsequent rounds of mitosis. Note that in the oogonia of a female, both X chromosomes are active; each has an equal chance of being passed on through the egg. (B) The coat of a calico cat has a mixture of white patches and two other colors, often orange and black as in this case. Like a tortoiseshell cat, it is heterozygous at an X-linked coat color locus, and in the cat represented here one allele specifies a black coat color, the other orange. The different color patches reflect clones in which different X chromosomes are inactivated. The white patches are the result of an unrelated coat color gene. (Adapted from Migeon BR [1994] *Trends Genet* 10:230–235; PMID 8091502. With permission from Elsevier.)

Because natural monoallelic expression occurs for a significant fraction of genes, the maternal and paternal genomes are not functionally equivalent in mammals. As a result, unlike several vertebrate species, mammals cannot naturally reproduce by *parthenogenesis* (reproduction without fertilization; a haploid chromosome set simply duplicates within the oocyte). It is possible to manipulate mammalian eggs artificially to make a diploid embryo with two maternal genomes, but embryonic lethality always ensues: the maternal genome cannot by itself support development—both a maternal and a paternal genome are required. Parthenogenesis fails in mammals because a subset of developmentally important genes is expressed only if inherited paternally; a different subset of genes is expressed only if inherited maternally.

As detailed below, imprinting patterns in genes are established by *cis*-acting regulatory sequences located at an *imprinting control region* that carries a type of imprint, often being methylated to very different extents in sperm DNA and egg DNA. The same differentially methylated region (DMR) can behave very differently when hypomethylated or extensively methylated.

A single allele can behave differently according to the parent of origin, but within an individual the pattern of transcriptional activity or inactivity is maintained through mitosis when somatic cells divide. The alleles are not intrinsically maternal or paternal, however: imprints need to be reversible. A man can inherit an allele from his mother that is inactive. But when he transmits that same allele in his sperm to the next generation the imprint needs to be erased. The allele is reactivated, which can result in reversal of imprints between generations (see [Figure 6.18](#)).

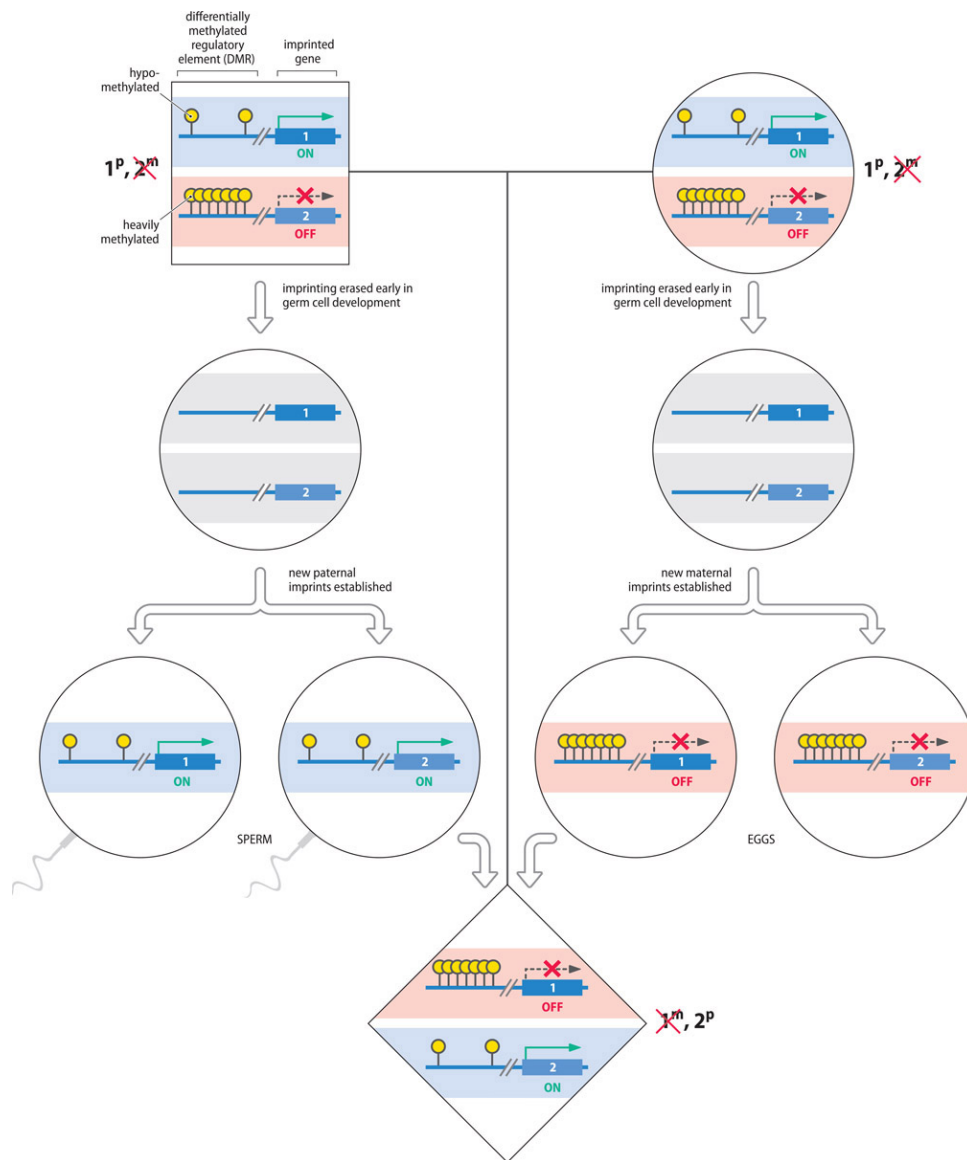


Figure 6.18 How imprints can be reversed between generations. In this example, a differentially methylated regulatory element (DMR) is heavily methylated when inherited maternally, so that a neighboring imprinted gene is silenced (OFF; red cross indicates transcriptional inactivation). When inherited paternally, the DMR is hypomethylated, and the gene is expressed (ON). The gene has two alleles, 1 and 2, and the man and woman at the top are both heterozygotes. Each of them has inherited an active allele 1 on a paternal chromosome (1^P) which is illustrated in pale blue shading; and an inactive allele 2 on a maternal chromosome (2^m) which is illustrated by pink shading. During early germ cell development, rapid demethylation of DNA occurs, and paternal and maternal epigenetic settings are erased (see left part of [Figure 6.16](#)). The loss of parental imprints, including DNA methylation marks, means that the chromosomes mostly lose their parental epigenetic marks and are now represented in neutral gray shading (middle of figure). Later in germ cell development, new imprints are established according to the sex of the individual. Each sperm from the man has an active allele 1 or an active allele 2; each of the woman's eggs has an inactive allele 1 or an inactive allele 2. Fertilization of an egg bearing allele 1 by a sperm carrying allele 2 can generate a child with the same genotype (1,2) as the parents, but the imprint in this child has been reversed: allele 1 is now the inactive allele (having been maternally inherited) and allele 2 is functional.

Extent and significance of genome imprinting

More than 140 mouse genes have been experimentally shown to be imprinted genes, and a smaller number of human imprinted genes have been validated. Catalogs of imprinted genes in different species are available at the geneimprint database at <https://www.geneimprint.com/site/genes-by-species>.

Many known imprinted genes have a role in embryonic and placental growth and development, and a popular theory attributes imprinting to a conflict of evolutionary interest between mothers and fathers. Propagation of paternal genes would be favored if the offspring were all very robust, even at the expense of the mother (potentially, a man can father children by very many different mothers). Enhanced propagation of maternal genes, however, depends on the mother's being healthy enough to have multiple pregnancies.

Mammalian development is unusual in that the zygote gives rise to both an embryo and also extra-embryonic membranes (including the trophoblast; these membranes act to support development, giving rise to the placenta). From the arguments above, paternal genes might be expected to promote the growth (and general robustness) of the fetus by maximizing the nutrients it can extract from the mother via the placenta. Paternal genes might therefore have a vested interest in supporting the development of the extra-embryonic membranes and placenta. Maternal genes, by contrast, might seek to limit the nutrient transfer so that it does not compromise the mother's health and future reproductive success. Some support for the paternal-maternal conflict theory comes from rare cases of uniparental diploidy in humans ([Figure 6.19](#)) and from artificially induced uniparental diploidy in mice.

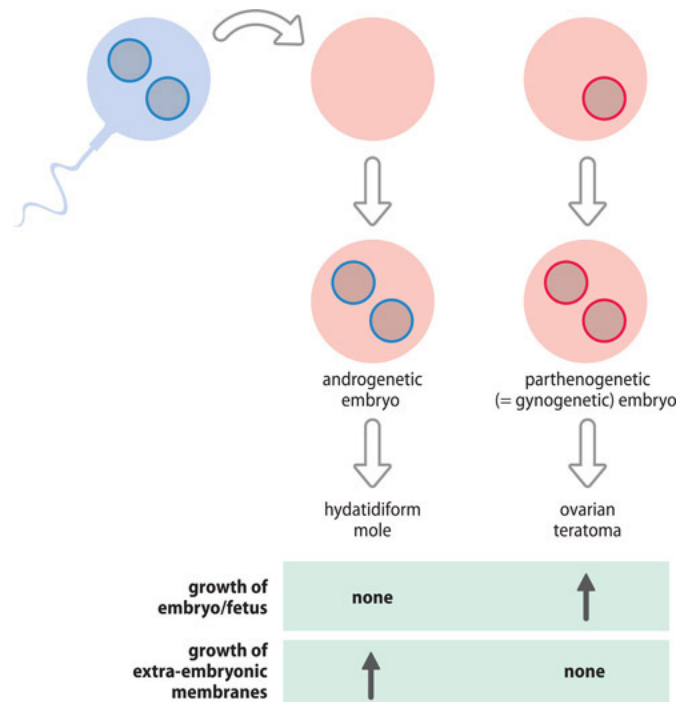


Figure 6.19 Uniparental diploidy and the divergent expression of paternal and maternal genomes. On rare occasions, a zygote is formed with a genome composed of paternal DNA only, producing an *androgenetic embryo* (this usually occurs when a diploid sperm fertilizes a faulty egg that lacks chromosomes, as shown here). Development produces an abnormal conceptus known as a hydatidiform mole, with widespread hyperplasia (overgrowth) of the trophoblast but no fetal parts. The reverse situation, where the zygotic genome is composed of maternal DNA only—producing a *parthenogenetic embryo*—gives rise to an ovarian teratoma that consists of disorganized embryonic tissues without the vital extra-embryonic membranes.

The paternal-maternal conflict theory can only be a partial explanation of why imprinting evolved in mammals. Not all imprinted genes have roles in intrauterine growth, and not all are imprinted in the direction predicted by the parental conflict theory. The theory also doesn't explain why imprinting is tissue-specific for many genes. The insulin-like growth factor gene *IGF2*, for example, is maternally imprinted in many tissues but biallelically

expressed in brain, adult liver, and so on, and the *UBE3A* gene, which is implicated in Angelman syndrome, is paternally imprinted in neurons but biallelically expressed in glial cells and other tissues.

Establishing sex-specific imprints by differential methylation

Imprinted genes are often found in clusters and under the control of a common differential methylation region known as an imprinting control region (ICR). During germ cell development, parental imprints are erased. Thereafter, imprinting is established when sex-specific patterns of *de novo* DNA methylation are created within CG dinucleotides in the ICRs.

Although the sperm and egg have very different overall DNA methylation patterns, large-scale demethylation in the early embryo removes the vast majority of the DNA methylation differences in the paternal and maternal genomes. An important exception is in the ICRs, where the sex differences in DNA methylation are retained in somatic cells (but will be erased in primordial germ cells).

Much of our knowledge of how ICRs regulate the expression of imprinted genes comes from studying certain imprinted gene clusters that are associated with developmental disorders in humans. We consider these mechanisms in the context of disorders of imprinting in [Section 6.3](#); as we will see, long noncoding RNAs are also often associated with clusters of imprinted genes and are important in imprinting control.

X-chromosome inactivation: compensating for sex differences in gene dosage

As described in [Section 7.5](#), a change in constitutional chromosome number (aneuploidy) is usually lethal, or it results in significant abnormalities. Chromosome loss is particularly damaging, and monosomy is lethal in the early embryo with just one exception: loss of an X chromosome from a female embryo is viable and results in Turner syndrome (45,X).

Aneuploidies cause problems because of abnormal **gene dosage**. We have elaborate gene interaction systems, and the products of genes across the genome sometimes can work together in ways where the relative amounts of participating gene products need to be very tightly controlled (changes in the amounts of an individual component can wreak havoc with regulatory systems, for example). An average chromosome has multiple genes in which a change in copy number (producing abnormal gene dosage) is positively harmful. And yet there is one glaring difference that is somehow tolerated in mammals: females have two X chromosomes, but males have only one X chromosome and a Y chromosome.

Whereas Y-specific genes are rare and largely devoted to male-specific functions, the X chromosome has more than 800 protein-coding genes and many RNA genes that work in all kinds of important cell functions. As first proposed by Mary Lyon, a gene dosage compensation mechanism equalizes X-chromosome gene expression in male and female cells by causing one of the two X chromosomes in female cells to be heterochromatinized (**X-chromosome inactivation**).

X-chromosome counting and inactivation choices

Early in embryogenesis, our cells somehow count how many X chromosomes they contain, then permanently inactivate all except one randomly selected X. At very early stages in development, both X chromosomes are active, but X-inactivation is initiated as cells begin to differentiate, occurring at the late blastula stage in mice, and most probably also in humans. Inactive X chromosomes remain in a highly condensed heterochromatic state throughout the cell cycle and can be seen as the Barr body (sex chromatin) on the periphery of the cell nucleus (see [Figure 5.4](#)).

The choice to inactivate the maternal or the paternal X is made randomly. But whichever of the parental X chromosomes is chosen for inactivation within a cell, that same X is inactivated in all daughter cells ([Figure 6.20A](#)). An adult female is thus a mosaic of cell clones, each clone retaining the pattern of X-inactivation that was established in its progenitor cell early in embryonic life. X-chromosome inactivation is strikingly revealed by

mixed coat colors in tortoiseshell and calico cats, both of which are almost always females (a small minority are XXY males). Tortoiseshell cats have a combination of two coat colors other than white (often orange and black); calico cats have additional white patches, as shown in [Figure 6.20B](#).

X-inactivation is stable through mitosis but not across the generations. A woman's maternal X can equally well have been the active or inactive one in her mother and has the same chance as her paternal X of being inactivated in her own cells.

***XIST* RNA and initiation of X-inactivation**

Inactivation of a human X chromosome is initiated at an X-inactivation center (XIC) at Xq13. It then propagates along the whole length of the chromosome in what may be an extreme example of the tendency of heterochromatin to spread (we consider heterochromatin spreading in more detail in the context of disease in [Section 6.3](#)).

The transient pairing of the two XIC sequences is probably the mechanism by which the X chromosomes are counted. Within this region, the *XIST* gene encodes a 17 kb spliced and polyadenylated noncoding RNA, an X-inactivation-specific transcript expressed exclusively from the *inactive* X chromosome.

XIST is centrally involved in spreading heterochromatinization outward from the XIC: both *XIST* RNA and the Polycomb proteins it recruits seem to spread along the inactive X to initiate gene silencing along the length of the chromosome. As a result, the inactive X has epigenetic marks typical of heterochromatin (H3K9me2, H3K9me3, H3K27me3, unmethylated H3K4, deacetylated H4, and frequent replacement of histone H2A by the macro-H2A histone variant). In differentiated cells that have already undergone X-inactivation, loss of *XIST* does not cause reactivation. That is, *XIST* is needed to establish X-inactivation but not to maintain it.

The mechanism of X-inactivation remains poorly understood. In addition to *XIST*, there are multiple other longer ncRNAs within the XIC. Several of them are known to have roles in the X-inactivation mechanism in mouse, but the organizations of the human XIC and its mouse counterpart, *Xic*, are rather different.

Escaping X-inactivation

A few genes on the X have active counterparts on the Y, notably in the terminal pseudoautosomal regions (but also in some other areas—see [Figure 5.7](#)). X-inactivation is therefore not a blanket inactivation of the entire chromosome, because no dosage compensation is needed for genes on the X that have functional equivalents on the Y. However, unlike in mouse, in which only a small number of genes escape X-inactivation and are not coated by *XIST* RNA, about 15 % of genes on the human X somehow escape inactivation.

6.3 ABNORMAL EPIGENETIC REGULATION IN MENDELIAN DISORDERS AND UNIPARENTAL DISOMY

Abnormal regulation of how our genes and other functional DNA sequences work can arise in different ways. Changes in DNA sequence may affect how DNA sequences work without necessarily causing any great change in their chromatin environment. In [Chapter 7](#) we look at how disease arises directly as a result of altered base sequences and copy number variation. In this section we are largely concerned with abnormal epigenetic regulation. That occurs in rare cases when two copies of the same chromosome are abnormally inherited from just one parent. In addition, some genetic disorders show abnormal epigenetic regulation.

Principles of epigenetic dysregulation

An abnormal epigenetic change (**epimutation**) at one or more loci can be the immediate cause of pathogenesis in Mendelian disorders that show abnormal epigenetic regulation. However, the *primary* event is often a genetic mutation at a defined locus that may be one of several types. It may be a gene that makes a protein or RNA that

controls epigenetic modifications at other genes located elsewhere in the genome. It may be a *cis*-acting regulatory sequence that regulates epigenetic modifications of neighboring genes. In each case the epimutations determine the disease phenotype; because they lie downstream of a primary genetic event, they are often classified as *secondary epimutations* ([Figure 6.21](#)).

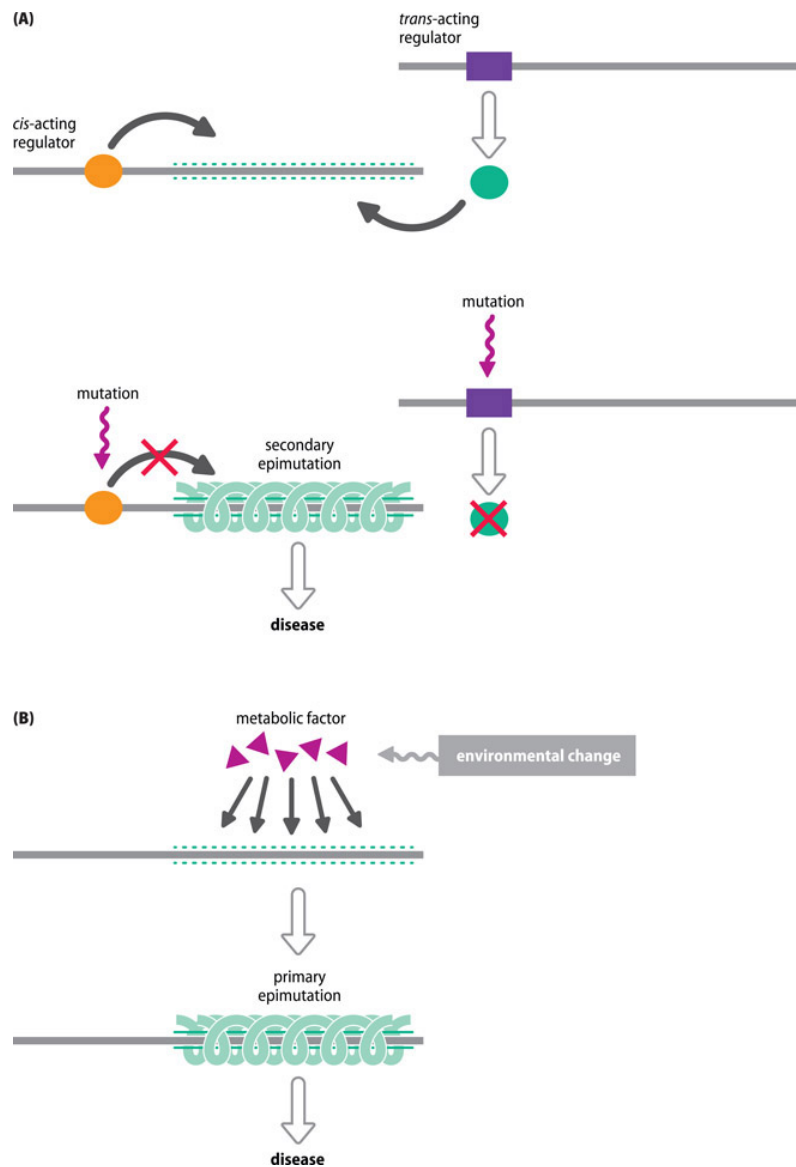


Figure 6.21 Primary and secondary epimutations. (A) Secondary epimutations arise through a change in chromatin state initiated by mutation at a *cis*-acting or *trans*-acting epigenetic regulator. Alterations to the regulators cause a change in chromatin state—in this case from a transcription-permissive environment (upper panel) to a repressive heterochromatic environment (lower panel). (B) Primary epimutations can effect a change in chromatin state without any change in the base sequence. Here, we imagine that an environmental change has changed the concentration of some metabolic factor that is important in DNA or chromatin modification, causing the change in chromatin state.

By contrast, *primary epimutations* can arise without any change to the base sequence. Instead, a chromatin state is reprogrammed, for example by some environmentally induced change, in a way that changes epigenetic controls (changes in metabolic factors may affect DNA methylation or histone modification states, and so on). Primary epimutations may be important in complex disease, and in [Chapter 8](#) we examine the roles of both genetic and

epigenetic factors in complex disease. Epigenetic factors have a particularly important role in cancer, and we consider these separately in [Chapter 10](#). Here we focus on epi-genetic dysregulation that arises through abnormal chromosome segregation or that is a feature of certain Mendelian disorders.

“Chromatin diseases” due to mutations in genes specifying chromatin modifiers

As detailed above, epigenetic marks are inscribed by a series of chromatin “writers”, enzymes that methylate DNA or add different types of chemical group to defined amino acids on core histones. They can be recognized and bound by specific protein “readers” or be removed by specific enzymes (“erasers”).

Individual genes that produce a chromatin writer, eraser, or reader can potentially regulate very many different genes across the genome, and a mutation in a gene of this type may result in heritable abnormal chromatin organization at multiple loci. In some cases the disruption to normal gene regulation might be incompatible with life because many of the target genes of chromatin modifiers are important in early development. Some disorders, however, do arise through mutations at chromatin modifier loci. These so-called “chromatin diseases” typically result in developmental disorders that can vary in phenotypes but are usually accompanied by mental retardation ([Table 6.7](#)). Affected individuals typically do not reproduce, usually presenting as sporadic (isolated) cases.

TABLE 6.7 EXAMPLES OF CHROMATIN DISEASES, DISORDERS THAT ARISE FROM MUTATION IN A CHROMATIN MODIFIER GENE

Class and type of chromatin modifier		Gene	Associated disease (reference)	Phenotype		
				Developmental	MR	Other
Writers	DNA methyl transferase	<i>DNMT3B</i>	ICF syndrome (OMIM 242860)	facial anomalies	variable	immunodeficiency; centromeric instability
	histone acetyltransferase	<i>CREBBP</i> or <i>EP300</i>	Rubinstein-Taybi syndrome (PMID 20301699; OMIM 180849)	characteristic facial features; digit anomalies	yes	
Erasers	histone lysine demethylase	<i>KDM5C</i>	Claes-Jensen type of syndromic, X-linked mental retardation (OMIM 300534)	variable—often mildly dysmorphic facial features; microcephaly	yes	
Readers	meCG-binding protein	<i>MECP2</i>	Rett syndrome (PMID 20301670; OMIM 312750)	see Clinical Box 2 on next page	variable	see Clinical Box 2 on next page
	chromatin remodeler	<i>ATRX</i>	α -thalassemia X-linked mental retardation syndrome (PMID 20301622; OMIM 301040)	cranial, facial, skeletal, and genital abnormalities; developmental delay and microcephaly	yes	thalassemia (<i>ATRX</i> regulates the α -globin genes among many others)

In [Clinical Box 2](#) we give a case study of a representative chromatin disease, Rett syndrome, an X-linked progressive neurodevelopmental disorder that results from mutations in the *MECP2* gene and affects girls almost exclusively. The regulatory protein MECP2 recognizes and binds 5-methylcytosine and is especially important in neuron maturation; the lack of a normal MECP2 protein in cells leads to an inability to recognize DNA

methylation. The phenotype is partly determined by the X-inactivation pattern: inactivating the normal X in a relatively high proportion of neurons would be expected to result in a particularly severe phenotype. Failure to produce any functional MECP2 protein was initially expected to be lethal and would explain why affected males are so rarely seen. (Affected males can occur as a result of a post-zygotic inactivating mutation or have certain missense MECP2 mutations and severe neonatal encephalopathy.)

CLINICAL BOX 2 A CASE STUDY: RETT SYNDROME

Evie was referred to the Genetics Clinic when she was 30 months old because of delayed development and recent loss of skills. She was born by emergency Caesarian section because of meconium staining and abnormal CTG (cardiotocograph) trace (used for early detection of fetal distress in the third trimester), but she did not require resuscitation. She breast fed well but was always floppier and less interactive than her two older sisters. She sat unsupported at nine months but had a tendency to flop her head forward when sitting.

Her parents described a definite loss of skills from around two years of age. She started bottom shuffling at 18 months but by the age of two she had stopped attempting to move. She stopped babbling and saying “bye” and being able to hold toys or feed herself finger foods. She had episodes suggestive of absence seizures where she became unresponsive for several seconds. Around this time there was also a change in her behavior, and she interacted less with others and became increasingly bad-tempered. Evie showed stereotypic hand wringing movements and frequently brought her hands to her mouth and chewed them. She had generalized low muscle tone and was able to weight bear (with hyperextension at the knees), but not to crawl, or pull to stand. Her head circumference, on the 75th centile at birth, was now on the 9th centile, suggesting slowing of head growth.

Evie had already been extensively investigated for a metabolic cause for her developmental regression and had had a normal cranial MRI scan and EEG. Methylation studies for Angelman syndrome were normal. Her developmental regression and hand stereotypies (repetitive hand movements) were, however, suggestive of Rett Syndrome. *Rather than show a photo of Evie, we refer readers to a YouTube video describing another girl with Rett syndrome where the hand stereotypies can clearly be seen:* <https://www.youtube.com/watch?v=H2iKz1Cx-HQ>.

To check if Evie had Rett syndrome, Sanger sequence analysis was carried out on the MECP2 gene and revealed a heterozygous pathogenic variant c.455C>G p.(Pro152Arg). This variant has been reported many times in girls with Rett syndrome, and functional studies have shown it to affect MECP2 protein function. Evie’s mother was tested and did not carry the variant; it was thus assumed to be *de novo*.

Evie was reviewed when she was four years old. She had not shown any further regression but had made very little developmental progress. She had developed three different types of seizure: nocturnal, tonic clonic, and absences, and was on two different anticonvulsant drugs. She had also developed a disordered breathing pattern, with periods of hyperventilation and breath-holding. She was only able to manage mashed food and had had several choking episodes.

Disease resulting from dysregulation of heterochromatin

Epigenetic regulation causes distinctive patterns of heterochromatin and euchromatin to form in our cells. Heterochromatin is first formed at nucleation sites, consisting of either repetitive DNA or silencer elements, and can then expand across long distances on a chromosome, even a whole chromosome, as in X-chromosome inactivation. Heterochromatin spreading involves converting open chromatin to condensed transcriptionally silent chromatin and is facilitated by communication between nucleosomes.

To avoid silencing essential genes, cells have evolved different mechanisms to limit the spread of heterochromatin. One such mechanism depends on **barrier elements**, a type of boundary element, that are able to protect genes from their surrounding environment. Barrier elements can include sequences that are selected to be comparatively nucleosome-free to provide a break in the nucleosome chain.

Altered heterochromatic states can impair normal gene expression in two quite different ways. Sometimes active genes are inappropriately exposed to heterochromatin controls and are silenced. An alternative form of dysregulation involves a reduction in heterochromatin and loss of gene silencing.

Inappropriate gene silencing

Aberrant heterochromatin regulation can silence genes that are normally meant to be expressed. A long-range **position effect** can mean that a gene is relocated to a position very close to constitutive heterochromatin (by a chromosome translocation or inversion, for example). In these cases, the boundary between euchromatin and heterochromatin can be reset, and the gene is silenced by heterochromatin spreading (Figure 6.22).

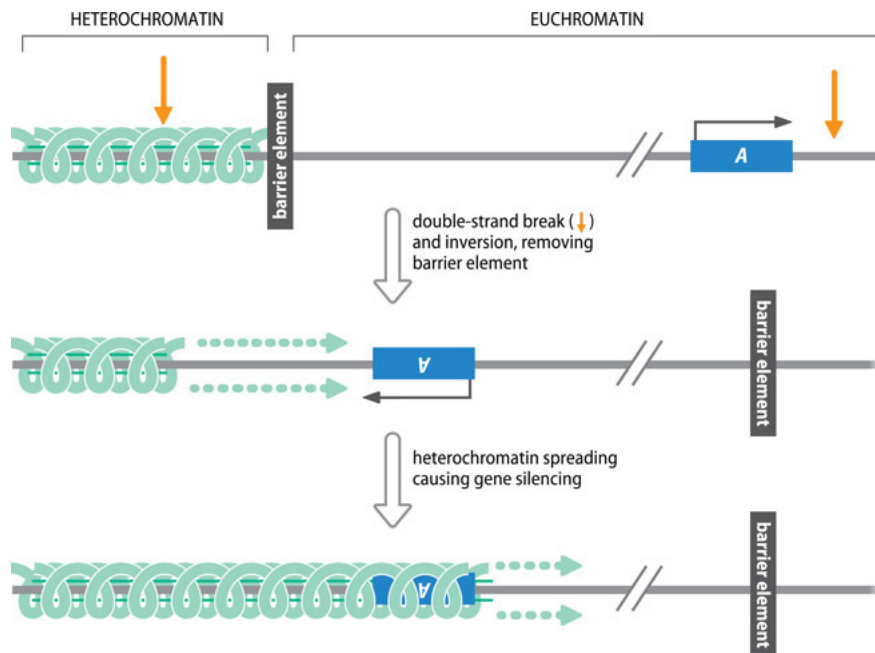


Figure 6.22 Heterochromatin spreading after an inversion causes displacement of a nearby barrier element. A barrier element protects genes in the euchromatin region from being silenced by adjacent heterochromatin (shown here in green). Large-scale rearrangements, such as the inversion shown here, can relocate the protective barrier element so that it no longer separates gene A from a neighboring heterochromatic region. This allows heterochromatinization of what had been a euchromatic region, resulting in gene silencing (a position effect).

Some special types of mutation can also induce heterochromatin formation within or close to a gene and so silence it. That can happen in the case of very large expansions of noncoding triplet repeats that occur in some recessively inherited disorders such as fragile X-linked mental retardation and Friedreich's ataxia. We consider this type of abnormal heterochromatin within the context of disease due to unstable oligonucleotide repeat expansion in [Section 7.3](#).

Heterochromatin reduction

A primary function of some tumor suppressor genes such as *BRCA1* is to maintain the integrity of constitutive heterochromatin. As described in [Chapter 10](#), mutations in these genes can result in a loss of heterochromatin organization; reduced centromeric heterochromatin leads to mitotic recombination and genome instability.

Some mutations causing heterochromatin reduction and inappropriate gene activation also result in inherited disorders. A classic example occurs in facioscapulohumeral dystrophy (FSHD), the third most common form of

muscular dystrophy. This dominantly inherited disorder occurs as a result of simultaneous inheritance of two genetic variants:

- a reduction in heterochromatin due to unequal crossover at an array of tandem macrosatellite repeats at 4q35, close to the telomere, with each repeat containing a transcriptionally repressed *DUX4* retrogene copy
- a variant that creates a polyadenylation site close to the most telomeric of the repeats, enabling the most telomeric repeat to become transcriptionally active.

The combination of the two variants allows inappropriate expression of the *DUX4* transcription factor, which is normally silenced in somatic cells and has been thought to be toxic to muscle cells—see [Figure 60.23A,B](#).

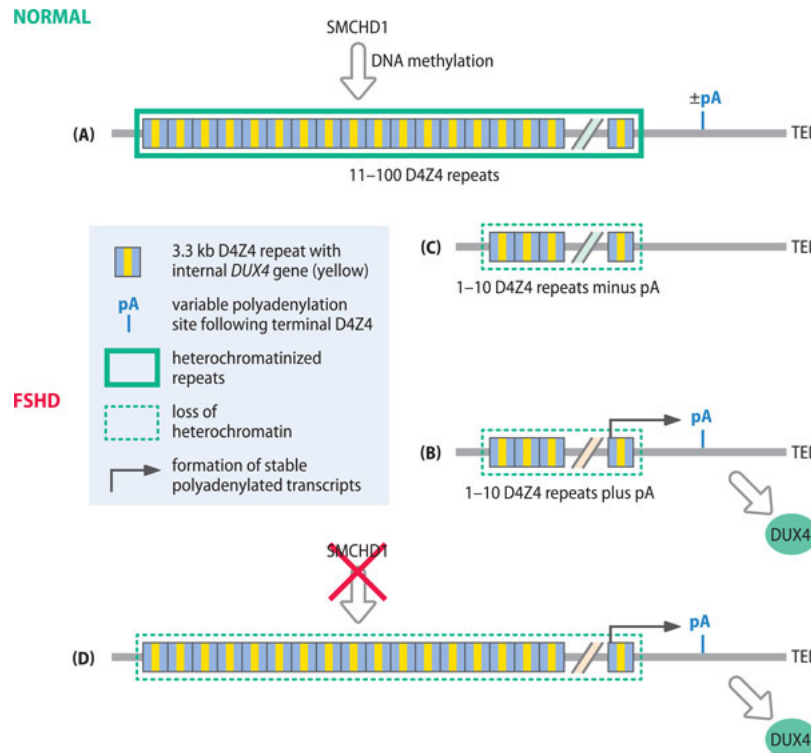


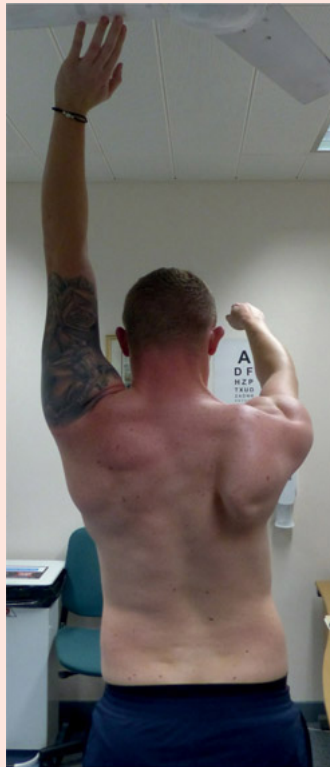
Figure 6.23 Heterochromatin reduction and inappropriate activation of the heterochromatic *DUX4* retrogene in facioscapulohumeral dystrophy (FSHD). (A) A normal chromosome 4 with a heterochromatinized array of 11–100 D4Z4 repeats, each ~3.3 kb long (the repeat copy number varies because of unequal crossover). A variable polyadenylation site is located adjacent to the last (most telomeric) D4Z4 repeat. (B) In FSHD1, the D4Z4 array is reduced in size to 1–10 repeats, causing a marked decrease in heterochromatin; the downstream polyadenylation site is present, allowing both transcription and translation of the last *DUX4* sequence. (C) In the absence of the downstream polyadenylation signal, the *DUX4* sequence cannot produce stable transcripts even when the heterochromatin is decreased. (D) In FSHD2 the downstream polyadenylation sequence is present, together with a long D4Z4 array that nevertheless has decreased heterochromatin because of a failure to produce the SMCHD1 methylation regulator.

One further complication is that the disorder is genetically heterogeneous: in the common FSHD1 form, illustrated by the case study shown in [Clinical Box 3](#), the reduction in heterochromatin is caused by significant reduction in the number of D4Z4 repeats within the macrosatellite array at 4q35. In a second form, FSHD2, the reduction in heterochromatin is caused by mutation in the *SMCHD1* gene on chromosome 18 that regulates DNA methylation ([Figure 6.23D](#)).

CLINICAL BOX 3 A CLINICAL CASE STUDY: FACIOSCAPULOHUMERAL DYSTROPHY

John was born at term from an uneventful pregnancy to nonconsanguineous parents. He had normal development through childhood and enjoyed playing regular sports. As an adult, John continued to be active, attending the gym four times a week and regularly playing rugby. When he was 20 years old, John developed difficulty raising his right arm above shoulder height. The arm was not painful but appeared to him to be weak. He was initially referred to an orthopedic surgeon, who passed him on to rheumatology. No cause was identified, and he was ultimately referred to the neuromuscular centre.

When first seen by the neuromuscular team at the age of 22 John had mild facial weakness, wasting of the pectoralis muscle on the right, and asymmetrical scapular winging, as shown in [Figure 1](#). Shoulder abduction was asymmetrically weak, scoring 3- on the MRC score on the right, but 5- on the left. Other shoulder movements including adduction and internal and external rotation were slightly impaired. His lower limbs were stronger and almost completely normal.



[Figure 1](#) Asymmetrical scapular winging.

The neuromuscular consultant suspected a diagnosis of facioscapulohumeral muscular dystrophy and arranged testing for FSHD1. The D4Z4 repeated region at 4q35 can be investigated using a P13E11 probe on Southern blots of *Eco*RI-digested genomic DNA. When this was done, John's DNA sample revealed an abnormally small *Eco*RI fragment of 24 kb, suggesting that one of his chromosome 4s has only around six D4Z4 repeats and the contraction in the normal size of the D4Z4 repeat region has allowed the terminal *DUX4* gene to become, inappropriately, active. (The *DUX4* gene is strongly expressed in germ cells where the DUX4 protein works as a transcriptional activator but normally it is transcriptionally repressed in somatic cells, as a result of heterochromatinization.)

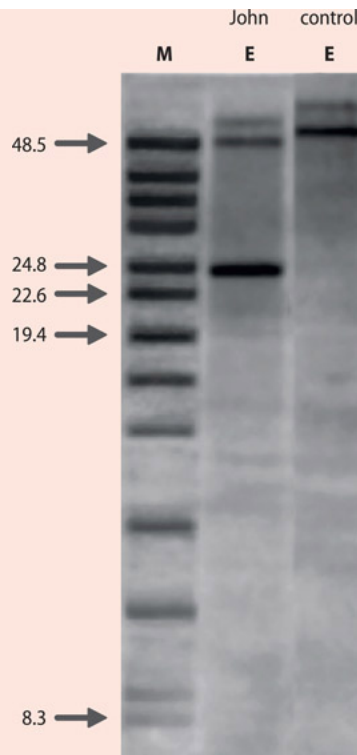


Figure 2 Southern blot analysis after pulsed field gel electrophoresis. Hybridization of a D4Z4 repeat region DNA probe, P13E11, to *EcoRI*-digested genomic DNA normally reveals a band of >40 kb, as shown in the unaffected control at right, but in John's DNA, the detected *EcoRI* fragment is only 24 kb. M, marker DNA lane with 13 size standards, five of which are identified by arrowheads. E, *EcoRI*-digested genomic DNA. Image kindly provided by Sarah Burton-Jones, South West Genomics Hub, Southmead Hospital, Bristol.

Uniparental disomy and disorders of imprinting

Recall that the sperm and egg genomes each carry epigenetic marks that are rather different from each other. Although the great majority of the gametic epigenetic marks are then erased in the early embryo, remaining imprints are retained in our somatic cells. We have more than 100 classically defined imprinted genes, and many of them have important roles in early development. In some cases, the maternal allele is consistently silenced or preferentially silenced (that is, monoallelic expression in some cell types, but biallelic expression in others); for other genes, it is the paternal allele that is consistently or preferentially silenced.

Occasional cases of uniparental diploidy (producing an androgenetic or gynogenetic embryo as shown in [Figure 6.19](#)) can occur. They are invariably lethal in the early embryo (each embryo of this kind fails to express multiple imprinted genes needed for fetal development). Sometimes, however, abnormal regulation of imprinted genes is confined to genes on a single chromosome and results in a developmental disorder. This can arise by a change in DNA sequence at or near the imprinted gene locus or by abnormal epigenetic regulation of the imprinted gene that may result as a downstream effect, often because of other genetic changes.

Uniparental disomy arises when a zygote develops in which both copies of one chromosome originated either from the father or from the mother. It occurs most often after a trisomic conceptus is first formed with two chromosome homologs from one parent and a single chromosome copy from the other parent; loss of the latter chromosome very early in development results in a heterodisomy ([Figure 6.24A](#)). The alternative is monosomy rescue, which results in isodisomy (with two identical copies of a chromosome—see [Figure 6.24B](#)). As described in the next section, uniparental disomy can result in a disorder of development if the chromosome happens to contain imprinted genes that are important in development.

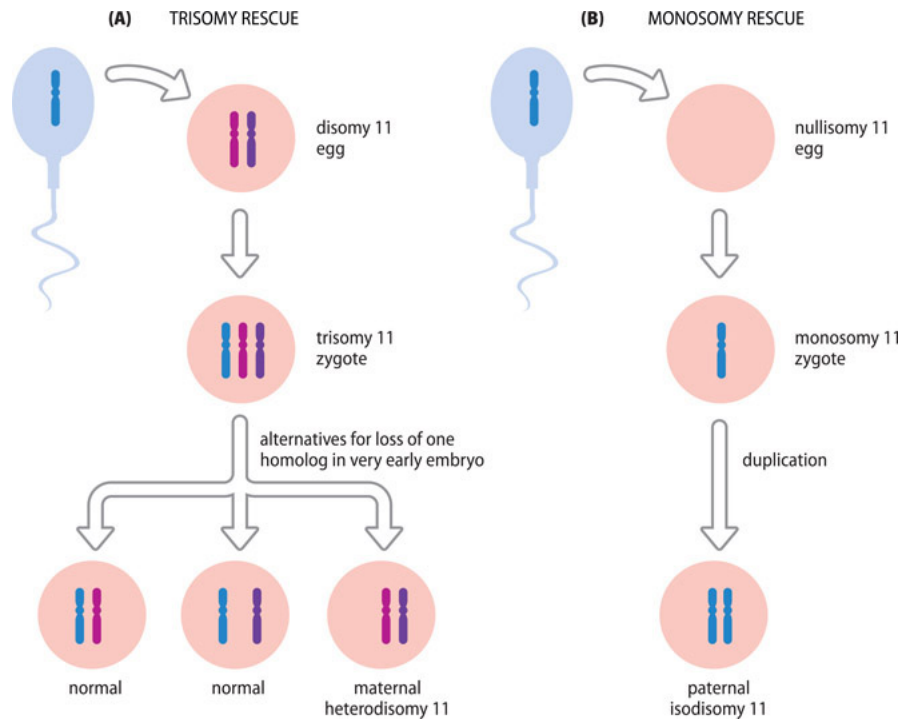


Figure 6.24 Uniparental disomy can arise by post-zygotic trisomy rescue or monosomy rescue. (A) Shown in the center is one type of trisomic zygote (with two maternal homologs plus one paternal chromosome, in this case chromosome 11). Trisomy 11 is lethal, but the trisomy can be corrected in the very early embryo by the loss of a chromosome 11 from one embryonic cell that then has a growth advantage and goes on ultimately to form an individual with the correct number of chromosomes. The disomic cell (and individual) may be normal (one paternal 11 plus one maternal 11) or have the two maternal chromosome 11 homologs (uniparental *heterodisomy*). (B) Monosomy rescue can occur by chromosome duplication, but in this case the result is uniparental *isodisomy* (the two chromosomes are identical, not homologs).

Abnormal gene regulation at imprinted loci

At an imprinted locus only one of the two parental alleles is consistently expressed (in at least some tissues), and alteration of the normal pattern of monoallelic expression can result in disease. Certain imprinted genes are important in fetal growth and development, so in these cases abnormal expression often results in recognizable developmental syndromes. Sometimes the normally expressed allele is not present or is defective, and deficiency of a gene product causes disease. In other cases, disease can be due to overexpression of a dosage-sensitive gene ([Table 6.8](#)). Analysis of human imprinting disorders has helped us understand the underlying gene regulation, and we provide here a background to two such imprinted gene clusters, as detailed below.

TABLE 6.8 EXAMPLES OF IMPRINTED DISORDERS OF EARLY CHILDHOOD

Disorder	Diagnostic clinical features	Molecular basis of imprinting disorder	Cause	
			UPD**	Others include
PATHOGENESIS DUE TO UNDEREXPRESSION OF GENES AT IMPRINTED LOCI				
Prader-Willi syndrome	DD*; low birth weight; hypotonia; hyperphagia	silencing/lack of active allele for imprinted genes at 15q11.2, including multiple <i>SNORD116</i> (<i>HB11-85</i>)	Mat.15, ~25%	Δ pat.15q11-13, ~70%

* DD, developmental delay; IUGR, intrauterine growth retardation.

** UPD, uniparental disomy; pat., paternal; mat., maternal.

*** ICR1, ICR2, imprinting control regions 1 and 2.

Disorder	Diagnostic clinical features	Molecular basis of imprinting disorder	Cause	
			UPD**	Others include
		snoRNA genes (see Figure 6.26)		
Angelman syndrome	DD (severe); no speech; epilepsy; ataxia	silencing/lack of active allele for imprinted <i>UBE3A</i> gene at 15q11.2 (see Figure 6.26)	Pat.** 15, ~5%	Δ mat.15q11-13,~75%
Silver-Russell syndrome	IUGR*; faltering growth; short stature	silencing/lack of active allele for <i>IGF2</i> at 11p15.5 (see Figure 6.25) or for <i>MEST(PEG1)</i> at 7q31-32 (<i>IGF2</i> and <i>MEST</i> are maternally imprinted)	Mat.** 11	loss of pat. ICR1 methylation, ~35-50%
			Mat.7, ~8%	
PATHOGENESIS DUE TO OVEREXPRESSION OF GENES AT IMPRINTED LOCI AND/OR FETAL GROWTH PROMOTION				
Beckwith-Wiedemann syndrome	macrosomia/overgrowth; macroglossia; umbilical defect	biallelic expression of <i>IGF2</i> at 11 p15.5 (normally silenced on maternal 11) and/or biallelic expression of a ncRNA that suppresses a growth-restricting gene, <i>CDKN1C</i> ; see Figure 6.22)	Pat.11	loss of mat. ICR2*** methylation, ~50 %; gain of mat. ICR1*** methylation, ~5%; <i>CDKN1C</i> mutation, ~5 %
Transient neonatal diabetes	IUGR*; neonatal diabetes with remission	biallelic expression of <i>PLAGL1</i> , a regulator of insulin secretion, at 6q24 (normally silenced on maternal chromosome 6)	Pat.6	30%

* DD, developmental delay; IUGR, intrauterine growth retardation.

** UPD, uniparental disomy; pat., paternal; mat., maternal.

*** ICR1, ICR2, imprinting control regions 1 and 2.

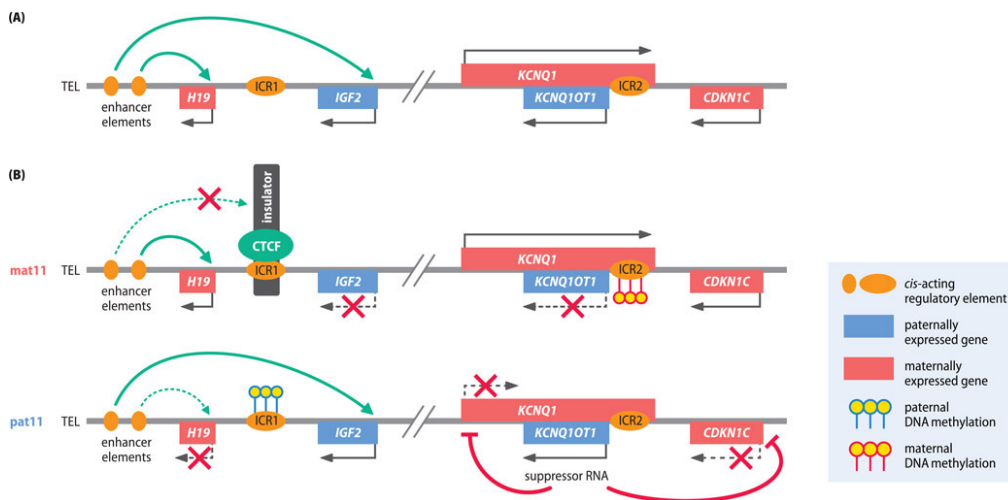


Figure 6.25 Imprinting control mechanisms in the 11p15.5 imprinted gene cluster. (A) Terminal genes and imprinting control regions (ICR1, ICR2) in the human 11p15.5 imprinted gene cluster. The ~600 kb gap (//) between the insulin-like growth factor 2 gene (*IGF2*) and *KCNQ1* contains at least five other imprinted genes. Arrows indicate the direction of transcription. TEL, telomere. (B) Regulation of

imprinted genes in the 11p15 cluster on maternal and paternal chromosome 11 (mat11 and pat11). ICR1 acts as an insulator. It is hypomethylated on mat11 and bound by the CTCF protein, which recruits other proteins to act as a barrier, blocking enhancer elements close to *H19* from activating the distant *IGF2* gene. On pat11, ICR1 is extensively methylated and is not bound by CTCF, allowing the enhancer elements to preferentially activate *IGF2* instead of *H19*. ICR2 is located in an intron of *KCNQ1* and acts as a promoter for the antisense RNA gene *KCNQ1OT1*, which encodes a *cis*-acting suppressor RNA. ICR2 is hypomethylated on pat11, allowing transcription of the *KCNQ1OT1* suppressor RNA, which inhibits the transcription of the neighboring genes *KCNQ1OT1* and *CDKN1C*. On mat11, ICR2 is extensively methylated, blocking transcription of *KCNQ1OT1* and allowing the expression of the neighboring genes.

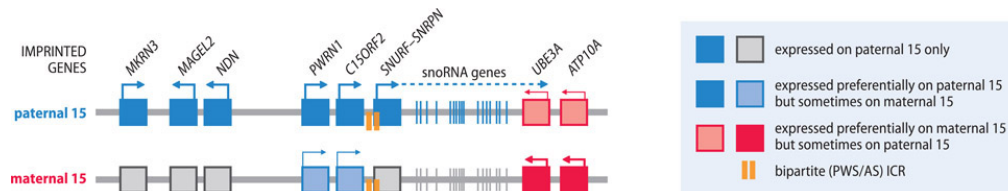


Figure 6.26 The imprinted gene cluster at 15q11–q12 associated with Prader-Willi syndrome (PWS) and Angelman syndrome (AS).

Arrows show the direction of transcription of the indicated genes. The prominent long dashed blue arrow signifies that there is a long transcription unit with multiple noncoding exons that has been proposed to overlap the *UBE3A* gene. Numerous snoRNA genes (individual vertical blue lines) are found in introns of the long transcription unit, two of which are present in multiple copies: *SNORD116* (previously HBII-85) and *SNORD115* (previously HBII-52). A bipartite imprinting control region (ICR) is located near the promoter region of *SNURF-SNRPN*.

The imprinted gene cluster at 11p15.5

A cluster of at least 10 imprinted genes in the subtelomeric 11p15.5 region has been well studied because of associations with Beckwith-Wiedeman syndrome (PMID 20301568) and many cases of Silver-Russell syndrome (PMID 20301499). The gene cluster has two different imprinting control regions, ICR1 and ICR2.

Figure 6.25A shows some key 11p15.5 genes regulated by ICR1 and ICR2. Both ICR1 and two nearby enhancer elements regulate *IGF2* (insulin growth factor type 2; paternally expressed) and *H19* (which makes a maternally expressed ncRNA). ICR2 regulates the *KCNQ1* gene (which makes a maternally expressed potassium channel), the *KCNQ1OT1* antisense RNA transcript (*KCNQ1* opposite strand transcript 1; paternally expressed), and *CDKN1C* (a suppressor of cell proliferation; maternally expressed).

ICR1 and ICR2 are each activated when hypomethylated and suppressed when extensively methylated. But they have opposite parental imprints: a maternally inherited chromosome 11 has a hypomethylated ICR1 and extensively methylated ICR2, but a paternal chromosome 11 has an extensively methylated ICR1 and a hypomethylated ICR2. They also use rather different control mechanisms (see **Figure 6.25B**).

Disease can result from significant changes in the methylation patterns or the base sequences of the ICRs and key imprinted genes. The most frequent cause of Silver-Russell syndrome is hypomethylation of ICR1 on both chromosome 11s so that both *IGF2* alleles are silenced. This can often happen by maternal 11 disomy or duplication of maternal 11p (see **Table 6.8**).

Beckwith-Wiedemann syndrome is marked by fetal overgrowth. It can occur when ICR1 is extensively methylated on both chromosome 11s so that both *IGF2* alleles are expressed, causing excessive growth. It can also occur when ICR2 is hypomethylated on both chromosome 11s, causing silencing of both alleles of the growth-restricting gene *CDKN1C*. Paternal disomy 11 is a common cause.

The imprinted gene cluster at 15q11-12

Another well-studied imprinted gene cluster located at 15q11–12, is associated with two neurodevelopmental disorders: Angelman syndrome and Prader-Willi syndrome. Angelman syndrome phenotypes include severe

intellectual disability and microcephaly, and affected individuals are prone to frequent laughter and smiling. In Prader-Willi syndrome (PWS), affected individuals show mild intellectual disability and hyperphagia leading to obesity.

A single, bipartite imprinting control region regulates the whole cluster, which contains very many imprinted genes that are expressed only or preferentially on the paternal chromosome 15, including many small nucleolar RNA (snoRNA) genes, but just two imprinted genes that are preferentially expressed on the maternal chromosome 15 ([Figure 6.26](#)). The imprinted snoRNA genes seem to be located within introns of an extended transcription unit that includes a few exons making two proteins (SNURF and SNURPN) plus a large number of noncoding exons. By overlapping *UBE3A* and possibly *ATP10A*, the very long transcripts might silence these two genes on paternal chromosome 15.

Angelman syndrome and PWS are both caused by genetic deficiency. In the former case, the key problem is loss or inactivation of the maternal *UBE3A* allele, which makes a ubiquitin-protein ligase (both *UBE3A* alleles are normally expressed in most tissues, but in neurons only maternal *UBE3A* is active). The PWS phenotype, however, is attributable to the deficient expression of different genes normally expressed only on the paternal chromosome 15, including *NDN*, which regulates adipogenesis, and *SNORD116* / *HBII-85* genes, which make a type of snoRNA that, in addition to its standard snoRNA role, also acts to regulate alternative splicing in some target genes.

Angelman syndrome and PWS are primarily due to large deletions that remove the same 5 Mb region of DNA including the genes shown in [Figure 6.26](#), either from maternal 15q11-q12 (Angelman syndrome) or paternal 15q11-q12 (PWS). This region is flanked by low-copy-number repeat sequences that make it inherently prone to instability. However, some cases may result from point mutation—see [Clinical Box 4](#) for an example.

CLINICAL BOX 4 A CASE STUDY: ANGELMAN SYNDROME

Keira was referred to the Genetics Clinic by her Pediatrician at 20 months of age. She was her parents' first child and her mother, Marie, was 20 weeks pregnant at the time of referral. Born after a normal pregnancy and delivery, Keira had difficulty breast feeding and was therefore bottle-fed. She was a colicky baby who frequently vomited feeds and at three months was diagnosed as having gastroesophageal reflux. This resolved by one year of age. Her parents were concerned that her motor development was delayed by the time she was 6 months old. She started crawling at 12 months and by 18 months was able to pull to stand and cruise. At 20 months she was not yet pointing and had no single words. She was an extremely happy baby who was always smiling and laughing, but had occasional episodes, each lasting a few seconds, where she would go quiet and be unresponsive.

On examination, the geneticist noticed that she had jerky movements and walked holding hands with feet turned out and an unsteady wide based gait. She drooled copiously, laughed a lot and clapped her hands together frequently. She had a relatively small head circumference in comparison to height and weight and a wide mouth. The geneticist suspected that Keira had Angelman syndrome and requested a chromosomal microarray and methylation studies of the Angelman ICR. There was no evidence of a microdeletion of chromosome 15q11-13 or loss of the unmethylated (maternal) *SNRPN* allele. Targeted clinical exome analysis of a panel of 96 genes associated with Syndromic Intellectual Disability was performed and showed a heterozygous pathogenic SNV in the *UBE3A* gene: c.1749dup p.(Glu584Ter). Parental testing showed that Keira had inherited the pathogenic *UBE3A* variant from her mother, Marie ([Figure 1](#)).

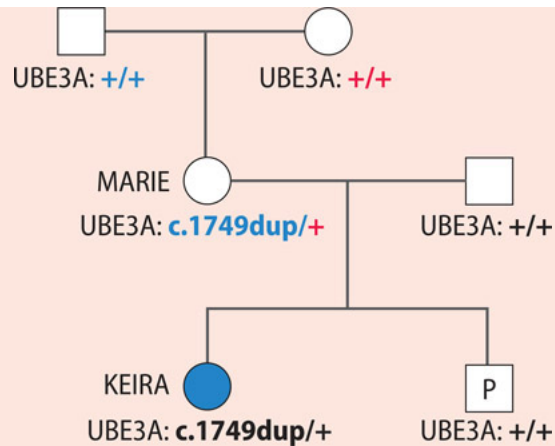


Figure 1

The pregnancy had therefore been at 50 % risk of having Angelman syndrome. Analysis of the cord blood of Keira’s baby brother did not detect the pathogenic *UBE3A* variant and he has shown no clinical features suggestive of Angelman syndrome. Marie has been offered prenatal diagnosis in a future pregnancy. Marie’s parents both tested negative for the *UBE3A* variant. The *UBE3A* variant is a *de novo* variant in Marie, but she does not have Angelman syndrome, having inherited the variant from her father. *UBE3A* is paternally imprinted in the brain, and so the maternal *UBE3A* allele is active but the paternal allele is not expressed. The c.1749dup p. (Glu584Ter) variant is not uncommon—see [Figure 2](#) for an unrelated boy with Angelman syndrome carrying this variant *de novo*.



Figure 2

Imprinting and assisted reproduction

Another aspect of imprinting disorders relates to concerns about apparently increased frequencies of these disorders in births in which assisted reproductive technology (ART) has been employed. *In vitro* fertilization is now well accepted in economically advanced societies, where it accounts for 1–4 % of births. Because early embryogenesis is a critical time for epigenetic regulation and is sensitive to environmental factors, ART might impose added stress on embryos that can result in altered epigenetic profiles.

Imprinting disorders are very rare, and statistical support for increased incidence in assisted conception is difficult to achieve. But studies in mice have shown that although intracytoplasmic sperm injection does introduce

primary epimutations, they are normally corrected by epigenetic reprogramming in the germline and are therefore not transmitted to subsequent generations.

SUMMARY

- Genetic regulation of gene expression is dictated by the base sequence; epigenetic regulation of gene expression is independent of the base sequence.
- Epigenetic regulation of gene expression is routinely achieved by controlling chromatin structure. According to need, “open” chromatin structures form in some regions (allowing access to transcription factors); in other regions the chromatin is highly condensed, and transcriptionally inactive.
- *Cis*-acting regulatory sequences are located on the same *individual* DNA or RNA molecule as the sequences they regulate.
- *Trans*-acting gene regulators migrate in the cell to bind target sequences on DNA or RNA molecules. *Trans*-acting regulatory proteins bind to targets by using nucleic acid-binding domains; *trans*-acting regulatory RNA molecules bind by base pairing.
- Gene promoters are composed of multiple short sequences. Core promoter elements are bound by ubiquitous transcription factors; other *cis*-acting regulatory elements are often bound by tissue-specific or developmental-stage-specific regulators.
- RNA splicing is largely dependent on recognition of *cis*-acting RNA sequence elements at splice junctions. Additional splice enhancer and splice suppressor sequences can be located in both introns and exons.
- Multiple different transcripts are produced for almost all of our genes and can give rise to alternative protein isoforms, increasing functional variation.
- Post-transcriptional regulation is often performed by microRNAs. An individual miRNA binds to partly complementary sequences in untranslated regions of multiple target mRNA sequences and downregulates expression.
- Open and condensed chromatin structure depends heavily on the extent of chemical modification of both DNA (via methylation of certain cytosines) and histones (notably by acetylation, methylation and phosphorylation of the side chains of certain amino acids).
- DNA methylation and histone modification patterns are prominent examples of epigenetic marks that can be stably inherited from one cell generation to the next. The appropriate chemical groups are added or removed by dedicated enzymes known, respectively, as “writers” or “erasers”.
- Methylated CG dinucleotides and chemically modified amino acids on histones are bound and interpreted by specific proteins (“readers”) that induce structural and functional changes in chromatin.
- ATP-dependent chromatin remodeling complexes can slide nucleosomes along the DNA to increase or decrease the spacing between nucleosomes, respectively allowing or denying access to transcription factors.
- *Cis*-acting long noncoding RNAs also work in epigenetic gene regulation in mammalian cells. They remain attached to the DNA strand from which they were transcribed and can serve as antisense RNAs (inhibiting expression of sense RNAs transcribed from the complementary DNA strand) or by forming scaffolds for binding certain *trans*-acting protein complexes.
- Genomic imprinting in mammals is an epigenetic phenomenon whereby certain genes are either expressed or silenced according to whether they reside on a paternally transmitted chromosome, or a maternal one.

- X-inactivation means that one of the two X chromosomes in women (and female mammals) is heterochromatinized.
- Barrier elements separate heterochromatin from neighboring euchromatin regions. If they are deleted or relocated by inversions or translocations, the neighboring euchromatin region can be heterochromatinized, causing gene silencing (a position effect).
- Uniparental disomy means that a pair of homologous chromosomes has been inherited from one parent. Disease can result when both chromosomes carry one or more imprinted genes.

QUESTIONS

Questions can be downloaded by visiting the following link, under Support Materials: www.routledge.com/9780367490812.

FURTHER READING

Enhancers, silencers, insulators, and general gene regulation

Ali T (2016) Insulators and domains of gene expression. *Curr Opin Genet Dev* 37:17–26; PMID 26802288.
 Kolovos P (2012) Enhancers and silencers—an integrated and simple model for their function. *Epigen Chromatin* 5:1; PMID 22230046.
 Latchman DS (2015) *Gene Control*, 2nd ed. Garland Science.

Alternative splicing and RNA editing

Kim E (2008) Alternative splicing: current perspectives. *BioEssays* 30:38–47; PMID 18081010.
 Tang W (2012) Biological significance of RNA editing in cells. *Mol Biotechnol* 52:1–100; PMID 22271460.

MicroRNAs and competing endogenous RNAs

Baek D (2008) The impact of microRNAs on protein output. *Nature* 455:64–71; PMID 18668037.
 Poliseno L (2010) A coding-independent function of gene and pseudogene mRNAs regulates tumour biology. *Nature* 465:1033–1038; PMID 20577206.

Epigenetics in gene regulation (general)

Bonasio R (2010) Molecular signals of epigenetic states. *Science* 330:612–616; PMID 21030644.
 Portela A & Esteller M (2010) Epigenetic modifications and human disease. *Nat Biotechnol* 28:1057–1068; PMID 20944598.

Histone modifications and DNA methylation

Chen Z & Riggs AD (2011) DNA methylation and demethylation in mammals. *J Biol Chem* 286:18347–18353; PMID 21454628.
 Deaton AM & Bird A (2011) CpG islands and the regulation of transcription. *Genes Dev* 25:1010–1022; PMID 21576262.
 Smallwood SA & Kelsey G (2012) De novo DNA methylation: a germ cell perspective. *Trends Genet* 28:33–42; PMID 22019337.

Suganuma T & Workman JL (2011) Signals and combinatorial functions of histone modifications. *Annu Rev Biochem* 80:474–499; PMID 21529160.

Long noncoding RNA in genetic and epigenetic regulation

Gil N & Ulitsky I (2020) Regulation of gene expression by *cis*-acting long non-coding RNAs. *Nature Rev Genet* 21:102–117; PMID 31729473.

Mercer TR & Mattick S (2013) Structure and function of long noncoding RNAs in epigenetic regulation. *Nat Struct Mol Biol* 20:300–307; PMID 23463315.

Statello L (2021) Gene regulation by long non-coding RNA and its biological functions. *Nat Rev Mol Cell Biol* 22:96–118; PMID 33353982.

Genomic imprinting, X-inactivation, heterochromatin spreading

Barkess G & West AG (2012) Chromatin insulator elements: establishing barriers to set heterochromatin boundaries. *Epigenomics* 4:67–80; PMID 22332659.

Barlow DP & Batolomei MS (2014) Genomic imprinting in mammals. *Annu Rev Genet* 45:379–403; PMID 21942369.

Galupa R & Heard E. (2018) X-chromosome inactivation: a crossroads between chromosome architecture and gene regulation. *Annu Rev Genet* 52:535–566; PMID 30256677.

Haig D (2000) Kinship theory of genomic imprinting. *Annu Rev Ecol Syst* 31:9–32.

Epigenetic regulation and disease

Demars J & Gicquel C (2012) Epigenetic and genetic disturbance of the imprinted 11p15 region in Beckwith-Wiedemann and Silver-Russell syndromes. *Clin Genet* 81:350–361; PMID 22150955.

Hahn M (2010) Heterochromatin dysregulation in human diseases. *J Appl Physiol* 109:232–242; PMID 20360431.

Lemmers RJLF (2010) A unifying genetic model for facioscapulohumeral muscular dystrophy. *Science* 329:1650–1653; PMID 23143600.

Lemmers RJ (2012) Digenic inheritance of an *SCMD1* mutation and an FSHD-permissive D4Z4 allele causes facioscapulohumeral muscular dystrophy type 2. *Nat Genet* 44:1370–1374; PMID 20724583.

Sahoo T (2008) Prader-Willi phenotype caused by paternal deficiency for the HBII-85 C/D box small nucleolar RNA cluster. *Nat Genet* 40:719–721; PMID 18500341.

Soellner L (2017) Recent advances in imprinting disorders. *Clin Genet* 91:3–13; PMID 27363536.

7

How genetic variation in DNA and chromosomes causes disease

DOI: [10.1201/9781003044406-7](https://doi.org/10.1201/9781003044406-7)

CONTENTS

[7.1 AN OVERVIEW OF HOW GENETIC VARIATION RESULTS IN DISEASE](#)

[7.2 PATHOGENIC NUCLEOTIDE SUBSTITUTIONS AND TINY INSERTIONS AND DELETIONS](#)

[7.3 PATHOGENESIS DUE TO VARIATION IN SHORT TANDEM REPEAT COPY NUMBER](#)

[7.4 PATHOGENESIS TRIGGERED BY LONG TANDEM REPEATS AND INTERSPERSED REPEATS](#)

[7.5 CHROMOSOME ABNORMALITIES](#)

[7.6 MOLECULAR PATHOLOGY OF MITOCHONDRIAL DISORDERS](#)

[7.7 EFFECTS ON THE PHENOTYPE OF PATHOGENIC VARIANTS IN NUCLEAR DNA](#)

[7.8 A PROTEIN STRUCTURE PERSPECTIVE OF MOLECULAR PATHOLOGY](#)

[7.9 GENOTYPE-PHENOTYPE CORRELATIONS AND WHY MONOGENIC DISORDERS ARE OFTEN NOT SIMPLE](#)

[SUMMARY](#)

[QUESTIONS](#)

FURTHER READING

In [Chapter 4](#), we outlined some basic principles of genetic variation. We covered the different types of genetic variation in our genome: both large-scale changes that make up structural variation and point mutations. And we related how sequence variation at the level of gene product relates to sequence variation at the level of DNA. Here we begin to consider the small fraction of human genetic variation that causes disease.

In this chapter we describe the different mechanisms that cause pathogenic DNA changes in genes and their consequences for the phenotype. The emphasis is on rare, highly penetrant variants that cause single-gene disorders, and on chromosome abnormalities. We do however take a broad view of protein dysregulation in disease. In [Chapter 8](#) we go on to consider variants of low penetrance that confer susceptibility to common complex disorders, and in [Chapter 10](#) we examine how genetic variation, predominantly somatic mutations, contributes to cancer.

In [Section 7.1](#) we give an overview of how genetic variation results in disease. According to the number of nucleotides that are changed (and the number of genes affected), we will consider the pathogenic changes as occurring at different levels (different mutation mechanisms can be involved, depending on the size of the DNA change). We describe in [Section 7.2](#) different kinds of pathogenic point mutations. They typically change just a single nucleotide (the most common mutation), or a very few nucleotides, and primarily affect the expression of a single gene.

As described in [Sections 7.3](#) and [7.4](#), various genetic mechanisms also produce moderate-to large-scale mutations (affecting from tens of nucleotides to a few megabases of DNA); they are often triggered by inappropriate pairing of repetitive DNA sequences. Some very large-scale mutations and chromosome breaks produce recognizable changes at the chromosome level as detected by light microscopy. They will be covered in [Section 7.5](#) together with other types of chromosome abnormality resulting from errors in chromosome segregation and recombination.

As initially considered in [Chapter 5](#), the link between deleterious mutations and disease phenotype is often not straightforward. A deleterious mutation that might be expected to result in a single-gene disorder, according to the principles described in [Chapter 5](#), may produce different degrees of disease severity, or no disease at all. We examine factors that complicate the link between genetic variation and disease phenotype, beginning in [Section 7.6](#) with the tiny mitochondrial genome: it has just

37 genes but is nevertheless important in disease, and has unique properties that complicate the molecular pathology. Then in [Section 7.7](#) we go on to describe factors affecting phenotypes that result from variation in the nuclear genome.

In [Section 7.8](#) we examine downstream effects at the protein structure level, notably how altered protein folding, and protein aggregation contribute to disease phenotypes. Finally, in [Section 7.9](#) we consider the difficulties in correlating genotypes and phenotypes, and we give examples of how the phenotype of a monogenic disorder can be influenced by various factors including genetic variation at other gene loci and environmental factors.

7.1 AN OVERVIEW OF HOW GENETIC VARIATION RESULTS IN DISEASE

The great majority of variation in our DNA appears to be without consequence. For the most part, that happens because just a small percentage of our genome is functionally important (the great majority of nucleotides within introns and in extragenic DNA can be changed by small mutations without any obvious effect on the phenotype). A second, and minor, reason is genetic redundancy: some genes are present in multiple, almost identical copies—an inactivating mutation in a single ribosomal RNA gene in nuclear DNA has no effect because each type of cytoplasmic rRNA is made by hundreds of extremely similar gene copies.

Pathogenic mutations do not occur haphazardly in our DNA. For example, single nucleotide substitutions, the most common type of pathogenic mutation, are not random: certain types of DNA sequence are more vulnerable to point mutation (mutation *hotspots*). And, as detailed below, different arrangements of repetitive DNA also predispose to different classes of mutation, including many large-scale DNA changes. The genetic variation that causes disease may do so by causing two broad changes at the level of gene product, as listed below.

- A change in the **sequence** of the gene product. The result may be a total loss of function—the mutant gene product is not produced, or is incapable of carrying out its normal task. Or it may have a significantly reduced ability to work normally (as a result of a hypomorphic mutation). Sometimes, it may acquire an altered function causing it to inhibit the working of a normal gene product produced by the other parental allele. Rarely, mutation causes the gene product to have a new function that is harmful in some way (causing

cells to die or to behave inappropriately). As we will see, loss of function and gain of function quite often involve a change in protein structure.

- A change in the **amount** of the gene product. This can happen in three major ways, as listed below.

a. *Change in gene copy number.* Whole gene deletion or duplication can result from different mechanisms, including short-range unequal crossover ([Section 7.4](#)), and also errors in chromosome segregation and recombination ([Section 7.5](#)). In addition, gene amplification (see [Figure 7.1](#)) is common in many cancers, as explained in [Chapter 10](#).

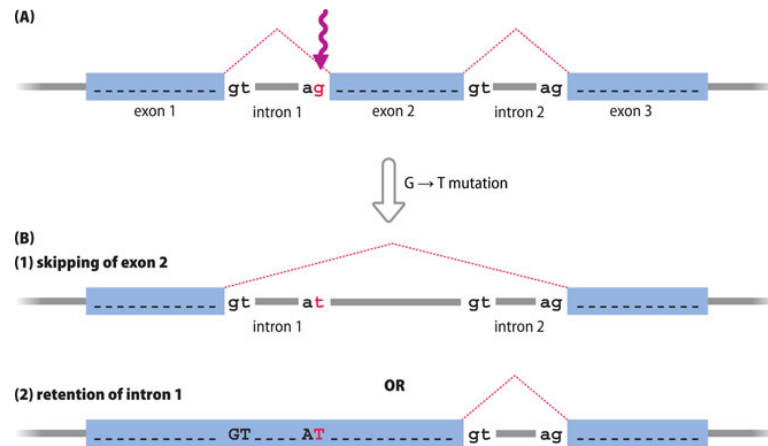


Figure 7.3 Splice-site mutations can cause exon skipping or intron retention. Exon sequences enclosed within blue boxes are represented by dashed lines, with upper-case letters indicating key individual nucleotides. Intron sequences are represented as gray horizontal lines with lower-case letters indicating key nucleotides. Dashed red lines show the positions where splicing will occur later *at the RNA level* to bring together transcribed exon sequences within RNA transcripts. (A) A normal situation with exons separated by introns with conserved 5' GT and 3' AG terminal dinucleotides. (B) Alternative outcomes of a G → T mutation at the conserved 3' terminal nucleotide of intron 1, inactivating the splice acceptor site. Outcome (1): *exon skipping*. Using the next available splice acceptor site (at the 3' end of intron 2) causes skipping of exon 2 and a frameshift if the number of nucleotides in exon 2 is not a multiple of three. Outcome (2): *intron retention*. Alternatively, splicing of intron 1 is abandoned and the intron 1 sequence is retained in the

mRNA, forming part of a large exon that also contains the original exon 1 and exon 2 sequences.

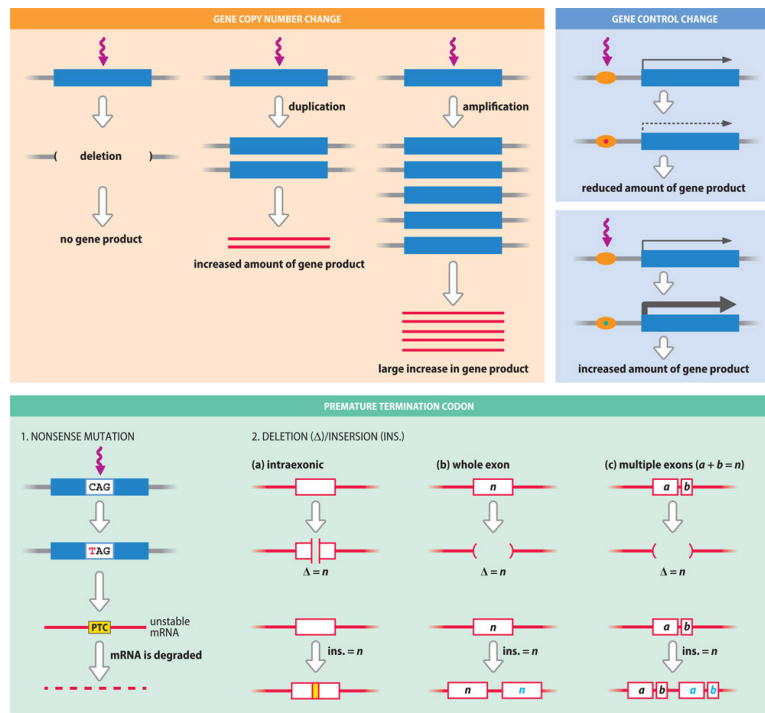


Figure 7.1 Classes of mutations that cause disease by altering the amount of a gene product. Mutations leading to premature termination codons (PTCs) frequently activate a pathway that leads to mRNA destruction (as explained in [Box 7.1](#)), and failure to make a protein. They are usually caused by nonsense mutations or by small or large frameshifting deletions and insertions (note: large insertions are often caused by duplications of one or more exons). If we denote the total number of inserted or deleted nucleotides of coding DNA as n , then a frameshift occurs when n is not exactly divisible by three. Occasionally (not shown here), a PTC may occur as a by-product of mutation that causes certain types of aberrant RNA splicing (exon skipping or intron retention, as described in [Figure 7.3](#)).

- b. *Change in gene regulation.* Gene expression can be very significantly impacted by mutations affecting important upstream or intragenic regulatory sequences (or by epigenetic changes such as imprinting and position effects as described in [Sections 6.2](#) and [6.3](#))
- c. *Premature termination codon.* Premature termination codons often result in unstable mRNA and no protein product. They arise principally from two classes of mutation: nonsense mutations and

frame-shifting insertions and deletions ([Figure 7.1](#) gives an overview).

Although most pathogenic mutations affect individual genes, some mutations (and chromosome abnormalities) can *simultaneously* affect multiple genes. Large-scale deletions and duplications, for example, result in a simultaneous change in the copy number of multiple genes, with adverse effects. Additionally, mutations in some genes that produce *trans*-acting regulators can indirectly have consequences for multiple different target genes that they regulate.

The importance of repeat sequences in triggering pathogenesis

As detailed in [Section 2.5](#), our genome has many types of repetitive DNA sequences, and various types of DNA duplication have been important in shaping the genome to allow the development of biological complexity. But despite the evolutionary advantage they confer, repeat sequences have a downside: they often predispose DNA molecules to undergo inappropriate pairing and subsequent sequence changes resulting in disease, and some repeats can cause disease by spontaneously and unpredictably inserting into the genome.

Pairing of DNA sequences in proper register is important at two different levels: (i) the paired individual strands of each DNA double helix and (ii) the paired double-stranded DNAs of chromatids of the same chromosome. *Tandem repeats*—those that are neighbours in a head to tail fashion (® ® ®)—can cause misalignment of the two single strands of a DNA helix and of the two double-stranded DNAs of paired chromatids. Resulting misaligned DNA strands in a double helix can undergo small deletions and duplications; misaligned chromatids can trigger moderate to large-scale deletions and duplications.

Interspersed repeats that may reside on the same nuclear DNA or mtDNA molecule, or on different DNA molecules, can also pair inappropriately with other repeats of the same type, predisposing to abnormal sequence exchanges. And certain families of interspersed *transposon repeat* can make copies of themselves that are able to insert elsewhere in the genome, including into genes. We cover the different mechanisms in later sections, but [Table 7.1](#) gives an overview.

TABLE 7.1 AN OVERVIEW OF HOW DNA REPEATS FREQUENTLY PREDISPOSE TO PATHOGENESIS

DNA repeats		Mechanisms (major mechanisms in italics)	Examples
Class	Involvement		
Tandem repeats	Short tandem repeats stabilize local mispairing of the two DNA	<i>Replication slippage</i> (DNA replication after slipped strand mispairing, causing strands to have fewer or more repeats)	Simple short tandem repeat variation causing insertions or deletions (Section 7.3).
	strands of a single DNA helix (<i>slipped strand mispairing</i>)	Mechanism poorly understood but involves slipped strand mispairing and DNA repair	Unstable short tandem repeat expansion (Section 7.3)
	Large tandem repeats stabilize local mispairing of two sister chromatids or two non-sister chromatids of a chromosome (Section 7.4).	Mispairing of large tandem repeats on opposing chromatids predisposes to: <i>unequal crossover (UEC)</i> ; <i>unequal sister chromatid exchange (UESCE)</i> ; and <i>gene conversion</i> , thereby producing deletions, duplications, or mutant genes	99% of cases of 21-OH deficiency are due to sequence exchanges occurring between mispaired large repeats containing the 21-hydroxylase gene and a related pseudogene (Clinical Box 6)
Interspersed repeats	Abnormal pairing of interspersed repeats in nuclear DNA	<i>UEC</i> , <i>UESCE</i> , or <i>intrachromatid recombination</i> causing deletions, duplications or inversions	Inversion in <i>F8</i> gene causing hemophilia A (Figure 7.11)
	Insertional inactivation by transposon repeat	<i>Retrotransposition</i> (cDNA of retro-transposon repeat inserts into genome)	
	Abnormal pairing of interspersed repeats in mtDNA (Section 7.6)	Mispairing of almost identical short repeats in mtDNA followed by	Deletions in mtDNA (Table 7.13)

DNA repeats		Mechanisms (major mechanisms in <i>italics</i>)	Examples
Class	Involvement		
		cleavage and rejoining of fragments	

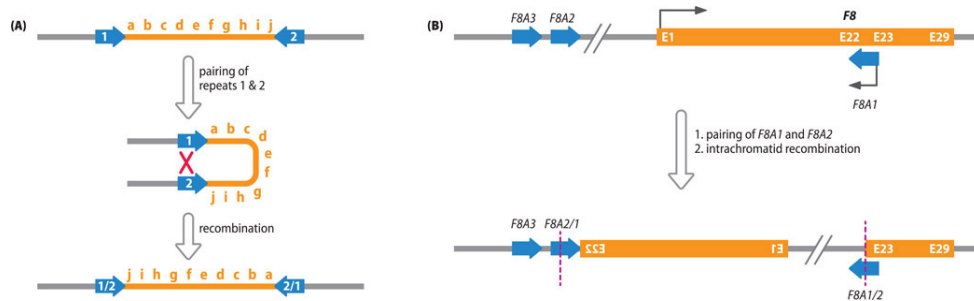


Figure 7.11 Intrachromatid recombination between inverted repeats produces inversions and is a common cause of hemophilia A. (A) Inverted repeats 1 and 2 on the same DNA strand can mispair by inducing looping of the intervening DNA. Subsequent recombination within the paired repeats produces hybrid repeat sequences (1/2 and 2/1) and inversion of the intervening DNA. (B) In about 50 % of cases with hemophilia A, the mutation is a large inversion that disrupts the blood clotting factor VIII gene (*F8*). The 191 kb *F8* gene has 29 exons, and within the large intron 22 is a small gene, *F8A1*, that is transcribed from the opposite strand. *F8A1* is a member of a family of low-copy-number repeats that includes two closely related sequences, *F8A2* and *F8A3*, located upstream of *F8*. Mispairing between either of these repeats and *F8A1* can induce an inversion, by looping out of the intervening DNA to allow recombination between the mismatched repeats (such as *F8A1* and *F8A2*, as shown here; the dashed red vertical lines mark the boundaries of the inversion). The resulting inversion disrupts the *F8* gene, splitting it into two oppositely oriented fragments, one containing exons 1–22 and the other from exons 23–29.

7.2 PATHOGENIC NUCLEOTIDE SUBSTITUTIONS AND TINY INSERTIONS AND DELETIONS

Pathogenic single nucleotide substitutions within coding sequences

A single nucleotide substitution within a coding sequence has the effect of replacing one codon in the mRNA by another codon. There is, however, substantial redundancy in the universal genetic code: as explained below, all amino acids other than methionine and tryptophan are specified by multiple codons (from two to six).

As a result of the redundancy in the genetic code, a mutated codon quite often specifies the same amino acid as the original codon. A coding sequence substitution

such as this—one that does not change an amino acid—is known as a **synonymous substitution** (sometimes called a **silent mutation**). Because there is no change in amino acid, no change in phenotype might be expected. Nevertheless, as discussed below, a minority of synonymous substitutions nevertheless cause disease, almost always by simultaneously altering RNA splicing.

The alternative is a **nonsynonymous substitution**. There are different classes ([Table 7.2](#)), but the predominant one causes one amino acid to be replaced by another, a **missense mutation**. Missense mutations sometimes have minimal effects on the phenotype, but some of them have adverse effects as described below.

TABLE 7.2 CLASSES OF NONSYNONYMOUS MUTATION

Class	Definition	Example and comments
Missense mutation	an amino-acid-specifying codon is replaced by a codon for a different amino acid	GGA (glycine) is replaced by CGA (arginine). The effect is greatest when the replacement amino acid has very different physiochemical properties
Nonsense (stop-gain) mutation	an amino acid-specifying codon is replaced by a premature stop codon	a G → T substitution may result in codon GGA (glycine) being replaced by UGA (stop). Results in unstable mRNA or production of a truncated protein (see Box 7.1)
Stop-loss mutation	a natural stop codon is replaced by an amino acid-specifying codon	a T → G substitution may result in UGA (stop) being replaced by GGA (glycine), with translational <i>read-through</i> (the first part of the 3' untranslated region is translated and the protein has an extended C-terminus [*])

* The 3' untranslated region will have termination codons in all three reading frames. As a result, in the case of a stop-loss mutation the “read-through” past the normal stop codon usually picks up another termination codon quite quickly, so that often the extended C-terminus is not long. The effect on protein function may often not be large unless the extended C-terminus causes problems for protein folding or protein stability.

Relative frequencies of silent and amino-acid-replacing substitutions

The relative frequencies of single nucleotide substitutions that are silent and those that cause an amino acid to be replaced (missense mutations) vary according to the

base position in a codon. If we first consider the genes in nuclear DNA, 61 codons can specify an amino acid ([Figure 7.2](#)). If we take an average, two out of every three substitutions at the third base position in a codon are silent; by contrast, 100 % of substitutions at the second base position and 184 out of 192 (about 96 %) of substitutions of the first base are nonsynonymous.

	AAA		Lys		CAA		Gln		GAA		Glu		UAA		STOP
	AAG				CAG				GAG				UAG		
	AAC		Asn		CAC		His		GAC		Asp		UAC		Tyr
	AAU				CAU				GAU				UAU		
	ACA				CCA				GCA				UCA		
	ACG		Thr		CCG		Pro		GCG		Ala		UCG		Ser
	ACC				CCC				GCC				UCC		
	ACU				CCU				GCU				UCU		
STOP	AGA		Arg		CGA				GGA			Trp	UGA		STOP
	AGG				CGG		Arg		GGG		Gly		UGG		Trp
	AGC		Ser		CGC				GGC				UGC		Cys
	AGU				CGU				GGU				UGU		
Met	AUA		Ile		CUA				GUA				UUA		Leu
	AUG		Met		CUG		Leu		GUG		Val		UUG		
	AUC		Ile		CUC				GUC				UUC		Phe
	AUU				CUU				GUU				UUU		

Figure 7.2 The genetic code. Pale gray bars to the right of codons identify 60 codons interpreted in the same way for nuclear and mitochondrial mRNA. Four codons—AGA, AGG, AUA, and UGA—are interpreted differently. Flanking blue bars and lettering to the right show the interpretation for nuclear genes; pale pink bars on the left show the interpretation for genes in mtDNA. The “universal” genetic code (for nuclear genes) has 61 codons specifying 20 different amino acids, with different levels of redundancy: from unique codons (Met, Trp) to sixfold redundancy (Arg, Leu, Ser). The remaining three codons—UAA, UAG, and UGA—normally act as stop codons (however, UGA can occasionally specify a 21st amino acid, selenocysteine, and UAG can occasionally specify glutamine). For genes in mtDNA, 60 codons specify an amino acid, and there are four stop codons (AGA, AGG, UAA, and UAG).

Consider, for example, a G → A substitution that results in replacement of codon GGG by GGA. The genetic code shows that both the original GGG codon and the replacement GGA codon specify the amino acid glycine. If the substitution had been G → C or G → T (to give GGC or GGU codons), the altered codon again would have specified glycine. Like glycine, several other amino acids are exclusively, or largely,

determined by the first two bases of a codon—there is flexibility in how the third base of a codon pairs with the 5¢ base of the anticodon (*base wobble*).

Genetic redundancy at the first base position is responsible for silent substitutions in some arginine and leucine codons. Thus, codons **AGA** and **AGG** specify arginine, as do codons **CGA** and **CGG** (an A ↔ C or C ↔ A change at the first base position is silent in these cases). Similarly, codons **CUA** and **CUG** specify leucine, as do codons **UUA** and **UUG**.

Conservative substitution: replacing an amino acid by a similar one

Nucleotide substitutions that change an amino acid can have different effects, according to the degree to which the replacement amino acid differs from the original amino acid (based on properties such as polarity, molecular volume, and chemical composition—see below). Perhaps fewer than 30 % of substitutions have no, or very little, functional significance; the remainder are roughly equally split into those with weak to moderate negative effects on protein function, and those with strongly negative effects.

A nucleotide substitution that replaces one amino acid by another of the same chemical class is a **conservative substitution** and often has minimal consequences for how the protein functions. [Table 7.3](#) lists different chemical classes of amino acids plus some distinguishing features of individual amino acids (see [Figure 2.2](#) for the chemical structures).

TABLE 7.3 AMINO ACIDS CAN BE GROUPED INTO SIX CLASSES ACCORDING TO THE CHEMICAL PROPERTIES OF THEIR SIDE CHAINS

Common feature of side chain		Amino acids*	Comments
Polar	Basic (positively charged)	Arg (R); Lys (K); His (H)	arginine and lysine have simple side chains with an amino ion ($-NH_3^+$); histidine has a more complex side chain with a positively charged imido group

* See [Figure 2.2](#) for structures of amino acids.

Common feature of side chain		Amino acids*	Comments
	Acidic (negatively charged)	Asp (D); Glu (E)	simple side chains ending with a carboxyl ion (-COO-)
	Amide group	Asn (N); Gln (Q)	simple side chains ending with a -CONH ₂ group
	Hydroxyl group	Ser(S); Thr(T); Tyr(Y)	serine and threonine have short simple side chains with a hydroxyl group; tyrosine has an aromatic ring
	Polar with sulfhydryl (-SH) group	Cys(C)	disulfide bridges (-S-S-) can form between <i>certain</i> distantly spaced cysteines in a polypeptide and are important in protein folding
Nonpolar		Gly (G); Ala (A); Val (V); Leu (L); Ile (I); Pro (P); Met (M); Phe (F); Trp (W)	glycine has the simplest possible side chain—a single hydrogen atom. Phenylalanine and tryptophan have complex aromatic side chains

* See [Figure 2.2](#) for structures of amino acids.

Nonconservative substitutions: effects on the polypeptide/protein

Replacing one amino acid by another belonging to a different chemical class may be expected to have more significant consequences. A key factor is whether the individual amino acid has an important role in the function of the protein. It might play a vital role at the activation site of an enzyme, for example, be a critical part of a specific recognition sequence used to bind some interacting molecule, or have a side chain that needs to be chemically modified in a specific way for the protein to be functionally active.

Additional factors include the potential effects on protein folding and protein structure (for a brief summary of protein structure, see [Section 2.1](#) and [Box 2.2](#)).

Thus, for example, for thermodynamic reasons, globular proteins usually fold so that nonpolar, uncharged amino acids are buried in the interior and polar amino acids are on the outside, exposed to what is usually a hydrophilic aqueous environment; substitutions that change this pattern may induce incorrect protein folding.

Some amino acids are not tolerated in certain structural elements. Thus, for example, proline cannot be accommodated in an α -helix: if an amino acid is substituted by a proline the α -helix is disrupted. Conversely, certain amino acids have specific structural roles. Glycine (with the smallest possible side chain—a single hydrogen atom), and proline (the only amino acid in which the side chain loops back to rejoin the polypeptide backbone) are important in allowing the polypeptide backbone to bend sharply. They often have important roles in protein folding. The triple-stranded helical structure of collagens, for example, requires glycine at about every third residue; prolines (and hydroxyprolines) are also extremely frequent in collagens.

Cysteine has a unique role in protein folding. The sulfhydryl ($-\text{SH}$) groups on *certain* distantly located cysteines on the same polypeptide may interact to form a disulphide bridge ($-\text{S}-\text{S}-$); this can be important in establishing globular domains (such as for the immunoglobulin superfamily proteins in [Figure 4.10](#)). Replacing either cysteine by any other amino acid breaks the intrachain disulphide bond; as a result, cysteine is the most conserved amino acid in protein evolution.

In addition to causing simple loss of normal function or incorrect protein folding, missense mutations can also result in some new protein property that is damaging to cells and tissues in some way, or alters their behavior. We consider this aspect in more detail when we discuss the effects of genetic variants in [Section 7.7](#).

Mutations that result in premature termination codons

A natural termination (stop) codon in mRNA triggers the ribosome to dissociate from the mRNA, releasing a polypeptide. However, many pathogenic mutations in coding DNA cause an in-frame *premature* termination codon to be inserted into a coding sequence, either directly or indirectly.

Nonsense mutations are nonsynonymous substitutions that directly replace an amino-acid-specifying codon by a stop codon. For nuclear DNA, that means a substitution that produces one of three stop codons UAA, UAG, or UGA in the corresponding mRNA. Note that the genetic code for mitochondrial DNA is different—as shown in [Figure 7.2](#).

A **frameshift mutation** may indirectly lead to a premature termination codon. Deletion or insertion of a sequence of n nucleotides in coding DNA produces a shift in the translational reading frame when $n/3$ is not an integer ([Figure 1](#) in [Box 2.1](#) on page 26 gives the principle). If a different reading frame is used, an in-frame premature termination codon is quickly encountered. Frameshifts often involve deletions or insertions at the DNA level, but they may also result from mutations producing altered splicing (exon skipping, intron retention), as described below.

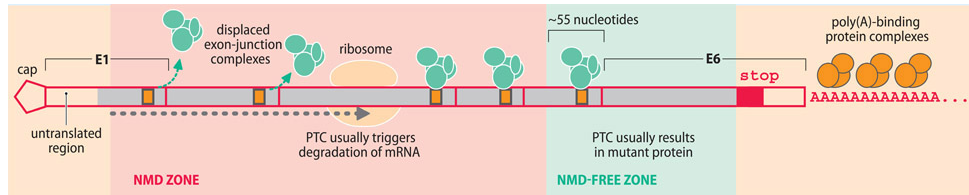


Figure 1 Nonsense-mediated decay. In mammalian cells, the primary nonsense-mediated decay (NMD) pathway is splicing-dependent. Certain components of the splicing machinery, called exon-junction complexes (EJCs), bind about 20–24 nucleotides upstream of the 3' end of each transcribed exon sequence and remain bound to the mature RNA. Here we illustrate a mature mRNA (with coding sequence in gray) formed from six exons (separated by thin vertical red lines; the first and last exons are labelled E1 and E6). Vertical orange boxes superimposed on the gray coding sequence show EJC-binding sites. The first ribosome to bind and then move along the mRNA will displace each EJC in turn until it reaches the stop codon and disengages. A premature termination codon (PTC) occurring up to 55 nucleotides before the end of the last exon (large pink box) leaves the mRNA with usually several EJCs attached, which normally triggers destruction of the mRNA. However, a PTC that occurs late in the mRNA, in the last exon and up to 55 nucleotides before it (large green box) usually means that the mRNA is translated to make a truncated protein that may sometimes give rise to a stronger phenotype than obtained by mRNA degradation.

At the DNA level, deletions or insertions of one or two nucleotides in coding DNA are a quite common cause of disease. As described in [Section 7.3](#), short tandem repeats with mononucleotide or dinucleotides predispose to 1- to 2-bp insertions and deletions. In addition, intragenic deletions that remove one or more exons or exon duplications often cause frameshifts ([Figure 7.1](#)). And transposons can occasionally accidentally insert into coding DNA (we consider large deletions and insertions like these in [Section 7.4](#)).

The usual result of nonsense mutations and frameshifting mutations is that the mRNA is degraded by a mechanism known as nonsense-mediated decay ([Box 7.1](#)). Occasionally, however, translation occurs to give a truncated protein. Truncated

proteins produced in this way can sometimes interfere with wild-type proteins produced from the normal allele and so affect their function (a *dominant-negative effect*). We consider the implications in [Section 7.7](#).

BOX 7.1 NONSENSE-MEDIATED DECAY AS AN mRNA SURVEILLANCE MECHANISM

Various RNA surveillance mechanisms monitor RNA integrity, checking for splicing accidents (such as when transcribed intron sequences are inappropriately retained in the mRNA—see [Figure 7.3B](#)[2]) and occasional errors in base incorporation during transcription. These errors frequently give rise to in-frame premature termination codons (PTCs) that could be dangerous—the aberrant transcripts could give rise to truncated proteins that might have the potential to interfere with the function of normal proteins.

To protect cells, an mRNA surveillance mechanism known as **nonsense-mediated decay (NMD)** degrades most mRNA transcripts that have an in-frame PTC. The primary NMD pathway is dependent on RNA splicing (single exon genes escape NMD because they do not undergo RNA splicing). Multisubunit protein complexes, called exon-junction complexes, are deposited shortly before the 3' end of each transcribed exon during RNA splicing and remain bound at positions close to the exon-exon boundaries in mature RNA ([Figure 1](#)).

The first ribosome to bind and move along the mRNA displaces each of these complexes in turn before disengaging from the mRNA at the natural stop codon. If there is an in-frame PTC, however, the ribosome detaches from the mRNA at an early stage; some exon junction complexes remain bound to the RNA, which usually signals mRNA destruction. However, in-frame PTCs within or just before the last exon often escape NMD and are translated to give truncated proteins ([Figure 1](#)).

Nonsense mutations, frameshifting insertions and deletions, and certain splicing mutations (such as those that result in retention of an intron) can activate nonsense-mediated decay.

Pathogenic splicing mutations

Many disease-causing mutations affect RNA splicing. The great bulk of them are in DNA sequences specifying *cis*-acting RNA elements that regulate how a specific gene undergoes RNA splicing, and this will be the focus here. Note, however, that disease can occasionally be caused by mutations in genes encoding *trans*-acting regulators of splicing.

As illustrated in [Figure 6.4A](#), fundamental *cis*-acting regulatory elements that control RNA splicing are located at or close to splice junctions. Point mutations in these sequences often have marked effects on RNA splicing, especially if they change highly conserved nucleotides such as the GT (GU in RNA) and AG dinucleotides at the extreme 5' end and 3' end, respectively, of an intron. That can result in abnormal splicing patterns such as omission of an exon (**exon skipping**), or failure to splice out an intron (*intron retention*)—see [Figure 7.3](#).

Pathogenic mutations can also occur in additional *cis*-acting splice regulatory elements, including splice enhancer and splice silencer sequences in exons and introns (see [Figure 6.4B](#)). Mutations like this may be less readily identified as pathogenic mutations and can explain why some mutations causing synonymous substitutions are pathogenic ([Figure 7.4A](#)).

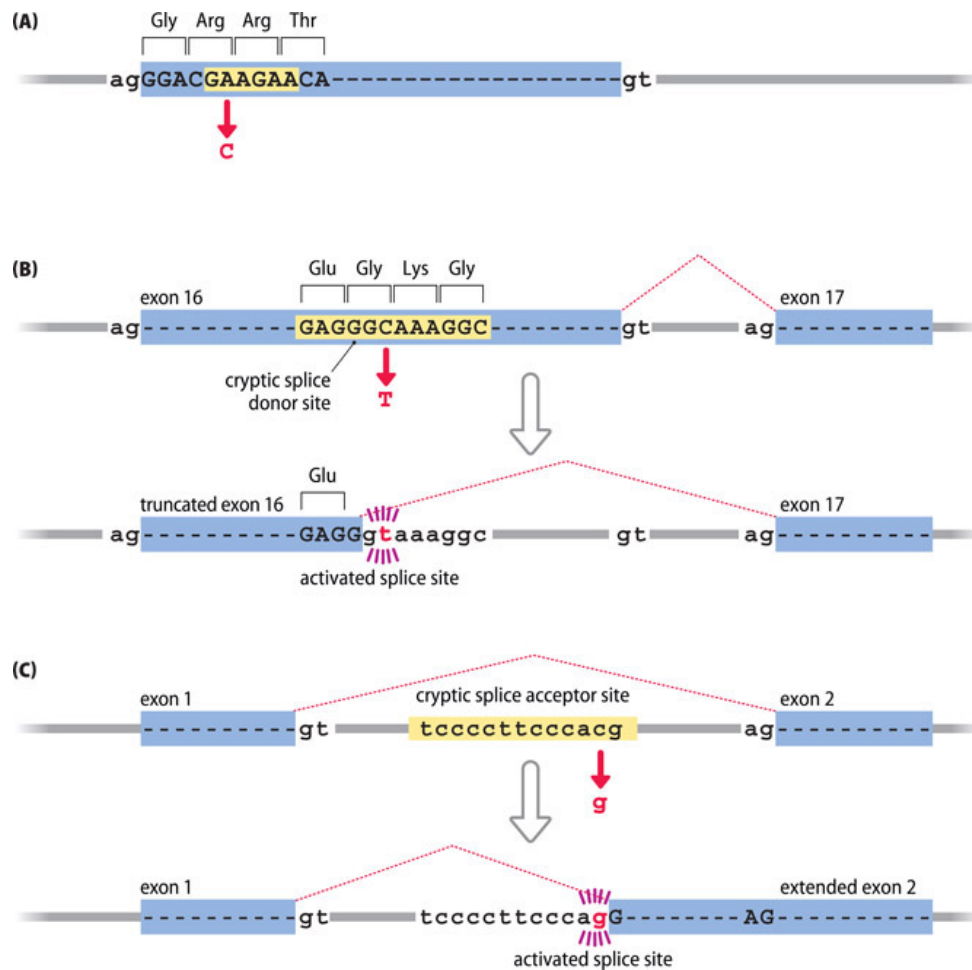


Figure 7.4 Apparently harmless synonymous and intronic substitutions can be pathogenic.

Within exons (blue boxes) significant nucleotides are shown in capital letters; important intronic nucleotides are in lower case. Dashed red lines indicate splicing of exon sequences (occurring *at the RNA level*). (A) The A \rightarrow C mutation leads to replacing one arginine-specifying triplet (CGA) by another (CGC), but causes disease by changing an exonic splice enhancer sequence (highlighted in yellow) at the beginning of the exon. (B) A homozygous synonymous C \rightarrow T substitution in exon 16 of the calpain 3 gene replaces one glycine-specifying triplet GGC by another (GGT); see PMID 7670461. It causes limb girdle muscular dystrophy by simultaneously activating a cryptic splice site (GAGGGCAAAGGC), to become a functional splice donor. As a result, exon 16 is truncated (the final 44 nucleotides are not included in RNA transcripts). The resulting shift in the translational reading frame produces a premature termination codon early in exon 17. (C) An apparently harmless single-nucleotide intronic substitution causes disease by activating a cryptic splice site closely resembling a 3' splice site (*splice acceptor*; the mutation results in the terminal AG dinucleotide required in splice acceptor sequences). Aberrant RNA splicing causes a sequence from the 3' end of intron 1 to be included in exon 2, extending it (causing disease by introducing a frameshift or by disturbing protein structure/protein folding).

By chance, sometimes a sequence may be almost identical to a genuine splice donor or splice acceptor site (a *latent* or **cryptic splice site**), and changing a single nucleotide can cause it to become a novel splice site. Activation of cryptic splice sites produces truncated or extended exons (see [Figure 7.4B,C](#) for examples).

Abnormal splicing can have variable consequences. For a protein-coding gene, a loss of coding exon sequences (by exon skipping or exon truncation) or a gain of coding exon sequences (by exon extension or intron retention) may result in a frameshift in the translational reading frame at the RNA level. In that case, the introduction of a premature termination codon might induce RNA degradation or produce a truncated protein (as described in [Box 7.1](#)).

If there is no shift in the translational reading frame, pathogenesis may nevertheless occur because of a loss of key amino acids, or, for exon extension, by the inclusion of extra amino acids that might destabilize a protein or impede its function. Intron retention in coding sequences would be expected to introduce a nonsense mutation (because of the comparatively high frequency of termination codons in all three reading frames).

Genesis and frequency of pathogenic point mutations

As detailed in [Section 4.1](#), single nucleotide substitutions often arise as a result of spontaneous chemical degradation of DNA that has not been repaired effectively. Some types of single nucleotide substitution are especially frequent in human DNA. C → T transitions are particularly common because the cytosine in CG dinucleotides is a hotspot for mutation in human (and vertebrate) cells. That happens because the CG dinucleotide is a target for cytosine methylation, and methylated cytosines tend to be deaminated to give thymines (as previously shown in [Figure 4.3](#)), which are not easily identified as altered bases by DNA repair systems.

Very small insertions and deletions are often produced by **replication slippage**, an error that typically occurs in DNA replication when a single nucleotide or short oligonucleotide is tandemly repeated. During DNA replication the nascent strand occasionally mispairs with the parent DNA strand (the mispairing is stabilized by base pairing between misaligned repeats on the two strands; the result is that the DNA polymerase either stutters at a tandem repeat or skips forward—see [Figure 4.6](#) on page 91, for the mechanism). Arrays with multiple tandem repeats are particularly susceptible to replication slippage; simply by chance, coding sequences occasionally have sequences with such repeats. For example, in about one out of four occasions, on average, two consecutive lysines in a protein are specified by the

hexanucleotide AAAAAA. Any run of consecutive nucleotides of the same type means a significantly increased chance of replication slippage—in this case the daughter strands are liable to have five or seven A's, causing a frameshifting deletion or insertion.

Mutation rates in the human genome

Comprehensive genome sequencing in family members indicates that the genomewide germ line nucleotide substitution rate is 10^{-8} per nucleotide per generation. That equates to about 30 *de novo* nucleotide substitutions on average in the 3 Gb haploid genomes inherited from each parent.

The mutation frequency varies across chromosomes and genes. Some gene-associated features make them more likely mutation targets. Genes are GC-rich and have a higher content of the CG dinucleotide; the cytosine in CG dinucleotides is a mutational hotspot, undergoing C → T transitions at a rate that is more than 10 times the background mutation rate.

The mitochondrial genome is extremely gene-rich, and the mutation rate is many times higher than in the nuclear genome. The mitochondrial genome is vulnerable, possibly because the great majority of reactive oxygen species are produced in mitochondria. Close proximity to these dangerous radicals results in much more frequent damage to the DNA, which is devoid of a protective chromatin coating, unlike nuclear DNA. Unrepaired DNA replication errors can also be significant in mtDNA.

Total pathogenic load

Only a small fraction of the novel changes that arise in the genomes we inherit from our parents is likely to be pathogenic, but because our parents are carriers of previously generated mutations our genomes contain many deleterious mutations. As yet there is no easy way to identify the total pathogenic load—all pathogenic mutations—within a genome.

Population genomics projects have shown that, depending on our ethnic background, each of us carries about 100 mutations that would be expected to result in loss of gene function (with an average of 20 genes that are homozygously inactivated), plus about 60 missense variants that severely damage protein structure.

One prediction is that the average person might have over 400 damaging DNA variants. That might seem an impossibly high load of pathogenic mutations but many of these mutations are common variants in non-essential genes, such as the *ABO* blood group gene (which is homozygously inactivated in people with blood group O).

Effect of parental age and parental sex on germ line mutation rates

Increased parental age often correlates with increased frequency of genetic disorders. We consider the maternal age effect in trisomy 21 in [Section 7.5](#). For small-scale mutations there is often a higher frequency of *de novo* mutation in the male germ line, and paternal age effects can be apparent, as described below.

The frequency of *de novo* mutation can be expected to be high in gametes that have undergone many cell divisions since originating from the zygote through a **primordial germ cell** (the cells that are set aside in the early embryo to give rise to the germ line). That happens because DNA replication precedes each cell division, and mutations often arise as a result of uncorrected errors in DNA replication. Two meiotic cell divisions are required to form oocytes and sperm cells, but the number of preceding mitotic cell divisions required to produce the first meiotic cells is very different between the two sexes. All the egg cells that will be available to a woman are formed before birth. By contrast, after the onset of puberty in men, sperm are continuously being formed by the division of spermatogonial stem cells. The number of cell divisions required to produce gametes is therefore higher in men, and especially so in older fathers ([Figure 7.5](#)).

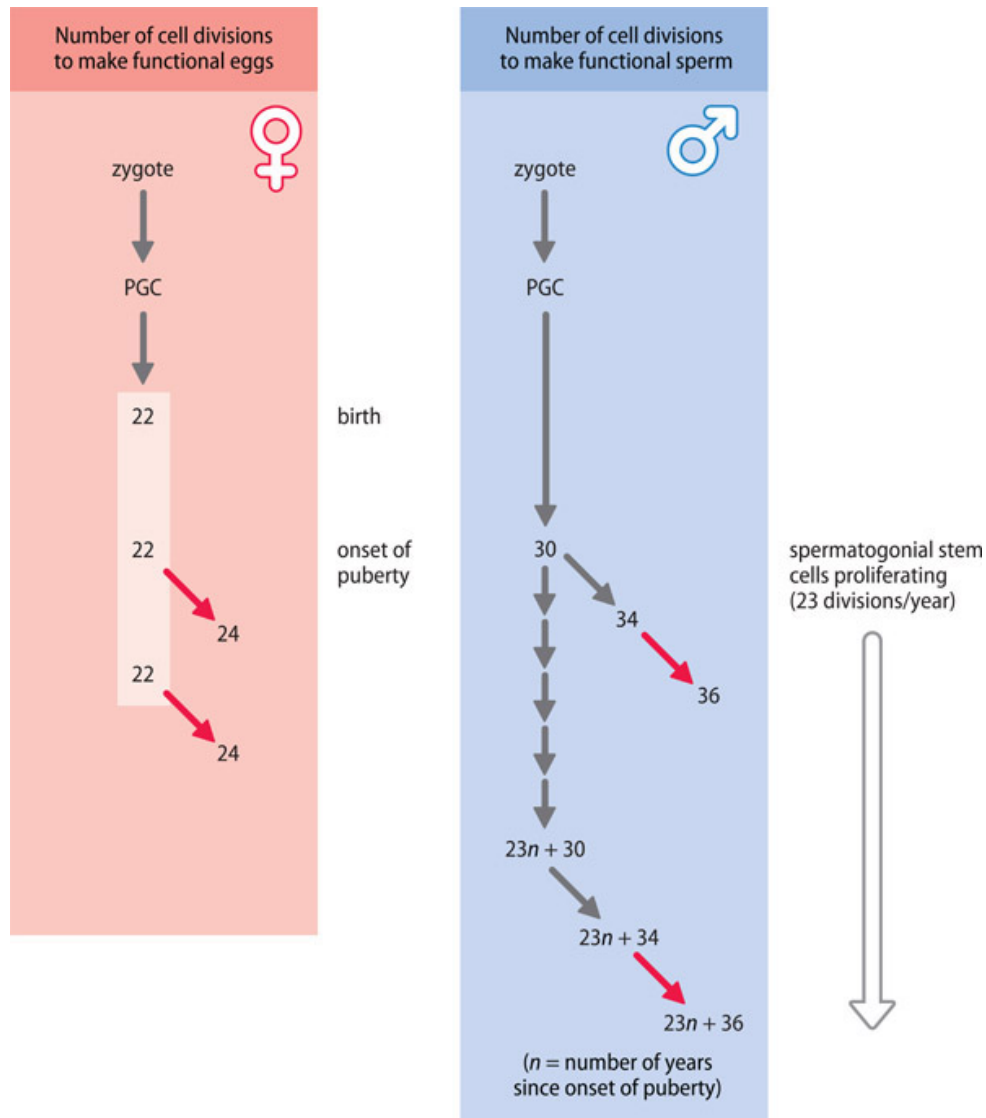


Figure 7.5 Sex differences and paternal age differences in the number of cell divisions

required to make gametes. Numbers represent completed cell divisions en route from a human zygote to gametes. Sperm and egg cells are formed by two sequential meiosis (red arrows) preceded by multiple mitoses (gray arrows). In females, all the ~22 mitoses required to get to the first meiotic cell are accomplished before birth (part of meiosis I has been completed by then, but then suspended until activated by ovulation). No matter how old mothers are, a total of ~24 cell divisions separate zygote from egg cells. Males are different: gametogenesis continues throughout adult life. About 30 cell divisions separate zygote from the spermatogonial stem cell used to make the first sperm cells at the onset of puberty. From spermatogonial stem cell to gamete takes four mitoses and then two meioses. So, sperm cells produced at the onset of male puberty have gone through $30 + 4 + 2 = 36$ cell divisions. Thereafter, spermatogonial stem cells divide about every 16 days (or about 23 times per year). If we take an average age of 14 years, say, for the onset of male

puberty, sperm from a 64-year-old man have been produced in a process requiring $36 + (23 \times [64 - 14])$ mitoses plus two meioses, or a total of 1186 cell divisions. PGC, primordial germ cell.

Paternal-age-effect disorders and selfish spermatogonial selection

A small group of exceptional congenital disorders occur spontaneously at remarkably high apparent rates, reaching 1 in 30 000 births for achondroplasia, with marked paternal age effects and paternal germ line transmission (see [Table 7.4](#) for some examples). That might suggest exceptional germ line mutation rates (up to 1000-fold higher than the average rate). However, studies of mutation rates in sperm have suggested that the underlying mutations (which are mis-sense mutations that change a single amino acid) do not occur at especially high frequencies. Instead, the mutations have been thought to result in mutant proteins that cause the dysregulation of spermatogonial stem cell behavior, thereby conferring a selective growth advantage on any spermatogonial stem cell that contains them. Stem cells containing these mutations might proliferate to reach high frequencies, explaining the paternal origin of many mutations and the paternal age effect. In each case the mutant proteins are fibroblast growth factor receptors or other proteins that work in the growth factor receptor-RAS signal transduction pathway. The underlying mutations belong to a class of gain-of-function mutations described in [Section 7.7](#).

TABLE 7.4 EXAMPLES OF PATERNAL-AGE-EFFECT DISORDERS SUGGESTED TO BE ASSOCIATED WITH SELFISH SPERMATOGONIAL SELECTION

Disorder	Gene	Mutation/amino acid change	Estimated birth prevalence for new mutations	Parental origin of mutation
Apert syndrome	FGFR2	p.Ser252Trp p.Pro253Arg	~1/65000	100% paternal
Crouzon/Pfeiffer syndrome	FGFR2	>50 mutations	~1/50 000 to 1/100000	100% paternal
Achondroplasia	FGFR3	p.Gly380Arg*	1/30000	100% paternal

* The G380R change in the FGFR3 (fibroblast growth factor receptor 3) protein in achondroplasia is caused by either a G → C or a G → A change at nucleotide number 1138 in the reference cDNA sequence. For further information see the paper by [Goriely & Wilkie \(2012\)](#) in Further Reading.

Disorder	Gene	Mutation/amino acid change	Estimated birth prevalence for new mutations	Parental origin of mutation
Muenke syndrome	FGFR3	p.Pro250Arg	~1/30000	100% paternal
Noonan syndrome	PTPN11	many mutations	~1/10000	100% paternal

* The G380R change in the FGFR3 (fibroblast growth factor receptor 3) protein in achondroplasia is caused by either a G → C or a G → A change at nucleotide number 1138 in the reference cDNA sequence. For further information see the paper by [Goriely & Wilkie \(2012\)](#) in Further Reading.

Surveying and curating point mutations that cause disease

Point mutations are the most frequent contributors to disease, and data on known or suspected pathogenic point mutations have been curated in a variety of different databases. We consider how point mutations are identified as being pathogenic in [Chapter 11](#), when we consider diagnostic DNA approaches. For now, we give some brief points in the subsections below.

Point mutations in coding DNA

The vast majority of pathogenic point mutations have been recorded in protein-coding genes. For coding DNA, identifying some types of disease-causing mutation is comparatively easy. Nonsense mutations and insertions or deletions that change a known translational reading frame almost scream “Pick me!” at the investigator.

Making the correct call for missense mutations and small non-frameshifting deletions or insertions is harder. Evolutionary and population genetic studies can often help here: substitution or deletion of an amino acid is much more likely to be pathogenic if that specific amino acid has been strongly conserved during evolution. As detailed in [Chapter 11](#), various computer programs can be used. One important application is to assess the likelihood of a missense mutation being pathogenic on the basis of predicted differences in the physicochemical properties of the original amino acid and the substituted one. Querying databases of previously recorded mutations can also be very useful.

Point mutations in RNA genes and other noncoding DNA

Although we have significantly more RNA genes than protein-coding genes, causative mutations in monogenic disorders have been identified almost exclusively in protein-coding genes (see [Table 7.5](#) for some examples of the very few RNA genes that have been implicated). Explanations for this anomaly may include the general difficulty in identifying pathogenic mutations in noncoding DNA (which is generally poorly evolutionarily conserved and lacking a reading frame), and genetic redundancy for some genes (such as rRNA and tRNA genes in nuclear DNA).

TABLE 7.5 EXAMPLES OF RNA GENES MUTATED IN SINGLE-GENE DISORDERS

RNA class	Locus*	Disorder	PMID
miRNA	<i>MIR96</i>	Autosomal deafness, type 50 (OMIM 613074)	19363479
	<i>MIR184</i>	EDICT syndrome (OMIM 614303)	21996275
	<i>MIR204</i>	Retinal dystrophy with ocular colomba (OMIM 616722)	26056285
snRNA	<i>RNU4ATAC</i>	microcephalic osteodysplastic primordial dwarfism type 1 (OMIM 210710)	21474760
long noncoding RNA	<i>TERC</i>	Type 1 Autosomal dominant dyskeratosis congenital (OMIM 127550)	11574891
MT-rRNA	<i>MT-RNR1</i>	Inherited nonsyndromic hearing loss	7689389
MT-tRNA	Many	Various—covered in Section 7.6	33655490

* Interested readers can find recent reviews at PMID 24007299 and 32741549. PMID, PubMed Identifier. MT, mitochondrial.

Arguably, the most important explanation is that proteins are the major functional endpoints of cells (even though there are impressive layers of gene regulation by noncoding RNAs). Many RNA genes have not been well studied, but more recent studies have emphasized the importance of various noncoding RNAs in the pathogenesis of other disorders, such as cancers.

Databases of human pathogenic mutations

Human mutation databases range from large general databases to more specific ones ([Table 7.6](#)). Locus-specific databases focus on a specific gene (or sometimes genes) associated with an individual disorder. The submitted data include both pathogenic mutations and normal variants, and so the databases can be of help in evaluating whether newly identified mutations are likely to be pathogenic, as described in [Chapter 11](#).

TABLE 7.6 EXAMPLES OF DIFFERENT TYPES OF DATABASES THAT CURATE DISEASE-ASSOCIATED MUTATIONS

Database	Description	Website
GENOMEWIDE DATABASES		
Human Gene Mutation Database	comprehensive data on germ line mutations in nuclear genes associated with human inherited disease	http://www.hgmd.cf.ac.uk/ac/index.php
COSMIC	comprehensive catalog of somatic mutations in cancer	https://cancer.sanger.ac.uk/cosmic
MITOMAP	mitochondrial genome database	https://www.mitomap.org/MITOMAP
LOCUS-SPECIFIC DATABASES (see also http://www.hgvs.org/locus-specific-mutation-databases and PMID 21540879)		
CFTR2	specific <i>CFTR</i> cystic fibrosis gene variants	https://cftr2.org/
MUTATION CATEGORY DATABASES		
SpliceDisease database	disease-associated splicing mutations	http://cmbi.bjmu.edu.cn/sdisease

The Human Genome Variation Society also maintains links to many other useful mutation databases at <http://www.hgvs.org/content/databases-tools>. For further background see PMID 17893115.

7.3 PATHOGENESIS DUE TO VARIATION IN SHORT TANDEM REPEAT COPY NUMBER

Our genome is littered with short tandem repeats of mononucleotides to oligonucleotides. Very few, notably TTTAGG repeats in telomere DNA, are functionally important. The great majority are there simply as a result of statistical inevitability. In our genome, which has a 41 % GC/59 % AT base composition, the odds that a hexanucleotide chosen at random has the sequence AAAAAA is 1 in $(0.295)^6$ or 1 in 1,517. In an average human chromosome (with 133 Mb of DNA) the sequence AAAAAA occurs over 87 000 times; nearly 16 000 examples of the decanucleotide ATATATATAT would be expected across the whole genome.

Short tandem repeats cause problems for aligning the two opposing DNA strands of a double helix. They increase the chances of *slipped strand mispairing*, the local mispairing of repeats on opposing DNA strands, causing the strands to slip slightly out of alignment. For an ATATATATAT array, for example, the third AT on one strand might mistakenly base pair with the fourth AT on the opposing strand, with consequent looping out of an AT. If this happens during DNA replication, then the newly synthesized DNA strand will have fewer or more AT repeats. We previously covered the general principle of **replication slippage** when we explained the basis of simple tandem repeat micropolymorphism ([Figure 4.6](#) on page 91). DNA strands may also be produced with fewer or more tandem repeats when DNA repair occurs at a DNA region where there is slipped strand mispairing.

The two main classes of pathogenic variation in short tandem repeat copy-number

Length variation in arrays of short tandem repeats results in disease in two ways, as listed below.

- *Frameshifting expansion/contraction in coding DNA.* Replication slippage can cause very short frameshifting insertions or deletions within arrays of short tandem repeats in coding DNA (the great majority are 1-or 2-nucleotide insertions/deletions). They account for a very significant component of pathogenic insertions and deletions in coding DNA.
- *Non-frameshifting expansion beyond safe limits.* Disease can occur when the number of repeats in some short tandem repeats expands beyond some safe limit, causing moderately long to quite large insertions at the DNA level. The expansions occur for certain triplet repeats in coding DNA and various types of repeat in noncoding DNA,

As an example of how very short pathogenic insertions and deletions are produced through replication slippage in short tandem repeats, consider the tandem mononucleotide sequence AAAAAA occurring on the sense strand of a coding DNA, and replication slippage leading to a single A being lost or gained (to give five or seven adenines, respectively). The loss (deletion) or gain (insertion) of an adenine would produce a frameshift in the translational reading frame, often resulting in a premature termination codon and failure to make a protein. Loss or gain of repeats within arrays of tandem di-, tetra- and pentanucleotides in coding DNA can similarly cause disease by introducing frameshifting and early premature termination codons.

Trinucleotide repeats in coding DNA are different: loss or gain of a single repeat does not affect the translational reading frame, and often the effect of the changed length is inconsequential. But non-frameshifting expansion of certain triplet repeats can nevertheless be pathogenic, sometimes by inactivating a gene, or by causing proteins to behave abnormally or by producing toxic RNAs, as described below.

Non-frameshifting pathogenic expansion in short tandem repeat number

More than 40 diseases are caused by expansion of the number of short tandem repeats (from trinucleotide to dodecanucleotide repeats) beyond safe limits. Some occur in coding DNA. Others are located in 5' or 3' untranslated sequences or in introns within a protein-coding gene. In the former case certain types of tandem triplet repeats are involved that ultimately specify alanine or glutamine (polyalanine and polyglutamine tracts are found in a considerable number of human proteins). Pathogenic short tandem repeat expansion in noncoding DNA involves different types of tandem repeats, from triplet to dodecanucleotide repeats. In the sections that follow we detail some specific examples. The list below provides an overview.

- *Polyalanine expansion.* Nine congenital disorders of development are known to be caused by expansion of alanine-specifying triplet repeats. (The resulting polyalanine tracts typically occur in transcription factors, serving as flexible nonpolar linkers between two folded domains.) Individual arrays often contain different types of alanine-specifying triplets, and are not polymorphic, while quite small, stable expansions are seen in affected individuals ([Table 7.7](#)). Disease results after the expansions pass a certain safe-size limit: proteins with a sufficiently extended polyalanine tract can

aggregate to cause problems for cells. We examine protein aggregation more generally in [Section 7.8](#).

TABLE 7.7 EXAMPLES OF DISEASES RESULTING FROM THE THREE CLASSES OF NON-FRAMESHIFTING PATHOGENIC EXPANSION OF SHORT TANDEM REPEATS

Class (Stability [*])	Disease examples	Repeat unit	Associated tandem repeats		
			Location	Copy number	
				Normal	Disease
Polyalanine expansion (STABLE)	Oculopharyngeal muscular dystrophy (OPMD)	GCG, GCT, GCA (specifying alanine)	Coding sequence	10	11-17
	Synpoldactyly type II			15	22-29
	Hand-foot genital syndrome			18	24-26
Polyglutamine expansion (UNSTABLE)	Huntington disease	CAG (specifying glutamine)	Coding sequence	6-35	36-250
	Spinal bulbar muscular atrophy (Kennedy disease)			4-34	35-72
	Dentatorubropallidoluysian atrophy			3-38	49-88
	Spinocerebellar ataxia type 7			7-41	43-51
Noncoding DNA expansion (UNSTABLE)	Friedreich ataxia	GAA	Intronic	6-32	200- 1700
	Fragile X syndrome (with intellectual disability)	CGG	5' - UTR	5-54	>200 to several ×1000
	Myotonic dystrophy type I	CTG	3' - UTR	5-37	50- 10000
	Myotonic dystrophy type II	CCTG	3' - UTR	10-26	75-11 000
	Spinocerebellar ataxia type 10	ATTCT	Intronic	10-29	500- 4500

* Some short tandem repeat expansions are unstable (see text).

Class (Stability [*])	Disease examples	Repeat unit	Associated tandem repeats		
				Copy number	
			Location	Normal	Disease
	Frontal dementia and/or amyotrophic lateral sclerosis	GGGGCC	Intronic	2-22	700- 1600

* Some short tandem repeat expansions are unstable (see text).

- *Polyglutamine expansion.* Nineteen disorders with neurodegenerative or neuromuscular phenotypes show expansion of CAG repeats specifying polyglutamine. Polyglutamine tracts are highly flexible, but unlike polyalanine are highly polar. Unlike the triplet repeats associated with polyalanine expansion, those associated with polyglutamine expansion are often homogenous CAG repeats showing length polymorphism in the general population but moderate to sometimes quite large expansions in affected individuals (see [Table 7.7](#)). The pathogenic expansion in polyglutamine tracts can be quite unstable, increasing in size after both mitotic and meiotic cell division (the underlying mutations are described as *dynamic mutations*—see next section).
- *Pathogenic expansion of noncoding short tandem repeats.* More than 20 human disorders are due to moderate to large expansion of short tandem repeats in introns or untranslated sequences of protein-coding genes. The repeat units are mostly 3 to 6 nucleotides in length—see [Table 7.7](#) for examples. These expansions can show very significant instability in mitotic and meiotic cell division (see next section).

Dynamic disease-causing mutations due to unstable expansion of short tandem repeats

We are accustomed to thinking that mutations are stable. When a disease-causing mutation is transmitted from one generation to the next, we expect to see the same mutation in affected individuals from different generations of a family. However, pathogenic expansion of polyglutamine-specifying CAG repeats and various types of short tandem repeats in noncoding DNA can be unstable. They are sometimes described as **dynamic mutations** because the repeat length can increase from one

generation to the next (and sometimes from mother cell to daughter cell in one individual). In marked contrast, in polyalanine expansion disorders the triplet repeat expansions are stable.

Increasing expansion of the short tandem repeats leads to increasing severity of the disease. Because the expansions can increase from one generation to the next, the phenotype can become increasingly severe from one generation to the next, a phenomenon known as **anticipation**. We previously gave an example of anticipation in myotonic dystrophy wherein the disease progressed from mild features in a grandmother to moderate features in her daughter and then to severe congenital muscular dystrophy in the grandson (see [Figure 5.17](#) on page 129). And because disorders arising from unstable expansion of short tandem repeats are often neurodegenerative, pronounced differences in the age of onset of symptoms can be attributable to the extent of repeat expansion, as shown strikingly in the case of Huntington disease ([Figure 7.6](#)).

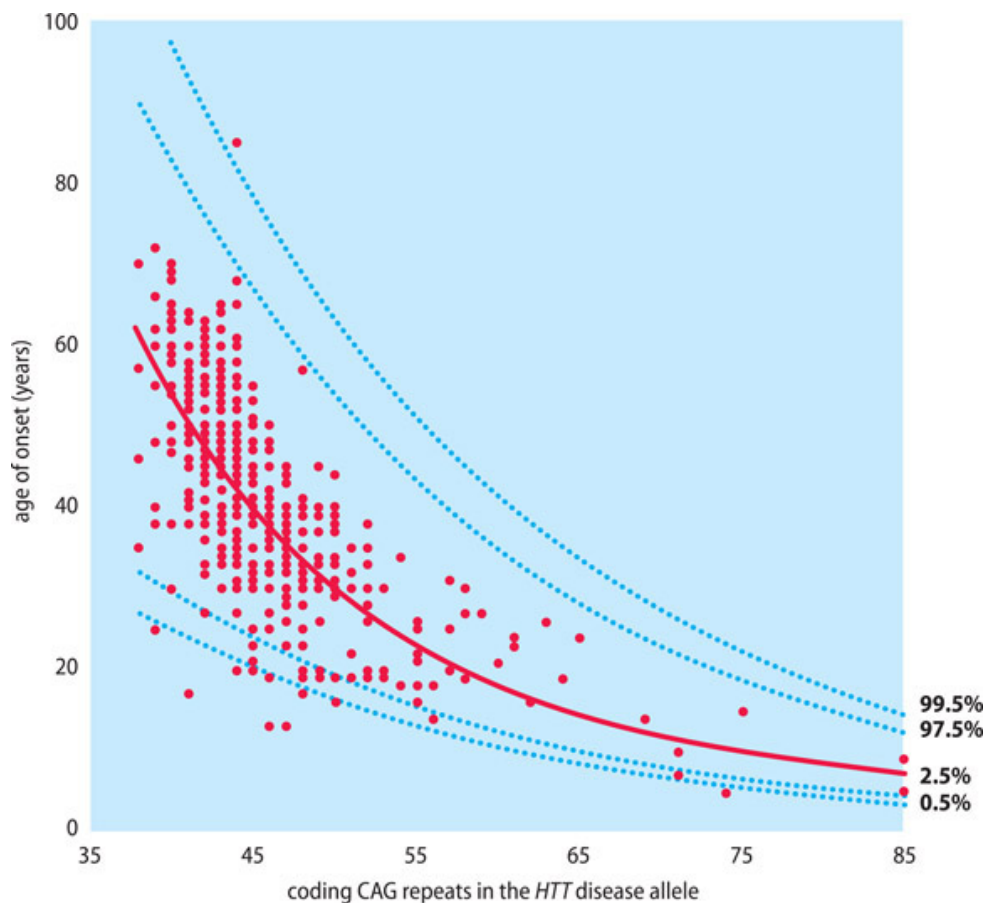


Figure 7.6. Inverse correlation between the extent of CAG repeat expansion and age of disease onset in Huntington disease. An increased CAG repeat number means increased loss of

neurons and earlier evidence of symptoms. Reproduced from Budworth H, McMurray CT (2013) A brief history of triplet diseases. In: Kohwi Y, McMurray C (Eds) *Trinucleotide repeat protocols. Methods in molecular biology (methods and protocols)*, 1010. Humana Press, Totowa, NJ. With permission from Springer Nature.

The unstable and large expansion of polyglutamine-specifying CAG repeats and certain noncoding tandem repeats is in marked contrast to the limited and very stable expansion of polyalanine-specifying triplet repeats. The difference is likely to be due to the comparative heterogeneity of polyalanine-specifying triplet repeats (individual arrays are heterogeneous, having two or more different triplets in arrays, hindering slipped strand mispairing and so limiting expansion). Arrays of polyglutamine-specifying CAG repeats and noncoding tandem repeats generally show very high repeat homogeneity.

The mechanism underlying unstable tandem repeat expansion is unclear. The arrays of tandem repeats are prone to forming abnormal secondary structures, including slipped strands with extrahelical loops, and hairpins, and also certain other unusual structures, including triplex and quadruplex DNA. Secondary structure elements such as these can impede DNA functions (replication, transcription, and so on), provoking DNA repair mechanisms to remove them. From studies of both mouse models and human genomewide association, components of the mismatch DNA repair system are known to be involved in the unstable repeat expansions.

The size of unstable repeat expansions can be very large for some disorders, notably for myotonic dystrophy (types 1 and 2) and Fragile X. As a result, for these disorders, a specialized laboratory analysis, known as triplet-repeat primed PCR is used to track the repeat size (see [Clinical Box 5](#) for a case study of myotonic dystrophy, showing an example of this analysis).

The complication of repeat-associated non-AUG translation

At the DNA level, unstable tandem repeat expansions might seem broadly similar, but at the cell level the pathogenic consequences can differ remarkably, as outlined in the subsection following this one. There is the added complication of unorthodox repeat-specific expression mechanisms as well as orthodox expression. Expanded CAG repeats can be conventionally expressed to give large polyglutamine tracts, but are also subject to unorthodox *repeat-associated non-AUG translation* (RAN

translation) in multiple reading frames to also produce polyserine (specified by recurring AGC) and polyalanine (specified by recurring GCA).

RAN translation across transcripts from expanded tandem DNA repeats does not require an AUG start codon and applies to expanded tandem noncoding DNA. Both sense and antisense transcripts from expanded noncoding repeats can be RAN-translated in different reading frames (see [Figure 7.7](#)). The contribution of RAN-translation expression products to pathogenesis is an area of active investigation.

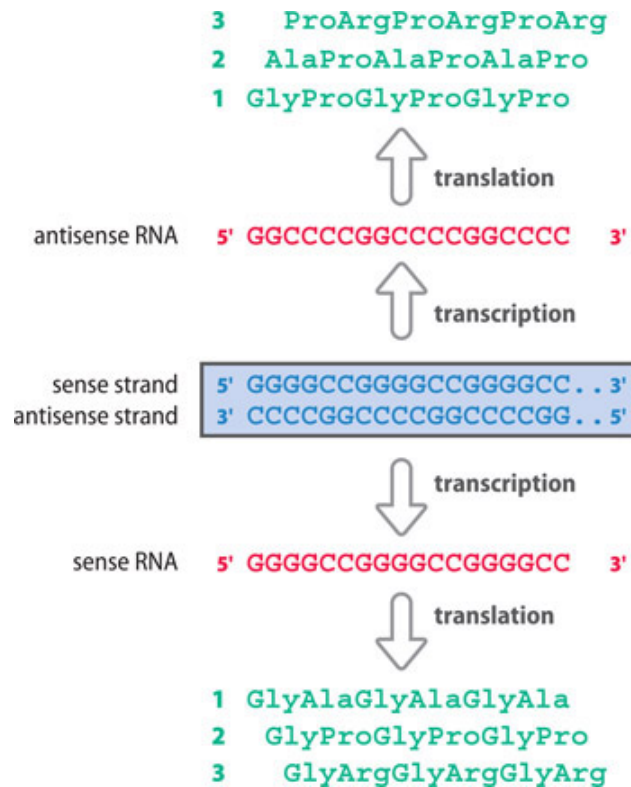


Figure 7.7. RAN translation produces five different polypeptides from tandem GGGGCC repeats associated with frontal dementia and amyotrophic lateral sclerosis. Transcription of both strands of tandem GGGGCC repeats (also called G⁴C² repeats) produces both sense and antisense RNA transcripts for each of which translation can occur in all three reading frames to give five different polypeptide products: poly (Gly-Ala); poly (Gly-Pro); poly (Gly-Arg); poly (Pro-Arg); and poly (Ala-Pro).

CLINICAL BOX 5 A CLINICAL CASE STUDY: MYOTONIC DYSTROPHY TYPE I

Janna was born at term from an uneventful pregnancy to non-consanguineous parents. She had a normal development through childhood. In early adolescence

she had difficulty in relaxing her hand muscles after making a fist, and in opening her mouth after the first bite of a meal. Janna also had difficulty in keeping focused on one task for more than an hour. She felt continuously tired, and would regularly sleep 12 hours a night or longer when not using an alarm. Soon after, she started having bouts of diarrhea that would last for some days, followed by episodes of constipation and abdominal pain. As an adult, Janna developed difficulties walking; she tripped constantly when walking on uneven surfaces outdoors and experienced several falls per month. Her voice changed, becoming nasal. She increasingly had difficulties swallowing, and eventually experienced some choking episodes. She was initially referred to an orthopedic surgeon who requested an EMG that identified myotonic discharges in several muscles ([Figure 1](#)). She was then referred to the neuro-muscular center for further follow-up.

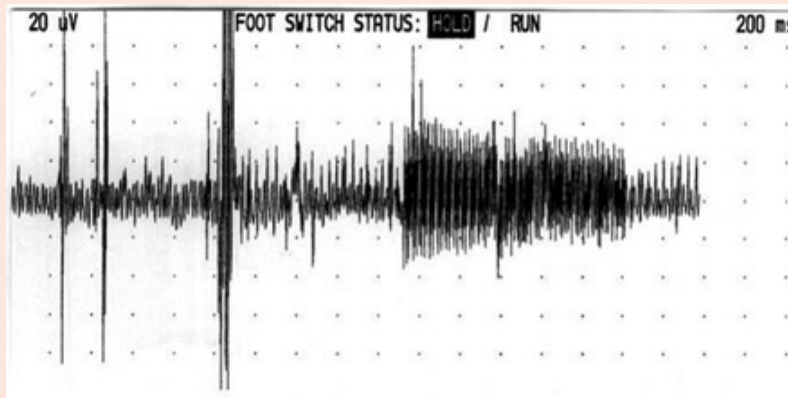


Figure 1 Electromyography (EMG) showing myotonic discharges. Several of Janna's muscles showed spontaneous discharges of muscle fibers that waxed and waned in both amplitude and frequency.

First seen by the neuromuscular team at the age of 25, Janna presented with mild ptosis, facial weakness and atrophy of temporal and sternocleidomastoid muscles ([Figure 2](#)). She had mild weakness of finger flexor muscles, myotonia in the hands, and distal muscle weakness in her lower limbs, affecting both tibialis anterior (3-/5 on the MRC scale).



Figure 2 The typical myopathic face in myotonic dystrophy. Janna had an elongated typical myopathic face, temporal muscle atrophy, ptosis, perioral weakness, a high arched palate and bilateral atrophy of the sternocleidomastoid muscles.

Janna was diagnosed with myotonic dystrophy type 1 in the genetic lab after TP-PCR analysis showed she had an abnormal expansion of >350 CTG repeats in the *DMPK* gene (**Figure 3**). Follow-up testing of Janna's parents and siblings confirmed a genetic diagnosis in her father, John (who had only minor symptoms, and did not show myotonia or muscle weakness). Janna's elder brother, Mark, was found to have myotonia in his hands and jaw, but no muscle weakness, but her younger brother, George, was unaffected.

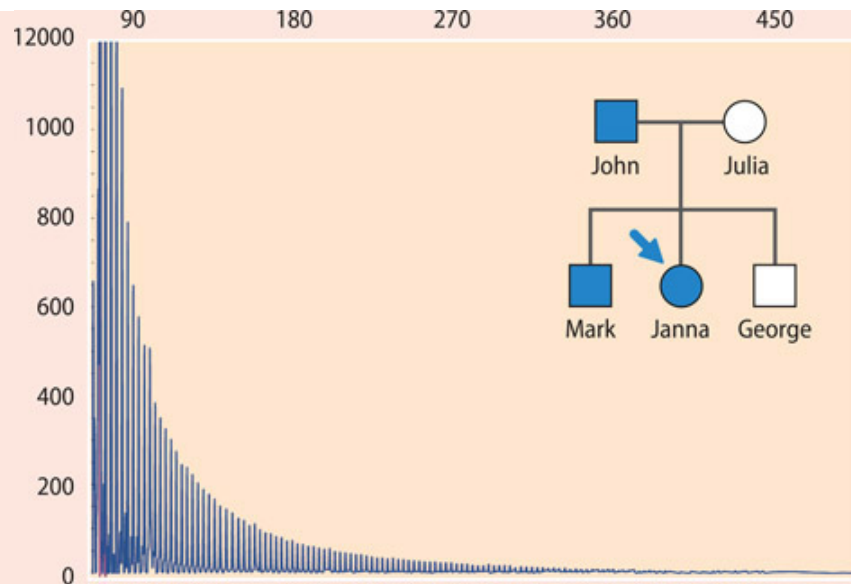


Figure 3 Triplet repeat-primed PCR (TP-PCR) result. It shows Janna has >350 CTG repeats in exon 15 of the *DMPK* gene (compared to 50 repeats or less in unaffected controls). The inset shows the pedigree obtained after following up Janna's family.

Janna underwent counselling and was informed that there was a 50 % risk of passing her condition to her off-spring. She was advised that due to the phenomenon of anticipation, an affected child would be at high risk of presenting with the most severe form of myotonic dystrophy, congenital myotonic dystrophy. A baby affected in this way would be expected to be floppy and might experience severe breathing and feeding difficulties after birth. And the instability of the repeat expansion would be expected to increase during life, and be most pronounced in post-mitotic cells, such as in muscle, brain, and heart.

Unstable expansion of short tandem repeats can cause disease in different ways

Unstable short tandem repeat expansion can cause loss of function in recessive disorders (carrier individuals with one normal allele are unaffected). In other cases, a mutant gene bearing the expanded repeat is expressed and the problem is a gain of function: the gene makes an aberrant RNA or protein product that is harmful to cells and causes disease in heterozygotes (the presence of a normal allele is insufficient to prevent the disease phenotype)—see [Table 7.8](#).

TABLE 7.8 THREE MAJOR PATHOGENESIS ROUTES FOR UNSTABLE TANDEM REPEAT EXPANSION DISORDERS

Mutation class	Cause of disease	Example of disorder	Comments
Loss of function	Gene expression is inhibited; not enough gene product is made	Fragile X syndrome	Expansion of a tandem CGG repeat in exon 1 of the <i>FMR1</i> gene > 200 repeats triggers methylation of the promoter region and prevents transcription.
Gain of function	Harmful RNA transcript	Myotonic dystrophy (types 1 and 2)	Abnormally expanded CUG or CCUG repeats in the 3'-UTR of mRNA sequester regulatory proteins that control alternative splicing (see text).
	Cytotoxic protein	Huntington disease	The CAG expansion causes production of a huntingtin protein with an abnormally large number of glutamines that is especially toxic to neurons

For both myotonic dystrophy types 1 and 2, the problem is that RNA transcripts containing expanded tandem repeats bind to, and sequester, MBNL1 and other members of the muscleblind-like family of regulatory proteins that control alternative RNA splicing. These proteins have binding sites for both CUG and CCUG, and so expanded CUG repeats and expanded CCUG repeats in transcripts from the disease gene (in patients with myotonic dystrophy type 1 and type 2 respectively), bind to large quantities of the MBNL family of regulators, preventing their normal function. As a result, cells have a plethora of splicing defects at various gene loci.

The third pathogenic class, characterized by production of a toxic protein, is known to be common in the case of unstable CAG repeat expansion in coding sequences. Proteins with large polyglutamine tracts are unstable, and prone to aggregation in a way that is toxic to cells. Because neurons are intended to be extremely long-lived cells they are not readily replaced, and steady neuron depletion over time results in neurodegenerative and neuromuscular disorders.

Disease manifestation in Fragile X pre-mutation carriers

The CGG repeat expansion in exon of the *FMR1* gene is unusual because it results in different diseases according to the extent of the expansion. The “full mutation” that causes Fragile X syndrome is defined as expansions of over 200 CGG repeats. When this 200-repeat limit is passed, methylation of the nearby promoter is triggered, and the *FMR1* gene, which is important in brain function, is silenced, causing cognitive defects in males. (Note: because it is a loss-of-function phenotype, a small percentage of cases may occur by alternative point mutations and deletions in *FMR1*.)

Abnormal expansion of the CGG repeat array to lesser levels of 55–200 repeats (*pre-mutations*) also causes disease but by a different, poorly misunderstood gain-of-function mechanism in which *FMR1* mRNA is produced in excess, causing toxicity and mitochondrial dysfunction. 40 % of male carriers of this pre-mutation and 16 % of female carriers develop Fragile X-associated tremor/ataxia syndrome (FXTAS), with late onset (typically between 60 and 65 years of age) of progressive cerebellar ataxia and intention tremor, followed by cognitive impairment. In addition, 20 % of women with the premutation allele develop Fragile X-associated primary ovarian insufficiency (FXPOI) with hypergonadotrophic hypogonadism before the age of 40 years (compared to 1 % in the general population).

7.4 PATHOGENESIS TRIGGERED BY LONG TANDEM REPEATS AND INTERSPERSED REPEATS

Disease caused by various larger changes in DNA may involve large scale changes, facilitated by long tandem repeats and interspersed repeats. Inappropriate pairing up of non-allelic but closely homologous repeats in nuclear DNA and subsequent recombination can produce pathogenic duplications, deletions and inversions, and this is the area that most of this section is devoted to. (Of course, very large deletions, duplications and inversions, many of which may occur by the same mechanisms, have long been recorded under the microscope; we cover these chromosomal abnormalities together with chromosome segregation abnormalities in [Section 7.5](#)).

Pathogenic exchanges between repeats occurs in both nuclear DNA and mtDNA

Many of the repeats involved in triggering large-scale mutations occur within introns or outside genes. Some large repeats, however, contain one or more gene sequences. Pathogenesis may arise when large-scale mechanisms adversely change the structure or copy number of genes, or when they adversely alter gene expression. We detail the principal genetic mechanisms giving rise to these changes in individual sections below.

The non-allelic but closely homologous repeat sequences that predispose to moderate to large-scale deletions, insertions and inversions in nuclear DNA belong to two classes as listed below.

- *Long tandem DNA repeats.* Recall from [Section 2.5](#) that local DNA duplication events have frequently occurred during genome evolution. The most recent produced highly homologous tandem repeats many kilobases or megabases long and often containing multiple genes (*segmental duplication*). The classic example of involvement in pathology is provided by a tandem repeat of ~30 kb that includes the steroid 21-hydroxylase gene: pairing of nonallelic repeats at meiosis is responsible for 99 % of the pathology in 21-hydroxylase deficiency, as detailed below.
- *Interspersed repeats.* As well as the very high copy-number interspersed Alu and LINE repeats, many families of highly homologous low copy-number repeats are found in noncoding DNA in and around genes. Large genes with many long introns have multiple interspersed repeats in the introns, making them more susceptible to non-allelic mispairing at meiosis. The dystrophin gene has 78 introns of average size >30 kb, and so it is not surprising that ~75 % of boys with Duchenne muscular dystrophy have large pathogenic deletions or duplications, many causing a shift in the translational reading frame (where the *net* effect is to delete or insert a number of coding nucleotides that is not a multiple of three; [Figure 2C](#) in [Box 2.1](#) on page 27 shows the general concept). Moderate insertions can also occur within a coding sequence to produce an extended protein that might be unstable, not fold properly, or be functionally disadvantaged). As detailed in [Section 7.6](#), deletions occur quite frequently in mtDNA, very often arising by sequence exchanges between interspersed repeats.

Non-allelic homologous recombination and transposition

Large-scale pathogenic mutations in nuclear DNA are often initiated by abnormal pairing between low-copy-number repeats with very similar sequences (*homologous repeats*), often occurring in noncoding DNA within genes, or close to the genes. Different families of low-copy-number repeats can be distinguished. They might be short interspersed sequences, or naturally duplicated sequences extending up to several hundreds of kilobases in length and containing multiple genes.

When two repeats with very similar sequences occur close to each other on the same chromosome arm, the high level of sequence identity between the repeats can lead to mispairing of chromatids. Non-allelic repeats can then pair up: repeat no. 1 on one chromatid might pair up with repeat no. 2 on the other chromatid. A subsequent recombination event occurring in the mispaired region produces a change in repeat copy number; the process is known as **nonallelic homologous recombination (NAHR)**.

Alternatively, recombination occurs between homologous repeats on the same DNA molecule; an *intrachromatid recombination* such as this is also a form of NAHR. We describe different NAHR mechanisms below, show how they can generate insertions, deletions, and inversions, and give examples of how they cause disease.

Pathogenic sequence exchanges between chromatids at mispaired tandem repeats

Many human genes and gene regions have significant arrays of long tandemly repeated DNA sequences. This can include the repetition of exons, whole genes, and even multiple genes. Tandem repeats within genes or spanning coding sequences can predispose to a type of nonallelic homologous recombination that can cause disease.

Normally, recombination between homologous chromosomes occurs after the chromosomes have paired up with their DNA sequences in perfect alignment. However, local misalignment of the paired chromosomes is more likely to occur in regions where there are highly similar tandem repeats—the DNA molecules of the two chromatids can line up out of register. That is, the alignment is staggered and one or more repeats on each chromatid do not pair up with their normal partner repeat on the other chromatid.

A subsequent recombination within the mismatched sequences is known as **unequal crossover (UEC)** and results in one chromatid with an insertion (more

tandem repeats) and one with a deletion (fewer tandem repeats) An equivalent process can also occur between sister chromatids, an *unequal sister chromatid exchange (UESCE)* ([Figure 7.8](#)). UEC and UESCE cause reciprocal exchanges between misaligned chromatids: one chromatid gains an extra DNA sequence, and the other loses an equivalent sequence. Disease may result from a change in gene copy number, or through the formation of hybrid genes that lack some of the functional gene sequence.

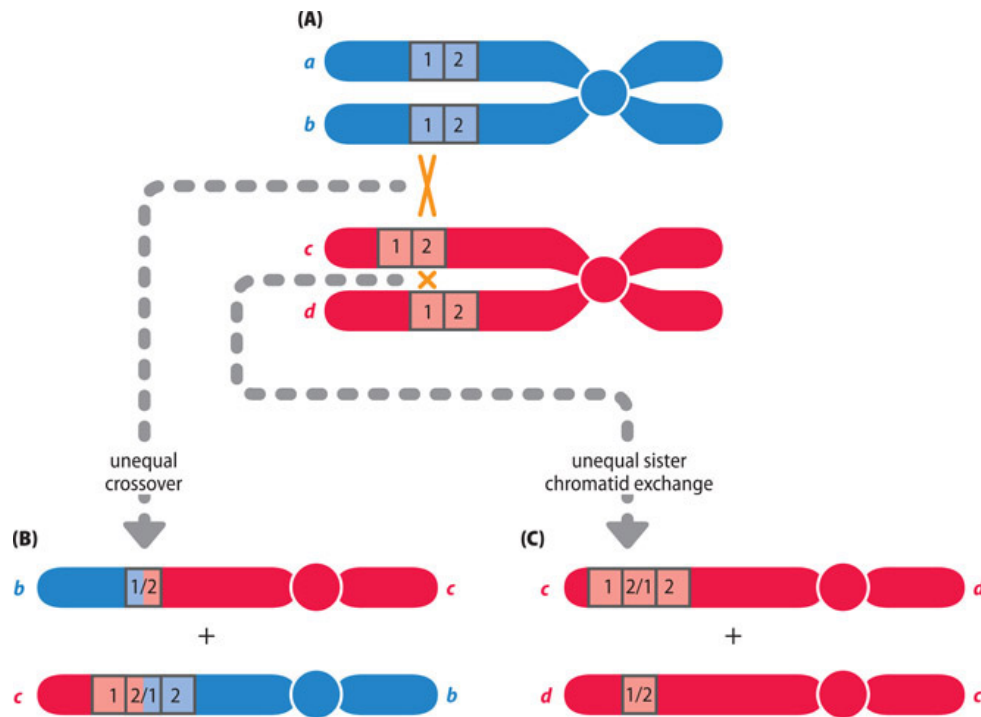


Figure 7.8 Unequal crossover and unequal sister chromatid exchange cause deletions and insertions. (A) Mispairing of chromatids with two very similar tandem repeats (1, 2). The very high sequence identity between these repeats can facilitate misalignment between the DNA of aligned chromatids so that repeat 1 on one chromatid aligns with repeat 2 on another chromatid. The misaligned chromatids can be on non-sister chromatids of homologous chromosomes, such as chromatids *b* and *c*, in which case recombination (large orange X) in the misaligned region results in an unequal crossover (B). Alternatively, there can be a recombinationlike unequal sister chromatid exchange (small orange X) between misaligned repeats on sister chromatids (*c*, *d*) of a single chromosome (C). In either case, the result is two chromatids, one with three repeat units and the other with a single repeat unit (a hybrid of sequences 1 and 2—shown here as 1/2 or 2/1).

Misalignment of repeats on paired chromatids can also cause disease by *non-reciprocal* sequence exchange. Here, one of the interacting sequences remains unchanged, but the other is mutated (**gene conversion**—see [Figure 7.9](#)). See

Clinical Box 6 for a common single-gene disorder, steroid 21-hydroxylase deficiency, in which the pathogenesis is due almost entirely to sequence exchanges between misaligned long tandem repeats, resulting in deletion or gene conversion.

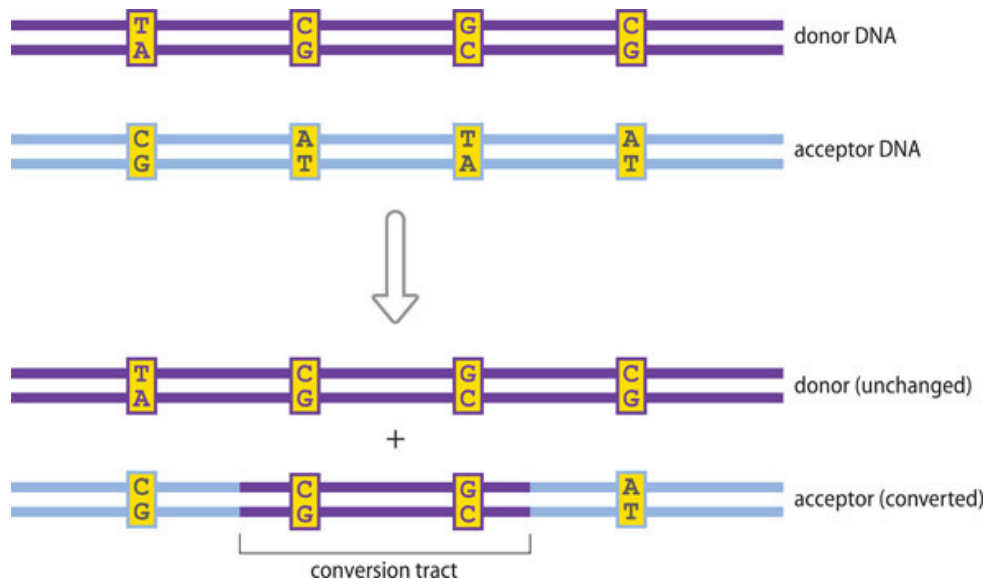


Figure 7.9 Principle of gene conversion. Gene conversion is a *nonreciprocal* sequence exchange between two related sequences that may be alleles or non-allelic (such as misaligned repeats on opposing non-sister chromatids). Sequence information is copied from one of the paired sequences, the donor sequence (which will remain unchanged) to replace an equivalent part of the other sequence, the acceptor sequence. The size of the sequence that is converted—the conversion tract—is often a few hundred nucleotides long in mammalian cells.

CLINICAL BOX 6 DISEASE PROFILE: STEROID 21-HYDROXYLASE DEFICIENCY, A DISORDER CAUSED BY GENE-PSEUDOGENE SEQUENCE EXCHANGES

Steroid 21-hydroxylase is a cytochrome P450 enzyme needed by the adrenal gland to produce the glucocorticoid hormone cortisol and aldosterone (which regulates sodium and potassium levels). Genetic deficiency in this enzyme is much the most common cause of congenital adrenal hyperplasia. In classical (congenital) forms of the disorder, excessive adrenal androgen bio-synthesis results in virilization of affected individuals, so that girls are often born with masculinized external genitalia. Classically affected individuals may have the “simple-virilizing” form of the disorder, but some also excrete large amounts of sodium in their urine, which

leads to potentially fatal electrolyte and water imbalance (“salt-wasting” phenotype).

Steroid 21-hydroxylase is encoded by a gene, *CYP21A2*, located in the class III region of the HLA complex. *CYP21A2* resides on an approximately 30 kb segment of DNA that is tandemly duplicated, with about 98 % sequence identity between the tandem 30 kb repeats. As a result, there is a closely related copy of the *CYP21A2* gene sequence on the other repeat, a pseudogene called *CYP21A1P*. The pseudogene has multiple deleterious mutations distributed across its length ([Figure 1A](#)) and does not make a protein.

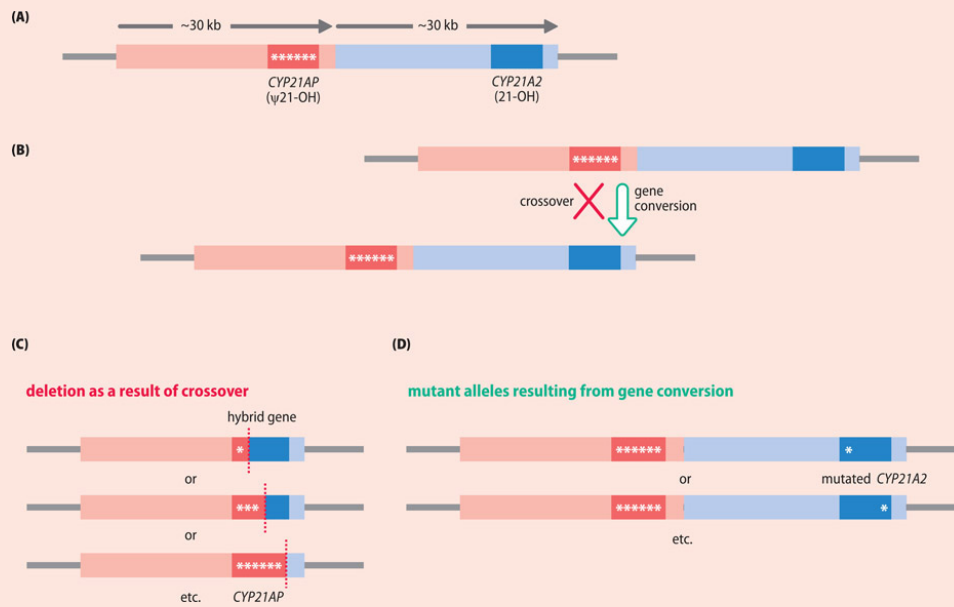


Figure 1 Tandem duplication of repeats containing the *CYP21A2* gene and a closely related pseudogene predispose to pathogenic sequence exchanges. (A) The steroid 21-hydroxylase gene, *CYP21A2*, and a closely related pseudogene, *CYP21A1P* (with multiple inactivating mutations shown by white asterisks), are located on tandemly duplicated ~30 kb repeats that also contain other genes (not shown here for simplification). (B) Mispairing of the non-allelic repeats is facilitated by the 98 % sequence identity between them. A subsequent reciprocal exchange occurring by recombination (shown by the red X) or a nonreciprocal exchange (gene conversion) can result in a loss of functional sequence. (C) Unequal crossover can produce chromosomes with a single 30 kb repeat containing a hybrid gene, part *CYP21A1P* and part *CYP21A2*, or just the *CYP21A1P* pseudogene (dashed vertical red lines indicate crossover points). (D) Gene conversion replaces a segment of the *CYP21A2* gene by equivalent sequence from the pseudogene. The conversion tract, often only a few hundred nucleotides long, can occur at different regions, introducing a copy of an inactivating mutation from any part of the pseudogene.

The very high sequence identity between the tandem repeats makes paired chromatids liable to local misalignment: a repeat containing the functional *CYP21A2* gene on one chromatid mispairs with the repeat containing the *CYP21AP* pseudogene on the other chromatid ([Figure 1B](#)).

In more than 99 % of cases with 21-hydroxylase deficiency, mispairing between the two repeats and subsequent sequence exchanges is thought to be responsible for pathogenesis. About 75 % of the mutations that cause disease are roughly 30 kb deletions caused by unequal crossover or unequal sister chromatid exchange. If the crossover point occurs between gene and pseudogene, a single nonfunctional hybrid 21-hydroxylase gene results; or if it is located just beyond the paired gene and pseudogene, it leaves just the 21-hydroxylase pseudogene ([Figure 1C](#)).

The remaining 25 % or so of pathogenic mutations are point mutations, but in the vast majority of cases the point mutation is introduced into the *CYP21A2* gene by a gene conversion event that copies a sequence containing a deleterious mutation from the pseudo-gene, replacing the original sequence ([Figure 1D](#) and [Table 1](#)).

TABLE 1 PATHOGENIC POINT MUTATIONS IN THE STEROID 21-HYDROXYLASE GENE ARE COPIED FROM A CLOSELY RELATED PSEUDOGENE

Mutation class and location	Normal 21 -OH gene sequence (<i>CYP21A2</i>)	Pathogenic point mutation	Equivalent <i>CYP21A2P</i> pseudogene sequence
Intron 2, splicing mutation	CCCACCCTCC	CCCAGCCTCC	CCCAGCCTCC
Exon 3, deletion of 8 ntds. within codons 111-113	GGA GAC TAC TCx Gly Asp Tyr Ser	G-----TCx V)al	G-----TCx
Exon 4, missense mutation I173N	ATC ATC TGT Ile Ile Cys	ATC AAC TGT Ile Asn Cys	ATC AAC TGT
Exon 4, clustered missense mutations in codons 237-240	ATC GTG GAG ATG Ile Val Glu Met	AAC GAG GAG AAG Asn Glu Glu Lys	AAC GAG GAG AAG

The gene conversion tract (the region copied from the pseudogene sequence) is usually no longer than a few hundred nucleotides.

Mutation class and location	Normal 21 -OH gene sequence (<i>CYP21A2</i>)	Pathogenic point mutation	Equivalent <i>CYP21A2P</i> pseudogene sequence
Exon 7, missense mutation: V282L	CAC GTG CAC His Val His	CAC T TG CAC His Leu His	CAC T TG CAC
Exon 8, nonsense mutation: Q319X	CTG CAG GAG Leu Gln Glu	CTG TAG GAG Leu S TOP	CTG T AG GAG
Exon 8, missense mutation: R357W	CTG CGG CCC Leu Arg Pro	CTG T GG CCC Leu T rp Pro	CTG T GG CCC

The gene conversion tract (the region copied from the pseudogene sequence) is usually no longer than a few hundred nucleotides.

Disease arising from sequence exchanges between distantly located repeats in nuclear DNA

Homologous repeats that are separated by a sizable intervening sequence can also predispose to nonallelic homologous recombination in which sequence exchanges occur between misaligned repeats. The repeats may be **direct repeats** (oriented in the same 5' → 3' orientation); in that case, the intervening sequence between the repeats can be deleted or duplicated. That may mean loss or duplication of exons, which can be frameshifting mutations, or loss or duplication of multiple genes (both of which can be pathogenic, as described below). Exchange between *inverted repeats* (repeats oriented in opposite 5' → 3' directions) can also cause inversion of the intervening sequence; this can also be pathogenic, as illustrated in the last subsection below.

Chromosome microdeletions and microduplications

Just as with mispairing of tandem repeats, distantly spaced direct repeats can mispair when chromatids are aligned within homologous chromosomes. Subsequent recombination at mismatched direct repeats results in deletions or duplications of the intervening sequence ([Figure 7.10A](#)). An equivalent type of exchange can also

occur between mispaired short direct repeats on the *same* DNA strand (a form of intrachromatid recombination), and this can also lead to deletion ([Figure 7.10B](#)).

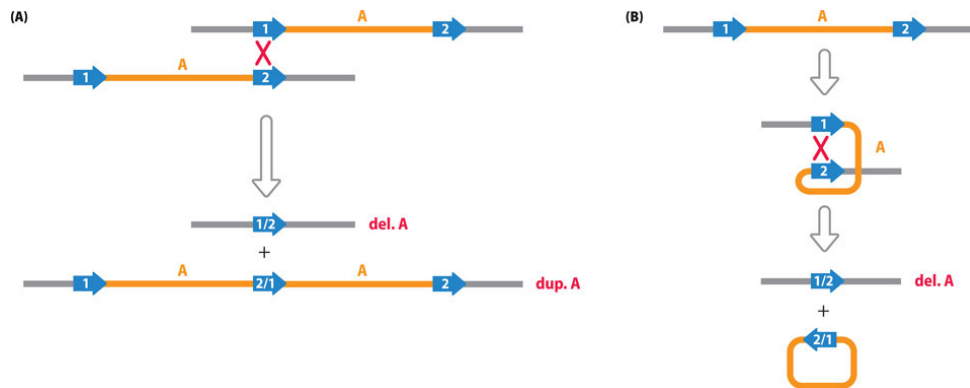


Figure 7.10 Deletion/duplication events due to nonallelic homologous recombination between low-copy-number direct repeats. Block arrows indicate highly similar low-copy-number *direct repeats* (oriented in the same direction). (A) DNA sequence A is flanked by low-copy-number repeats 1 and 2 that have identical or highly similar nucleotide sequences. Chromatid misalignment can occur so that repeat 1 on one chromatid pairs with repeat 2 on the other chromatid, and subsequent crossover can result in a chromatid with a deletion of sequence A (del. A) and one with two copies of sequence A (dup. A). (B) An intrachromatid “recombination” between direct repeats on the same DNA molecule can also produce a deletion of sequence A. If the other product, a circular DNA containing sequence A, lacks a centromere, it will be lost after cell division.

The sizes of the duplication/deletions produced are frequently <6 Mb (see [Table 7.9](#)), and because they are not detectable under the light microscope using standard chromosome staining they have been described as chromosome *microdeletions* or *microduplications*. As shown in [Table 7.9](#) several examples are known to be associated with clinical phenotypes. We consider the effects on gene expression and the clinical impact in [Section 7.5](#).

TABLE 7.9 EXAMPLES OF CLINICAL PHENOTYPES ARISING THROUGH RECOMBINATION BETWEEN DISPERSED REPEATS ON A CHROMOSOME

Disorder	Location	Length of recombining repeats	Deletion (Δ) / duplication (dup.) size	Key disease locus
----------	----------	-------------------------------	---	-------------------

Abbreviations are: HNPP—hereditary neuropathy with liability to pressure palsies; VCFS—velocardiofacial syndrome.

Disorder	Location	Length of recombining repeats	Deletion (Δ) / duplication (dup.) size	Key disease locus
Azoospermia type AZFc	Yq11.2	230 kb	Δ 3.5 Mb	<i>DAZ</i> family
Angelman syndrome	15q11–q13	400 kb	paternal Δ 5 Mb	<i>UBE3A</i>
Prader-Willi syndrome			maternal Δ 5 Mb	<i>SNRPN</i>
HNPP	17p12	24 kb	Δ 1.4 Mb	PMP22
Charcot-Marie-Tooth 1A			dup. 1.4 Mb	
DiGeorge syndrome/VCFS	22q11.2	225–400 kb	Δ 3 Mb or 1.5 Mb	<i>TBX1</i>
Smith-Magenis syndrome	17p11.2	175–250 kb	Δ 4 Mb	RAI1
Potocki-Lupski syndrome			dup. 4Mb	
Williams-Beuren syndrome	7q11.2	300–400 kb	Δ 1.6 Mb	<i>ELN</i>
Sotos syndrome	5q25	400 kb	Δ 2 Mb	NSD1

Abbreviations are: HNPP—hereditary neuropathy with liability to pressure palsies; VCFS—velocardiofacial syndrome.

Intrachromatid recombination between inverted repeats

Inverted repeats on a single chromatid can also mispair by looping out the intervening sequence. Subsequent recombination at the mispaired sequences will produce an inversion of the intervening sequence ([Figure 7.11A](#)). Disease may result, for example by relocating part of a gene, by disrupting the gene, or by separating a gene from important *cis-acting* control elements.

An instructive example is provided by hemophilia A: in about 50 % of cases the cause is a large inversion that disrupts the *F8* gene, which makes blood clotting factor VIII. A low-copy-number repeat, *F8A1*, within intron 22 of the *F8* gene can

mispair with either of two very similar repeat sequences, *F8A1* and *F8A2*, that are located upstream of *F8* and in the opposite 5' to 3' orientation to *F8A1*. Subsequent recombination between mispaired *F8A* repeats produces an inversion of about 500 kb of intervening sequence, causing disruption of the *F8* gene ([Figure 7.11B](#)).

7.5 CHROMOSOME ABNORMALITIES

Many large-scale changes to our DNA sequences that cause diseases are more readily studied at the level of chromosomes, as are changes in chromosome copy number resulting from errors in chromosome segregation. In standard cytogenetic **karyotyping** suitable metaphase or prometaphase chromosome preparations are chemically stained to reveal a pattern of alternating light and dark bands under light microscopy, which are examined to identify chromosome abnormalities. See [Box 7.2](#) for some details of the techniques and relevant nomenclature. Alternative approaches to identify or screen for chromosome abnormalities are detailed in [Chapter 11](#).

BOX 7.2 HUMAN CHROMOSOME BANDING AND ASSOCIATED NOMENCLATURE

CHROMOSOME PREPARATION AND CHROMOSOME BANDING METHODS

To study chromosomes under the light microscope, the chromosomes must be suitably condensed—metaphase (or prometaphase) chromosome preparations are required. A peripheral blood sample is taken and separated white blood cells are stimulated to divide by using a mitogen such as phytohemagglutinin. The white blood cells are grown in a rich culture medium containing a spindle-disrupting agent (such as Colcemid) to maximize the number of metaphase cells (cells enter metaphase but cannot progress through the rest of M phase). Prometaphase preparations can also be obtained; they have slightly less-condensed chromosomes, making analysis easier.

Chromosome banding involves treating chromosome preparations with denaturing agents; alternatively they are digested with enzymes and then exposed to a dye that can bind to DNA. Some dyes preferentially bind to AT-rich sequences; others bind to GC-rich sequences. The dyes show differential binding to different

regions across a chromosome that will reflect the relative frequencies of AT and GC base pairs.

The most commonly used method in human chromosome banding is **G-banding**. The chromosomes are treated with trypsin and stained with Giemsa, which preferentially binds AT-rich regions, producing alternating dark bands (Giemsa-positive; AT-rich) and light bands (Giemsa-negative; GC-rich). Because genes are preferentially associated with GC-rich regions, dark bands in G-banding are gene-poor; light bands are gene-rich.

HUMAN CHROMOSOME AND CHROMOSOME BANDING NOMENCLATURE

Human chromosome nomenclature is decided periodically by the International Standing Committee on Human Cytogenetic Nomenclature; see under Further Reading for the most recent ISCN report published in 2020. The nomenclature assigns numbers 1–22 to the autosomes according to perceived size, and uses the symbols p and q to denote, respectively, the short and long arms of a chromosome. Depending on the position of the centromere, chromosomes are described as *metacentric* (centromere at or close to the middle of the chromosome), *submetacentric* (centromere some distance from the middle and from telomeres), or *acrocentric* (centromere close to a telomere).

Each chromosome arm is subdivided into a number of regions, according to consistent and distinct morphological features (depending on the size of the chromosome arm, there may be from one to three regions). Each region is in turn divided into bands, and then into sub-bands and sub-sub-bands, according to the banding resolution ([Figure 1](#)).

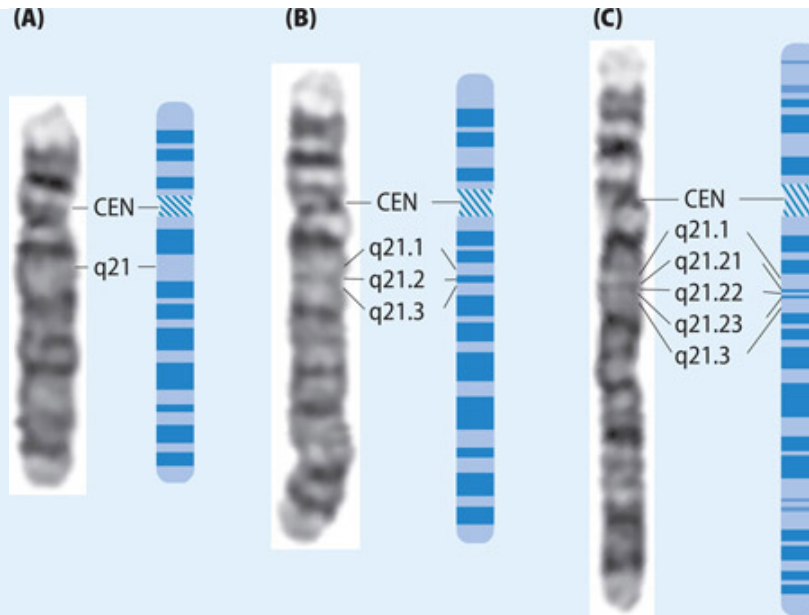


Figure 1 Chromosome banding resolutions can resolve bands, sub-bands and sub-sub-bands. G-banding patterns for human chromosome 4 (with accompanying ideogram to the right) are shown at increasing levels of resolution. The levels correspond approximately to (A) 400, (B) 550, and (C) 850 bands per haploid set, allowing the visual subdivision of bands into sub-bands and sub-sub-bands as the resolution increases. CEN, centromere. For an example of a full set of banded chromosomes, see [Figure 2.8](#). (Adapted from Cross I & Wolstenholme J [2001] in *Human cytogenetics: constitutional analysis*, 3rd ed (Rooney DE, Ed). With permission from Oxford University Press.)

The numbering of regions, bands, sub-bands, and sub-sub-bands is done according to relative proximity to the centromere. If a chromosome arm has three regions, the region closest to the centromere would be region 1, and the one closest to the telomere would be region 3. For example, the band illustrated in [Figure 1A](#) would be known as 4q21 for these reasons: it is located on the long arm of chromosome 4 (= 4q); it resides on the second (of three regions) on this chromosome arm (= 4q2); within this region it is the nearest band (band 1) to the centromere. The last two digits of band 4q21 are therefore pronounced two-one (not twenty-one) to mean region two, band one. Similarly, in [Figure 1C](#) the numbers following 4q in the sub-sub-band 4q21.22 are pronounced two-one-point-two-two.

Note also that in chromosome nomenclature the words **proximal** and **distal** are used to indicate the relative position on a chromosome with respect to the centromere. Thus, proximal Xq means the segment of the long arm of the X that is

closest to the centromere. Similarly, distal 3p means the portion of the short arm of chromosome 3 that is most distant from the centromere (= closest to the telomere).

According to their distribution in cells of the body, chromosome abnormalities can be classified into two types. A **constitutional** abnormality is present in all nucleated cells of the body and so must have been present very early in development. It can arise as a result of an abnormal sperm or egg, through abnormal fertilization, or through an abnormal event in the very early embryo. A somatic (or acquired) abnormality is present in only certain cells or tissues of a person, who is therefore a genetic **mosaic** (by possessing two populations of cells with altered chromosome or DNA content, each deriving from the same zygote).

Chromosomal abnormalities, whether constitutional or somatic, can be subdivided into two categories: structural abnormalities (which arise through chromosome breakage events that are not repaired), and numerical abnormalities (changes in chromosome number that often arise through errors in chromosome segregation). See [Table 7.10](#) for a guide to the nomenclature of human chromosome abnormalities.

TABLE 7.10 NOMENCLATURE OF CHROMOSOME ABNORMALITIES

Type of abnormality	Examples	Explanation/notes
NUMERICAL		
Triploidy	69,XXX, 69,XXY, 69,XYY	a type of polyploidy
Trisomy	47,XX,+21	gain of a chromosome is indicated by +
Monosomy	45,X	a type of aneuploidy; loss of an autosome is indicated by -
Mosaicism	47,XXX/46,XX	a type of mixoploidy
STRUCTURAL		
Deletion	46,XY,del(4)(p16.3)	terminal deletion (breakpoint at 4p16.3)

This is a short nomenclature; a more complicated nomenclature is defined by the current ISCN report that allows complete description of any chromosome abnormality; see [McGowan-Jordan J et al. \(2020\)](#) under Further Reading.

Type of abnormality	Examples	Explanation/notes
	46,XX,del(5) (q13q33)	interstitial deletion (5q13-q33)
Inversion	46,XY,inv(11) (p11p15)	paracentric inversion (breakpoints on same arm)
Duplication	46,XX,dup(1) (q22q25)	duplication of region spanning 1 q22 to 1 q25
Insertion	46,XX,ins(2) (p13q21q31)	a rearrangement of one copy of chromosome 2 by insertion of segment 2q21—q31 into a breakpoint at 2p13
Ring chromosome	46,XY,r(7) (p22q36)	joining of broken ends at 7p22 and 7q36
Marker	47,XX,+mar	indicates a cell that contains a marker chromosome (an extra unidentified chromosome)
Reciprocal translocation	46,XX,t(2;6) (q35;p21.3)	a balanced reciprocal translocation with breakpoints at 2q35 and 6p21.3
Robertsonian translocation (gives rise to one derivative chromosome)	45,XY,der(14;21) (q10;q10)	a balanced carrier of a 14;21 Robertsonian translocation. q10 is not really a chromosome band, but indicates the centromere; der— derivative—is used when one chromosome from a translocation is present
	46,XX,der(14;21) (q10;q10),+21	an individual with Down syndrome possessing one normal chromosome 14, a Robertsonian translocation 14;21 chromosome, and two normal chromosome 21s

This is a short nomenclature; a more complicated nomenclature is defined by the current ISCN report that allows complete description of any chromosome abnormality; see [McGowan-Jordan J et al. \(2020\)](#) under Further Reading.

Structural chromosomal abnormalities

As detailed in [Section 4.1](#), abnormal chromosome breaks (caused by double-strand DNA breaks) occur as a result of unrepaired damage to DNA or through faults in the recombination process. Chromosome breaks that occur during the G2 phase (after the DNA has replicated) are really *chromatid* breaks: they affect only one of the two sister chromatids. The breaks occurring during the G1 phase that are not repaired by S phase (when the DNA replicates) become chromosome breaks (both sister chromatids are affected). A cell with highly damaging chromosome breaks may often be removed by triggering cell death mechanisms; if it survives with unrepaired breaks, chromosomes with structural abnormalities can result.

Errors in recombination that produce structural chromosome abnormalities can occur at meiosis. Paired homologs are normally subjected to recombination mechanisms that ensure the breakage and rejoining of non-sister chromatids, but if recombination occurs between mispaired homologs, the resulting products may have structural abnormalities. Intrachromatid recombination can also be a source of structural abnormalities.

A form of somatic recombination also occurs naturally in B and T cells in which the cellular DNA undergoes programmed rearrangements to make antibodies and T cell receptors. Abnormalities in these recombination processes can also cause structural chromosomal abnormalities that may be associated with cancer (described in [Chapter 10](#)).

Structural chromosome abnormalities are often the result of incorrect joining together of two broken chromosome ends. Different mechanisms are possible, as detailed in the following subsections.

Microdeletions and microduplications

As detailed in [Section 7.3](#), exchanges between mispaired repeats on opposing chromatids, or even in the same chromatid can produce duplications and deletions several Mb long within a chromosome arm. Such microdeletions and microduplications result in simultaneous change in copy number of genes in the affected region, and clinical phenotypes can result if any of these genes are especially *dosage-sensitive* (just as they do in the case of whole chromosome duplication, as in trisomy 21).

Microduplications can cause disease by increasing the copy number of a single dosage-sensitive gene. For example, Charcot-Marie-Tooth syndrome 1A (OMIM 118220) can be caused by duplications at 17p12 that include the dosage-sensitive

PMP22 gene (which makes peripheral myelin protein 22). Having three copies of *PMP22* instead of the normal two copies is sufficient to cause problems for cells, as are activating point mutations in *PMP22* that cause overexpression. Microdeletions can cause disease in different ways, as shown by the examples listed in [Table 7.11](#).

TABLE 7.11 SOME OF THE DIFFERENT WAYS IN WHICH CHROMOSOMAL MICRODELETIONS CAUSE DISEASE

Cause of disease	Examples	Key disease gene(s) / allele
Reduced copy number of a single dosage-sensitive gene	Deletions at 16p13.3 associated with Rubinstein-Taybi syndrome (OMIM 180849)	dosage-sensitive <i>CBP</i> gene
	Deletions at 20p12.1 associated with Alagille syndrome type 1 (OMIM 118450)	dosage-sensitive <i>JAG1</i> gene
Reduced copy number of >1 dosage-sensitive gene (<i>segmental aneuploidy</i>)	Deletions at 11 p13 associated with WAGR (Wilms tumor, aniridia, genitourinary abnormalities and mental retardation—OMIM 194072)	dosage-sensitive <i>PAX9</i> , <i>WT1</i> genes
Loss of the active allele of one or more imprinted genes	Deletions of maternal 15q 11 -q 13 associated with Angelman syndrome (OMIM 105830)	maternal <i>UBE3A</i> allele (the only active allele)
	Deletions of paternal 15q 11 -q 13 associated with Prader-Willi syndrome (OMIM 176270)	paternal <i>SNRPN</i> allele (the only active allele)
Loss of the only allele for genes on the male X-chromosome	Contiguous gene syndrome causing Duchenne muscular dystrophy, chronic granulomatous disease and retinitis pigmentosa (PMID 4039107)	DMD, <i>CYBB</i> , <i>RPGR</i> (single alleles in males due to hemizyosity)

Large-scale duplications, deletions, and inversions

Still larger changes can occur when breaks occur in both arms of a chromosome. If a single chromosome sustains two breaks, incorrect joining of fragments can result in chromosome material being lost (deletion), switched round in the reverse direction (inversion), or included in a circular chromosome (a ring chromosome) ([Figure 7.12](#)).

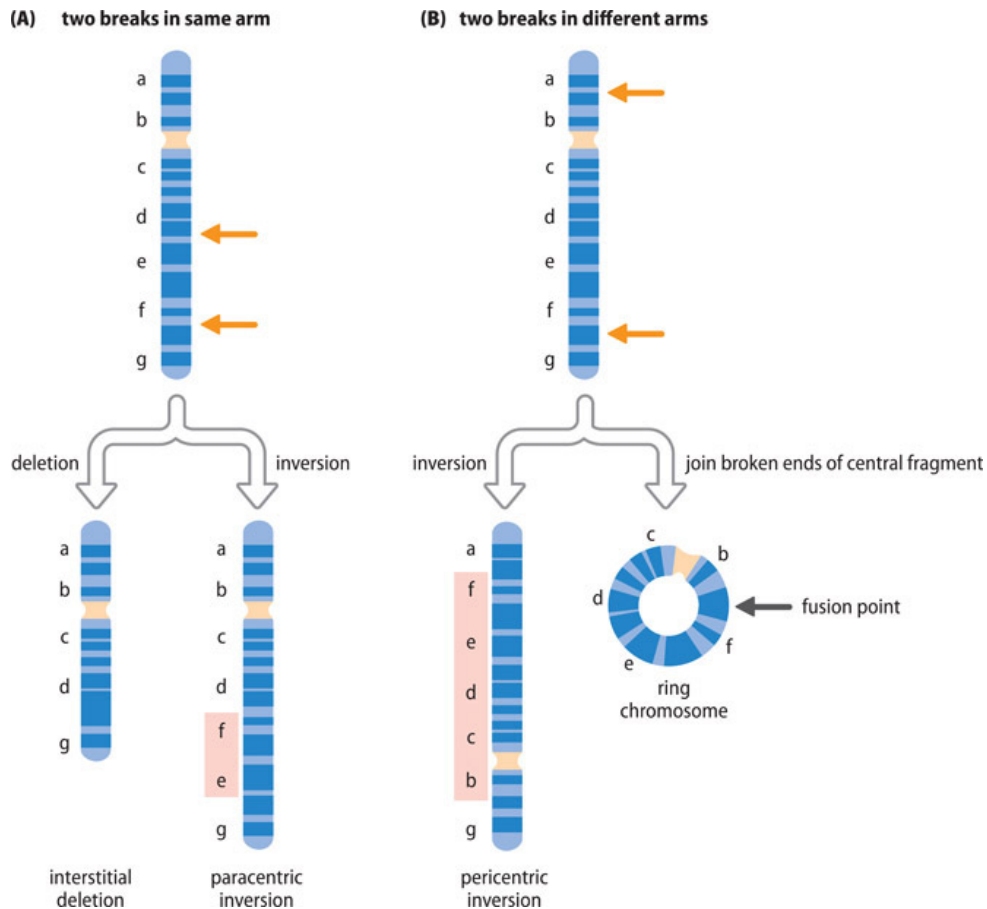


Figure 7.12 Stable outcomes after incorrect repair of two breaks on a single chromosome. (A)

Incorrect repair of two breaks (orange arrows) occurring in the same chromosome arm can involve loss of the central fragment (here containing hypothetical regions e and f) and rejoining of the terminal fragments (deletion), or inversion of the central fragment through 180° and rejoining of the ends to the terminal fragments (called a paracentric inversion because it does not involve the centromere). (B) When two breaks occur on different arms of the same chromosome, the central fragment (encompassing hypothetical regions b to f in this example) may invert and rejoin the terminal fragments (pericentric inversion). Alternatively, because the central fragment contains a centromere, the two ends can be joined to form a stable ring chromosome, while the acentric distal fragments are lost. Like other repaired chromosomes that retain a centromere, ring chromosomes can be stably propagated to daughter cells.

Structurally abnormal chromosomes with a single centromere can be stably propagated through successive rounds of mitosis. However, any repaired chromosome that lacks a centromere (an *acentric* chromosome) or possesses two centromeres (a *dicentric* chromosome) will normally not segregate stably at mitosis, and will eventually be lost.

Chromosomal translocations

If two different chromosomes each sustain a single break, incorrect joining of the broken ends can result in the movement of chromosome material between chromosomes (**translocation**). A *reciprocal translocation* is the general term used to describe an exchange of fragments between two chromosomes ([Figure 7.13A](#)). If an acentric fragment from one chromosome (one that lacks a centromere) is exchanged for an acentric fragment from another, the products each have a centromere and are stable in mitosis. Structurally rearranged chromosomes like this that have a centromere are known as **derivative chromosomes**. Exchange of an acentric fragment for a centric fragment results in acentric and dicentric chromosomes that are normally unstable in mitosis (but see below for an exceptional class of translocations in which dicentric products are stable).

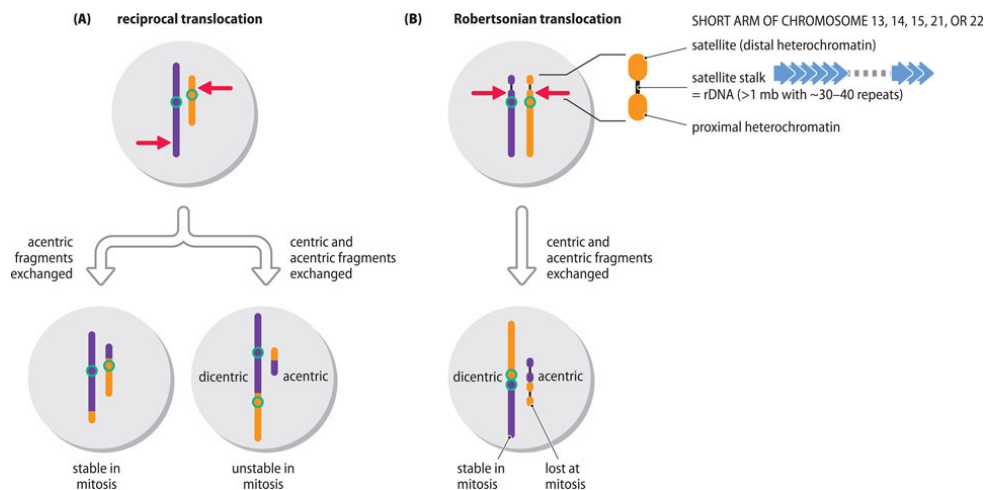


Figure 7.13 Reciprocal and Robertsonian translocations. (A) Reciprocal translocation. The *derivative* chromosomes produced by the translocation are stable in mitosis when one acentric fragment is exchanged for another, but when a centric fragment is exchanged for an acentric fragment, the derivative chromosomes are usually unstable. (B) Robertsonian translocation (centric fusion). This is a highly specialized reciprocal translocation in which exchange of centric and acentric fragments produces a dicentric chromosome that is nevertheless stable in mitosis. It occurs

exclusively after breaks in the short arms of two of the acrocentric chromosomes 13, 14, 15, 21, and 22. As illustrated, the short arms of the human acrocentric chromosomes have a common structure. A region of distal heterochromatin (called a *satellite*) is joined to a proximal heterochromatic region by a thin *satellite stalk* made up of *ribosomal DNA* (rDNA), an array of tandem DNA repeats that each make three types of rRNA. Breaks that occur close to the centromere can result in a dicentric chromosome in which the two centromeres are so close that they can function as a single centromere. The loss of the small acentric fragment has no phenotypic consequences.

For a chromosomal translocation to occur, the regions containing double-strand DNA breaks in the two participating chromosomes must be in close proximity (allowing incorrect joining prior to double-strand break repair). The spatial distribution of chromosomes in the nucleus is not random, and chromosomes tend to occupy certain “territories”. Chromosomes that tend to be physically closer to each other are more likely to engage in translocation with each other. For example, human chromosomes 4, 13, and 18 are preferentially located at the periphery of the nucleus and frequently translocate with each other but not with physically distant chromosomes localized in the interior of the nucleus. Specific types of translocations are common in certain cancers, and may reflect close physical association of the two chromosomal regions that participate in translocation.

One exceptional form of chromosome association occurs between the very small short arms of the five human acrocentric chromosomes (chromosomes 13, 14, 15, 21, and 22). Each of the short arms of these chromosomes has about 30–40 large tandem DNA repeats, each containing sequences for making three ribosomal RNAs: 28S, 18S, and 5.8S rRNAs; the five ribosomal DNA (rDNA) regions congregate at the nucleolus to produce these rRNAs. The close physical association of these five chromosome arms is responsible for a specialized type of translocation, called Robertsonian translocation or centric fusion, that involves breaks in the short arms of two different acrocentric chromosomes followed by exchange of acentric and centric fragments to give acentric and dicentric products ([Figure 7.13B](#)).

The acentric chromosome produced by a Robertsonian translocation is lost at mitosis without consequence (it contains just highly repetitive noncoding DNA plus rRNA genes that are also present at high copy number on the other acrocentric chromosomes). The other product is an unusual dicentric chromosome that is stable in mitosis: the two centromeres are in close proximity (centric fusion) and often function as one large centromere so that the chromosome segregates regularly. (Nevertheless, such a chromosome may present problems during gametogenesis.)

More complex translocations can involve multiple chromosome breakages. Insertions typically require at least three breaks: fragment liberated by two breaks in one chromosome arm inserts into another break located in another region of the same chromosome or in a different chromosome.

Isochromosomes

An additional rare class of structural abnormality is a symmetrical **isochromosome** in which the arms of the chromosome are mirror images of one another, having either two long arms or two short arms ([Figure 7.14](#)). The centromere *appears* to divide transversely instead of longitudinally. Isochromosomes can form by a type of misdivision of the centromere but more commonly arise through the breakage and fusion of sister chromatids. Overall the effect equates to a combined deletion-duplication event (deletion of one chromosome arm and duplication of the other). Human isochromosomes are rare, except for i(Xq) and also i(21q), an occasional contributor to Down syndrome.

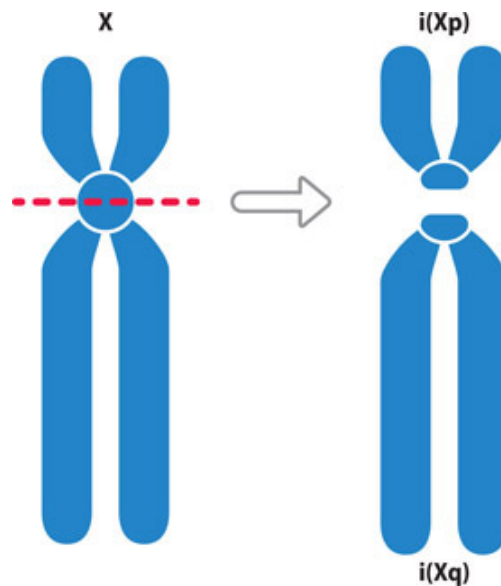


Figure 7.14. Isochromosome formation. This can sometimes occur by a type of misdivision of the centromere as shown here (for alternative origins, see text).

Chromosomal abnormalities involving gain or loss of complete chromosomes

Three classes of numerical chromosomal abnormalities can be distinguished: polyploidy and aneuploidy (summarized in [Table 7.12](#)), plus mixoploidy (described below).

TABLE 7.12 CLINICAL CONSEQUENCES OF NUMERICAL CHROMOSOME ABNORMALITIES

Abnormality	Clinical consequences
POLYPLOIDY	
Triploidy (69,XXX or 69,XYY)	1 -3 % of all conceptions; almost never born live and do not survive long
ANEUPLOIDY (AUTOSOMES)	
Nullisomy (missing a pair of homologs)	lethal at pre-implantation stage of embryonic development
Monosomy (one chromosome missing)	lethal during embryonic development
Trisomy (one extra chromosome)	usually lethal during embryonic [*] or fetal [*] stages, but individuals with trisomy 13 (Patau syndrome) and trisomy 18 (Edwards syndrome) may survive to term; those with trisomy 21 (Down syndrome) may survive beyond age 40
ANEUPLOIDY (SEX CHROMOSOMES)	
Additional sex chromosomes	individuals with 47,XXX, 47,XXY, and 47,XYY all experience relatively minor problems and a normal lifespan
Lacking a sex chromosome	while 45,Y is never viable, in 45,X (Turner syndrome), about 99 % of cases abort spontaneously; survivors are of normal intelligence but are infertile and show minor physical diagnostic characteristics

^{*} In humans, the embryonic period spans fertilization to the end of the eighth week of development; fetal development then begins and lasts until birth.

Polyploidy

Three per cent of recognized human pregnancies produce a triploid embryo ([Figure 7.15A](#)). The usual cause is two sperm fertilizing a single egg (dispermy), but triploidy is sometimes attributable to fertilization involving a diploid gamete. With three copies of every autosome, the dosage of autosomal genes might be expected to be balanced, but triploids very seldom survive to term, and the condition is not compatible with life (but diploid/triploid mosaics can survive). The lethality in triploids may be due to an imbalance between products encoded on the X chromosome and autosomes, for which X-chromosome inactivation would be unable to compensate.

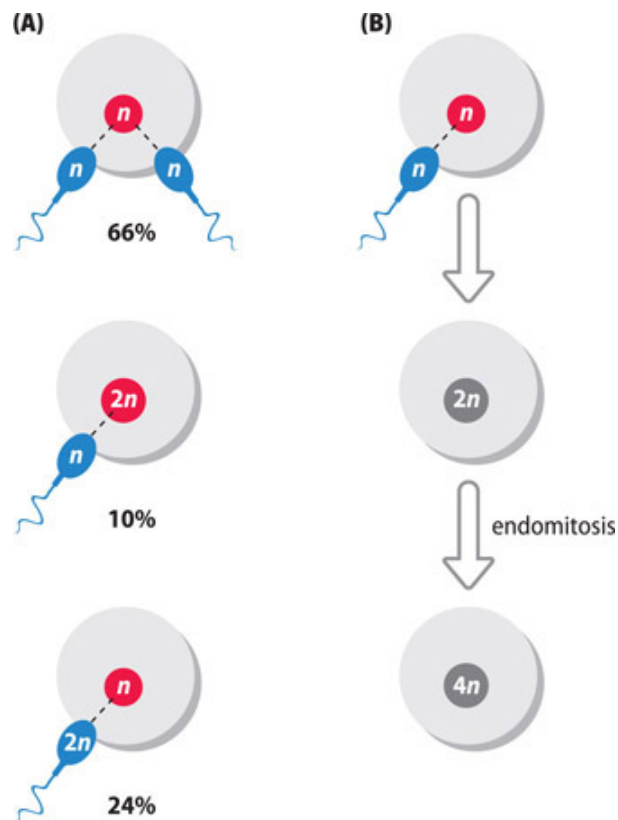


Figure 7.15 Origins of triploidy and tetraploidy. (A) Origins of human triploidy. Dispermy (top) is the principal cause, accounting for 66 % of cases. Triploidy is also caused by diploid gametes that arise by occasional faults in meiosis, such as nondisjunction (see [Figure 7.16](#)); fertilization of a diploid ovum (middle) and fertilization by a diploid sperm (bottom) account for 10 % and 24 % of cases, respectively. (B) Tetraploidy involves normal fertilization and fusion of gametes to give a normal zygote. Subsequently, however, tetraploidy arises when DNA replicates without subsequent cell division (*endomitosis*).

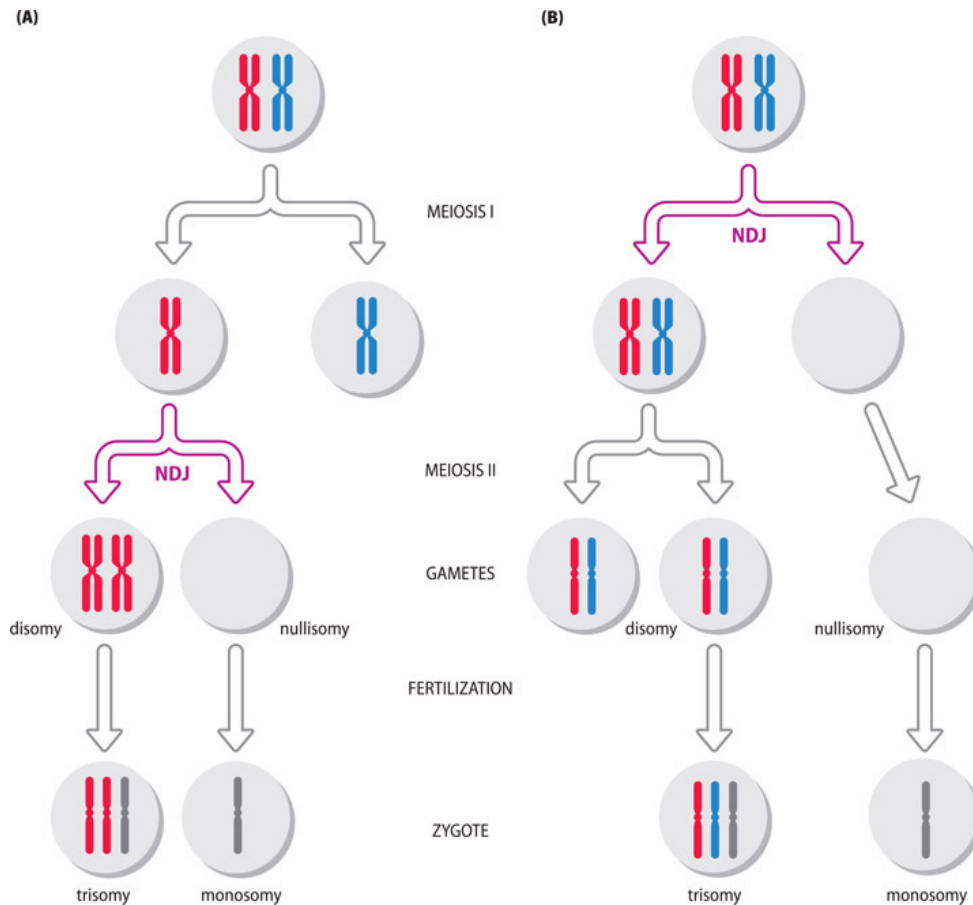


Figure 7.16 Meiotic nondisjunction and its consequences. In meiotic nondisjunction (NDJ) a pair of homologous chromosomes fail to disjoin and migrate to the same daughter cell, instead of being individually allocated to each daughter cell. That can happen at the second meiotic division (A) or at the first meiotic division (B). The result is that a person produces abnormal gametes, with either no copies of a chromosome (nullisomy) or two copies of a chromosome (disomy). In the latter case, the two chromosome copies may originate from a single parent (for NDJ at meiosis I) or individually from each parent (for NDJ at meiosis I). Fertilization with a normal gamete containing one copy of the chromosome (shown in gray) results in trisomy or monosomy.

Tetraploidy ([Figure 7.15B](#)) is much rarer and always lethal. It is usually due to failure to complete the first zygotic division: the DNA has replicated to give a content of 4C, but cell division has not then taken place as normal. Although constitutional polyploidy is rare and lethal, some types of cell are naturally polyploid in all normal individuals—for example, our muscle fibers are formed by recurrent cell fusions that result in multinucleate syncytial cells.

Aneuploidy

Normally our nucleated cells have complete chromosome sets (euploidy), but sometimes one or more individual chromosomes are present in an extra copy or are missing (**aneuploidy**). In trisomy, three copies of a particular chromosome are present in an otherwise diploid cell, such as trisomy 21 (47,XX,+21 or 47,XY,+21) in Down syndrome. In monosomy a chromosome is lacking from an otherwise diploid state, as in monosomy X (45,X) in Turner syndrome. Cancer cells often show extreme aneuploidy, with multiple chromosomal abnormalities.

Aneuploid cells arise through two main mechanisms. One is **nondisjunction**, in which paired chromosomes fail to separate (disjoin) during meiotic anaphase I and migrate to the same daughter cell, or sister chromatids fail to disjoin at either meiosis II or mitosis. Nondisjunction during meiosis produces gametes with either 22 or 24 chromosomes, which after fertilization with a normal gamete produce a trisomic or monosomic zygote ([Figure 7.16](#)). If nondisjunction occurs during mitosis, the individual is a mosaic with a mix of normal and aneuploid cells.

Aneuploidy can also occur by **anaphase lag**. If, during anaphase, a chromosome or chromatid is delayed in its movement and lags behind the others, it may not be incorporated into one of the daughter nuclei. Chromosomes that do not enter a daughter cell nucleus are eventually degraded.

Having the wrong number of chromosomes has serious, usually lethal, consequences ([Table 7.12](#)). Even though the extra chromosome 21 in a person with trisomy 21 (Down syndrome) is a perfectly normal chromosome, inherited from a normal parent, its presence causes multiple abnormalities that are present from birth (congenital). Embryos with trisomy 13 or trisomy 18 can also survive to term but both result in severe developmental malformations, respectively Patau syndrome and Edwards syndrome. Other autosomal trisomies are not compatible with life. Autosomal monosomies have even more catastrophic consequences than trisomies and are invariably lethal at the earliest stages of embryonic life. We consider in [Section 7.7](#) how gene dosage problems in aneuploidies result in disease.

Maternal age effects in Down syndrome

In principle, the nondisjunction that causes a gamete to have an extra copy of chromosome 21 could occur at either meiotic division in spermatogenesis or oogenesis, but in practice in about 70 % of cases it occurs at meiosis I in the mother.

This is almost certainly a consequence of the extremely long duration of meiosis I in females (it begins in the third month of fetal life but is arrested and not completed until after ovulation). That means this one meiotic division can take decades; by contrast, male meiosis occurs continuously in the testes from puberty to old age. Not only is there a sex difference in the origin of the extra chromosome 21 but there is also a very significant *maternal age effect*. Thus the risk of having a Down syndrome child in a 20-year-old pregnant woman is about 1 in 1500, but that increases to about 1 in 25 for a 45-year-old woman.

Mixoploidy: mosaicism and chimerism

Mixoploidy means having two or more genetically different cell lineages within one individual. Usually, the different cell populations arise from the same zygote (*mosaicism*). More rarely, a person can have different cell populations that originate from different zygotes and is described as a **chimera**; spontaneous chimerism usually arises by the aggregation of fraternal twin zygotes or immediate descendant cells within the very early embryo. Abnormalities that would otherwise be lethal (such as triploidy) may not be lethal in mixoploid individuals.

Aneuploidy mosaics, with a proportion of normal cells and a proportion of aneuploid cells, are common. This type of mosaicism can result when non-disjunction or chromosome lag occurs in one of the mitotic divisions of the early embryo (any monosomic cells that are formed usually die). Polyploidy mosaics (such as human diploid/triploid mosaics) are occasionally found. As the gain or loss of a haploid set of chromosomes by mitotic nondisjunction is extremely unlikely, human diploid/triploid mosaics most probably arise by fusion of the second polar body with one of the cleavage nuclei of a normal diploid zygote.

7.6 MOLECULAR PATHOLOGY OF MITOCHONDRIAL DISORDERS

Up until now we have focused on disease resulting from changes to the nuclear DNA where the vast majority (99.9 %) of our genes reside. The very small (16.6 kb) circular mitochondrial genome is thought to have originated after an anaerobic proto-eukaryotic cell engulfed an aerobic proteobacterium in a type of co-operative symbiosis about two billion years ago, when the amount of oxygen in the

atmosphere was increasing rapidly. The internalized cell (endosymbiont) was not destroyed; instead, its genome and protein synthesis capacity were retained.

Over time, the genome of the engulfing cell expanded to become the large nuclear genome. Many of the sequences of the internalized cell's genome—and their functions—were inserted into the nucleus, and the genome of the endosymbiont became depleted (through genetic redundancy and sequence loss). Nuclear insertion of mtDNA sequences from degraded mitochondria is an evolutionarily ongoing process that continues to this day. It has resulted in a family of defective mtDNA sequences scattered across the nuclear genome, previously called mitochondrial pseudogenes but now known as NUMT (**n**uclear-**m**itochondrial) sequences. Some NUMTs have very high sequence similarity to regions of mtDNA and may hinder mtDNA analyses.

Our mitochondrial genome has a total of 37 RNA genes, producing 22 mitochondrial tRNAs and two mitochondrial rRNAs. It makes just 13 proteins, all components of the multisubunit protein complexes needed for oxidative phosphorylation. However, the close to 140 other proteins needed for oxidative phosphorylation (plus another >1100 proteins needed for mitochondrial function) are made by nuclear genes, translated on cytoplasmic ribosomes and imported into mitochondria. Because mitochondria act as a cell's power source, "mitochondrial disease" is the generic term for a clinically heterogeneous group of disorders that directly affect oxidative phosphorylation. Multiple systems may be involved, but the organs and tissues most often affected are usually those with high energy demands.

Mitochondrial disorders arising from mutation of nuclear genes show Mendelian inheritance: autosomal dominant, autosomal recessive and X-linked forms are all seen (plus occasional *de novo* mutations). Although there are just 37 genes in mtDNA, they all play vital roles in energy production in our cells. As a result, defects in mitochondrial genes cause, or contribute to, a very wide variety of genetic conditions.

In this section we consider disorders resulting from mtDNA variants of strong effect. This clinically diverse grouping includes many disorders that result from large deletions (often affecting multiple genes), or to point mutations. The disorders can be genetically heterogeneous: a single disorder may arise from mutations in any one of several possible genes in mtDNA, and in some cases mutations in nuclear genes may also produce the same phenotype. Because of some peculiar features of mitochondria, disorders due to mutations in mtDNA show some unique properties.

Pathogenic mtDNA variants can also make important contributions to complex genetic diseases, as described in [Chapter 8](#).

Mitochondrial disorders due to mtDNA mutation show maternal inheritance and variable proportions of mutant genotypes

During fertilization, a sperm injects its nuclear DNA into an oocyte. Paternal mtDNA is not normally transferred; even if it were, it would be recognized as being foreign and be degraded within the fertilized oocyte. As a result, mitochondrial genes appear to be exclusively maternally inherited—but NUMTs are inherited from both parents.

An important difference between nuclear DNA and mtDNA is the control of DNA replication and segregation of genetic material into daughter cells. For nuclear DNA sequences, DNA replication is very tightly constrained to ensure precise doubling of the amount of genetic material. And segregation of the genetic material into daughter cells is tightly controlled to ensure equal division of maternal and paternal sequences into the two daughter cells. As a result, if a person inherits a normal allele of gene *X* from one parent and a mutant allele from the other parent, the ratio of normal allele to mutant allele will be uniformly 1:1 in all diploid cells (barring any copy number changes), and gametes would contain either a normal or a mutant allele.

Unlike the nuclear genome, mtDNA turns over continuously and independently of the cell cycle, in both dividing and non-dividing cells (such as muscle cells and neurons). The mtDNA undergoes a type of *relaxed replication*: individual mtDNA molecules replicate at random, making sometimes multiple copies at a time (*clonal expansion*). Overall, an approximately constant total number of mtDNA molecules is maintained in the cell (but the total number per cells can vary according to the cell type, somewhere in a range of from 1000–10 000 copies in most cell types, and >100 000 copies in an egg cell). Different mtDNA variants can co-exist in the cells of a person—a situation known as **heteroplasmy**. But unlike for nuclear DNA, the ratio of a normal mtDNA sequence and a mutant one is not fixed, and can be unpredictably variable for different reasons, as listed below.

- *Differential replication of mutant and normal mtDNA*. From studies of model organisms we know that some mutant mtDNA sequences with large mtDNA deletions or point mutations in the major control sequence are able to replicate more rapidly than normal mtDNA copies. The proportion of mutant mtDNA in cells can then increase over time.

- *Vegetative segregation.* Unlike nuclear DNA, mtDNA molecules in a dividing cell are often segregated unequally to the daughter cells. The dividing cell splits unequally so that one daughter cell acquires the majority of both the cytoplasm and the mitochondria.
- *Genetic bottleneck in female germ line development.* Before giving rise to an egg cell, primary human oocytes originate from diploid primordial germ cells (PGCs) through a series of mitotic divisions. In very early PGCs a genetic bottleneck occurs whereby only a very small number of mitochondria (and mtDNA molecules) are transmitted, randomly, to daughter cells; as a result, the resulting cells can have very different proportions of mutant mtDNA. Subsequently, a large increase in mtDNA copy number occurs (to give ultimately >100 000 mtDNA copies in egg cells), but eggs from a woman can continue to show wide differences in the proportion of mutant mtDNA ([Figure 7.17](#)).

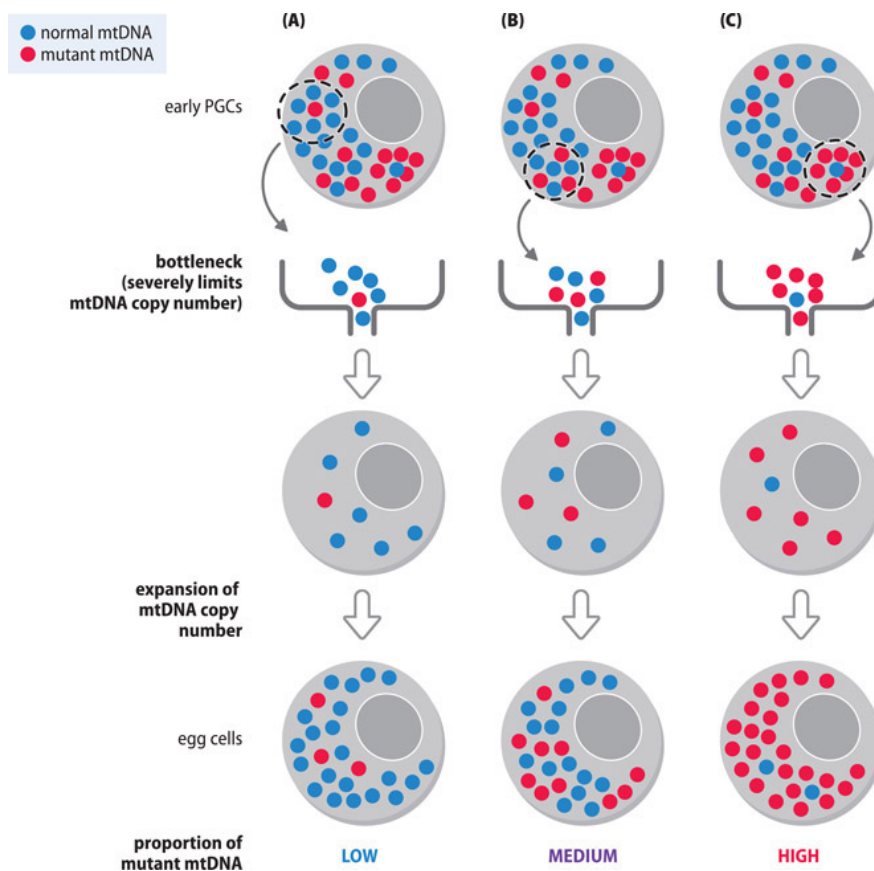


Figure 7.17. The mitochondrial genetic bottleneck: how a heteroplasmic woman can give rise to egg cells with quite different mutation loads. Mammalian egg cells are derived by two successive meiotic divisions from primary oocytes (not shown) that in

turn originate by sequential mitotic divisions from primordial germ cells (PGCs). In this example we consider how eggs might be produced from a mother who is approximately 50 % heteroplasmic for a pathogenic mtDNA mutation. The early PGCs might show 50 % heteroplasmy, but then because of the mitochondrial genetic bottleneck a very limited number of the mtDNAs in these cells are passed to daughter cells. Depending on which mtDNAs are transmitted through the bottleneck, later PGCs may have different proportions of mutant mtDNA, low (A), intermediate (B) or high (C). After the subsequent large increase in mtDNA copy numbers, such mutant frequencies may persist in the more mature egg cells.

The variable heteroplasmy in maternal eggs illustrated in [Figure 7.17](#) poses problems for genetic counselling. A woman affected by a mtDNA disorder has generally high levels of mutant mtDNA and because the disorder is maternally transmitted, there is a risk of an affected child every time she conceives. However, the different egg cells that she produces can show a wide range of heteroplasmy from small proportions of mutant mtDNA to high mutation loads. As we explore in the next section, the development of a disease phenotype is dependent on a threshold value of heteroplasmy that can vary according to the type of mitochondrial disorder.

The two major classes of pathogenic DNA variant in mtDNA: large deletions and point mutations

Associated pathogenic mutations have been identified in virtually all the genes in the mitochondrial genome. Although the frequency of pathogenic mtDNA mutations in the population is high (about 1 in 200 people), the majority of people carrying them may be unaffected (because of low proportions of mutant mtDNA; note that common mtDNA variants of weak effect may also play a role in common genetic disorders, such as Parkinson disease). Here, we focus on rare mtDNA variants of strong effect that come in two common types: those affecting single genes through point mutations, and large deletions that typically affect multiple genes. Because of variable heteroplasmy, the mutation load in cells and tissues can vary. When mutant mtDNAs reach a certain proportion of the total mtDNA, a *threshold* level, a biochemical phenotype resulting from inefficient oxidative phosphorylation results and clinical symptoms develop. As we show below, threshold heteroplasmy values can vary according to the nature of the mutation.

The connection between large mtDNA deletions and short repeats

Large-scale mtDNA deletions are frequent, and the region spanning nucleotide positions 5 700 to 16 500, almost two-thirds of the mitochondrial genome, has been viewed as a high frequency deletion zone. The deletions can be large—often from 4 kb to approaching 8 kb in length.

The high frequency of large deletions may be yet another case of repeat sequences predisposing to pathogenesis (previously summarized in [Table 7.1](#)). Examination of large pathogenic mtDNA deletions reveals that in many cases the sequence deleted in mtDNA is normally flanked by short, often perfect, direct repeats (see [Table 7.13](#) for examples). Physical association of the direct repeats may predispose to deletions, either during mtDNA replication (when the single-stranded repeats are exposed), or more likely during repair of double-strand breaks. Some mtDNAs with large deletions can replicate more rapidly than normal mtDNA copies. Large deletions may result in certain clinical phenotypes, as listed below:

- Kearns-Sayre syndrome: a multisystem disorder with onset before 20 years of age with progressive external ophthalmoplegia (PEO) and often pigmentary retinopathy; additional features may include cerebellar ataxia, impaired growth, hypoadrenalism, diabetes mellitus and cardiomyopathy and occasionally sensorineural hearing loss and cognitive impairment.
- Around one-third of cases have the common 4.977 kb deletion (see [Table 7.13](#)).
- Pearson syndrome: sideroblastic anemia and pancreas deficiency (commonly exocrine pancreas deficiency, but endocrine deficiency also often involved). Frequently a devastatingly fatal condition with profound anemia, thrombocytopenia, and lactic acidosis in the neonatal-infantile period.
- PEO: ptosis, impaired eye movements, oropharyngeal weakness, and variably severe proximal limb weakness.

Mitochondrial deletion disorders are rarely inherited: a best estimate is a 1 in 24 risk of inheritance. Usually, they originate by *de novo* deletion in the maternal germ line or in the very early embryo. Often, there is a quite low threshold of ~50 % to 60 % heteroplasmy—if the proportion of mutant mtDNA exceeds this level the bioenergetic capacity of tissues declines to a level where symptoms are evident. Interested readers can find a detailed review on mitochondrial deletion disorders at <https://www.ncbi.nlm.nih.gov/books/NBK1203/>)

TABLE 7.13 DIRECT REPEATS ARE HOTSPOTS FOR PATHOGENIC DELETIONS IN mtDNA *

Deletion size	Sequence and location of repeats	
	Repeat 1	Repeat 2
4.420 kb	10942 - 10951 AACAAACCCCC	15362 - 15371 AACAAACCCCC
4.977 kb	8470 - 8482 ACCTCCCTCACCA	13447 - 13459 ACCTCCCTCACCA
7.521 kb	7 975 - 7 982 AGGCGACC	15496 - 15503 AGGCGACC
7.664 kb	6325 - 6341 CCTCCGTAGACCTAACC	13989 - 14004 CCTCCTAGACCTAACC
7.723 kb	6076 - 6084 TCACAGCCC	11964 - 11972 TCACAGCCC

* Numbers adjacent to sequences are nucleotide co-ordinates in mtDNA—see [Figure 2.12](#) on page 45 for the mtDNA gene map. The pairs of direct repeats are identical except for a 1 bp difference in those causing the 7.664kb deletion. The 4.977 kb deletion is particularly common—see text.

Mitochondrial disorders arising from mtDNA point mutations

Pathogenic point mutations have been reported in virtually all genes in mtDNA. Whereas genetic redundancy offers protection against mutation in nuclear rRNA and tRNA genes (which are all present in multiple copies), the single-copy rRNA and tRNA genes in mtDNA are more exposed to harmful mutations, and pathogenic mutations have been identified in the *MT-RNR1* gene (making 12S rRNA), and all mitochondrial tRNA and protein-coding genes. Note, however, that as of January 2021, the MITOMAP database had confirmed pathogenic variants in 17 of the 22 tRNA genes.

Most of the associated clinical phenotypes are heteroplasmic, with varying thresholds beyond which biochemical and clinical phenotypes manifest. Most pathogenic mutations in mitochondrial tRNA genes show thresholds of >90 % heteroplasmy. Some mitochondrial disorders, notably Leber hereditary optic neuropathy, are said to be homoplasmic: the proportion of mutant mtDNA appears to be virtually 100 %.

Various mtDNA disorders show very considerable genetic heterogeneity, such as Leber hereditary optic neuropathy (LHON) and Leigh syndrome. The former disorder typically presents in young adults as bilateral, painless, subacute visual failure and the great majority of those who lose their vision do so before 50 years of

age. Leigh syndrome shows a similar phenotype. A case study, presented in [Clinical Box 7](#), shows how investigation of a mtDNA disorder initially believed to be LHON, revealed the cause to be a pathogenic point mutation associated with Leigh syndrome.

CLINICAL BOX 7 A CASE STUDY: LEIGH SYNDROME

Joanne, a healthy woman aged 41, was about to undergo her first round of IVF when her sister Carole told her that her grandson Archie had been diagnosed with Leigh syndrome (OMIM #256000), a genetically heterogeneous disorder (see [Figure 1](#) for the family tree). Joanne was advised by family members that the Leigh syndrome in her family was maternally inherited, and that she should postpone her plans to start a family until she had been tested. Joanne’s mother was in her 70s and generally well, but Carole had developed severe visual impairment at four years of age and remained registered blind.

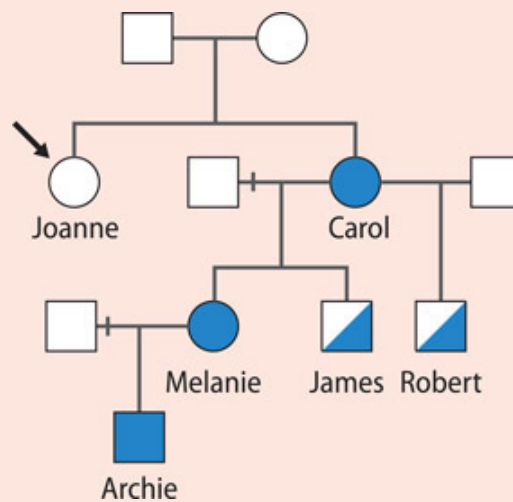


Figure 1 Family tree.

Carole’s daughter, Melanie, had also developed profound visual impairment (at age 23), presenting to the ophthalmology department with subacute onset of asymmetric painless central visual loss that became bilateral within days. Fundoscopy had revealed a pattern of optic disc pallor compatible with that observed in Leber hereditary optic neuropathy (LHON; OMIM 535000). However, investigations of the three common mtDNA variants associated with LHON (m.3460G>A, m.11778G>A and m.14484T>C, located respectively in the *MT-ND1*, *MT-ND4* and *MT-ND6* genes) had been negative.

Follow-up whole mtDNA sequencing revealed a pathogenic variant, m.13051G>A, known to be associated with Leigh syndrome, rather than LHON. This variant, occurring in the *MT-ND5* gene (which makes the NADH:ubiquinone oxidoreductase core subunit 5 protein), was present at 92 % heteroplasmy (that is, 92 % of the mtDNA molecules had the pathogenic variant; 8 % were normal). Melanie became pregnant at the age of 24 years, just a month before she received this genetic information. Melanie's brothers, James and Robert, remained well and both refused formal testing, but were counselled regarding maternal inheritance of this disease.

Melanie had a normal pregnancy and delivered her son Archie at 40 weeks and 5 days by caesarean section following a failed induction of labor. No resuscitation was required and Archie's early developmental milestones were considered normal, though he didn't walk independently until 17 months and remained unsteady with a broad-based gait when assessed at 2.5 years. Acute onset of strabismus and visual impairment at the age of 2.5 years, in the context of continued gait instability prompted further investigation of Archie including a cranial MRI scan. Bilateral, symmetrical foci of hypointense T1 and hyperintense T2 signal in the nigrostriatal pathways of the upper brainstem and floor of the 4th ventricle ([Figure 2](#)) raised the clinical suspicion of Leigh syndrome and, given Melanie's diagnosis, rapid analysis for the m.13051G>A variant was requested in blood DNA from Archie. This genetic testing confirmed a heteroplasmy of 96 %. Archie remains visually impaired with optic disc pallor and ataxia, but his developmental trajectory is comparable with that of his peers at present.

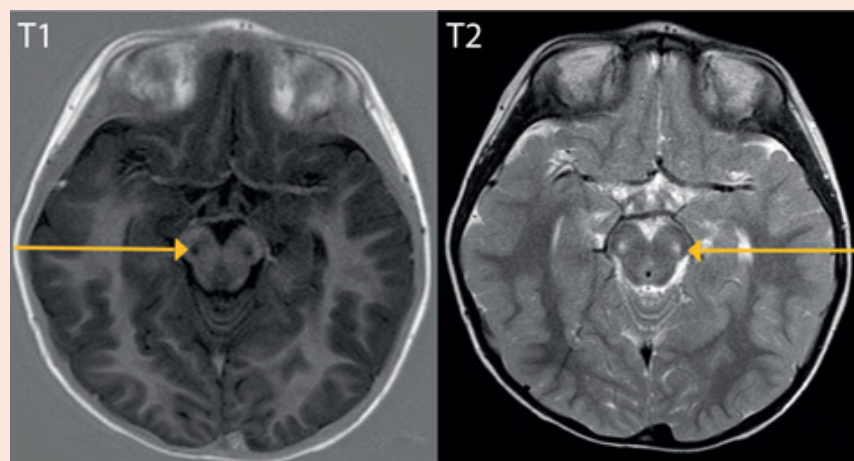


Figure 2 Cranial MRI scan of Archie's optic disc. Yellow arrows indicate the hypointense T1/hyperintense T2 signal change in the substantia nigra of the midbrain.

Further genetic testing within the family confirmed that Carole also harbored the m.13051G>A pathogenic variant but heteroplasmy in blood and urinary sediment DNA was considerably lower than that recorded in Melanie and Archie. Much to her relief, Joanne tested negative for m.13051G>A in blood and urinary sediment. The possibility that this pathogenic variant had arisen spontaneously in Carole and had been maternally transmitted thereafter could not be investigated because the mother of Joanne and Carole was not available for testing. Joanne has now resumed her plans for IVF.

The optic disc appearance and associated central scotoma that gave rise to the clinical diagnosis of LHON in Melanie, and her mother Carole, has been observed with a number of mtDNA pathogenic variants affecting the activity of complex I of the mitochondrial respiratory chain. Unlike LHON, the clinical features resulting from these variants in *MT-ND5* often extend beyond the optic nerve to involve the brain and other organ systems. Generally, the extent and severity of clinical features in mtDNA-related mitochondrial disease correlate with heteroplasmy—high levels being associated with earlier onset of more severe disease—though the threshold for specific cellular and organ dysfunction may be high and will vary between individuals and between different pathogenic variants. It is also clear that heteroplasmy is not alone in determining pheno-type and other epigenetic and nuclear genetic factors exert substantial influence on outcome. Smoking and excessive alcohol consumption are environmental factors associated with clinical presentation of LHON in those individuals harboring the three common mtDNA variants. Given the preponderance of males affected by LHON (typically a 5:1 ratio for affected males to females) an X-linked nuclear modifier has long been suspected, but not yet confirmed.

7.7 EFFECTS ON THE PHENOTYPE OF PATHOGENIC VARIANTS IN NUCLEAR DNA

[Sections 7.2](#) to [7.5](#) described how disease-causing genetic changes arise in DNA and chromosomes, and [Section 7.6](#) covered the special case of pathogenesis due to mutations in mtDNA. Here we consider the effects on the phenotype of pathogenic variants in nuclear DNA. There are several considerations, as listed below.

- *The effect of a pathogenic variant on gene function.* The simplest situation occurs when the variant affects how a single gene functions, which we consider in the section below. Some variants, however, simultaneously affect, directly or indirectly, how multiple genes work. A large-scale mutation, for example, can directly change the sequence or copy number of multiple neighboring genes. In addition, a simple mutation in a regulatory gene, such as an miRNA gene, can indirectly affect the expression of multiple target genes, and can potentially cause complex phenotypes.
- *The extent to which normal copies or different copies of the mutant gene are also available.* For diploid nuclear genes, that means considering how a mutant allele and a normal allele work in the presence of each other, or estimating the combined effect of two mutant alleles. The situation for mitochondrial DNA mutants is quite different, and rather unpredictable, partly because we have so many copies of mtDNA in each cell, and partly because, unlike nuclear DNA, mtDNA replication is not governed by the cell cycle.
- *The effect of other interacting factors.* Affected genes do not work in isolation: many factors—non-allelic genetic factors (other loci), epigenetic factors, and environmental factors—affect the extent to which an individual pathogenic variant affects the phenotype. The situation becomes even more complicated when multiple genetic variants at different loci are involved—we examine this in [Chapter 8](#), within the context of complex disease.

Mutations affecting how a single gene works: an overview of loss of function and gain of function

Mutations that affect how a single gene works can have quite different effects on how the gene makes a product. Some affect expression levels only, often causing complete failure to express the normal gene product, or causing a substantial reduction in expression. Mutations that increase gene copy number and occasional activating point mutations result in overexpression (which can be a problem for *dosage-sensitive genes*). Other mutations can result in an altered gene product that lacks the normal function, or that has an altered or new function.

One broad way of classifying the overall effect of a mutation is to consider whether the mutation results in a loss of function or a gain of function, as described below. As described in the section after this one, the effect of a loss-of-function

mutation is often minimal in the presence of a normal allele, but a gain-of-function mutation has a harmful effect even in the presence of a normal allele.

Loss-of-function mutations

Loss-of-function mutations can have different consequences for how a gene is expressed. Sometimes, the final gene product is simply not produced or is nonfunctional (in which cases, the mutant gene is said to be a **null allele**). Different types of mutation can produce null alleles, notably large-scale deletions (eliminating the entire gene or a significant portion of it). Various types of mutation in protein-coding genes introduce early premature termination codons, causing the mRNA to be degraded so that no protein is made (as detailed in [Box 7.1](#)). They include many nonsense mutations, and various frameshifting mutations operating at either the DNA level (insertions or deletions), or at the RNA level, either via exon skipping or intron retention (as shown in [Figure 7.3](#)), or via exon truncation or exon extension ([Figure 7.4](#)).

For some loss-of-function mutations, there is some residual activity (the gene is expressed at abnormally low levels because of mutations in a regulatory sequence), or it is expressed normally but works poorly (as a result of a small non-frameshifting insertion, for example). A small minority of loss-of-function mutations are missense mutations that replace a key amino acid by a rather different amino acid. Recall that specific amino acids can have critical functional roles (including in post-translational processing), and structural roles (certain cysteines participate in disulphide bonding, for example).

Gain-of-function mutations

Gain-of-function mutations typically give rise to products that are positively harmful in some way. They are common in cancers, but in inherited disorders they are usually much less common than loss-of-function mutations. An outstanding exception to this general rule is provided by the paternal age effect disorders previously encountered in [Table 7.4](#). Each of these disorders is caused by missense mutations that activate a member of the growth factor receptor–RAS signal transduction pathway so as to confer a selective growth advantage on spermatogonial stem cells. The end result is clonal expansion of spermatogonial

stem cells containing the mutant gene—see [Figure 7.18](#) for the example of mutations in the *FGFR3* (fibroblast growth factor receptor 3) gene, causing bone dysplasia.

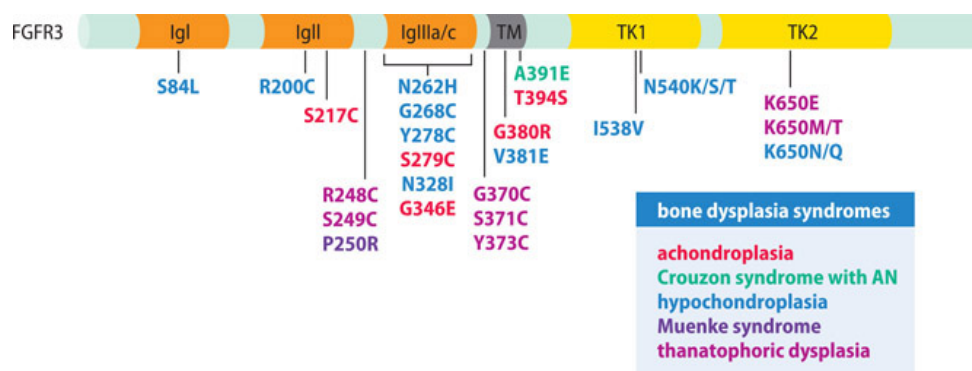


Figure 7.18 Pathogenic amino acid substitutions in the FGFR3 protein due to gain-of-function missense mutations that confer a selective growth advantage on spermatogonial stem cells.

The fibroblast growth factor receptor 3 (FGFR3) protein has three immunoglobulin-like domains (Igl, IglI, and IglIIa/c) in the extracellular region (shown on the left), a single hydrophobic transmembrane (TM) domain, and two cytoplasmic tyrosine kinase domains (TK1, TK2). The indicated amino acid substitutions shown below cause a type of bone dysplasia syndrome, according to the color scheme in the box. They arise from gain-of-function mutations that appear to confer a selective advantage on spermatogonial stem cells containing the mutant allele. AN, acanthosis nigricans. (Adapted from [Goriely A & Wilkie AO \[2012\] Am J Hum Genet 90:175–200](#). With permission from Elsevier.)

For inherited monogenic disorders other than the examples showing paternal age effect, gain-of-function mutations can work in different ways. Some produce a radically altered product. Recall the examples in [Section 7.3](#) where unstable dynamic mutations give rise to toxic proteins (which kill cells), or harmful *trans*-dominant RNAs (which cause disease after inappropriately binding to proteins made by other genes).

Most gain-of-function mutations do not produce a radically new product. Instead, they tend to make products that are structurally abnormal (and cause aberrant protein folding or aggregation), or are expressed inappropriately in some way. For example, a cell-surface receptor might be inappropriately expressed in the wrong cells, so that a particular signaling pathway is available when it should not be. As a result, the ectopically or inappropriately expressed product is able to interact with other cell components that it normally might not interact with, causing disease.

Sometimes it is a question of mutation causing altered specificity in some protein. A good example is the Pittsburgh variant of α_1 -antitrypsin (α_1 -AT), a member of the serpin family of serine protease inhibitors, which cleave its target proteins, potentially dangerous serine proteases, at specific sites. An important function of α_1 -antitrypsin (α_1 -AT) is to protect normal tissues from high levels of elastase, a protease expressed by neutrophils during inflammatory responses (high elastase levels trigger a compensatory increased α_1 -AT production to suppress elastase). Elastase cleaves a specific peptide bond in the α_1 -AT backbone that activates a radical conformational change in the α_1 -AT molecule, leading to elastase destruction. In the Pittsburgh variant of α_1 -AT a missense mutation changes the specificity of the enzyme, and the mutant α_1 -AT now attacks a different serine protease, the blood clotting factor thrombin, causing a lethal bleeding disorder (interested readers can find a detailed account at PMID 11778003).

Gain-of-function mutations are particularly common in cancer; many of these arise from chromosomal translocations and other rearrangements that create chimeric genes. Chromosomal translocations cause major problems in meiosis and so they are rarely responsible for inherited disease. But cancers arise from a mitotic division of mutant somatic cells in which the chromosomal rearrangements can be readily propagated from mother cell to daughter cells.

The effect of pathogenic variants depends on how the products of alleles interact: dominance and recessiveness revisited

Most of the genes in our diploid cells are present in two copies—one inherited from the mother and one from the father. For heterozygotes, therefore, we need to consider how the expression of a pathogenic variant (mutant allele) might affect the normal allele. To what extent might the effect of the mutant allele be reduced by—or compensated for by—having a normal allele? And, secondly, can a mutant allele have an adverse effect on how the normal allele is expressed?

Loss-of-function versus gain-of-function mutations in recessive and dominant disorders

Recall that in dominant conditions the disease phenotype is somehow expressed in the heterozygote, but in recessive conditions heterozygotes are unaffected. In autosomal recessive conditions, therefore, the disorder is expressed only when both

alleles are pathogenic variants. Heterozygous carriers cannot have a gain-of-function mutation (otherwise they would be expected to be affected) and so have one loss-of-function mutation, and affected individuals have two alleles with loss-of-function mutations.

Dominant conditions are less uniform. There is one mutant allele and, according to the disorder, the mutant allele may be a gain-of-function or a loss-of-function mutation. It is easy to imagine how a gain-of-function mutation might work in a heterozygote if we think of it as being positively harmful, causing damage even if the other allele is pumping out the correct product. Think of the toxic products produced by many unstable short tandem repeat expansions—the presence of functional product made by the normal allele is not going to stop the toxic effects of the mutant allele. (Potential gene therapies that rely on providing normal alleles could never work; instead we would have to inhibit the harmful effects of the mutant allele in some way.)

But how does just one loss-of-function mutation cause a dominant disorder? Why is the normal product made by the unaffected allele not enough? In a few cases, such as for imprinted loci ([Section 6.3](#)), only one of the two alleles is normally expressed, and the unaffected allele happens to be the one that is silenced. Thus, for example, a single loss-of-function mutation of the maternal *UBE3A* allele is enough to cause Angelman syndrome—the paternal *UBE3A* allele is not expressed (at least, not in the brain).

In most cases, both alleles of a diploid gene locus are normally expressed, but a single loss-of-function mutation can nevertheless cause disease if the gene concerned shows exceptional dosage sensitivity ([Box 7.3](#)). (Note that for all dominantly inherited disorders caused by a loss-of-function mutation, the phenotype is recessive at the cellular level: the normal phenotype can be restored by introducing a functioning allele or by reactivating a silenced allele.)

BOX 7.3 DOSAGE-SENSITIVE GENES AND HAPLOINSUFFICIENCY

A small number of our genes are not essential (people with blood group O, for example, have two inactive alleles at the *ABO* gene locus without any harmful effects). For very many single-copy genes, however, homozygous inactivation is a problem, and complete absence of a gene product often results in disease or is

lethal. The amount of product made by most genes can, however, show significant variation without harmful effects, if above some critically low level.

For a diploid locus, an occasional gene duplication—resulting in a total of three gene copies and an expected 50 % increase in gene product—often makes no obvious difference to the phenotype. According to circumstances, overproduction can even be advantageous: Western populations that traditionally have diets rich in starch have many copies of the starch-processing α -amylase gene, *AMY1A* (see [Figure 4.8](#)).

Similarly, for many genes, a reduction to 50 % of gene product is often inconsequential. In recessive loss of function, one null allele at a diploid locus typically causes no harm (provided the other allele works normally). Even if the second allele also has a hypomorphic mutation leaving it only partly functional, and the combined output of the two alleles is, say, 30 % of the normal amount of gene product, there may be little evidence of pathogenesis ([Figure 1](#)). For these genes, the product needs to drop to rather low levels before disease becomes apparent. (This especially applies to products such as enzymes that can be used over and over again.)

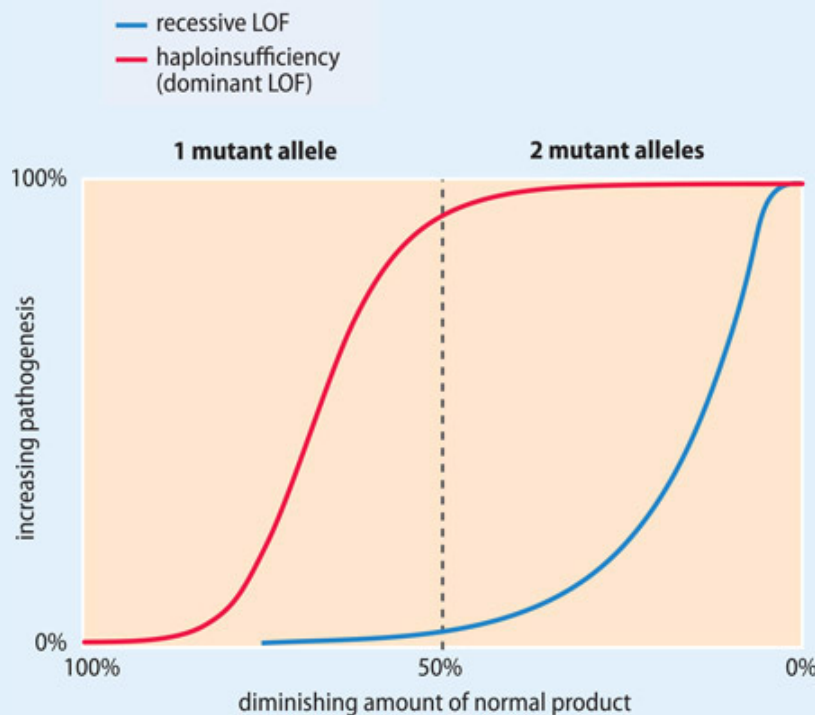


Figure 1 The relationship of disease susceptibility to diminishing amounts of gene product in dominant and recessive disorders due to loss-of-function mutations. The vertical dotted line marks the point at which an individual has a fully functional normal allele and a null allele

(such as a gene deletion) and might be expected to make about 50 % of the normal amount of gene product. If the allele is not null but partly functional, the amount of product made increases (as shown on left). If the second allele also has a loss-of-function (LOF) mutation, the amount of gene product made diminishes, depending on the severity of that mutation (as shown on the right). For dosage-sensitive genes, the reduction from 100 % to 50 % gene product is sufficient to result in disease (haploinsufficiency). Increasing the amount of gene product (by having three gene copies) can also induce disease. In recessive disorders, the disease is normally manifested only when both alleles have loss-of-function mutations; pathogenesis increases rapidly as the amount of normal product made approaches zero. There is some variability in recessive diseases—for some disorders, heterozygotes may show limited pathogenesis. Note that the idealized curves shown here do not take into account other factors such as modifier genes or environmental factors.

A minority of our genes are especially **dosage-sensitive**—the amount of product made is critically important. Changes in gene copy number (gene *dosage*) can cause disease by changing the amount of gene product beyond normal limits, and certain types of point mutation can have the same effect by reducing or amplifying gene expression. Disease can occur when too much of a product is made; sometimes that happens when the increase is only 50 % above the normal amount. For example, as described below, type 1A Charcot–Marie–Tooth disease (a hereditary motor and sensory neuropathy) is often caused by a duplication giving rise to three copies of the *PMP22* gene, or by point mutations that lead to the overexpression of *PMP22*.

More commonly, disease is due to a loss-of-function mutation in one allele of a dosage-sensitive gene (the other allele can be normal and expressed). If the mutant allele is a null, such as a gene deletion, the amount of normal gene product might be expected to be reduced by about 50 % (but pathogenesis can sometimes be observed when the mutant allele retains partial function). Because heterozygotes are affected, this type of loss of function is dominantly inherited, and is known as **haploinsufficiency** (see [Figure 1](#) in [Box 7.3](#)). Heterozygotes for a loss-of-function mutation in a dosage-sensitive gene are rare, and so having two loss-of-function alleles—usually by being a compound heterozygote—is extremely rare. When observed, the phenotype is often slightly more severe than that for a heterozygote.

Dosage-sensitive genes typically make products that need to be calibrated against the level of some other interacting or competing gene product. In many cases, the

products have roles in quantitative signaling systems or other situations in which precisely defined ratios of the products of different genes are important for them to work together effectively. Genes that regulate other genes are likely candidates: they might do so by making transcription factors, signaling receptors, splicing regulators, or chromatin modifiers, for example. Or the different gene products may be antagonistic, competing with each other to ensure that some critical reaction is carried out that is important in development or metabolism. Because chromosomes usually have multiple dosage-sensitive genes, constitutional aneuploidies are often lethal, but some are viable (as shown in [Table 7.12](#)).

Striking loss of function produced by dominant-negative effects in heterozygotes

In heterozygotes a null allele causing a loss of function does not normally affect the function of the normal allele. Sometimes, however, a mutation results in a mutant protein that cannot perform the normal function and also inhibits the function of protein produced by the normal allele. A mutant allele making a protein that antagonizes the protein made by the normal allele in a heterozygote is sometimes known as an *antimorphic* allele, and provides an example of a **dominant-negative** effect.

A common example occurs when the normal protein is part of a multimer that is inactive if it incorporates any of the mutant protein. Imagine the simplest possible case, when the multimer is a homodimer. A heterozygote for a null allele might be expected to make 50 % of the normal homodimer. However, a heterozygote who makes equal quantities of the normal monomer and a mutant monomer (which can only form inactive dimers) might be expected to make only 25 % of the normal amount of functional dimer ([Figure 7.19A](#)).

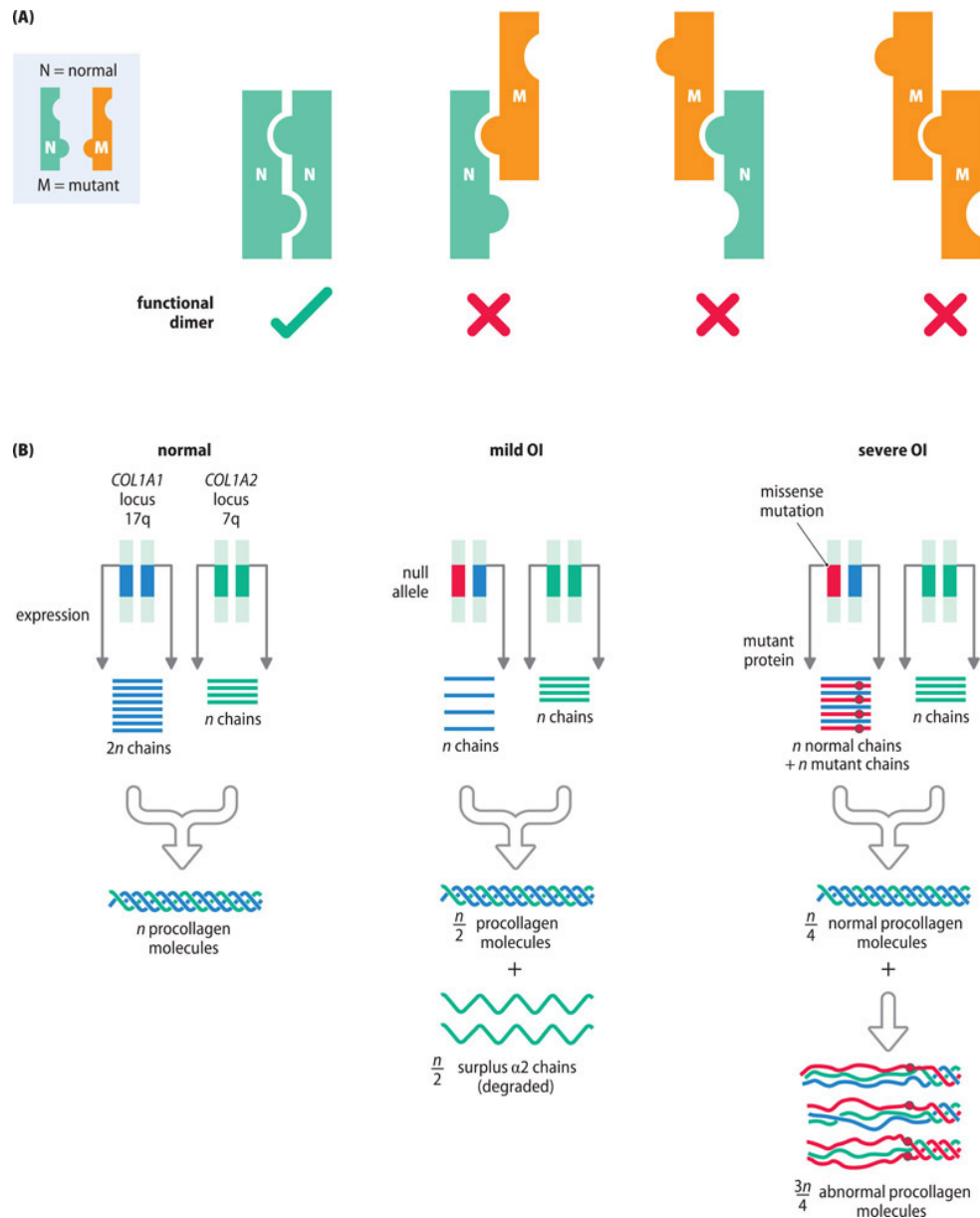


Figure 7.19 Dominant-negative effects: when a heterozygous missense mutation is more harmful than a null allele. (A) A hypothetical example: the product of a gene forms a homodimer, and the mutated allele produces a protein at normal quantities that can only form nonfunctional dimers. As a result, only one-quarter of the normal amount of functional dimers is made (there are two ways of forming nonfunctional heterodimers and one type of nonfunctional mutant homodimer). (B) A clinical example: a dominant-negative mutation causing a severe type of osteogenesis imperfecta (OI). Two *COL1A1*-encoded polypeptides and one *COL1A2*-encoded polypeptide are required to make a triple-helical type I procollagen. A null allele at *COL1A1* simply reduces the amount of type I procollagen by half and results in a mild form of OI. However, mutations that replace a structurally important glycine with any other amino acid usually have

strong dominant-negative effects because they disrupt the packing of the three chains in the triple helix. The mutant *COL1A1*-encoded polypeptide is included within three-quarters of the type I procollagen molecules made, making them nonfunctional. With only 25 % of the normal type I procollagens, affected individuals develop a severe type of OI.

Disorders of structural proteins that work as multimers often provide disease phenotypes arising from dominant-negative proteins, including examples in osteogenesis imperfecta (collagens), Marfan syndrome (fibrillins), and epidermolysis bullosa (keratins). As an illustration, consider collagens in osteogenesis imperfecta (also called brittle bone disease). Collagens are initially synthesized as procollagens in which three polypeptide chains are wound round each other, beginning at the C-terminal end, to form a stiff, rope-like triple helix. Each helical polypeptide chain undergoes a remarkable one turn every three amino acids, which is made possible because collagen polypeptides have a unique structure based on tandem repeats of a three-residue sequence with the general formula Gly–X–Y. The glycines, with a single hydrogen atom in the side chain, provide the flexibility; X and Y are often (but not always) proline and hydroxyproline, respectively, which have unique side chains that loop back and connect to the polypeptide backbone, stabilizing the helical structure.

In type I procollagen, the precursor of the most common type of collagen, two of the three polypeptide chains are made by the *COL1A1* gene and the third is made by another gene, *COL1A2*. A null mutation in *COL1A1* (or *COL1A2*) results in a mild form of osteogenesis imperfecta—in each case the amount of procollagen made might be expected to decrease to 50 % ([Figure 7.19B](#)). By contrast, a mutation that replaces a glycine in a collagen polypeptide by any other amino acid usually has dominant-negative effects: it causes abnormal packing of the collagen polypeptides into a triple helix for any collagen molecule into which it is incorporated. Such a mutation in the *COL1A1* gene causes a severe form of osteogenesis imperfecta, the type IIA form, because the amount of procollagen produced would be expected to decrease to just 25 % (see [Figure 7.19B](#)).

The antagonistic effect of an antimorph in preventing wild-type protein from performing its function might be considered a gain of function, but as shown in [Figure 7.19B](#) the net effect is to create a greater loss of function than a null allele.

Gain-of-function and loss-of-function mutations in the same gene can produce different phenotypes

Phenotypes arising from a loss of function are typically associated with mutational heterogeneity (there are many different ways of inactivating a gene—frameshifting mutations, nonsense mutations, major splice-site mutations, missense mutations, and whole gene deletions). Gain-of-function mutations are less common in inherited disorders, and the resulting phenotypes are typically associated with mutational homogeneity. They may be a class of unstable oligo-nucleotide expansions, for example, or specific activating missense mutations, or a mutation that results in overexpression, but not a great range of different types of mutation.

Gain-of-function mutations and loss-of-function mutations in the same gene quite often produce very different phenotypes. In some cases, an inherited disorder is caused by the loss-of-function mutation, but a gain-of-function mutation in the same gene produces a cancer. For example, loss-of-function mutations in the *RET* gene (which makes a tyrosine kinase) result in susceptibility to Hirschsprung's disease in which affected individuals have a congenital absence of ganglia in the gut. But different, very specific, kinds of activating missense mutations in the same gene result in different types of cancer: either medullary thyroid carcinoma or multiple endocrine neoplasia types 2A or 2B.

Loss-of-function mutations in the androgen receptor gene (*AR*) cause androgen-insensitivity syndrome (also called testicular feminization syndrome). Affected individuals have a 46,XY karyotype but because their androgen receptors do not work normally the end organs are insensitive to androgens, resulting in an X-linked recessive form of pseudohermaphroditism. Exon 1 of the *AR* gene also happens to have a tandem CAG repeat that can undergo unstable expansion to produce androgen receptor proteins with an expanded polyglutamine tract. The resulting proteins (and possibly RNA transcripts) are toxic to vulnerable cells and lead to spinal and bulbar muscular atrophy (also called Kennedy disease). In this case there is degeneration of lower motor neurons affecting certain muscles in the arms and legs and also some muscles in the face and throat (bulbar muscles).

Another case of divergent gain-of-function and loss-of-function phenotypes in one gene is provided by the *PMP22* gene, which makes a peroxisomal membrane protein. Because the great majority of pathogenic mutations are duplications or deletions of a 1.4 Mb region at 17p11.2 that contains multiple genes in addition to *PMP22*, we consider the pathogenesis in the next section.

Multiple gene dysregulation resulting from aneuploidies and mutations in regulatory genes

Some genes produce regulatory proteins or RNAs that regulate many different target genes. Examples include genes encoding high-level gene regulators (making master transcription factors, microRNAs or splicing regulators) and genes involved in global epigenetic regulation, such as those that make chromatin modelers or DNA methyltransferases. These genes are typically dosage-sensitive, and heterozygotes with a loss-of-function mutation can often show complex phenotypes (see [Table 6.7](#) on page 167 for some examples).

Whole-chromosome aneuploidies and large-scale intrachromosomal deletions and duplications (*segmental aneuploidies*) also directly affect multiple genes simultaneously, in this case by changing their copy number. However, because most genes are comparatively insensitive to dosage effects, the phenotype in affected heterozygotes is due to the combined effects of a comparatively small number of dosage-sensitive genes.

Whole chromosome aneuploidies

Because the number of dosage-sensitive genes over a whole chromosome is often large, several dosage-sensitive genes might be present; reducing the gene copy number by 50 % could be expected to have severe consequences. Monosomies are almost always lethal because the cumulative effect of deleting one copy of each dosage-sensitive gene on a chromosome is simply too much to support embryonic or fetal development.

One monosomy can sometimes be viable. The 45,X genotype leads to spontaneous abortion in about 99 % of cases, but occasionally leads to Turner syndrome, a comparatively mild condition: short stature with certain minor physical abnormalities (webbed necks, low-set ears) and gonadal dysfunction causing sterility. X-chromosome inactivation means that 45,X women are normally functionally monoallelic for most X-linked genes, but a few genes on the X chromosome, including genes in the pseudoautosomal regions, are not subject to X-inactivation, however.

Providing an extra gene copy for dosage-sensitive genes might be expected to have less harmful effects, but across a whole chromosome the combined effects of dosage imbalance in multiple genes means that most autosomal trisomies are also lethal. Chromosomes with few genes have fewer dosage-sensitive genes, and the three autosomal trisomies compatible with life—trisomies 13, 18, and 21—each involve chromosomes with relatively few genes.

Having extra sex chromosomes has far fewer ill effects than having an extra autosome because of X-inactivation (inactivating all X chromosomes except one) and the scarcity of genes on the Y chromosome. People with 47,XXX and 47,XYY karyotypes often function within the normal range, and in comparison with people with an autosomal trisomy, men with 47,XXY (Klinefelter syndrome) have relatively minor problems, notably hypogonadism and reduced fertility.

Segmental aneuploidies

Large-scale subchromosomal deletions and duplications also cause disease by simultaneously changing the copy number of multiple linked genes. If they occur in chromosomal regions that are constitutionally hemizygous, or have genes showing monoallelic expression (via imprinting, X-inactivation and so on), the functional copy number of some genes will be reduced to zero, so that no gene product is made at all. A profound effect can therefore often be expected for a few of our gene loci, however, a complete absence of gene product does not result in a clinical phenotype (people with blood group O, for example, have two inactive *ABO* alleles).

Males are constitutionally hemizygous for X- and Y-specific regions; large deletions in these regions therefore result in a complete absence of multiple gene products. The Y chromosome has few genes and they are very largely devoted to male-specific functions; accordingly, large deletions here are associated with azoospermia and infertility. Large deletions within the X chromosome in males are often lethal because of the high density of genes that perform a wide range of important functions. Certain regions such as Xp21 are comparatively gene-poor, however, and large deletions here can result in disease phenotypes.

Large-scale deletions and duplications on autosomes can cause disease by changing the copy number of comparatively rare dosage-sensitive genes. Deletions reduce the functional gene copy number to one (so that disease results from haploinsufficiency in dosage-sensitive genes), and duplications increase the gene copy number to three, resulting in the overexpression of multiple genes.

7.8 A PROTEIN STRUCTURE PERSPECTIVE OF MOLECULAR PATHOLOGY

Until now we have looked at pathogenesis mostly from the perspective of altered gene expression or changes in protein sequence that either result in loss of function or produce a gain of function that does not necessarily involve a major change in protein structure. However, for many disorders the pathogenesis is ultimately due to major changes in protein *structure* that are typically induced by single nucleotide substitutions or other point mutations.

We briefly described elements of protein structure and folding in [Chapter 2](#). In the next two sections we consider disease caused when proteins adopt altered structures and when changes in structure can predispose proteins to form aggregates that can cause disease. Understanding the basis of major changes in protein structure is important for understanding molecular pathology and developing novel therapeutic strategies.

Pathogenesis arising from protein misfolding

We have previously considered various aspects of how the expression of protein-coding genes is regulated to give a functional product, dwelling mostly on transcriptional, post-transcriptional, and translational control in [Chapter 6](#), and touching on mRNA surveillance in [Box 7.1](#). However, for proteins to function correctly they also need to fold properly to assume the correct three-dimensional conformation so that they can bind the appropriate interacting molecules. They need to function correctly within the right environment (in a hydrophilic environment, proteins fold up with hydrophobic amino acids located in the interior, and hydrophilic amino acids on the surface). Proteins also need to be able to interact correctly with other proteins when forming multimers.

Regulation of protein folding

Protein folding is not straightforward: there are various different paths that can be taken by an unfolded or partly folded protein to arrive at the final conformation, and natural errors in protein folding are common. Some proteins can fold correctly without help, but many proteins require assistance in folding from specific molecular chaperones such as Hsp60 or Hsp70 (chaperones are sometimes called *heat-shock proteins*, with their names prefixed by *Hsp*, because their expression is drastically increased when cells are exposed to even moderate temperature increases,

such as from 37°C to 42°C, that nevertheless cause an increase in protein misfolding).

Chaperones can help to fold both incompletely folded proteins and some incorrectly folded proteins. However, when attempts to refold a protein are unsuccessful, the protein is shunted into a proteolytic pathway in which it is destroyed by the *proteasome*, a complex compartmentalized protease that is distributed in many copies throughout the cytosol.

Aberrant protein folding causing disease

Protein misfolding is a common cause of disease in many genetic disorders, such as cystic fibrosis and phenylketonuria, in which mutations that change a single amino acid are quite common. Thus, about 90 % of individuals with cystic fibrosis have one or two copies of the p.Phe508del allele in which deletion of a single phenylalanine residue is sufficient to cause aberrant protein folding that cannot be rectified by chaperones. Whereas the normal protein would continue its regular journey to take up residence in the plasma membrane, the mutant protein is rapidly subjected to degradation in the endoplasmic-reticulum.

Sometimes mutations destroy the ability to adopt highly specific structures required for assembly of multimers such as collagens, fibrillins, and keratins. Collagens, for example, require complex packaging of three collagen polypeptides into trihelical structures in which three individual collagen strands are wound round each other. As described above, missense mutations that replace glycines in collagen chains cause major packaging problems (as shown previously in [Figure 7.19B](#)).

The many different ways in which protein aggregation can result in disease

The pathogenesis of many disorders, both monogenic and common diseases, involves protein aggregation. Soluble oligomers and large, often insoluble complexes can form and the aggregated proteins are often found as cellular inclusions or pericellular deposits.

At present there is still some uncertainty about the significance of protein aggregates observed in many common diseases—are they a direct cause of disease, or are they more peripherally associated with pathogenesis? In monogenic disorders there can be more certainty, although sometimes even here the precise pathogenetic

process is still not clear. In some monogenic disorders, however, the evidence for mutation-induced protein aggregation is clear; we give two examples below, one in which proteins aggregate to form extremely long protein fibers, and one in which the damage is done by protein aggregates in inclusion bodies within cells.

We end this section with what used to be thought of as a bizarre protein aggregation disease mechanism, one involving a type of epigenetic information transfer *that does not involve nucleic acids* (and so is quite distinct from the epigenetic mechanisms described in [Section 6.3](#), which rely on heritable chromatin states). The first evidence came from studies of mutant prion proteins, but as described below similar pathogenetic mechanisms are now thought to occur in various disorders including some common neurodegenerative diseases such as Alzheimer and Parkinson disease.

Sickle-cell anemia: disruptive protein fibers

Normal adult hemoglobin is a tetramer with two α -globin chains and two β -globin chains. Individuals affected by sickle-cell anemia are homozygous for a specific missense mutation that replaces a charged, hydrophilic glutamate residue at position 6 in the β -globin chain by a hydrophobic valine residue. The resulting mutant hemoglobin S (HbS) has a strong tendency to aggregate when deoxygenated, resulting in fibers composed of 14 long strands of HbS tetramers ([Figure 7.20A,B](#)). The fibers cause red blood cells to be deformed so that they are crescent-shaped, like a sickle. The much shorter life span of the abnormal sickle cells (10–20 days, in contrast with the normal 90–120 days) means that the body cannot replace dead red blood cells fast enough, resulting in anemia. The HbS fibers also block small blood vessels, causing hypoxic tissue damage.

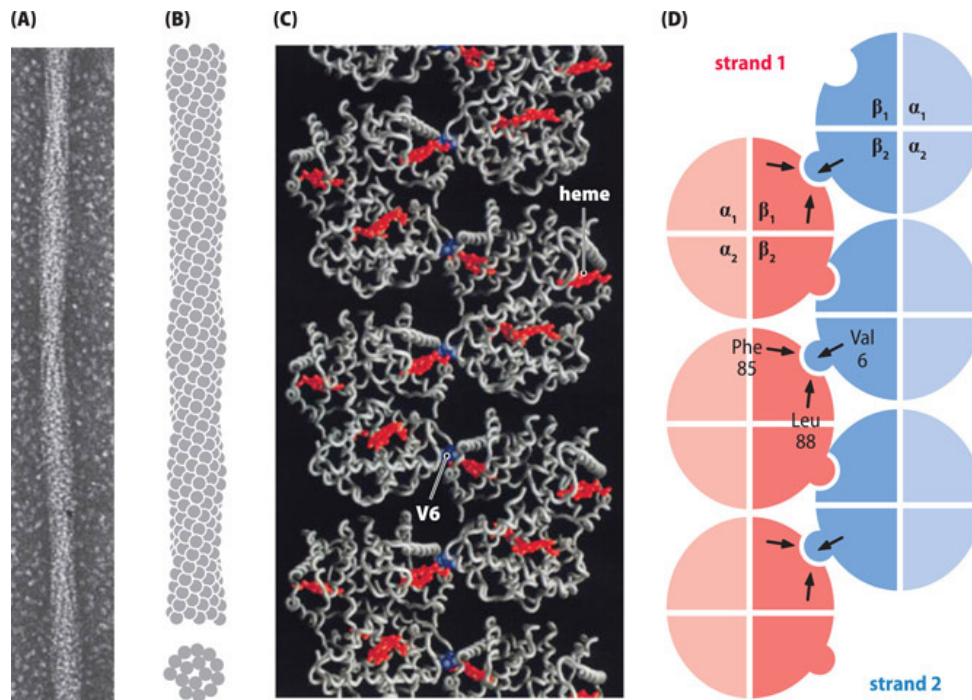


Figure 7.20 Aggregation of hemoglobin to form complex fibers in sickle-cell disease. (A,B) The 14-strand structure of deoxyhemoglobin S fibers including an electron micrograph in (A) of a stained fiber and the interpreted structure in (B), showing a lateral image at the top and a cross section at the bottom. The 14-strand structure is built from seven sets of paired strands. (C,D) The basic paired strand component of deoxyhemoglobin S fibers. (C) Structural model showing the mutant valine (V6; in blue color) located on the outside of the b-globin chains, facilitating lateral contact between b-globin chains on different HbS tetramers. Heme groups are shown in red. (D) Diagram illustrating how each double strand of hemoglobin tetramers is stabilized by lateral contacts involving the mutant valine on a b-globin chain from one strand interacting with a pocket formed between two helices on a b-globin chain on a Hb tetramer on the opposing strand. (A,B, From Dykes G et al. [1978] *Nature* 272:506–510; PMID 692655. With permission from Macmillan Publishers Ltd; C, from Harrington DJ et al. [1997] *J Mol Biol* 272:398–407; PMID 9325099. With permission from Elsevier.)

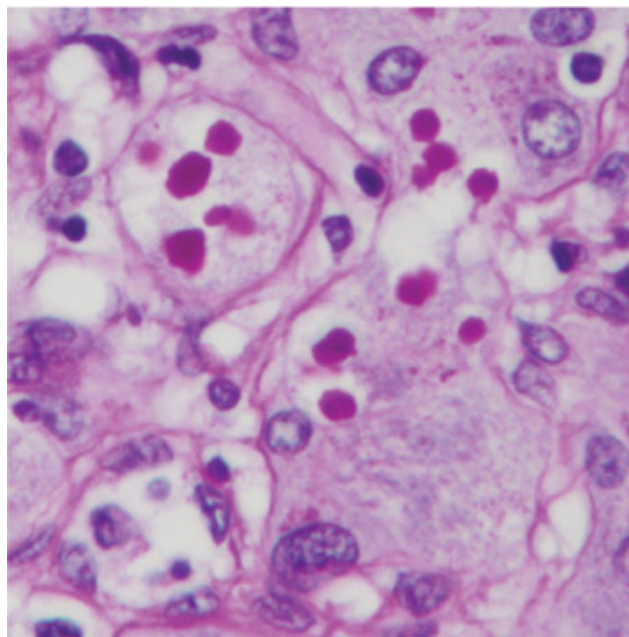
The 14-strand structure is built on the lateral association of seven sets of paired HbS tetramer strands. The side chain of mutant valines on the b-globin chains of one HbS strand can interact with a complementary pocket on b-globin residues of the neighboring HbS tetramer. That type of bonding drives the formation of paired strands of HbS tetramers ([Figure 7.20C,D](#)) that then form the higher-order structure shown in [Figure 7.20A,B](#) by additional lateral associations.

α 1-Antitrypsin deficiency: inclusion bodies and cell death

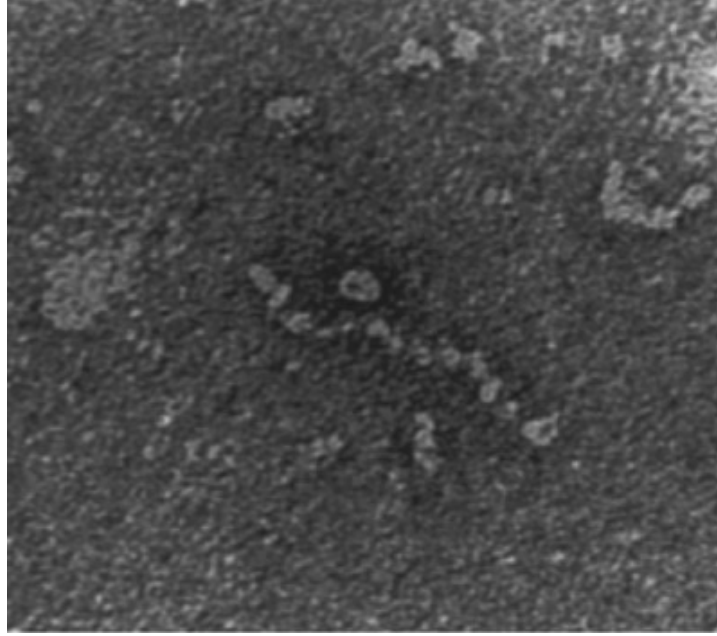
α 1-Antitrypsin (α 1-AT) is made by the liver and secreted to regulate the levels of certain serine proteases such as elastase (which can be overproduced by neutrophils during inflammation and might damage sensitive tissue, such as the alveoli of lungs, if not kept in check). α 1-Antitrypsin deficiency is common in Caucasian populations in which two missense mutations are especially common: the mild PI*S allele (E264V) and the severe PI*Z allele (E342K).

Plasma concentrations of α 1-AT in ZZ homozygotes (about 15 % of normal) and in SZ compound heterozygotes (about 40 %) are not high enough to protect lungs from damage by elastase over a lifetime, especially in people who smoke. Affected individuals often develop emphysema, a form of chronic obstructive lung disease in which tissues needed to support the shape and function of the lungs are destroyed. The low plasma α 1-AT concentrations typically do not result from failure of liver cells to make any of the protein; instead the problem is a blockage in α 1-AT processing and secretion from liver cells.

The retained α 1-AT proteins can aggregate in the endoplasmic reticulum of hepatocytes to form intracellular inclusions (*inclusion bodies*) that can be readily recognized by using suitable stains and can be seen to contain bead-like polymerases in ZZ homozygotes ([Figure 7.21](#)). The inclusion bodies cause hepatocytes to die and can result in eventual cirrhosis of the liver, especially in ZZ homozygotes.



(A)



(B)

Figure 7.21 Intracellular inclusion bodies and protein aggregates in α 1- antitrypsin deficiency.

(A) Staining of hepatocytes with a periodic acid-Schiff stain reveals inclusion bodies as bright pink globules (arrowed). (B) Electron microscopy showing bead-like polymers of Z-type α 1-antitrypsin.

(A, Courtesy of the National Society for Histotechnology; B, from Lomas DA et al. [1993] *J Biol Chem* 268:15333–5; PMID 8340361. With permission from The American Society for Biochemistry and Molecular Biology.)

Seeding of mutant protein using aberrant protein templates

In [Section 6.3](#) we explored epigenetic gene regulation in inherited disorders. That relies on heritable chromatin states rather than the DNA sequence, but yet another type of information that directs heritable changes in gene products is confined to the protein level. For some types of protein, mutant proteins that can aggregate are also able to direct normal forms of the same protein to adopt the mutant protein structure, allowing cellular spreading of protein aggregation. The first examples came from prion diseases in livestock, which garnered much public attention because of the danger to health in eating meat from infected cattle (“mad cow disease”). Similar mechanisms allow the cellular spreading of protein aggregation in certain other neurodegenerative diseases, including some monogenic disorders and also complex diseases including Alzheimer disease and Parkinson disease (see [Clinical Box 8](#)).

CLINICAL BOX 8 PRION DISEASES AND PRION-LIKE NEURODEGENERATIVE DISEASES: SEEDED SPREADING OF PROTEIN AGGREGATES BETWEEN ORGANISMS AND CELLS

Prion diseases—also known as transmissible spongiform encephalopathies—are progressive, fatal, and incurable neurodegenerative disorders that affect humans and other animals, in which holes develop in brain tissues, giving them a sponge-like texture. The disease can be spread from one organism to another by ingesting or internalizing affected tissue. For example, consumption of affected tissue from cows with bovine spongiform encephalopathy has led to outbreaks of variant Creutzfeldt–Jakob disease (vCJD; also known in this specific instance as “mad cow disease”). In addition to acquired prion protein disease, sporadic and hereditary forms exist. Creutzfeldt–Jakob disease (CJD), fatal familial insomnia, and Gerstmann–Straussler–Scheinker syndrome are dominantly inherited allelic disorders resulting from mutations in the *PRNP* prion protein gene at 20p13.

In prion disease, a normal cellular form of prion (PrP^C) is misfolded into an abnormal conformation (PrP^{Sc}) that is rich in β -pleated sheets and prone to aggregation (**Figure 1A**; the Sc superscript comes from scrapie, a sheep prion disease that was one of the first to be studied).

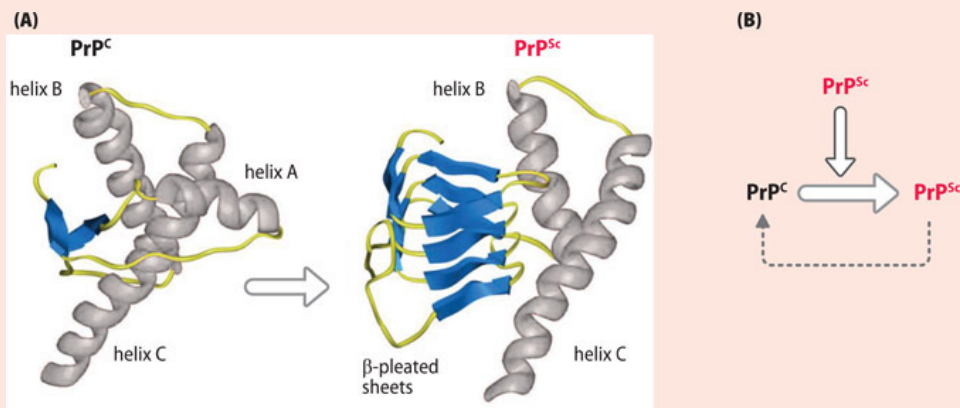


Figure 1 Conversion of normal prion protein (PrP^C) to the aggregation-prone isoform PrP^{Sc}. (A) The normal PrP^C protein has three α -helices (A, B, and C in gray) and two very short β -strands (blue), but it can misfold to give the PrP^{Sc} isoform; because it has a very high content of β -strands that form a β -pleated sheet, the PrP^{Sc} isoform is susceptible to aggregation. (B) A PrP^{Sc} isoform that may have been induced by mutation, arisen spontaneously, or be derived from another person or an animal with prion protein disease (for example after blood transfusion or the

ingestion of affected animal tissue) can induce a normal human PrP^C protein to change to the infectious PrP^{Sc} isoform. The latter can then induce other host PrP^C proteins to convert to PrP^{Sc}, spreading the disease between cells. (A, Adapted from Norrby E [2011] *J Intern Med* 270:1–14. With permission from John Wiley and Sons, Inc.)

The most striking characteristic of PrP^{Sc} is that when it comes into contact with normal PrP^C proteins it can induce them to switch conformation so that they, too, adopt the PrP^{Sc} structure. Thus, if our cells are exposed to abnormal prion proteins from an infected animal or person, the abnormal foreign prion proteins will induce host PrP^C proteins to adopt the PrP^{Sc} structure ([Figure 1B](#)).

The abnormal prion protein structure can effectively self-propagate by a form of replication that has nothing to do with nucleic acid sequences. In that respect the disease mechanism resembles classical epigenetic mechanisms (which typically involve chromatin modifications). The abnormal prion proteins are infectious because the misfolded protein can be acquired (by the ingestion of infected cells or tissue). Alternatively, prion proteins originate by a chance misfolding of a newly synthesized PrP^C protein in a sporadic case or develop as a result of a genetic mutation (in which the mutant sequence has a greater propensity to misfold).

The brain is the main target of prion toxicity—neurons, being extremely long-lived and not effectively replaced, are especially vulnerable to toxic protein aggregates. How prions enter the body and infect brain cells is an interesting question. Somehow, the abnormal protein aggregates can get past mucosal barriers, survive the attentions of innate and adaptive immunity, pass across the blood–brain barrier, and spread to different brain cells. Infection can be efficient. vCJD has been transmitted to hemophiliacs who were treated with a factor VIII extract isolated from blood samples provided by donors who included subclinically infected blood donors. Growth hormone deficiency and infertility have also been treated in the past with growth hormones or fertility hormones recovered from human cadaveric pituitary glands, but because the pituitary extracts had been contaminated by infected human brain tissue, more than 160 treated people died of vCJD.

AMYLOID DISEASES AND PRION-LIKE NEURODEGENERATIVE DISEASE

Prion proteins are members of the family of amyloid proteins that have a high content of β -sheets, making them prone to aggregation and the formation of elon-

gated, unbranched amyloid fibrils. The spines of the fibrils consist of many-stranded b-sheets, arranged in a cross-b structure ([Figure 2](#)).

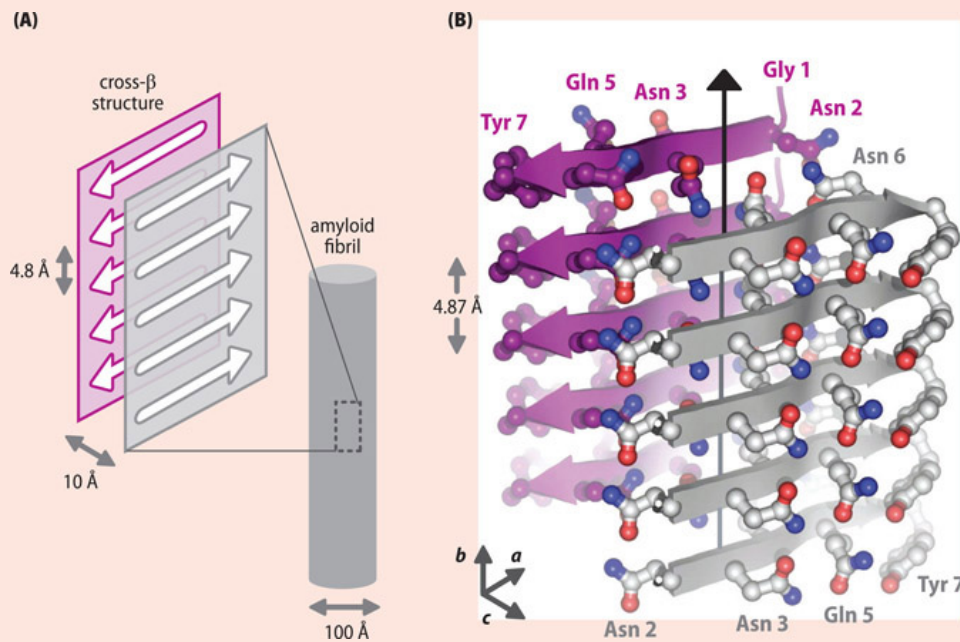


Figure 2 Characteristics of amyloid fibers. (A) Amyloid fibrils generally have a diameter of about 100 Å (10 nm) and form by protein aggregation. A characteristic property is the cross-beta spine, a set of b-strands that run perpendicular to the axis of the fibril. (B) Certain protein segments that are six to seven amino acids long bind to other copies with the same amino acid sequence to form two tightly interdigitating b-sheets. In the case of the prion protein, that sequence is GNNQQNY or Gly-Asn-Asn-Gln-Gln-Asn-Tyr, as shown here. (Adapted from Eisenberg D & Jucker M [2012] *Cell* 148:1188–1203. With permission from Elsevier.)

Amyloid proteins are frequently associated with disease ([Table 1](#)), and the aggregates may be extracellular (PrP^{Sc}; b-amyloid), nuclear (huntingtin), or cytoplasmic (SOD1; Tau; and synuclein, which forms *Lewy bodies*). Amyloid protein aggregation is seen in some common diseases not associated with neurodegeneration, such as type 2 diabetes (where aggregates of serum amyloid A protein are found in the pancreatic islets of Langerhans). However, neurodegeneration is the most striking clinical characteristic of many amyloid diseases ([Table 1](#)).

TABLE 1 EXAMPLES OF AMYLOID DISEASES

Disease	Amyloid protein (precursor)
---------	-----------------------------

Disease	Amyloid protein (precursor)
NON-NEURODEGENERATIVE DISORDERS	
Atherosclerosis	apolipoprotein AI
Rheumatoid arthritis	IAPP/amylin
Type 2 diabetes	serum amyloid A
NEURODEGENERATIVE DISORDERS	
Alzheimer disease	β -amyloid/A β (APP);Tau
Amyotrophic lateral sclerosis (motor neuron disease)	SOD1
Frontotemporal lobar degeneration (FTLD)-tau*	tau
Huntington disease	huntingtin
Parkinson disease	α -synuclein
Prion protein diseases	PrP ^{Sc} (PrPC)

IAPP, islet amyloid polypeptide; APP, amyloid precursor protein; SOD1, superoxide dismutase 1.

* The non-amyloid protein TDP-43 is also commonly found to be aggregated in frontotemporal lobar degeneration.

Neurodegenerative amyloid diseases, such as Alzheimer disease, Parkinson disease, amyotrophic lateral sclerosis, and frontotemporal disease, resemble prion protein diseases in many ways and are sometimes classified as prionoid diseases. The direct involvement of the aggregated proteins in disease is supported from familial forms of these disorders in which mutations in the relevant gene promote the formation of amyloid protein, including mutations in *APP* (Alzheimer disease), *SNCA*, encoding α -synuclein (Parkinson disease), *SOD1* (amyotrophic lateral sclerosis), and *MAPT*, the microtubule-associated protein tau (frontotemporal dementia), for example.

There is no evidence from animal studies that the aggregated proteins in these disorders are infectious like prion proteins. But there is quite strong evidence that the pathogenesis resembles prion protein disease in two respects. First, like prion proteins, misfolded amyloid proteins in these disorders can induce the formation of the amyloid state in the normal proteins so that they aggregate. Secondly, for several of the disorders there is quite strong evidence for cell-to-cell spreading of the disorder.

7.9 GENOTYPE–PHENOTYPE CORRELATIONS AND WHY MONOGENIC DISORDERS ARE OFTEN NOT SIMPLE

Assessing the effect of pathogenic variants on the phenotype is a component of the broader quest to understand genotype–phenotype correlations. If we know the genotype, to what extent can we predict the phenotype? This can be a difficult question to answer, even for monogenic disorders.

The effect of a pathogenic variant on the phenotype is not just dependent on the effect of the mutation on the ability of that allele to make its normal gene product and the interaction with the product made by the other allele. Genes from other loci can also influence the disease phenotype, as can environmental factors. It is now clear that monogenic disorders are often not as simple as sometimes described, and the division of genetic disorders into chromosomal, monogenic, and multifactorial disorders is a simplification.

The difficulty in getting reliable genotype–phenotype correlations

Interpreting the effect of a single pathogenic mutation, even in a well-defined fully characterized gene, is often not straightforward. Splicing mutations can be difficult to gauge; apparently harmless synonymous substitutions can be pathogenic; missense mutation effects are not easy to predict (loss of function, gain of function, or no clear effect?).

Even for nonsense and frameshifting mutations, the effects of the mutation may be hard to predict. Largely depending on where a premature termination codon is introduced within the mRNA, the effect may be to trigger mRNA destruction, with failure to make a protein, or to produce a mutant protein that may or may not have a gain of function (see [Box 7.1](#)). And who would have expected that deleting a single nucleotide in the 2.5 Mb dystrophin gene could produce the severe Duchenne form of muscular dystrophy, while deleting a 1 Mb region containing that same nucleotide along with a million others (including many coding exons) would result in a much milder form of muscular dystrophy? Part of the explanation lay in the differential effects of frameshifting and in-frame deletions, and the observation prompted a novel RNA therapy for Duchenne muscular dystrophy, as described below.

For individuals affected by an autosomal recessive disorder there is an added complication: the need to assess the combined effect of two mutant alleles. In such

situations, the mutant alleles are typically loss-of-function mutations and the degree of overall residual function is the major determining factor.

For some disorders, such as many enzyme deficiencies, there is a good correlation between product levels and severity of the phenotype. In steroid 21-hydroxylase deficiency, for example, individuals with non-classical forms (later onset, mild) typically have 10–15 % of residual enzyme activity, whereas in classical forms (congenital, severe) there is from about 2 % residual enzyme activity, which usually manifests as a “simple-virilizing” phenotype, to 0 % enzyme activity in the most severe (“salt-wasting”) form (the clinical phenotypes are given at the beginning of Clinical Box 6, on page 201). Phenotypes due to deficient X-linked hypoxanthine guanine phosphoribosyltransferase activity also show significant correlation with the amount of residual enzyme activity ([Figure 7.22](#)).

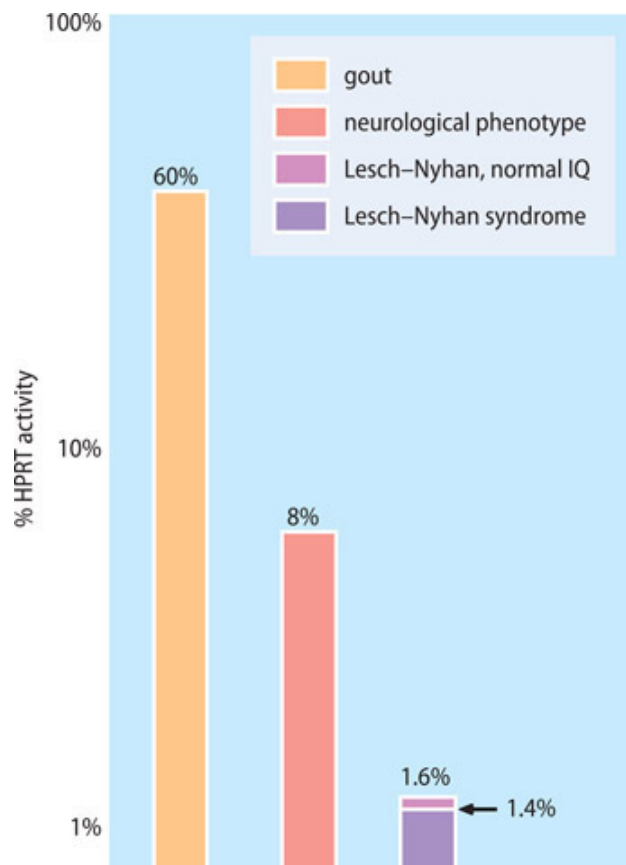


Figure 7.22 Different threshold levels for different phenotypes resulting from loss of activity of hypoxanthine guanine phosphoribosyltransferase (HPRT). Loss-of-function mutations in the X-linked *HPRT* gene can result in gout (which becomes manifest at less than 60 % of normal HPRT activity). If the HPRT activity falls to below 8 %, additional neurological features can begin to develop and are manifested as clumsiness and involuntary movements such as migrating

contractions (chorea) and twisting and writhing (athetosis). A decrease in HPRT activity to less than 1.4 % results in full Lesch–Nyhan syndrome (with choreathetosis and additional spasticity, self-mutilation, and mental retardation), but with an HPRT activity of about 1.4–1.6 %, individuals with Lesch–Nyhan syndrome can have normal intelligence.

Exceptional versus general reasons for poor genotype-phenotype correlations

For many monogenic disorders, genotype–phenotype correlations can be extremely complicated. Sometimes affected individuals who have identical mutant alleles (affected members of the same family; genotyped affected individuals within a population) show remarkable differences in phenotype.

We have already considered some exceptional factors that contribute to poor genotype–phenotype correlations in some Mendelian disorders: epigenetic factors (notably, parent-of-origin effects, as described in [Section 6.3](#)); dynamic mutations (due to an unstable expansion of short tandem repeats that can cause intergenerational differences in phenotype); and mosaicism (including differential X-chromosome inactivation that results in different effects for an X-linked mutation in women).

Disorders caused by mitochondrial mutations also show particularly poor genotype–phenotype correlation. Here, due to the exceptional problem of multiple copies of mtDNA per cell can be *homoplasmic* for a mutant allele (all mtDNA copies carry the mutation) or be *heteroplasmic* (with a mix of normal and mutant mtDNA molecules). Because egg cells typically contain more than 100 000 mtDNA molecules, every child of an affected heteroplasmic mother will inherit at least some mutant mtDNA molecules, but the proportion is difficult to predict and the ratio of mutant to normal mtDNA copies can change over time. As a result, mutations in mtDNA can have low penetrance and rather unpredictable effects.

Modifier genes and environmental factors: common explanations for poor genotype–phenotype correlations

In addition to the exceptional factors described above, two *general* factors can explain differences in the phenotype of a monogenic disorder in affected members of the same family (who can almost always be expected to have identical mutations in

the case of a monogenic disorder) and in affected individuals within a population who have been revealed to have identical genotypes.

One of these is genetic variation at other loci. Genes that interact with the disease locus to modify the disease phenotype are known as **modifier genes** (the interaction between a disease locus and a modifier gene locus is called *epistasis*). Different alleles at a modifier locus can have different effects on a disease phenotype—they may sometimes have a protective effect (resulting in a milder disease phenotype) or an aggravating effect (inducing a more severe phenotype). The second general factor that influences a disease phenotype comes from the environment, as described below.

Modifier genes: the example of β -thalassemia

Until recently, modifier genes were not easy to identify directly in humans. Instead, heavy reliance was often placed on carrying out various types of analyses in animal disease models, which could then suggest candidate modifier genes for human diseases. Here we consider how modifier genes can affect the phenotype of a well-studied blood disorder, β -thalassemia.

Individuals with β -thalassemia have a genetic deficiency in β -globin, a component of hemoglobin. Although monogenic, this disorder is far from simple. It is usually autosomal recessive, but in occasional individuals the phenotype is dominantly inherited (one allele is normal; the other has an exceptional gain-of-function mutation). Although mutation in the β -globin gene, *HBB*, is the predominant factor in causing the disease, affected individuals with identical *HBB* alleles can show very significant differences in phenotype. Genetic variation at several modifier loci is also very important.

Adult hemoglobin is a tetramer with two α -globin chains and two β -globin chains; the synthesis of α - and β -globin chains is normally tightly regulated to ensure a 1:1 production ratio. However, when mutation in *HBB* results in a reduced production of β -globin chains, there will be a relative excess of α -globin chains. The excess α -globin monomers, present at high concentration, aggregate and precipitate, causing the death of early hemoglobin-producing cells in the bone marrow, and ineffective production of red blood cells. Those red blood cells that reach the peripheral blood also contain excess α -globin, which induces the formation of inclusion bodies and increased production of reactive oxygen species, leading to membrane damage and

hemolysis. Because the anemia that results from lower numbers of red blood cells is life-threatening, current therapy is based largely on blood transfusions.

Genetic variation at other globin loci can affect the clinical severity of β -thalassemia. Thus, a mutation causing a reduced output of α -globin chains reduces the globin chain imbalance and allows the production of more red blood cells. Normal individuals usually have two tandemly repeated α -globin genes (*HBA1* and *HBA2*) on each chromosome 16, but as a result of unequal crossover the number of copies of the α -globin gene can vary: 0 (-); 1 (- α); 2 (aa); 3 (aaa); or 4 (aaaa). Large numbers of α -globin genes can further add to the excess of α -globin chains that results from reduced β -globin production ([Figure 7.23](#)). As evidence of the modifier effect, individuals who are heterozygous for a null β -thalassemia (β^0) allele but have a total of six or more α -globin genes (aaa/aaa or aa/aaaa) can have a disease phenotype that resembles homozygous β -thalassemia).

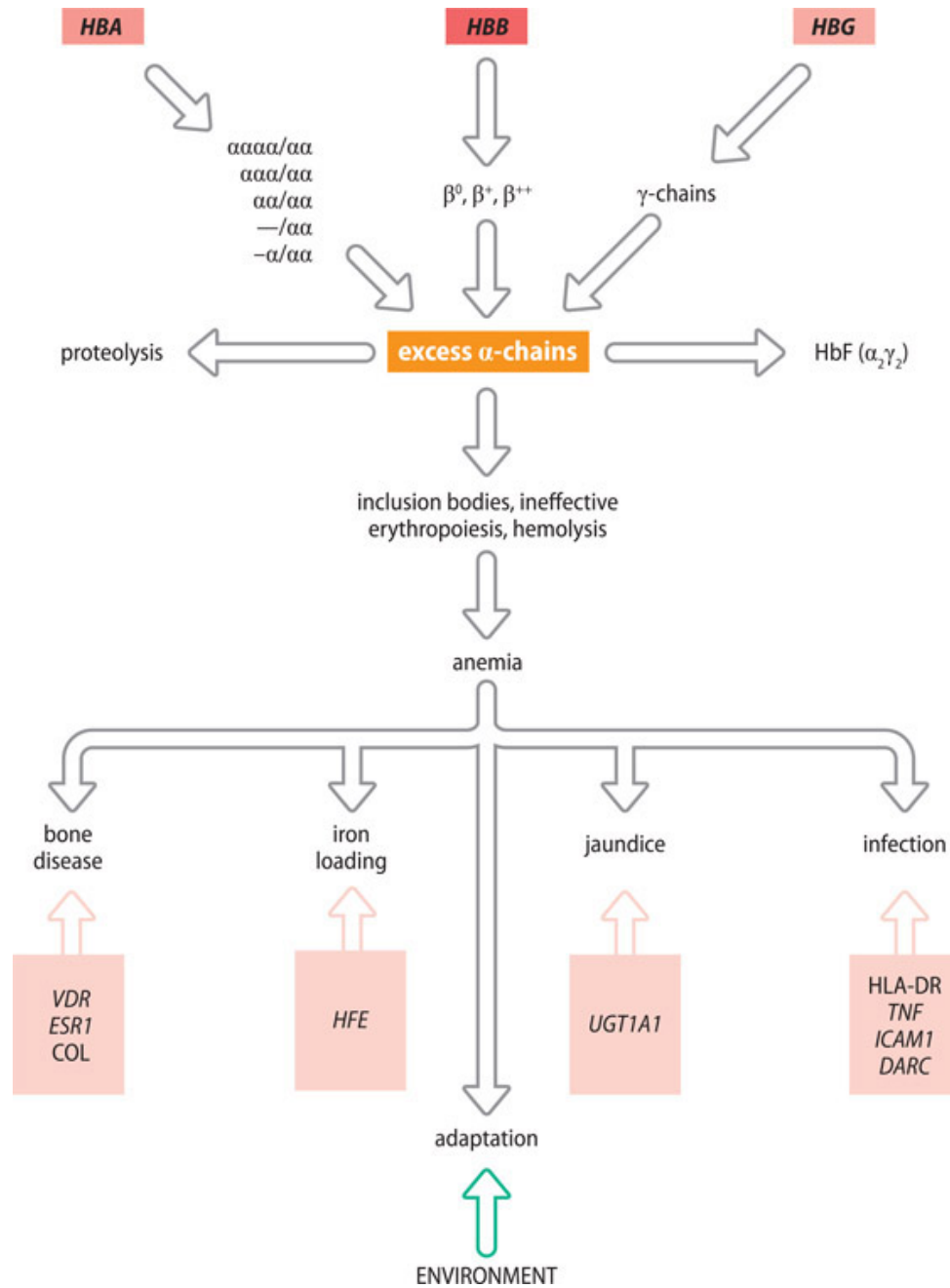


Figure 7.23 Multiple factors determine the β -thalassemia phenotype. The basic pathology of β -thalassemia results from excess α -globin chains damaging red blood cell precursors and red blood cells. Depending on the mutations at the disease locus *HBB*, there can be different levels of decrease in β -chain production (β^0 , null; β^+ , partial function; and so on) with a direct effect on the amount of excess α -globin chains. The phenotype can also vary as a result of variation in α -globin copy number, variation in the ability to produce HbF after birth (which uses up variable amounts of α -chains), and, possibly, through different rates of removal of α -chains by proteolysis. The many complications of the resulting anemia can also be modified by genetic variability, including

variation in the genes listed at the bottom. Genes or gene loci are: *HBB*, b-globin; *HBA*, a-globin loci; *HBG*, g-globin loci; *VDR*, vitamin D receptor; *ESR1*, estrogen receptor-1; *COL*, collagen loci; *HFE*, locus for hereditary hemochromatosis; *UGT1A1*, UDP glucuronyltransferase involved in bilirubin metabolism; *HLA-DR*, major histocompatibility complex loci; *TNF*, tumor necrosis factor; *ICAM1*, intercellular adhesion molecule-1; *DARC*, Duffy antigen receptor for chemokines. (Adapted from Weatherall D [2010] *Nature Med* 16:1112–1115. With permission from Macmillan Publishers Ltd.)

The β -thalassemia phenotype is also modified by genetic variants that control the production of hemoglobin F (HbF), which is composed of two a-globin chains and two g-globin chains. HbF is the dominant hemoglobin made during the fetal period (its high O₂-binding capacity makes it suited to working at the fetal stage), but there is a rapid decrease in HbF production at birth, and although it is still present at significant levels in infants it usually accounts for less than 1 % of hemoglobin in adults. However, HbF can account for 10–40 % of the hemoglobin in rare affected individuals with hereditary persistence of fetal hemoglobin, and there is significant variation between normal individuals in HbF levels. By forming more HbF, g-globin polypeptides compensate for reduced production of b-globin. The elevated HbF levels in infants is thought to be protective, explaining the delayed onset of symptoms in b-thalassemia, and comparatively high HbF levels at later stages may be partly protective.

Many of the complications of the disease are also modified by genetic variation at other loci (see [Figure 7.23](#)). There are also differences in the patterns of adaptation to anemia at different ages, and environmental factors, notably exposure to malaria, can also modify the phenotype.

Environmental factors influencing the phenotype of genetic disorders

In some disorders, expression of the disease phenotype depends very significantly on environmental factors that may act at different levels: at a distance (external radiation sources); by direct exposure of our cells to harmful or potentially harmful chemicals that we ingest (in food and drink) or inhale (such as tobacco smoke or atmospheric pollution); and by contact with microbes and toxins.

Especially important in triggering cancers, as described in [Chapter 10](#), environmental factors are also very important in other complex diseases, whether at the earliest stages of development (factors in the uterine environment) or at later

stages (such as exposure to chemicals and microbes). We consider some aspects in [Chapter 8](#).

Environmental factors are also known to be important in some single gene disorders. We illustrate the example of how dietary factors can influence disease with reference to phenylketonuria in [Clinical Box 9](#). In [Chapter 9](#), we also consider how differential sensitivity to drugs can influence other monogenic disorders, within the broader context of pharmacogenetics.

CLINICAL BOX 9 DISEASE PROFILE: PHENYLKETONURIA AS AN INBORN ERROR OF METABOLISM, A MULTIFACTORIAL CONDITION, AND AN EMBRYOFETOPATHY

The first genetic disorders that were investigated at the molecular level were *inborn errors of metabolism*. Affected individuals lacked a single enzyme that catalyzed one step in a metabolic pathway (usually consisting of a series of enzyme-catalyzed steps in which the product of one step becomes the substrate for the next step). Deficiency in one such enzyme would cause a metabolic block ([Figure 1A](#)). The resulting buildup of the substrate proximal to the block might drive an alternative pathway (red arrow). By analyzing blood and urine samples, pioneers in the field were able to obtain molecular clues as to the cause of a genetic disorder many decades before we knew about DNA structure and were able to study genes.

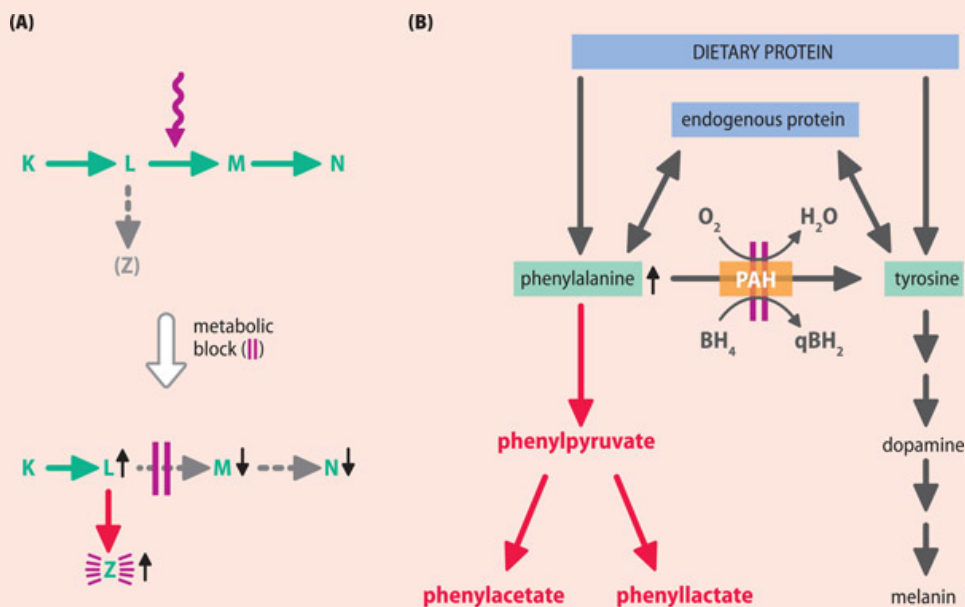


Figure 1 Metabolic blocks: principles and the example of phenylketonuria. (A) Principle of a metabolic block. Metabolites K, L, M, and N are linked by a series of enzyme-catalyzed reactions (green arrows) in which the product of an enzyme step serves as the substrate for the next enzyme. Here, as a result of genetic deficiency, there is a lack of the enzyme that converts L to M, leading to low concentrations of M, with a knock-on effect for the next step and a decreased concentration of N. The substrate L, proximal to the block, increases in concentration and that may lead to excessive production of metabolite Z, which becomes a biomarker of the disease. (B) Phenylalanine is converted to tyrosine by phenylalanine hydroxylase (PAH), which requires the cofactor tetrahydrobiopterin (BH₄). When mutations cause homozygous deficiency in PAH, the conversion of phenylalanine to tyrosine is blocked (double magenta bar). As a result, high levels of phenylalanine build up (hyperphenylalaninemia), driving the production of new phenylalanine metabolites (shown by red arrows). Three phenylketones are produced (phenylpyruvate, phenylacetate, and phenyllactate) and are excreted. Deficiency of different genes involved in BH₄ metabolism can also cause hyperphenylalaninemia.

Phenylketonuria was one of the earliest inborn errors of metabolism to be studied; it results from a deficiency of phenylalanine hydroxylase, a liver enzyme that converts phenylalanine to tyrosine ([Figure 1B](#)). Genetic deficiency in this enzyme results in elevated levels of phenylalanine (hyperphenylalaninemia) that can be sub-clinical (120–600 $\mu\text{mol/liter}$) or in untreated individuals result in mild phenylketonuria (600-1200 $\mu\text{mol/liter}$) or classical phenylketonuria (>1200 $\mu\text{mol/liter}$).

Elevated phenylalanine concentrations drive the production of phenylketone derivatives (see [Figure 1B](#)), which are excreted. The clinical symptoms of phenylketonuria are largely due to the toxic effects of very high phenylalanine levels in the brain—untreated children show progressively impaired brain development, leading to severe intellectual disability and various other symptoms including behavioral problems.

The standard treatment is really a form of prevention. Infants identified as having very high levels of blood phenylalanine are placed on a low-phenylalanine diet that is generally successful (although there may be problems with compliance in later years). The low-phenylalanine diet works because phenylketonuria is really a multifactorial disorder—two factors are absolutely required for the disease to manifest itself: a genetic factor (mutations at the *PAH* locus causing homozygous deficiency of phenylalanine hydroxylase) and an environmental factor (normal l-phenylalanine levels in dietary protein).

Phenylketonuria is classified as a monogenic disorder only because the vast majority of us are exposed to the environmental factor. Because affected sibs can show significant differences in clinical phenotype, modifier genes are also likely to be involved. The phenotype could be influenced by genetic variation in different processes (such as protein degradation, phenylalanine transport and disposal, transport of phenylalanine across the blood–brain barrier, brain sensitivity to phenylalanine toxicity).

Very high levels of phenylalanine in phenylketonuria can be teratogenic and can result in *embryofetopathy*. A homozygous mother (who might nevertheless have a mild phenotype that could go unrecognized) can have heterozygous offspring who go on to develop mental retardation. During pregnancy the placenta naturally selects for higher concentrations of amino acids; as a result, phenylalanine levels may double in fetal blood, causing serious damage to brain and some other organ systems during development. Again, this can be prevented or ameliorated if the expectant mother is placed on a low-phenylalanine diet from the earliest stages of pregnancy.

SUMMARY

- A small fraction of genetic variation causes disease, either by altering the amount of gene products (via a change in gene copy number or gene regulation, or by introducing premature termination codons), or by changing the sequence of gene products.
- The genetic code is redundant (most amino acids can be specified by multiple different codons), and universal for nuclear genes; mitochondria use a slightly different genetic code.
- Synonymous single nucleotide substitutions (silent mutations) replace one codon by another without changing the amino acid. They occasionally cause disease by altering RNA splicing.
- Nonsynonymous substitutions replace an amino-acid-specifying codon by a codon specifying a different amino acid (missense mutation) or by a stop codon (nonsense mutation).
- A missense mutation is likely to be pathogenic if the replacement amino acid is physiochemically rather different from the original amino

acid.

- Splicing mutations often alter important splice junction sequences. Additional splicing mutations change other important splice regulatory sequences within exons and introns, or activate a cryptic splice site to make a novel splice site.
- Insertions and deletions produce a translational frame shift if the resulting number of coding sequence nucleotides is not exactly divisible by three. Such mutations usually introduce an in-frame premature stop codon. RNA splicing mutations can also cause a translational frameshift.
- An in-frame premature termination codon often signals degradation of the mRNA (nonsense-mediated decay), but if the premature stop codon is close to the normal stop codon a truncated protein is usually produced that may sometimes result in a more severe phenotype.
- In vertebrate DNA, the CG dinucleotide is a target for cytosine methylation and a hotspot for C → T mutations; the resulting 5-methylcytosine is prone to deamination to give a thymine.
- DNA strands in a helix often pair out of register at runs of a short tandem repeat; replication in the mispaired region can cause pathogenic frameshifting insertions and deletions.
- Long arrays of CAG triplet repeats in coding DNA and various types of short tandem repeats in noncoding DNA can undergo unstable expansion. Such dynamic mutations show meiotic and mitotic instability and can cause disease by producing harmful proteins or RNAs.
- Nonallelic homologous recombination usually means a sequence exchange that occurs after pairing of nonallelic repeats with highly similar sequences.
- Reciprocal exchange between mispaired tandem repeats on chromatids (unequal crossover) or between distantly spaced direct repeats on the same DNA molecule or on paired chromatid DNAs can result in deletions or duplications. In alternative nonreciprocal exchanges the sequence of one copy is replaced in part by the sequence of another copy (gene conversion).

- Exchange between inverted repeats on the same strand can produce pathogenic inversions.
- Breaks in the DNA of one chromosome can result in subchromosomal deletions, inversions, and also ring chromosomes (formed after a chromosome has lost a terminal segment on each arm and the two broken ends of the centromere-containing fragment join up).
- Translocations occur when two chromosomes undergo breakages and then exchange fragments. A balanced translocation means that there has been no obvious net loss of DNA.
- Aneuploidy involves the gain or loss of whole chromosomes. The effects on the phenotype are due to a minority of genes that are especially dosage-sensitive.
- Disorders due to mtDNA mutations are maternally inherited and show variable ratios of mutant DNA to normal mtDNA (heteroplasmy). Clinical symptoms appear after heteroplasmy passes a threshold value that causes tissue damage due to defective oxidative phosphorylation.
- A mitochondrial genetic bottleneck occurs naturally in certain egg cell precursors, causing random severe reduction in the amount of mtDNA molecules passed to daughter cells. It can cause large variability in mutation load in egg cells produced by a heteroplasmic woman.
- Loss-of-function mutations can result in a null allele (complete absence of product, or complete functional inactivity) reduced expression, or an altered product with reduced functional activity.
- Haploinsufficiency means that a loss of function of one allele causes a phenotype in the presence of a working normal allele.
- Dominant-negative mutations result in a mutant gene product that somehow impairs the activity of the normal allele in a heterozygote.
- Gain-of-function mutations have a phenotypic effect in the presence of a normal allele that cannot be compensated for by producing more of the normal gene product.
- Loss-of-function and gain-of-function mutations in one gene can result in different phenotypes.
- Different components of a phenotype may be manifested at different threshold levels of gene function.

- Prion proteins and related proteins, can misfold to give a structure prone to self-aggregation that can induce other normally folded versions of the protein to misfold, thereby seeding protein aggregation to cause disease.
- Predicting the phenotype from the genotype is often difficult, even for a monogenic disorder. The effect of some types of mutation can be difficult to predict, and the phenotype is often influenced by genetic differences at other gene loci (modifier genes) and by environmental factors.

QUESTIONS

Questions can be downloaded by visiting the following link, under Support Materials: www.routledge.com/9780367490812.

FURTHER READING

Amino acid substitutions and silent mutations

Betts MJ & Russell R (2007) Amino-acid properties and consequences of substitutions. In *Bioinformatics for Geneticists*, 2nd ed (Barnes MR ed.), pp 311–341. Wiley-Blackwell.

Boyko AR (2008) Assessing the evolutionary impact of amino acid mutations in the human genome. *PLoS Genet* 4:e100083; PMID 18516229.

Grantham R (1974) Amino acid difference formulae to help explain protein evolution. *Science* 185:862–864; PMID 4843792. [Gives a matrix that quantifies the effect of amino acid substitutions.]

Ng PC & Henikoff S (2006) Predicting the effects of amino acid substitutions on protein function. *Annu Rev Gen Hum Genet* 7:61–80; PMID 16824020.

Sauna ZE & Kimchi-Sarfaty C (2011). Understanding the contribution of synonymous mutations to human disease. *Nat Rev Genet* 12:683–691; PMID 21878961.

Nonsense-mediated decay (NMD)

- Bhuvanagiri M (2010) NMD: RNA biology meets human genetic medicine. *Biochem J* 430:365–377; PMID 20795950.
- Holbrook JA (2004) Nonsense-mediated decay approaches the clinic. *Nat Genet* 36:801–808; PMID 15284851.

Splicing and regulatory mutations

- Cartegni L (2002) Listening to silence and understanding nonsense: exonic mutations that affect splicing. *Nat Rev Genet* 3:285–298; PMID 11967553.
- Jarinova O & Ekker M (2012) Regulatory variations in the era of next-gen sequencing: implications for clinical molecular diagnostics. *Hum Mutat* 33:1021–1030; PMID 22431194.
- Sterne-Weiler T (2011) Loss of exon identity is a common mechanism of human inherited disease. *Genome Res* 21:1563–1571; PMID 21750108.
- Wang GS & Cooper TA (2007) Splicing in disease: disruption of the splicing code and the decoding machinery. *Nat Rev Genet* 8:749–761; PMID 17726481.

Mutation rates, mutation load, and noncoding DNA mutations

- Conrad DF (2011) Variation in genome-wide mutation rates within and between human families. *Nat Genet* 43:712–714; PMID 21666693.
- French JD & Edwards SL (2020) The role of noncoding variants in heritable disease. *Trends Genet* 36:880–891; PMID 32741549.
- [Goriely A & Wilkie AOM](#) (2012) Paternal age effect mutations and selfish spermatogonial selection: causes and consequences for human disease. *Am J Hum Genet* 90:175–200; PMID 22325359.
- Keightley PD (2012) Rates and fitness consequences of new mutations in humans. *Genetics* 190:295–304; PMID 22345605.
- MacArthur DG (2012) A systematic survey of loss-of-function variants in human protein-coding genes. *Science* 335:823–828; PMID: 22344438.
- Makrythanasis P & Antonorakis S (2013) Pathogenic variants in nonprotein-coding sequences. *Clin Genet* 84:422–428; PMID 24007299.
- Xue Y (2012) Deleterious- and disease-allele prevalence in healthy individuals: insights from current predictions, mutation databases, and population-scale resequencing. *Am J Hum Genet* 91:1022–1032; PMID 23217326.

Pathogenic unstable expansion of short tandem repeats

Hannan AJ (2018) Tandem repeats mediating genetic plasticity in health and disease. *Nat Rev Genet* 19:286–298; PMID 29398703.

Khristich AN & Mirkin SM (2020) On the wrong DNA track: Molecular mechanisms of repeat-mediated genome instability. *J Biol Chem* 295:4134–4170; PMID 32060097.

Rodriguez CM & Todd PK (2019) New pathologic mechanisms in nucleotide repeat expansion disorders. *Neurobiol Dis* 130:104515; PMID 31229686.

Gene conversion and nonallelic homologous recombination

Chen JM (2007) Gene conversion: mechanisms, evolution and human disease. *Nat Rev Genet* 8:762–775; PMID 17846636.

Liu P (2012) Mechanisms for recurrent and complex human genomic rearrangements. *Curr Opin Genet Dev* 22:1–10; PMID 22440479.

Chromosome abnormalities and nomenclature

[McGowan-Jordan J, Hastings RJ & Moore S](#) (eds.) An International System for Human Cytogenomic Nomenclature (2020) *Cytogenet Genome Res* 160: Issues 7–8.

Nagaoka SI (2012) Human aneuploidy: mechanisms and new insights into an age-old problem. *Nat Rev Genet* 13:493–504; PMID 22705668.

Roukos V (2013) The cellular etiology of chromosome translocations. *Curr Opin Cell Biol* 25:357–364; PMID 23498663.

Weckselblatt B & Rudd MK (2015) Human structural variation: mechanisms of chromosome rearrangements. *Trends Genet* 31:587–599; PMID 26209074.

Molecular pathology of mitochondrial disorders

Craven L (2017) Recent advances in mitochondrial disease. *Annu Rev Genom Hum Genet* 18:257–275; PMID 28415858.

Gasparre G & Porcelli AM eds. (2020) *The Human Mitochondrial Genome – From Basic Biology to Disease*. Academic Press.

Gusic M & Porkisch H (2021) Genetic basis of mitochondrial diseases. *FEBS Lett* 595:1132–1158; PMID 33655490.

MITOMAP database at <https://www.mitomap.org> [a human mitochondrial genome database, with pathogenic mtDNA mutations and associated clinical characteristics etc].

Stewart JB & Chinnery PF (2021) Extreme heterogeneity of human mitochondrial DNA from organelles to populations. *Nat Rev Genet* 22:106–118; PMID 32989265.

Gain-of-function mutations, haploinsufficiency and molecular basis of genetic dominance

Lester HA & Karschin A (2000) Gain-of-function mutants: ion channels and G protein-coupled receptors. *Annu Rev Neurosci* 23:89–125; PMID 10845060.

Veitia RA & Birchler JA (2010) Dominance and gene dosage balance in health and disease: why levels matter! *J Pathol* 220:174–185; PMID 19827001.

Wilkie AOM (1994) The molecular basis of genetic dominance. *J Med Genet* 31:89–98; PMID 8182727.

Protein misfolding and protein aggregation in disease

Chiti F & Dobson CM (2017) Protein misfolding, amyloid formation, and human disease: a summary of progress over the last decade. *Annu Rev Biochem* 86:27–68; PMID 28498720.

Cox D (2020) Protein aggregation in cell biology: an aggregomics perspective of health and disease. *Semin Cell Dev Biol* 99:115–130; PMID 29753879.

Dobson CM (2020) The amyloid phenomenon and its significance in biology and medicine. *Cold Spring Harb Perspec Biol* 12:a033878; PMID 30936117.

O'Carroll A (2020) Prions and prion-like assemblies in neurodegeneration and immunity: The emergence of universal mechanisms across health and disease. *Semin Cell Dev Biol* 99:115–130; PMID 31818518.

Modifier genes

Drumm ML (2012) Genetic variation and clinical heterogeneity in cystic fibrosis. *Annu Rev Pathol* 7:267–282; PMID 22017581.

8

Identifying disease genes and genetic susceptibility to complex disease

DOI: [10.1201/9781003044406-8](https://doi.org/10.1201/9781003044406-8)

CONTENTS

[8.1 IDENTIFYING GENES IN MONOGENIC DISORDERS](#)

[8.2 APPROACHES TO MAPPING AND IDENTIFYING GENETIC SUSCEPTIBILITY TO COMPLEX DISEASE](#)

[8.3 ASPECTS OF THE GENETIC ARCHITECTURE OF COMPLEX DISEASE AND THE CONTRIBUTIONS OF ENVIRONMENTAL AND EPIGENETIC FACTORS](#)

[SUMMARY](#)

[QUESTIONS](#)

[FURTHER READING](#)

Until recently, the molecular identification of rare genes for monogenic disorders was laborious, often consuming many years of painstaking effort to identify even just one disease gene. Now, in the era of massively parallel DNA sequencing (next-generation sequencing), it has almost become routine. We cover the principles in [Section 8.1](#). Some difficulties remain, however, because for some single gene disorders the disease phenotypes do not have a very well-defined, distinctive pathology. And a good deal of follow-up work will be needed to dissect out all the factors in monogenic diseases, which are sometimes rather complex, as described in [Section 7.9](#).

The next big challenge has been to identify genes underlying complex (multifactorial) diseases, in which there is no obviously predominant disease locus (at least, not to the extent found in monogenic disorders). Instead, expression of the disease phenotype may be notably dependent on a few genes (oligogenic disorders) or many genes (polygenic disorders), with

variable (and sometimes very strong) contributions from environmental factors. We cover the background to complex disease and polygenic theory in [Section 8.2](#).

The genetic contribution to complex diseases differs according to the disease and between populations, and its overall impact can vary within a single population, depending on changeable environmental conditions. Investigations into the genetic susceptibility to multifactorial disease began decades ago but had limited success until the early 2000s. More recently, however, many DNA variants have been identified to confer susceptibility to complex diseases (genetic risk factors) and some have been shown to lower disease susceptibility (protective factors). We outline the different general approaches used to uncover the genetic susceptibility to complex diseases in [Section 8.2](#).

In [Section 8.3](#) we consider selected aspects of the genetic architecture of complex disease to illustrate the progress that has been made. We then follow up by considering gene-environment interactions and the contribution made by epigenetic factors. We consider investigations of common cancers separately in [Chapter 10](#).

8.1 IDENTIFYING GENES IN MONOGENIC DISORDERS

A historical overview of identifying genes in monogenic disorders

Identifying genes underlying monogenic disorders began with a very few exceptional cases in the 1970s and early 1980s. The underlying genes were able to be identified through a known protein product (*functional cloning*), or as a result of huge enrichment of corresponding mRNAs in certain cell types. In the former case, for example, hemophilia A was known to be due to a deficiency of blood clotting factor VIII. Enough factor VIII protein was purified from pig blood to obtain a partial amino acid sequence. After an optimal short sequence of amino acids was identified, a panel of different but related oligonucleotides was synthesized to cover all codon possibilities for the selected sequence of amino acids. The resulting oligonucleotides were used as probes to screen DNA libraries, identifying first homologous cDNA clones and eventually human gene clones.

An alternative approach was candidate gene testing. Candidate genes could be selected on the basis of knowing the biology of the condition or often from observed similarities between the disease phenotype and a highly related phenotype in humans or animals. After the *FBN1* fibrillin gene was shown to be a locus for Marfan syndrome, for example, the related *FBN2* fibrillin gene was quickly and successfully investigated as a candidate gene for a very similar disorder, congenital contractual arachnodactyly. And after the mouse *Sox10* gene was identified as the locus for the *Dominant megacolon (Dom)* phenotype, a mouse model of Hirschsprung disease, human *SOX10* was quickly shown to be mutated in Waardenburg–Hirschsprung disease.

Positional cloning strategies were needed to identify genes underlying diseases where little was known about the kind of gene product they might make. They relied on first

getting a subchromosomal position for the disease gene. (For X-linked conditions, at least the location had already been narrowed down to a single chromosome.) As listed below, two types of approach were used to identify subchromosomal locations for an underlying gene in monogenic disorders.

- *Linkage analysis.* This is a general method applicable to the great majority of monogenic disorders. Blood samples would be collected from multiple family members for each of many families with the disorder. DNA from the blood samples would then be used to assay genotypes for each of a collection of hundreds of DNA markers that had previously been mapped to specific subchromosomal areas from across the genome. If a particular DNA marker co-segregated with disease, the disease gene could be inferred to map to the same subchromosomal location as the marker.
- *Chromosome break mapping.* Linkage analysis is not suitable for some disorders, notably dominant disorders where affected individuals fail to reproduce. Scanning patient blood samples for chromosome breaks can be profitable because of the high chance that *de novo* chromosome breaks (from translocations, deletions, or inversions) could be disease-associated (an important gene might have been inactivated as a result of the chromosomal damage; if so, it must lie within the deleted area or close to a chromosome breakpoint).

Positional cloning versus positional candidate approaches

Early efforts at getting a subchromosomal position for a monogenic disorder ran into the problem that very little would be known about the DNA sequences and genes in that region; DNA sequences from the same subchromosomal area needed to be cloned, sequenced, and tested. Such positional cloning efforts could be hugely laborious. Interested readers can get an idea of what was involved in positional cloning of the cystic fibrosis gene from PMID 23378595.

The task became easier as the genome began to be deciphered, with data from multiple laboratories being used to map large numbers of protein-coding genes to subchromosomal regions that had previously been poorly studied. As a result, much of the slog of cloning DNA was avoided, and the projects became *positional candidate approaches*. Genes in the subchromosomal region of interest could now be identified by consulting gene, genome, and literature databases, further studied as required, and then prioritized for mutation screening, according to their known characteristics. Ultimately, detailed gene maps were obtained for each chromosome; they were enormously helpful for positional candidate approaches with which to identify disease genes.

The final step: mutation screening

The final step is to scan DNA samples from affected individuals for disease-associated mutations in candidate genes (by comparison with unaffected controls). For a highly penetrant dominant disorder, pathogenic mutations should normally be found in affected individuals only; for a recessive disorder, the pathogenic mutations will occasionally be found in normal individuals (who might then be suspected to be heterozygous carriers). The list of candidate genes mapping in the relevant subchromosomal region might often be daunting, and computer programs were developed to prioritize the most likely candidate disease genes.

Mutation screening typically involves amplifying individual exons and the immediately surrounding intronic sequence from individual candidate genes and sequencing them to identify mutations associated with disease in panels of DNA samples from affected individuals and controls. As we will see in the final part of [Section 8.1](#), however, genome-wide gene or exon sequencing can often dispense with the time-consuming need to first identify candidate disease genes.

Protein-coding genes have been the overwhelming choice for candidate disease genes. Loss-of-function mutations are expected in recessive conditions, and in dominant disorders resulting from haploinsufficiency. They are relatively easy to identify in protein-coding genes because they often occur in coding DNA sequences (where it is easy to spot mutations causing premature termination codons, changes to the reading frame, or amino acid substitutions) or close to exon–intron boundaries (causing splicing abnormalities). Gain-of-function mutations often involve specific missense mutations (which would not be expected in controls). A tiny number of RNA genes have, however, been identified as loci for monogenic disorders (see [Table 7.5](#) on page 191).

Linkage analysis to map genes for monogenic disorders to defined subchromosomal regions

Genetic markers (polymorphic loci) from across the genome can be used to track the inheritance of a gene by using **linkage analysis**. Different types of linkage analysis can be carried out, but success usually depends on having suitably informative families with multiple affected individuals. Because human family sizes are generally very small, multiple different families need to be investigated. Hundreds of genetic markers are needed, from defined locations distributed across the genome. That became possible with the development of human genetic maps.

Human genetic maps

Human genetic mapping has been a recent endeavor, unlike genetic mapping in model organisms where gene mutations causing readily identifiable phenotypes can easily be mapped. In *Drosophila*, for example, crosses can be set up to breed mutant white-eyed flies with flies that have abnormal curly wings; the progeny are then examined to see if the two mutant phenotypes segregate together or not. In humans, however, that kind of approach could never be applied—a different strategy was needed.

Instead of having a genetic map based on gene mutations, the solution to making a human genetic map was to identify *general* DNA variants (which rarely map within coding sequences, and usually have no known effects on the phenotype). Different types of variants were identified and mapped to specific genome locations, beginning with restriction fragment length polymorphisms (RFLPs) that created or destroyed a restriction site ([Figure 4.4](#)), followed by microsatellite polymorphisms that varied in the copy number of short tandem repeats ([Figure 4.5](#)).

The first comprehensive map of human genetic markers (polymorphisms) did not appear until 1994. Based on microsatellite and restriction site polymorphisms, it had a marker spacing of just over one marker per megabase of DNA. Microsatellite markers have the advantage that they are highly polymorphic (with multiple alleles, each having a different number of copies of the repeat), whereas restriction site polymorphisms often have just two alleles.

The most recent maps are based on single nucleotide polymorphisms (SNPs). They also have limited polymorphism, with often just two alleles. But they have two very strong advantages: they are extremely abundant in the human genome, and they are amenable to automated typing.

Data on individual SNPs can be accessed at the dbSNP database (<https://www.ncbi.nlm.nih.gov/snp/>). Identifying reference numbers are composed of a seven-digit to nine-digit number prefixed by rs (= reference SNP), such as rs1800588. The database can also be queried with a gene symbol to find SNPs in a specific gene; the resulting data can be filtered progressively to get human SNPs, and then SNPs that are of clinical interest or that have been recorded in the corresponding locus-specific database.

Principle of genetic linkage

One fundamental principle underlies genetic linkage: alleles at very closely neighboring loci on a DNA molecule are co-inherited because the chance that they are separated by recombination is very low. By extension, alleles at distantly spaced loci on the same DNA molecule are much more likely to be separated by recombination at meiosis. (During human meiosis, chromosomes are often split by recombination into between two and seven segments—see [Figure 8.1](#) for an example.)

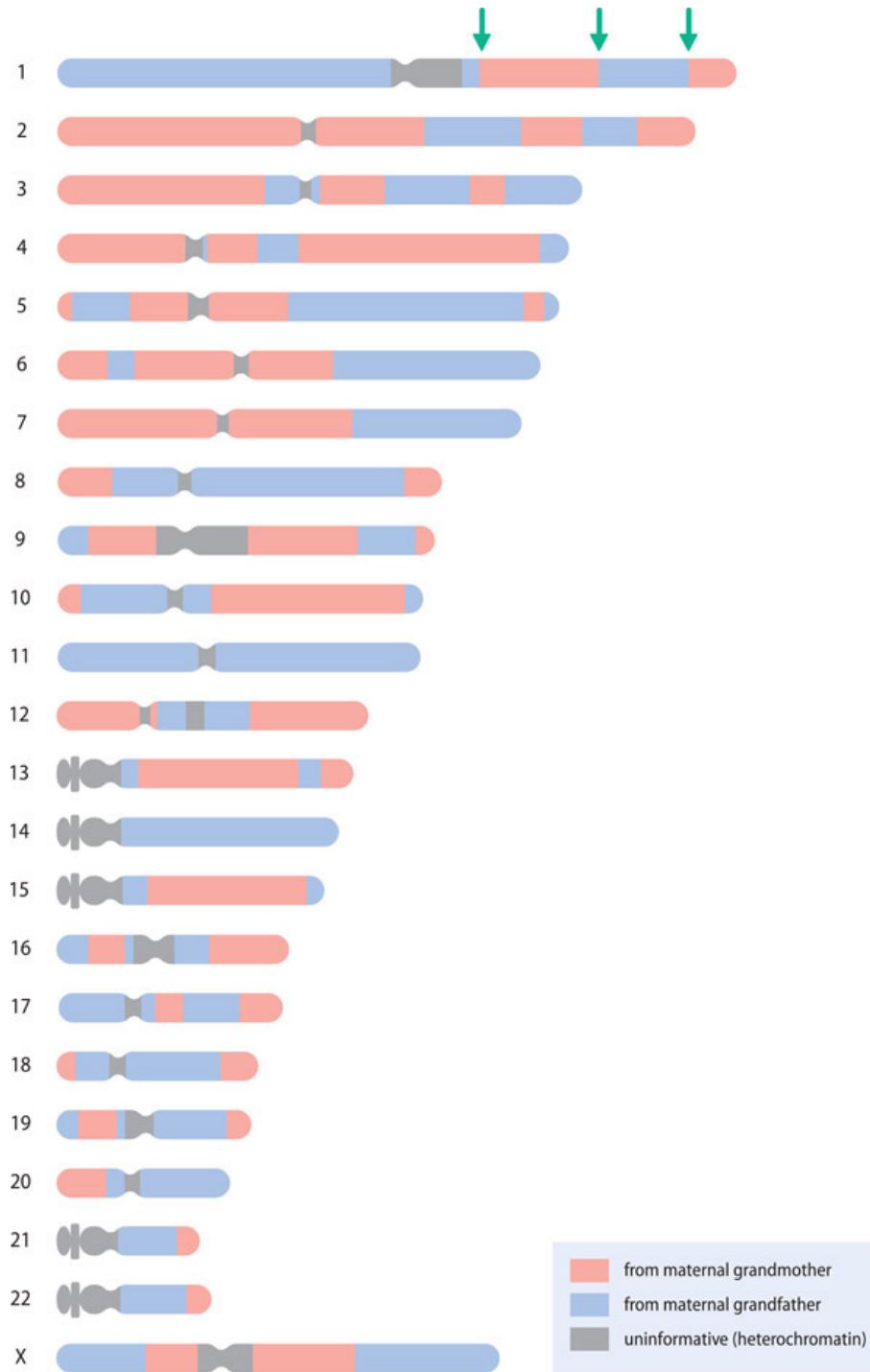


Figure 8.1 Mapping human meiotic crossovers by genome-wide SNP genotyping in families. This example shows the deduced pattern of meiotic recombination in the chromosomes passed on from a mother to her daughter, after whole genome SNP typing of the family members, including grandparents. The egg transmitted by the mother to her daughter was formed by meiotic cell division in which chromosomes from the maternal grandmother (pink) recombined with those from the maternal grandfather (blue) to give the alternating blue and pink patterns shown (as an illustration, the green arrows show the location of three

crossovers on chromosome 1). More than 50 crossovers can be detected, but none are apparent, unexpectedly, on chromosomes 11 and 14 (undetected crossovers might have occurred here in heterochromatic regions where no SNP markers were available; most of the short arm of chromosome 14 is composed of heterochromatin). (Courtesy of Rosemary J Redfield.) (CC BY-SA 2.5 CA).

A **haplotype** is a series of alleles at two or more neighboring loci on a *single* chromosomal DNA molecule. In human genetics, the term was first widely used within the context of the HLA system (readers who are unclear about haplotypes might wish to have a look at [Figure 2](#) in [Box 4.3](#) on page 106 to see how haplotypes are derived after obtaining genotypes from multiple family members).

We illustrate the principle of a *disease haplotype* in [Figure 8.2](#) within the context of a gene for an autosomal dominant disorder. The marker loci flanking the disease locus in [Figure 8.2](#) are imagined to be very close to the disease locus. Recombination within this region would be extremely rare—the marker loci would be said to be *tightly linked* to the disease locus.

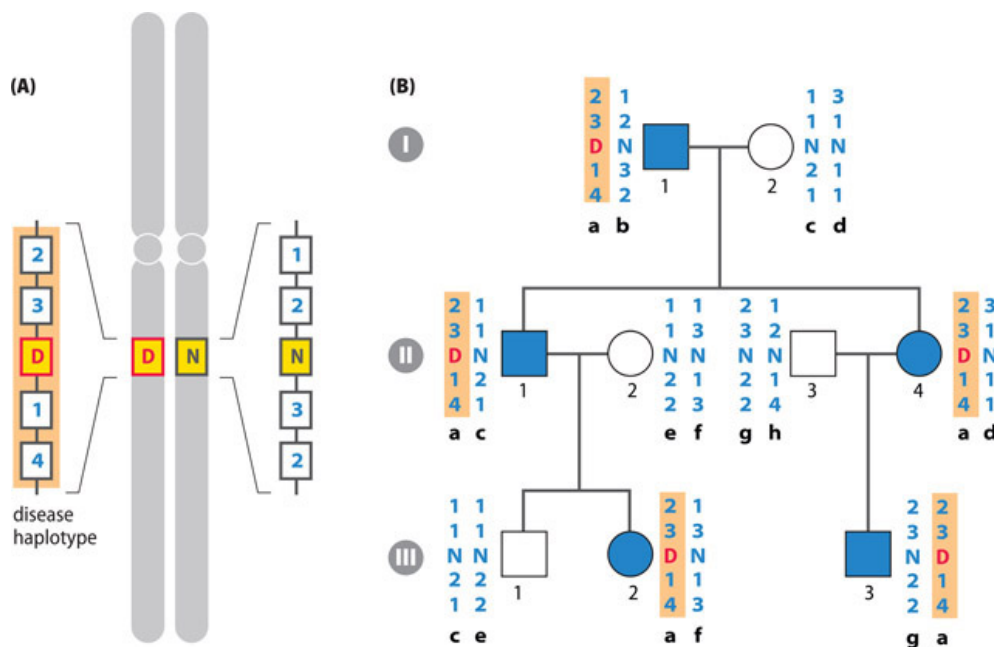


Figure 8.2 Inheritance of a disease haplotype in an autosomal dominant disorder. (A) The disease locus, highlighted by the yellow box, can be imagined to have two alleles: D (disease) and N (normal). Here we also consider alleles at two proximal marker loci and two distal marker loci that are physically located very close to the disease locus, for example within 0.5 Mb. The disease haplotype in the affected individual is defined by the sequence of *alleles* at consecutive neighboring marker loci that could be represented here as 2-3-1-4 (when read in the proximal to distal direction). (B) The haplotypes in (A) belong to the affected grandfather (I-1) in this pedigree. The highlighted disease haplotype (a) is transmitted without change to affected individuals in generations II and III.

In [Figure 8.2B](#) the disease haplotype is transmitted unchanged through four meioses: from the grandfather (I-1) to his affected son and daughter, and then to the two affected grandchildren (III-2 and III-3). For marker loci that are increasingly distant from the disease locus, the incidence of recombination between disease locus and marker locus becomes progressively greater.

As the distance separating a marker locus and the disease locus on the same chromosome increases, a point will be reached where the chance of recombination between the two loci would equal the chance of no recombination between them. The marker would then be said to be *unlinked* to the disease locus (it would be no different from a marker on a different chromosome, for which an allele would have a 50 % chance of segregating with disease, just by chance).

Human meiotic recombination frequencies

For any two loci, the chance of recombination is a measure of the distance between them. Loci separated by recombination in 1 % of meioses are said to be 1 *centimorgan* (*cM*) apart. Genetic distances are related to physical distances, but not in a uniform way: there is a rough correspondence between a genetic distance of 1 cM in humans and a physical distance of close to 1 Mb of DNA, but there are considerable regional variations across chromosomes.

Recombination is much more common at subtelomeric regions than in the middle of chromosome arms, for example, and is much less frequent in heterochromatic regions. At higher resolution, the majority (60 % or more) of crossovers occur at a number of short hotspots, about 1–2 kb long, across the genome.

Recombination frequencies in human meiosis also show significant sex differences. Using dense genome-wide SNP mapping in nuclear families, two large studies (by [Cheung et al. \[2007\]](#) and [Coop et al. \[2008\]](#)—see under Further Reading) looked at 728 and 557 meioses respectively; the overall mean scores averaged across the two studies was 25.2 crossovers in male meiosis and 39.1 crossovers in female meiosis. But there is also variation between individuals, and even between individual meioses within a single individual (as shown in [Figure 8.3](#)). Overall, therefore, there is no one correct human genetic map length: for any one subchromosomal region, the correspondence between the physical (DNA sequence) length and genetic map length will vary from one meiosis to another.

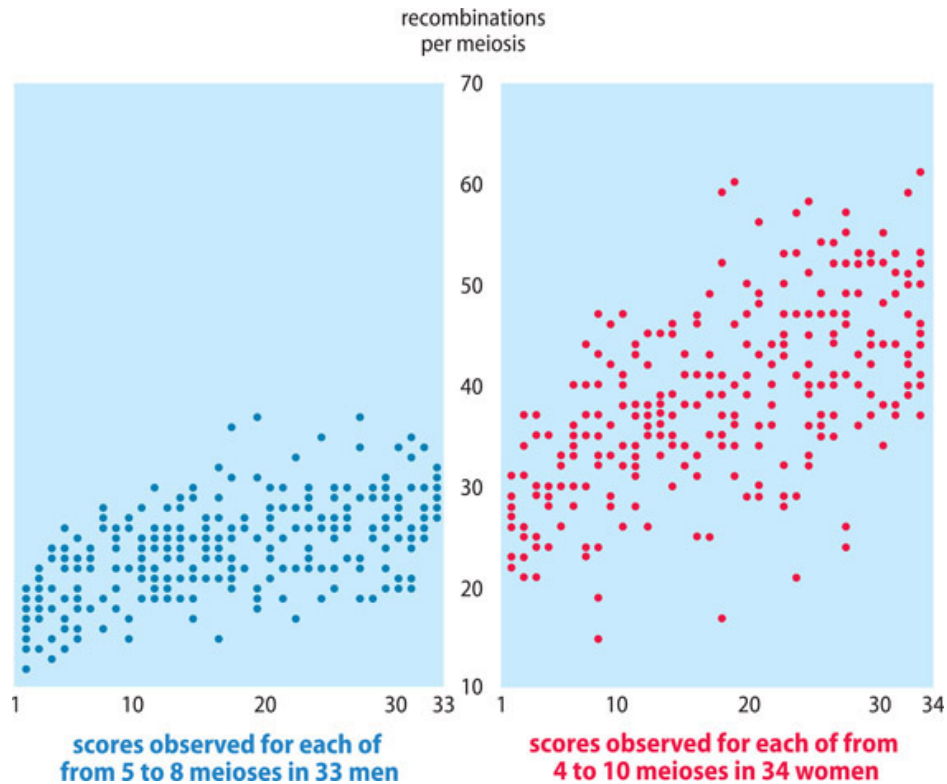


Figure 8.3 Individual differences in the numbers of recombinations per meiosis. Each dot represents the number of recombinations identified in an individual meiosis; each vertical line of dots represents the scores determined in each of multiple meioses in a single male or female, as shown. The number of recombinants per meiosis was determined by genotyping individuals in families with 6324 SNP markers. (From [Cheung VG et al. \[2007\]](#) *Am J Hum Genet* 80:526–530; PMID 17273974. With permission from Elsevier.)

Standard genome-wide linkage analyses

To map disease genes to specific chromosomal regions, genome-wide linkage analyses can be used. Usually, several hundred genetic markers from defined loci across the whole genome are genotyped in family members with the disease.

The results may show some marker loci that are tightly linked to the disease, thereby indicating a subchromosomal location for the disease gene.

Using, say, 400 markers for genome-wide linkage analyses would give a marker density of one every 7–8 Mb or so. Given that our chromosomes are about 50–250 Mb in length and are split by meiotic recombination into usually only two to seven segments ([Figure 8.1](#)), there is a high chance that one or more of a 400 genome-wide marker set will be sufficiently close to the disease locus for a marker allele to co-segregate with a disease allele.

The segregation of alleles from each marker locus is followed through a suitably large number of informative meioses (see [Figure 8.4](#) for examples of informative and uninformative meioses). In practice that means having access to samples from multiple affected and unaffected members usually drawn from several families.

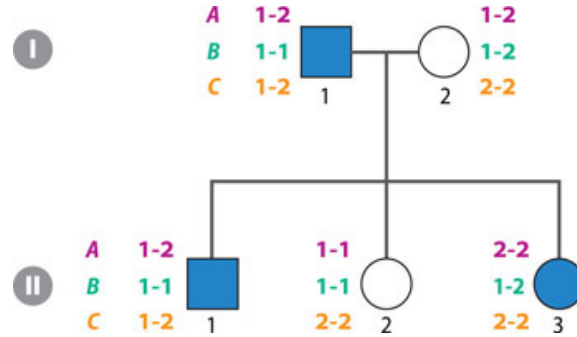


Figure 8.4 Informative and uninformative meioses. Let us assume full penetrance and autosomal dominant inheritance of the phenotype in the pedigree shown here. The disease allele has been transmitted from the father to the son and one daughter (II-3). Genotypes for three unlinked marker loci, A, B, and C, are shown by the respective colored figures. Consider marker A. For the affected son it is impossible to tell which parent contributed allele 1 and which contributed allele 2, but it is possible to infer that each parent contributed an allele 1 to II-2 and an allele 2 to II-3. Marker B is completely uninformative here because I-1 is homozygous and it is impossible to tell which of the two paternal alleles 1 was transmitted to each child. Marker C is informative in each case: the father transmitted allele 1 to his affected son, but transmitted allele 2 to both daughters.

In an idealized situation, **recombinants** and non-recombinants can be clearly identified. In the autosomal dominant pedigree in [Figure 8.5](#), the affected individual in generation II is heterozygous for a disease allele and he is also heterozygous for a marker, having a 1,2 genotype. In the highly unusual circumstances shown in this figure it is possible to identify recombinants and non-recombinants unambiguously. In practice, recombinants often cannot be identified unambiguously (linkage studies often use families that do not have such an ideal structure, and key meioses may often be uninformative).

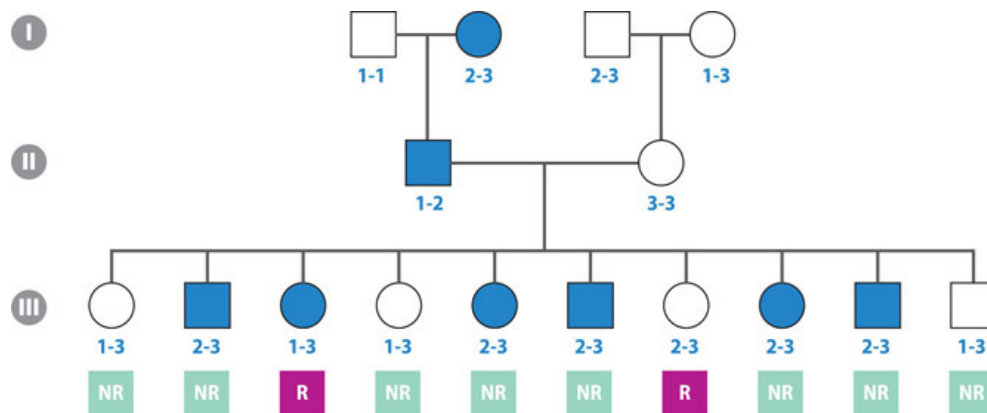


Figure 8.5 Unambiguous identification of recombinants and non-recombinants in an idealized pedigree.

Members of this autosomal dominant pedigree have been typed for a marker that has three alleles (1,2, and 3). The available data suggest that allele 2 of the marker is segregating with the disease. In that case, 8 out of the 10 children in generation III are non-recombinants (NR): either they have inherited both the disease and allele 2 from their affected father, or they have inherited paternal allele 1 and are unaffected. The other two are recombinants (R): either paternal allele 1 has segregated with disease, or paternal allele 2 is associated with a normal allele at the disease locus.

To get round the difficulties of identifying recombinants in human genetic mapping, sophisticated computer programs are needed. They do not attempt to identify individual recombinants. Instead, their job is to survey the linkage data and then calculate alternative probabilities for linkage and nonlinkage. They then express the ratio of these probabilities as a logarithmic value called a **lod score**, as described in [Box 8.1](#). Programs such as these are dependent on previous information on the mode of inheritance, disease gene frequency, and the penetrance of the genotypes at the disease locus (that is, the frequency with which the genotypes manifest themselves in the phenotype). For monogenic disorders, the mode of inheritance and disease gene frequency often do not present much difficulty; penetrance can be a more difficult problem.

BOX 8.1 LOD SCORES AND STATISTICAL EVIDENCE FOR LINKAGE

Computer-based linkage analysis programs calculate two alternative probabilities: (i) the likelihood of the marker data, given that there is linkage between the marker and the disease locus at specified recombination fractions; and (ii) the likelihood of the marker data, assuming that the marker is unlinked to the disease locus. The *likelihood ratio* is the ratio of likelihood (i) and likelihood (ii), and provides evidence for or against linkage.

The convention has been to use the logarithm of the likelihood ratio, called the **lod score** (logarithm of the odds). Individual lod scores are calculated for a defined recombination fraction (θ), and so for each marker the computer programs provide a table of lod scores for different recombination fractions ($\theta = 0, 0.10, 0.20, 0.30$, and so on). The reported recombination fraction is chosen to be the one where the lod score (Z) is at a maximum (Z_{\max}).

A lod score of +3 is normally taken to be the *threshold* of statistical significance for linkage between two loci. It means that the likelihood of the data given that the two loci are linked is 1000 times greater than if it is unlinked ($\log_{10}1000 = 3$). Linkage between two loci (say, a disease locus and a marker locus) can theoretically be achieved with 10 informative meioses. At each informative meiosis there are two choices: the two loci are linked (a specific marker allele segregates with disease), or they are not linked (the marker allele does not segregate with disease). If the same marker allele segregates with disease in each of 10 informative meioses in a large pedigree, the odds of that happening by chance

are 2^{10} to 1 against, or just over 1000:1 against. In practice, because of poor family structures and some uninformative meioses, 20 or more meioses are often needed for linkage to be successful; DNA samples are usually needed from affected and unaffected members of multiple families.

The ratio of 1000:1 might seem overwhelming evidence in favor of linkage, but it is required to offset the inherent improbability of linkage. With 22 autosomes, two randomly chosen loci are unlikely to be on the same chromosome. Even if the two loci are on the same chromosome, however, they may be well separated, and so unlinked. Factoring in both of these observations, the prior odds are about 50:1 *against* linkage, or 1:50 in favor of linkage. That means we need pretty strong evidence from linkage analysis data to counteract the low starting probability. A likelihood ratio of 1000:1 in favor of linkage multiplied by a prior odds of 1:50 in favor of linkage gives a final odds of only 20:1 in favor of linkage. That is, a single lod score of 3 is not proof of linkage; there is a 1 in 20 chance that the loci are not linked.

Higher lod scores provide greater support for linkage. A lod score of 5 is 100 times more convincing than a lod score of 3. In practice, therefore, genome-wide claims for linkage based on a single lod score less than 5 should be treated as provisional evidence for linkage. However, significant lod scores may often be obtained for several markers clustered in one subchromosomal region; if so, the combined data provide strong evidence of linkage. See [Table 1](#) for the example of a dominantly inherited skin disorder, Hailey–Hailey disease (OMIM 169600), in which four neighboring markers in the 3q21-q24 interval show significant evidence of linkage.

TABLE 1

PAIRWISE LOD SCORES FOR HAILEY-HAILEY DISEASE AND MARKERS AT 3Q21-Q24

Marker	Lod score (Z) at				Maximum likelihood estimates	
	0.00	0.10	0.20	0.30	Z _{max}	AT θ =
D3S1589	-0.99	2.29	1.90	1.14	2.29	0.09
D3S1587	4.54	3.80	2.83	1.73	4.54	0.00
D3S1292	2.62	4.98	3.84	2.41	5.32	0.04
D3S1273	3.36	5.52	4.12	2.54	6.10	0.03

The descending order of markers is from proximal to distal. Analyses were carried out in six disease families. The underlying disease gene was subsequently found by positional cloning to be *ATP2C1* and to map just proximal to D3S1587. (Data from Richard G et al. [1995] *J Invest Dermatol* 105:357-360; PMID 7665912).

Marker	Lod score (Z) at				Maximum likelihood estimates	
	0.00	0.10	0.20	0.30	Z _{max}	AT θ =
D3S1290	-2.81	3.83	3.05	1.94	3.90	0.07
D3S1764	-8.62	2.21	2.06	1.38	2.26	0.13

The descending order of markers is from proximal to distal. Analyses were carried out in six disease families. The underlying disease gene was subsequently found by positional cloning to be *ATP2C1* and to map just proximal to *D3S1587*. (Data from Richard G et al. [1995] *J Invest Dermatol* 105:357-360; PMID 7665912).

The threshold for excluding linkage is a lod score of -2 . *Exclusion mapping* can be helpful in excluding a candidate gene of interest, and in genome-wide studies the exclusion of a substantial fraction of the genome can direct extensive analysis of the remaining regions.

Linkage can theoretically be achieved with 10 informative meioses, but in practice linkage analysis is rarely successful when there are fewer than 20 or so meioses (see [Box 8.1](#)). A major confounding problem in linkage analysis is locus heterogeneity, when the same disease in different families under study may be caused by different genes—it is important to try to study families with extremely similar disease phenotypes.

After obtaining evidence of linkage, crossover points are deduced to identify a minimal subchromosomal region for a disease gene ([Figure 8.6](#)).

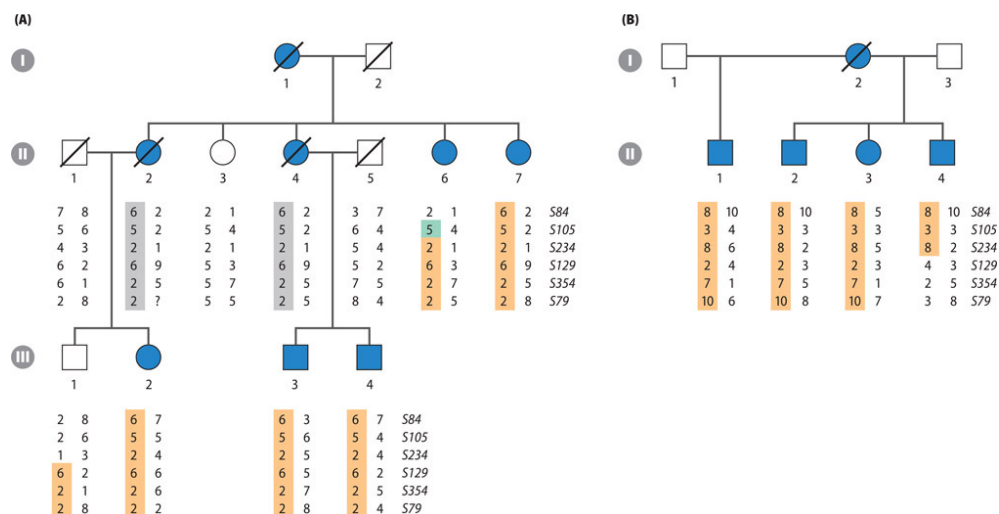


Figure 8.6 Defining the minimal candidate region by inspection of haplotypes. The two pedigrees show a dominantly inherited skin disorder, Darier–White disease, that had previously been mapped to 12q. The 12q marker haplotype segregating with disease is highlighted by orange shading. Gray boxes mark inferred haplotypes in deceased family members. In pedigree (A) the recombination in II-6 maps the disease gene

distal to marker *D12S84* (abbreviated to *S84* in the figure); the *D12S105* marker is uninformative because I-1 was evidently homozygous for allele 5—compare the genotypes of II-3 and II-7. The recombination shown in III-1 suggests that the disease gene maps proximal to *D12S129*, but this requires confirmation (the interpretation depends on the genotypes of II-1 and II-2 being inferred correctly, and on III-1 not being a non-penetrant gene carrier). The recombination in II-4 in pedigree B provides the confirmation. The combined data locate the Darier gene to the interval between *D12S84* and *D12S129*. (Adapted from Carter SA et al. [1994] *Genomics* 24:378–382; PMID 7698764. With permission from Elsevier.)

Autozygosity mapping in extended inbred families

The term **autozygosity** means homozygosity for markers that are *identical by descent*—the two alleles are copies of one specific allele transmitted to both parents from a recent common ancestor. In some societies, such as in the Middle East and parts of Asia, cousin marriages are quite common, and in extended inbred families there may be several individuals who are autozygous for an allele because of parental consanguinity. As illustrated in [Figure 1](#) of [Box 5.2](#) on page 115, second cousins share respectively 1/32 of their genes, and so their children would be autozygous at 1/64 of all loci.

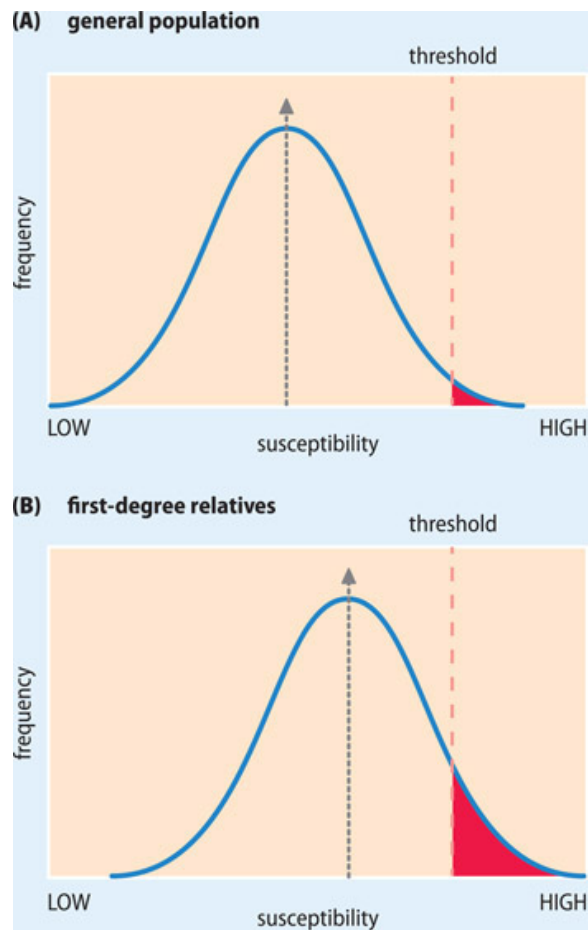


Figure 1 Distribution of susceptibility to a complex disease in the general population and in first-degree relatives of an affected person.

Homozygosity for a particular marker allele could be due to autozygosity. Alternatively, it might result from the inheritance of a second, independent copy of that allele that has been brought into the family at some stage; alleles such as this would be said to be *identical by state*. Homozygosity for a haplotype of marker alleles, however, is more likely to indicate autozygosity; if there are additional sibs who are homozygous for the same marker haplotypes, quite small consanguineous families can generate significant lod scores. Autozygosity mapping can therefore be a very efficient way of mapping a recessive monogenic disorder; see Goodship et al. (2000) under Further Reading for a successful application.

Chromosome abnormalities and other large-scale mutations as routes to identifying disease genes

Some affected individuals show a specific chromosome abnormality or other very large-scale mutation that can be detected quite readily (see below). Abnormalities such as these might occur coincidentally. That is, the abnormality might have nothing to do with disease (some parts of our genome do not contain critically important genes, and chromosome abnormalities affecting these regions might be found in a small percentage of normal people). Alternatively, the abnormality causes the disease by affecting the expression or structure of a gene or genes in that region. That would be more likely if the same DNA region were disrupted in two or more unrelated individuals with the same disease, or if the abnormality occurred *de novo* in a sporadic case (the affected individual has no family history of the disorder, and the abnormality is not present in either parent).

Chromosomal abnormalities occur rarely, so this approach can never be a general one for identifying disease genes. But it has been useful for some disorders, notably dominantly inherited disorders with a severe congenital phenotype (because they normally cannot reproduce, affected individuals occur as sporadic cases, making linkage analyses problematic). Metaphase or prometaphase chromosome preparations can be prepared readily from blood lymphocytes, and then stained with DNA-binding dyes to reveal altered chromosome banding patterns that indicate a chromosome abnormality such as a translocation, deletion, inversion, and so on ([Box 7.2](#) on p 204–5 gives the chromosome banding methodology).

Balanced translocations and inversions can be particularly helpful because, unlike large deletions, they may involve no net loss of DNA, and the underlying disease gene might be expected to be located at, or close to, a breakpoint that can readily be identified. See [Table 8.1](#) for some examples.

TABLE 8.1

EXAMPLES OF SUCCESSFUL GENE IDENTIFICATION PROMPTED BY THE IDENTIFICATION OF DISEASE-ASSOCIATED CHROMOSOMAL ABNORMALITIES

Disorder	Chromosome abnormality	Comments	PMID
Duchenne muscular dystrophy	an affected boy with a cytogenetically visible deletion at Xp21.3 and a woman with a balanced Xp21; 21 p12 translocation	positional cloning strategies identified genes within the deletion/translocation breakpoint, finally implicating the giant dystrophin gene	2993910; 3001530
Sotos syndrome	a girl with a <i>denovo</i> balanced translocation with breakpoints at 5q35 and 8q24.1	the disease gene, <i>NSD1</i> , was found to be severed by the 5q35 breakpoint	11896389

PMID, PubMed identifier of relevant publications, at <http://www.ncbi.nlm.nih.gov/pubmed/>—see glossary.

Genome-wide screens for large-scale duplications and deletions are also available through different modern methodologies involving DNA hybridization or DNA sequencing approaches. We will describe these methods in detail within the context of genetic testing in [Section 11.2](#).

Exome sequencing: let's not bother getting a position for disease genes!

Although many genes underlying monogenic disorders have been identified by the methods used above, some monogenic disorders have not been well studied because they are very rare, or because they are not readily identifiable. There may be difficulties, for example, in recognizing individual phenotypes within complex sets of overlapping phenotypes, such as intellectual disability. More than 7000 monogenic disorders are estimated to exist, and although the methods above have been very successful in gene identification, a new approach was needed to identify genes in the substantial proportion of monogenic disorders that are very rare or where there is difficulty in distinguishing the phenotype from related disorders.

The new approach? That would be massively parallel DNA sequencing (also called next-generation sequencing), which we introduced in [Section 3.3](#); details of two popular methods are given in [Chapter 11](#)). This new approach offers a vast increase in sequencing capacity, and the time taken to sequence a human genome has fallen from several years to a few days, or even a few hours. And the expense has dropped drastically. The inevitable consequence has been an explosion of genome sequencing.

The ability to sequence whole genomes (and therefore all genes) both rapidly and cheaply means that disease genes can often be identified without any need to first find a chromosomal position for the disease gene. Because the vast majority of disease genes are protein-coding genes and given that the great majority (perhaps 85 % or so) of currently known disease-causing mutations occur in the exons of protein-coding genes, whole-genome sequencing initially appeared a rather laborious and costly approach to identify a disease gene. However, all the exons of the protein-coding genes together account for only just over 30 Mb of our DNA; sequencing this fraction, just over 1 % of the genome, appeared an easier and cheaper option than genome sequencing.

Exome is the collective term for all exons in the genome. Operationally, exome sequencing has largely focused on the exons of protein-coding genes (RNA genes have been a low priority, mostly because they are viewed as very infrequent direct contributors to disease). Exome sequencing involves first capturing exons from the DNA of affected individuals, and then sequencing the captured DNA. In practice, exome capture is designed to capture exons with a little flanking intron sequence (to cover splice junctions) plus DNA sequences specifying some miRNAs; hybridization with a control set of cloned exon sequences allows capture of the desired exons, as shown in [Figure 8.7](#).

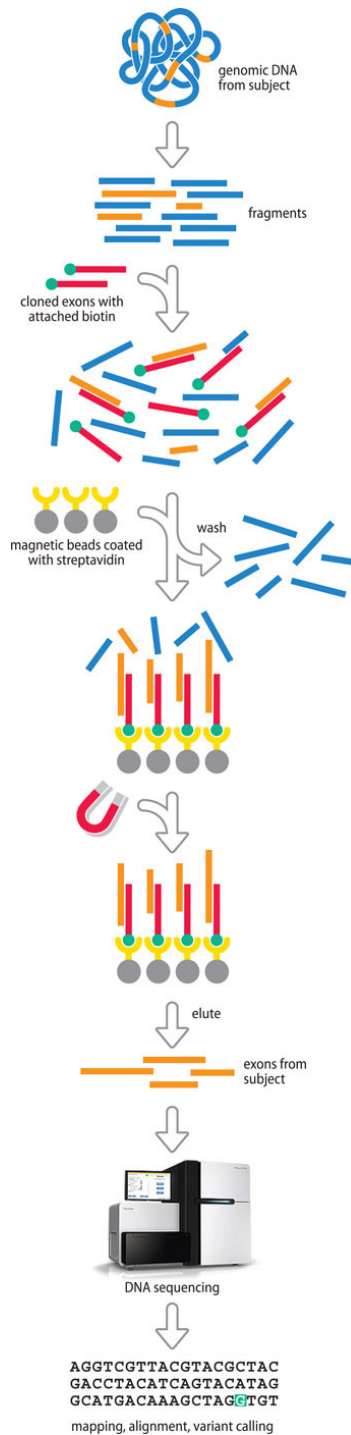


Figure 8.7 Exome capture and sequencing. A genomic DNA sample to be analyzed is randomly sheared and the fragments are used to construct a DNA library (library fragments are flanked by adapter oligonucleotide sequences; not shown). The DNA library is then enriched for exon sequences (orange rectangles) by hybridization to DNA or RNA baits that have been designed to represent human exon sequences (red rectangles). The baits have a biotin group (green circle) attached to their end. After capturing exon sequences from the test sample DNA by hybridization, the biotinylated baits can be selected by binding to magnetic

beads (gray) coated with streptavidin (yellow goblet shapes)—streptavidin has an extremely high affinity for binding to biotin, and the streptavidin–biotin–exon complexes can be removed using a magnet. The captured exon sequences from the sample DNA are subjected to massively parallel DNA sequencing and the data are interpreted as described in the text. (Adapted from [Bamshad MJ et al. \[2011\] Nat Rev Genet 12:745–755; PMID 21946919](#). With permission from Macmillan Publishers Ltd.)

To identify a gene for a rare disorder, sequencing of the exomes from just a few affected individuals is often sufficient. That is so because clearly deleterious mutations (frameshifting and nonsense mutations, and some types of nonconservative amino acid substitution) can often be easily identified in protein-coding sequences. Because we have some non-essential genes, each of us carries a surprising number of deleterious mutations like this, which are scattered throughout the genome and vary from individual to individual. However, unrelated individuals with the same single-gene disorder might be expected often to have causative mutations in the same gene. Where there is parental consanguinity, exome sequencing can sometimes be used to identify genes underlying an autosomal recessive condition by exome sequencing of affected members of a single family.

Since its first successful application in identifying disease genes in 2009, exome sequencing has been dramatically successful in identifying genes underlying very rare autosomal recessive and congenital dominant disorders (neither of which is amenable to linkage analyses because of a lack of suitable families); see [Table 8.2](#) for examples. It has also been important in the case of extremely heterogeneous phenotypes—in one early study, 50 novel genes were identified for recessive cognitive disorders ([Figure 8.8](#)). As sequencing costs drop, future studies may use whole genome sequencing.

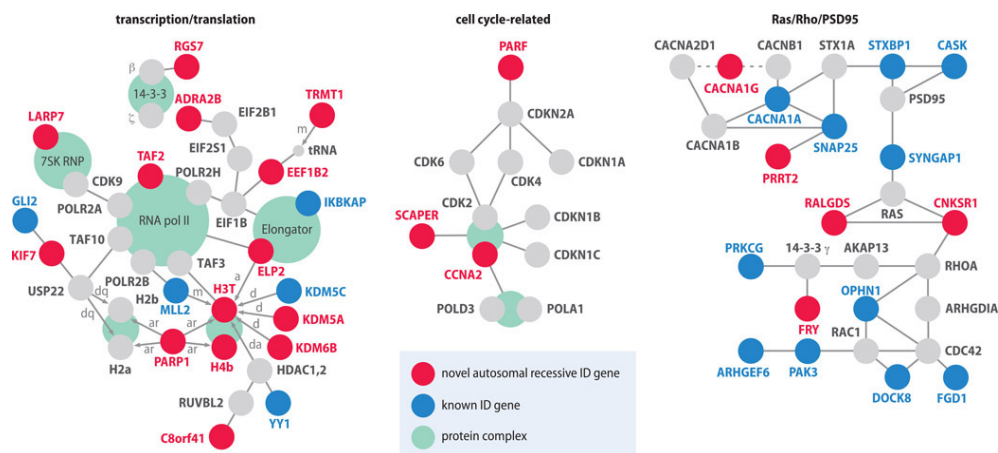


Figure 8.8 Known and novel genes for intellectual disability are part of protein and regulatory networks. In a single exome sequencing study, 50 novel genes were identified for recessive intellectual disability (ID). The novel genes were predicted to encode components of protein and regulatory networks that had been implicated by studies of known genes for intellectual disability, including the transcriptional/translational network, the cell cycle-related network, and the Ras/Rho/PSD95 network. (Adapted from [Najmabadi H et al. \[2011\] Nature 478:57–63; PMID 21937992](#). With permission from

TABLE 8.2**EXAMPLES OF SUCCESSFUL USE OF EXOME SEQUENCING TO IDENTIFY GENES CAUSING MONOGENIC DISORDERS**

Disease	Type of disorder	Origin of exomes	Mode of inheritance	OMIM number	Gene locus	PMID
Miller syndrome	congenital disorders of development	mostly unrelated sporadic cases	AR	263750	<i>DHODH</i>	19915526
Kabuki syndrome type1			AD	147920	<i>KMT2D (MLL2)</i>	20711175
Schinzle-Giedion syndrome			AD	269150	<i>SETBP1</i>	20436468
Osteogenesis imperfect type VI	connective tissue disorder	affected sibs born to consanguineous parents	AR	613982	<i>SERPINF1</i>	21353196
Spastic paraplegia 30	early-onset neuromuscular disorder		AR	610357	<i>KIF1A</i>	21487076

For a review of different strategies for exome sequencing to identify disease genes, and of successful applications, see Gilissen C et al. (2012) *Eur J Hum Genet* 20:490–497; PMID 22258526. AD, autosomal dominant; AR autosomal recessive. (PMID, PubMed identifier at <http://www.ncbi.nlm.nih.gov/pubmed/> as explained in the glossary.)

8.2 APPROACHES TO MAPPING AND IDENTIFYING GENETIC SUSCEPTIBILITY TO COMPLEX DISEASE

As outlined in [Section 8.1](#), various strategies, often initially using gene linkage analyses, have led to identification of numerous causative genes for Mendelian disorders. But Mendelian disorders are rare; complex common diseases that are so much more difficult to analyze represent the great bulk of genetic disease. Family-based genetic linkage analyses

have rarely been successful in these cases. Instead, there has been heavy reliance on mapping genetic susceptibility to complex disease by *association* analyses, notably genomewide association (GWA) studies. In this section we explain the background to these endeavors, the progress made, and the limitations of GWA studies. But first we provide some background on the complexity of common genetic disorders.

The polygenic and multifactorial nature of common genetic disorders

Investigations into monogenic disorders became a principal focus in medical genetics after the DNA cloning and sequencing revolutions began in the 1970s. We are easily accustomed to the idea of disease arising from causative mutation at a single gene locus. And up until now the kinds of genetic disease that we have described are single-gene disorders, mostly of neonatal or childhood onset, where genetic variants at a single gene locus are sufficient to cause disease, either in a nuclear gene (by Mendelian inheritance) or in a mitochondrial gene (by maternal inheritance).

To demonstrate Mendelian inheritance, a character must depend almost exclusively on the genotype at a single locus, virtually regardless of genotypes at every other locus and of the environment, history and lifestyle of the person bearing the trait. Monogenic disorders are inevitably therefore the rare exceptions and the way in which we view the inheritance mechanism for monogenic disorders is quite different for the great bulk of genetic disease.

The two quite different ways of viewing inheritance mechanisms

In the second half of the nineteenth century, two pioneering researchers independently laid down the foundations of modern genetics. Gregor Mendel focused on the independent combination of *discrete* characters (such as the famous yellow and green peas), while Francis Galton, Darwin's second cousin, used statistical principles to describe *continuous* phenotypes (height, weight, and so on).

Up until the present time the distinction of discrete versus continuous characters continues to reflect a fundamental dichotomy in the way we view mechanisms of inheritance. The dichotomy is somewhat reminiscent of the particle-wave duality of quantum physics, but unlike in quantum physics, the basic theory to reconcile the different Mendelian and Galtonian views of the mechanisms of inheritance was worked out relatively quickly. In the early twentieth century Ronald Fisher formulated the infinitesimal (= polygenic) model of inheritance. In this model, Fisher envisaged that continuous features were determined by an almost infinite number of Mendelian factors, and that the large variance among children of the same parents was due to the segregation of those factors that were heterozygous in parents.

Given the complex interactions between gene products in the biological pathways in our cells and the diverse interactions with environmental components, single-gene disorders simply have to be the rare exceptions. Common genetic disorders—such as type 1 and 2 diabetes, coronary artery disease, stroke, rheumatoid arthritis, Alzheimer disease, and so on—are *polygenic*. That is, genetic variants at multiple loci are important contributors to the phenotype. And environmental factors also may make very significant contributions to the phenotype, such as in the case of cigarette smoking for a range of common genetic diseases. Because of the genetic contributions of multiple genes plus environmental factors, common genetic diseases are said to be **multifactorial**.

The not quite clear-cut division between single-gene and multifactorial disorders

In reality, the division between single-gene disorders and multifactorial genetic disorders is not quite so clear cut. In virtually no case is the phenotype of a single-gene disorder entirely attributable to a single gene locus ([Figure 7.21](#) gives the example of b-thalassemia). We can see the effects of other loci when the phenotype in a “single-gene” disorder differs between affected members of the same family (as in [Figure 5.14](#)). And when single-gene knockouts are made in mice, the phenotype can vary very significantly in different strains of mice.

The difference in phenotype between affected family members with a single gene disorder, who would be expected to have the same disease allele(s), or in different strains of mice having identical gene knockouts, is due to having different alleles at one or more **modifier** loci. The product of a modifier gene interacts with the disease allele in some way: it may regulate expression of the disease allele, for example, or it may interact in the same pathway as the product of the disease allele, affecting its function (see [Section 7.9](#) for examples of modifier genes).

What distinguishes a single-gene disorder is that, although there may be minor effects from variants at other genetic loci, rare genetic variants at a primary gene locus have a very great effect on the phenotype. By contrast, it is usual to think of a polygenic disorder as being one in which the genetic susceptibility to disease risk is not dominated by a primary locus where individual variants can have extremely strong effects on the phenotype.

There may be a predominant genetic locus in a true polygenic disease, but its effect would not normally be large: variants at that locus would not be expected to be necessary or sufficient to cause disease. However, as described later, quite often problems arise out of heterogeneity: individual complex diseases can be a mix of related phenotypes, and in some of these diseases rare variants can be of quite strong effect.

The phenotypes of polygenic diseases are also influenced by nongenetic factors, including environmental factors (often working through epigenetic modification) and also chance (*stochastic*) factors. Environmental factors can sometimes have a strong influence in certain

monogenic disorders, but their effects are very important right across the spectrum of polygenic disorders.

Because a combination of multiple genetic susceptibility loci and environmental factors is involved, common genetic diseases have been considered to be complex or multifactorial diseases. Underlying polygenic theory is the idea that there is a *continuous susceptibility to the disease within the population*, and that disease manifests when a certain threshold is exceeded ([Box 8.2](#)).

BOX 8.2 POLYGENIC THEORY AND THE LIABILITY THRESHOLD CONCEPT TO EXPLAIN DICHOTOMOUS TRAITS

Human traits can be divided into two classes. Some, like diseases, are *dichotomous* (you either have the trait or you do not). Others, such as height or blood pressure, are *continuous* (or *quantitative*) traits—everybody has the trait, but to differing degrees. To explain quantitative traits, polygenic theory envisages the additive contributions of variable alleles from multiple loci. Many alleles at the underlying **quantitative trait loci (QTL)** might have subtle differences (causing modest changes in expression levels of certain genes, for example); different combinations of alleles at multiple loci might then produce continuous traits (adult height, for example, is known to be governed by genetic variants at a minimum of 180 loci).

Polygenic theory can be extended to also explain dichotomous traits using the concept of a *liability threshold*. The idea is that there is a continuous liability (susceptibility) to disease in the population, but only people whose susceptibility exceeds a certain threshold will develop disease. For each complex disease, the susceptibility curve for the general population will be a normal (bell-shaped) curve: most individuals will have a medium susceptibility, with smaller numbers of people having low to very low susceptibility or high to very high susceptibility ([Figure 1A](#)). Only a small percentage of the general population will have a susceptibility that exceeds the threshold so that they are affected by the disease (shown in red in [Figure 1](#)).

Close relatives of a person affected by a complex disease have an increased susceptibility to that disease ([Figure 1B](#)). The affected person must have a combination of different high-susceptibility genes; relatives will share a proportion of genes with the affected person and will therefore have an increased chance of having high-susceptibility genes.

By chance, some relatives may have only a few of the high-susceptibility genes, but others may have many high-susceptibility genes in common with the affected person. While disease susceptibility can show wide variation among first-degree relatives (parent and child, sibs), the overall effect is that the curve—and the median susceptibility (dashed vertical line)—is displaced to higher susceptibilities to the disease ([Figure 1B](#)). Because

the threshold remains the same, more individuals are affected, and the relatives that are most closely related to an affected person are more likely to be affected.

VARIABILITY IN THE LIABILITY THRESHOLD FOR AN INDIVIDUAL DISEASE

Thresholds of susceptibility to a complex disease are not absolutely fixed—they can show differences between the two sexes, for example. For many auto-immune disorders, women have significantly more disease risk than men. The reverse is true for some other conditions. For example, pyloric stenosis occurs in about 1 in 200 newborn males, but only in about 1 in 1000 newborn females. That is, a double threshold exists, one for females and one for males. The female threshold for pyloric stenosis is farther from the mean than that for the male. However, because it takes more deleterious genes to create an affected female, she has more genes to pass on to the next generation. Her male offspring are at a relatively high risk of being affected when compared with the population risk.

The threshold model accommodates environmental effects by postulating that such effects can reposition the threshold with respect to the genetic susceptibility. Protective environmental factors move the threshold to higher genetic susceptibilities; other environmental factors can increase the risk of disease by moving the threshold to lower genetic susceptibilities.

Complexities in disease risk prediction

For a common genetic disease, no single variant causes the disease. Instead, variants at multiple loci can act independently to increase or decrease the risk that a person will develop the disease. And non-genetic factors—environmental factors, lifestyle choices, and sometimes even chance—can make very significant contributions to the risk of disease.

For the reasons above, common genetic diseases do not give the typical dominant and recessive patterns seen in Mendelian diseases. Sometimes, however, a common genetic disease can show some evidence of running in families. A minority of families may have multiple affected members, and although rare, some families seem to show Mendelian or quasi-Mendelian transmission (such as in early-onset forms of Alzheimer disease, diabetes, Parkinson disease, and various types of cancer). Frequently, however, there is little evidence of family history; quite commonly an affected individual appears as a sporadic case.

Relatives share genes, and so may share variants predisposing to a common disease and may have a higher risk of developing the condition than the average risk across a population. The **relative risk** of developing a common genetic disease (also called the **risk ratio**) is the disease risk for a relative of an affected person divided by the disease risk for an unrelated

person in the general population and is denoted by the symbol l . Different values for l can be calculated for different degrees of relationship, such as l_s which expresses the relative risk for sibs (brother or sisters) of an affected person. [Table 8.3](#) gives illustrative values of l_s for certain monogenic and multifactorial conditions.

TABLE 8.3

EXAMPLES SHOWING CONTRASTING RELATIVE RISKS AND LIFETIME RISKS FOR MENDELIAN AND MULTIFACTORIAL DISORDERS

Disease	λ_s (relative risk)	Lifetime risk (to age 80)
Huntington disease	5000	0.01 %
Cystic fibrosis	500	0.05%
Alzheimer disease (l.o.) [*]	3	17%
Breast cancer, female	2	12%
Type 1 diabetes	15	1%
Type 2 diabetes	3	15%

Monogenic disorders are very rare. But if a person has an affected sibling their relative risk (λ_s) is very high, such as 5000 for Huntington disease (the 50% risk for a dominant disease is 5000 times greater than the incidence of 0.01 % in the general population). The lifetime risk for common diseases such as Alzheimer disease or breast cancer is high, and the relative risk is low.

^{*}

l.o. = late onset

The calculation of disease risk for complex diseases is therefore often quite different than for Mendelian disorders. In the latter case, disease risks are mostly based on theoretical calculations that remain quite stable. If a child is born to a couple with an autosomal recessive condition, for example, then one can normally assume that both parents are carriers and that all subsequent pregnancies carry a 1:4 risk of an affected fetus. (For some disorders, however, there can sometimes be complications, often because of low penetrance and variable expressivity.)

For some complex diseases, we are beginning to accumulate information on major predisposing genetic and environmental factors; this may ultimately lead to a more informed measure of disease risk than has previously been possible. Traditionally, a lack of knowledge

about the predisposing risk factors has meant that for complex diseases the disease risk has often been *empiric*: that is, it is based on observed outcomes in surveys of families and is often *modified according to past incidence of disease*. For example, a couple who have had one baby with a neural tube defect would be quoted a specific recurrence risk for a subsequent pregnancy that would be based on observed frequencies in the population; if they do go on to have a second affected child, the disease risk for a subsequent pregnancy would now be substantially higher. The real disease risk would not have changed; the birth of the second affected child helps us to recognize that the parents carry more high-susceptibility genes than was apparent after the birth of their first affected child.

Difficulties with lack of penetrance and phenotype classification in complex disease

Researchers seeking to identify the genetic susceptibility in multifactorial disorders are confronted by multiple challenges. In the sections below we cover some of the technological difficulties that had to be surmounted. Here we consider more intrinsic difficulties arising from the general lack of penetrance, or from problems with defining and classifying disease phenotypes.

Recall that reduced penetrance can be a feature of some monogenic disorders, such as imprinting disorders. But, in general, DNA variants at a Mendelian disease locus are of very strong effect and are therefore highly penetrant. As a result, affected people typically have a disease-associated genotype; unaffected people do not. For complex disease, however, the picture is different. If we discount rare Mendelian subsets of complex disease, reduced penetrance is the norm, simply because multiple genes make small contributions to the phenotype. That is, a DNA variant strongly associated with a complex disease is often at best a **susceptibility factor**: its overall frequency should be significantly increased in affected individuals compared with controls, but normal people can quite often possess the variant conferring disease susceptibility and affected people can quite often lack it. There is a clear contrast, therefore, between studying the genetics of monogenic disorders and multifactorial diseases: in the former we look for DNA variants that are fully responsible for causing the disease, and in the latter, we seek genetic susceptibility factors that predispose to disease.

Phenotype classification and phenocopies

In many complex diseases the phenotype can have many variable components. As detailed in [Section 5.3](#), monogenic disorders can also have variable phenotypes, but here the high penetrance and the range of phenotypic features in multiple affected individuals reported in the literature allow us to be clear about which aspects of a person's phenotype are disease

components, and which are not. In complex diseases, however, the situation is not so simple, and phenotype delineation and classification can be a major problem.

Some affected people, who do not have genotypes commonly associated with the disease, are **phenocopies** that have been wrongly classified as having the disease under study. For some phenocopies the disease is caused by different genetic factors than expected. For example, accurate diagnosis of Alzheimer disease has traditionally relied on *post-mortem* brain pathology. If we wish to study living patients, various clinical tests can be conducted, but a subsequent diagnosis of Alzheimer disease is often provisional (“*probably* Alzheimer disease”); post-mortem examination might subsequently show a different type of dementia, such as Lewy dementia, frontotemporal dementia, and so on. For other phenocopies, the phenotype might have an environmental origin.

According to the condition, defining the disease phenotype can be straightforward or very challenging. At one extreme are conditions in which there is a very recognizable and rather specific phenotype. For example, in primary biliary cirrhosis, the most common autoimmune form of chronic liver disease, affected individuals have markedly similar phenotypes, and have signature autoanti-bodies that are specifically directed against the E2 subunit of the mitochondrial pyruvate dehydrogenase complex. At another extreme are some behavioral and psychiatric conditions, in which even classifying individuals as affected and unaffected can sometimes not be straightforward. The pathology might not be well defined and there can be heavy reliance on interviews (and subjective information). Here, clear diagnostic criteria are of paramount importance.

Deciding which aspects of a person’s phenotype are components of a complex disease is not easy, and deciding how far we should lump together different, but clearly related, phenotypes is a significant issue in complex disease studies. If two first cousins have had different types of congenital heart malformation, for example, should we consider the phenotypes as independent occurrences, or lump them together and report two affected individuals in one family?

Estimating heritability: the contribution made by genetic factors to the variance of complex diseases

Phenotypes are determined both by genetic factors and by nongenetic factors (often described as environmental factors, but also comprising stochastic factors). The *variance* (V) of a phenotype is a statistical term defined as the square of the standard deviation. The total variance of a phenotype, $V_P = V_G + V_E$ (where V_G is the genetic variance and V_E is the environmental variance).

The genetic variance, V_G , is the sum of three components: (i) additive genetic effects (the combined contributions from different loci); (ii) dominance effects (interaction of alleles at a single locus); and (iii) interaction effects (interaction between genes at different loci; the effect of genes at several loci may not be simply additive if they interact with each other).

The **heritability** (h^2) is that proportion of the variance that can be attributed to genetic factors—that is, $h^2 = V_G/V_P$. Possible values for h^2 range from 0 (no genetic factors involved) to 1 (exclusively due to genetic factors). As described below, the ratio of genetic to environmental involvement in the etiology of a disease varies according to the disease class.

Despite its importance in the prediction of disease risk, heritability has often been misunderstood. First, it must be appreciated that heritability is not a fixed property. Secondly, it describes a population, not an individual. More accurately, it describes the genetic contribution to variance *within a population and in a specific environment at a certain time*. To estimate the heritability of a complex trait or disease, the incidence of disease is compared in genetically related individuals. To do this, three types of study have been undertaken: family studies, twin studies, and adoption studies.

Family studies

Having a relative with a complex disease increases your risk of developing that disease. The **risk ratio** (λ) is defined as the disease risk to a relative of an affected person divided by the disease risk to an unrelated person in the general population. The comparative disease risk for a sib (brother or sister) of an affected individual is designated as l_s . Unlike a monogenic disorder (for which the risks to family members are fairly precisely defined, according to simple theoretical calculations), the risks for complex diseases have quite often been empiric; that is, they have often been based on surveys of disease incidence in families.

Because l_s is a measure of how important genetic factors are in the etiology of the disease, extremely high l_s values are found in monogenic disorders ([Table 8.3](#)). For example, in populations of western European origin, the general lifetime risk of cystic fibrosis is about 1 in 2000, but the risk to a fetus is 1 in 4 if the parents have had a previously affected child. In that case, $l_s = 1/4$ divided by $1/2000 = 500$. For some complex diseases, l_s values can be quite high, such as 25 for Crohn's disease, 20 for multiple sclerosis and 15 for type 1 diabetes.

There is, however, a significant drawback in using family studies to infer genetic factors: family members would normally be exposed to some common environmental factors as a result of the shared family environment.

Twin studies

Twin studies measure how often the twins are concordant (both have the disease) or discordant (only one is affected). Monozygotic twins are genetically identical but dizygotic

twins, like any pair of sibs, share 50 % of their genes. Regardless of zygosity, twins should be exposed to rather similar environment factors.

TABLE 8.4 DEGREE OF CONCORDANCE BETWEEN TWIN PAIRS FOR SELECT COMPLEX DISEASES, AVERAGED FROM MANY STUDIES

Disease	Concordance(%)	
	In MZ twins	In DZ twins
Type 1 diabetes	42.9	7.4
Type 2 diabetes	34	16
Multiple sclerosis	25.3	5.4
Crohn’s disease	37	10
Ulcerative colitis	7	3
Alzheimer disease	32.2	8.7
Parkinson disease	15.5	11.1
Schizophrenia	40.8	5.3

MZ, monozygotic; DZ, dizygotic.

Table 8.4 lists observed concordance rates for monozygotic and dizygotic twins for various complex diseases. There are two major points to note. First, there are significant discordance rates between monozygotic twins—genetics is not everything! Secondly, it is clear that in certain diseases the concordance between monozygotic twins is much greater than that between dizygotic twins. Diseases in which genetic factors have a large role show a relatively high concordance in monozygotic twins and a much lower concordance rate in dizygotic twins: the greater this ratio, the greater is the genetic contribution. Thus, for

example, genetic factors would seem to be much more important in schizophrenia than in Parkinson disease.

The best type of study would seek to disentangle the contributions of genetic and environmental factors by tracking monozygotic twins separated at birth and raised in different environments, but this occurs so rarely that statistically valid sample sizes cannot be obtained.

Adoption studies

Adoption studies offer the best practical way of separating out the contributions made by genetic and environmental factors. Two types of study can achieve this. One type of study investigates children who were born to affected individuals but were then adopted at birth

into an unaffected family (the children might be expected to share genetic factors with their biological parents but be exposed to different environments).

A second type of adoption study starts with adopted people who suffer from a specific common genetic disease that can run in families. The object is to seek evidence for the disease running in the biological family or their adoptive family. A celebrated example is a Danish study of 14 427 adopted people aged 20–40 years, 47 of whom were diagnosed with chronic schizophrenia. The 47 schizophrenia cases were extensively matched (for age, sex, social status of adoptive family, and number of years in institutional care before adoption) with 47 control non-schizophrenic adoptees from the 14 427 adoptees. The results showed quite clearly that genes rather than the family environment increased the risk for the offspring. Interested readers can find the data in the publication by Kety SS et al. (1994) *Arch Gen Psychiatr* 51:442–455; PMID 8192547.

Adoption studies are the gold standard for teasing out the separate contributions of genetic factors and environmental factors, but they are difficult to carry out. Often, little is often known about the biological family, and intrusive approaches to the biological family may be undesirable. Efficient adoption registers may be lacking in many countries, and, in the interests of the child, adoption registers tend to choose a foster family resembling the biological family. Thus far, adoption studies have largely been carried out for psychiatric conditions (where there has been great interest in the contributions of nature and nurture to disease development).

Variation in the genetic contribution to disorders

Heritability studies have indicated that genetic factors make different contributions to different diseases. Monogenic disorders are primarily determined by genetic factors (but in some cases environmental factors can make very significant contributions). By contrast, infectious diseases are primarily determined by environmental factors (exposure to the infectious agent). Here, however, host genetic factors also make a contribution, so that people vary in their susceptibility to an infectious disease and a small number of individuals may be disease-resistant.

For the great bulk of other complex diseases, genetic and environmental factors both make large contributions to the phenotype. In some complex diseases, such as schizophrenia, autism spectrum disorder, Alzheimer disease, type 1 diabetes, multiple sclerosis, and Crohn's disease, there is a strong genetic contribution; in others, such as Parkinson disease and type 2 diabetes, genetic factors seem to be less important.

The heritability of an individual disease varies from one population to another. It can also vary in the same population in response to a changing environment. Consider phenylketonuria. As detailed in Clinical Box 9 on page 234, a deficiency of phenylalanine hydroxylase produces elevated phenylalanine and toxic by-products that can result in

cognitive disability. In the recent past the disease was almost wholly due to genetic factors, and so the heritability was extremely high. In modern times, neonatal screening programs in many countries allow early detection and treatment using low-phenylalanine diets. Now, in societies with advanced health care, phenylketonuria results mostly from environmental factors that lead to failure to deliver the treatment (inefficiency in health care systems, reluctance of families to seek out treatment, non-compliance with the diet, and so on).

The incidence of some complex diseases is very dependent on environments that can change very significantly with time. Thus, the huge recent increase in type 2 diabetes (mostly as a result of increasingly unhealthy diets and lack of exercise) means that the heritability of this disorder in many populations is now much reduced when compared with just a few decades ago.

The very limited success of linkage analyses in identifying genes underlying complex genetic diseases

The linkage analyses used to map genes for monogenic disorders are said to be *parametric*—the data can be analyzed only if a specific genetic model is assumed. The model needs to give details of certain key parameters: the mode of inheritance, disease gene frequency, and penetrance of disease genotypes. Parametric linkage analyses have been very successful in mapping genes for Mendelian disorders, but they have had very limited applicability in complex disease (because of the general difficulty in providing all the required parameters). They have, however, been of use when applied in two ways: (i) in analysing rare Mendelian subsets of complex diseases; (ii) when using certain non-parametric linkage analyses, usually to study affected sibs from large numbers of families.

Parametric linkage analyses in Mendelian subsets

Parametric linkage analyses are more readily applied when a complex disease shows very strong familial clustering. A near-Mendelian pattern may simply reflect the chance occurrence of several affected people in one family. If, by chance, most members of a family possess multiple genetic variants conferring high susceptibility to disease (not necessarily at the same loci in different affected family members), a single common disease-susceptibility allele (that has been transmitted in a Mendelian fashion within that family) might be enough to tip the balance past the susceptibility threshold.

For several complex diseases there are also subsets in which the disorder shows clear Mendelian inheritance. That is most obvious when there is dominant inheritance. In early-onset Alzheimer disease, for example, some large pedigrees, such as the one shown in [Figure 8.9](#), permitted identification of three

disease genes: the amyloid precursor protein gene (*APP*), and two presenilin genes involved in processing the APP protein. In some other complex diseases, such as Parkinson disease, a number of autosomal recessive pedigrees have led to gene identification. We detail the genes identified through Mendelian subsets of Alzheimer and Parkinson disease in [Section 8.3](#).

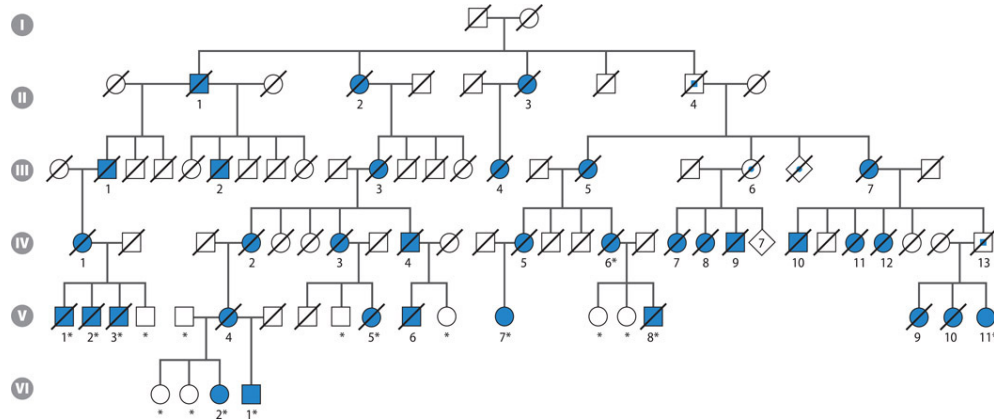


Figure 8.9 An exceptional pedigree showing dominantly inherited Alzheimer disease. Affected members of this pedigree had early-onset Alzheimer disease (average age at onset 46 ± 3.5 years). They were subsequently shown to have a mutation in the presenilin 1 gene, *PSEN1*. (From Campion D et al. [1995] *Neurology* 45:80–85; PMID 7824141. With permission from Wolters Kluwer Health.)

How might the phenotypes of a complex disease and one that segregates like a Mendelian disorder be so similar? The gene mutated in the Mendelian subset might also be a disease-susceptibility locus for the complex disease (with a rare, highly penetrant variant in the Mendelian subset, and a common variant of weak effect at the same locus in the complex disease). Or, if genes mutated in Mendelian subsets are not significant disease-susceptibility loci for a complex disease, the common pathogenesis might suggest that at least different genes associated with the two forms of the disease are part of a common biological pathway or process.

Affected sib-pair and other nonparametric linkage analyses

Nonparametric methods of linkage analysis do not require any genetic model to be stipulated, and so can generally be applied to analyzing complex disease. They rely on the principle that, regardless of the mode of inheritance, affected individuals in the same family would tend to share not just major disease-susceptibility genes but also the immediate chromosomal regions. That is, a major disease-susceptibility locus and a very closely linked marker would show a strong tendency to be co-inherited within affected individuals in the same family (because of the very low chance of recombination between the marker locus and the disease locus).

Nonparametric linkage studies occasionally use samples from all affected family members, but it is usually more convenient to simply use affected sib pairs. The aim here is to obtain genome-wide marker data in affected sibs from multiple families and then identify chromosomal regions that have been shared more often than would be predicted by random Mendelian segregation. As sibs share 50 % of their genes, affected sibs need to be studied in many families with the same complex disease. For marker loci that are not linked to a major disease susceptibility gene, sibs would be expected to share 50 % of alleles on average (some sib pairs might share 2, 1, or 0 alleles by chance, but the *overall average* across all sets of sibs would be 1 allele in common). For marker loci close to a major disease susceptibility gene, affected sibs would be expected to share significantly more than 50 % of alleles (see [Figure 8.10](#) for the principle).

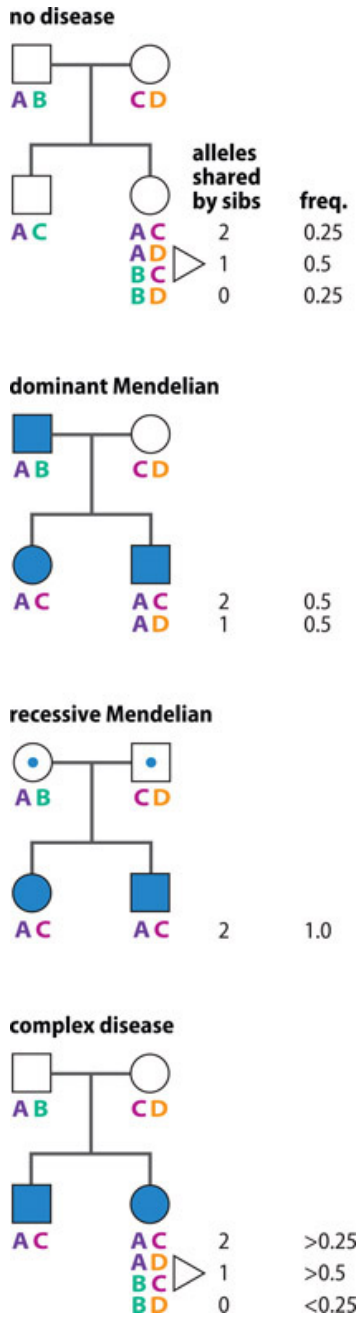


Figure 8.10 Principle of affected sib-pair analysis. By random segregation, any pair of sibs share 2, 1, or 0 parental haplotypes in the relative proportions 1:2:1. Pairs of sibs who are both affected by a dominant Mendelian condition must share the segment that carries the disease allele, and they may or may not (a 50:50 chance) share a haplotype from the unaffected parent. Pairs of sibs who are both affected by a recessive Mendelian condition necessarily share the same two parental haplotypes for the relevant chromosomal segment. For complex conditions, haplotype sharing greater than that expected to occur by chance may allow the identification of chromosomal segments containing susceptibility genes.

Affected sib-pair analyses are comparatively easy to carry out (they need samples from just a few people per family) and robust (the method makes few assumptions). But there are inevitable limitations. Because any individual susceptibility factor is neither necessary nor sufficient for a person to develop a complex disease, the underlying genetic hypothesis is weaker than for Mendelian conditions. This means that finding statistically significant evidence for a disease susceptibility factor is going to be harder.

The calculations in [Table 8.5](#) show that under ideal conditions affected sib-pair analyses can be carried out with reasonable numbers of samples (typical studies use a few hundred sib pairs). But if the effects are weak, unfeasibly large numbers of samples are needed to detect them, and the studies will be defeated if there is a high degree of heterogeneity (individual susceptibility factors will operate in just a small proportion of families).

TABLE 8.5 NUMBER OF AFFECTED SIB PAIRS NEEDED TO DETECT A DISEASE SUSCEPTIBILITY FACTOR				
Relative risk of disease (γ)	Probability of allele sharing by affected sibs (γ)		Number of affected sib pairs needed to detect the effect*	
	At $p = 0.1$	At $p = 0.01$	At $p = 0.1$	At $p = 0.01$
1.5	0.505	0.501	115481	7868358
2.0	0.518	0.502	9162	505272
2.5	0.536	0.505	2328	103007
3.0	0.556	0.509	952	33 780
4.0	0.597	0.520	313	7253
5.0	0.634	0.534	161	2529

Here p is the frequency of the disease susceptibility allele and q is the frequency of normal alleles at the disease susceptibility locus (so that $p+q = 1$). The relative risk of disease (γ) is a measure of how the disease risk changes when comparing persons with the susceptibility factor to those without. The calculations are based on the formulae derived by Risch & Merikangas in their [1996](#) paper (see Further Reading).

* Really the number of affected sib pairs with 80 % power to detect the effect. The take-home message is that unless the disease susceptibility factor is both quite common and confers a high disease risk, very many affected sib pairs are required to detect it.

Genome-wide nonparametric linkage scans require higher thresholds for statistical significance. Lod scores above 5.4 are considered highly significant evidence for linkage; scores between 3.6 and 5.4 are significant; and scores from 2.2 to 3.6 are suggestive. In practice, affected sib-pair analyses deliver typically modest lod scores that often do not reach statistically significant thresholds.

The review by [Altmüller et al. \(2001\)](#) under Further Reading reports an analysis of different nonparametric linkage studies for many diseases and underscores

the difficulties. Of 10 genome-wide linkage studies of schizophrenia analyzed in that review, four were unable to find any evidence of linkage, five found only suggestive evidence of linkage (lod scores between 2.2 and 3.5, but at many different regions on eight different chromosomes), and only one recorded more significant evidence of linkage. In addition to the lack of any great consistency in the results, getting independent replication of significant results proved very difficult. But in some cases, such as an 8p21 location, the initial finding was replicated in multiple populations. That finding eventually led to investigation of alleles of the neuregulin gene, *NRG1*, as a risk factor (the gene had been mapped to 8p21 and was known to be involved in synaptic transmission).

Identifying the disease-susceptibility gene

Even if a significant candidate chromosome region can be identified for a complex disease susceptibility locus, finding the implicated gene is often problematic in the absence of clues that suggest a candidate gene. That is so because sibs share large chromosome segments, and so candidate chromosome regions are generally very large (in Mendelian disorders, by contrast, the candidate chromosomal region can be progressively reduced by looking for rare recombinants between marker loci and the disease locus). To get closer to the disease susceptibility gene additional *linkage disequilibrium* mapping methods have sometimes been used. The same methods are regularly used in association analyses, and we will consider these in detail in the next section.

Despite the above difficulties, genome-wide linkage studies have had a measure of success in mapping susceptibility genes in complex disease to specific candidate regions that would then allow subsequent gene identification using other approaches. In addition to the schizophrenia-associated *NRG1* allele mentioned above, successes include mapping of genes conferring susceptibility to age-related macular degeneration to the 1q32 region, and genes conferring susceptibility to Crohn's disease at the 16q11-16q12 region. These advances allowed subsequent *association analyses* to be targeted to these regions as described below, ultimately allowing identification of the *CFH* (complement factor H) gene at 1q32 and the *NOD2* gene at 16q12 as novel disease-susceptibility factors.

The *CFH* gene had previously been well known, but the gene that came to be known as *NOD2* was identified only very shortly before being implicated in Crohn's disease; it would provide the first molecular insights into the pathogenesis of this disease. In Crohn's disease an abnormal immune response is directed

against various *nonself* antigens in the gut, including harmless (and often beneficial) commensal bacteria; the resulting accumulation of white blood cells in the lining of the intestines produces chronic inflammation.

The *NOD2* gene was finally implicated in Crohn's disease by identifying three comparatively common variants: two missense mutations and, notably, a frame-shift mutation that occurs near the end of the coding sequence and has a weak effect ([Figure 8.11](#)). In one survey 50% of 453 European patients were reported to have presumptive pathogenic mutations in the *NOD2* gene. The three common mutations accounted for 81 % of the mutations; homozygotes or compound heterozygotes for these mutations are not uncommon in Crohn's disease, but are very rare in the normal population. A heterogeneous set of rare missense mutations were suggested to account for the remaining causal variants.

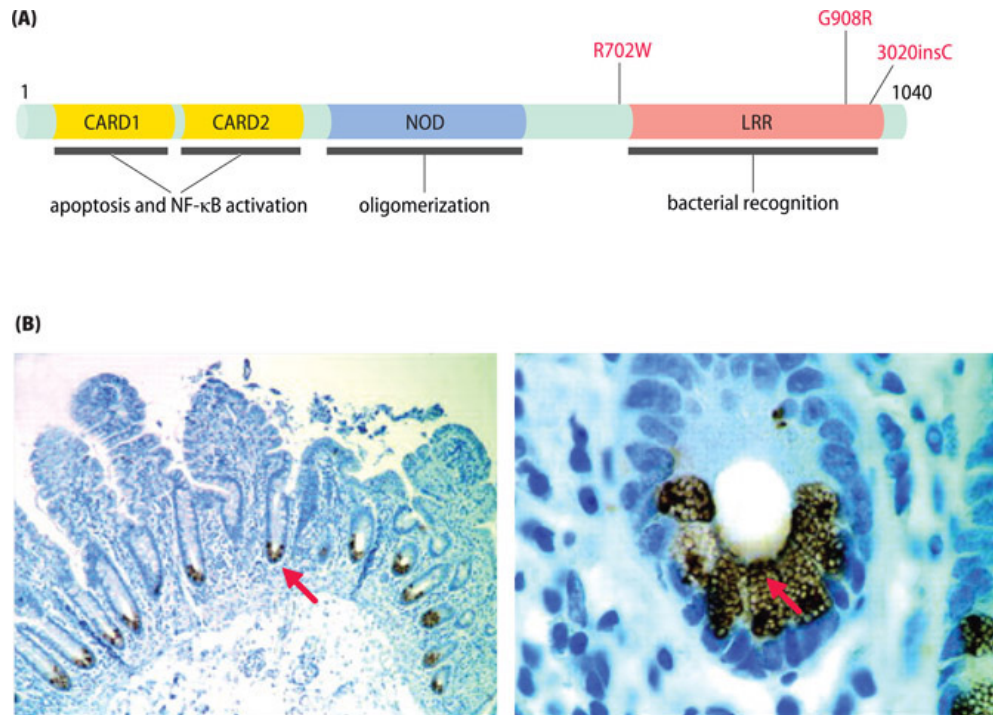


Figure 8.11 The Crohn's disease susceptibility factor *NOD2*: common variants and expression in Paneth cells. (A) Domain structure of the 1040-residue *NOD2* protein and corresponding location of common variants associated with Crohn's disease (in red). The 3020insC variant appears to be a mild frameshift mutation; it inserts a cytosine, causing a stop codon to be introduced at the next codon position (codon 1008), eliminating just the final 33 amino acids. Like 3020insC, the missense mutations G908R and R702W are located within or close to the LRR domain. Domains: CARD1, CARD2, caspase-activating recruitment domains; NOD, nucleotide-binding oligomerization domain; LRR, leucine-rich repeats domain. The LRR domain is now known to bind to specific breakdown products of peptidoglycan, a major component of bacterial cell walls. (B) The *NOD2* protein is predominantly expressed in Paneth cells, specialized secretory epithelial cells found at the base of intestinal crypts (arrows show examples of staining with a specific anti-*NOD2* antibody). Paneth cells secrete certain anti-microbial peptides, notably α -defensins. (B, From Ogura Y et al. [2003] *Gut* 52:1591–1597; PMID 14570728. With permission from the BMJ Publishing Group Ltd.)

The *NOD2* protein is now known to be part of the **innate immune system** (which produces the initial non-specific immune responses against pathogens). It has a C-terminal

domain that recognizes a specific peptide motif found in a wide variety of bacterial proteins (see [Figure 8.11](#)). The three common DNA variants in [Figure 8.11A](#) seem to be partial loss-of-function mutations that impair the ability of NOD2 to recognize bacterial protein.

Gut flora (*microbiota*) include many microbes that are of active benefit to us. They help us derive additional energy through the fermentation of undigested carbohydrates, help us break down xenobiotics, and synthesize vitamins (B and K) for us. Although they are foreign microorganisms they are therefore tolerated by suppressing the usual innate immune responses. NOD2 works in this area by down-regulating those innate immune responses that require the Toll-like receptors; when NOD2 is impaired, a strong immune response is launched in response to the gut flora, causing inflammation.

The fundamentals of allelic association and the importance of HLA-disease associations

Linkage analyses inherently have very limited power to detect susceptibility factors in complex disease. They have been profitably used in rare Mendelian subsets to detect susceptibility factors with sizeable effects, as in the case of early-onset Alzheimer disease (the amyloid precursor protein gene *APP* and the presenilin genes *PSEN1*, *PSEN2*); breast cancer (*BRCA1*, *BRCA2*); colon cancer (*APC*), and maturity onset type 2 diabetes (*GSK*, the glucokinase gene). When the alternative of affected sib pair analyses have been employed to get to a subchromosomal location, allelic association studies have often then been used to get to the gene (such as in helping to identify *NOD2* variants as susceptibility factors for Crohn's disease).

From the mid-2000s onwards, powerful genomewide association analyses supplanted linkage analyses as the route to identifying genetic factors in complex disease. We describe below how these are conducted, but first let us look into what exactly allelic association means.

The nature of allelic association

Whereas linkage is a *genetic* phenomenon, **association** is essentially a *statistical* property that is simply concerned with unexpected frequencies for the cooccurrence of alleles (and/or phenotypes) in individuals within a population. [Figure 8.12](#) gives a summary of the essential differences between genetic linkage and allelic association.

LINKAGE		ASSOCIATION
a specifically genetic relationship	NATURE	a statistical relationship based on a frequency of co-occurrence that deviates from random expectation
a relationship between loci ; needs comparison of family members	IN GENETICS	a population -based relationship between alleles or phenotypes
a marker locus and a disease locus are linked if they lie close together on chromosome, such as the HLA-B locus and the CYP21A2 (21-hydroxylase) locus	AN EXAMPLE	the HLA-B27 allele is associated with ankylosing spondylitis. In the UK >90% of patients exhibit it, but only about 8% of the general population

Figure 8.12 Genetic linkage and association compared.

In a population under study, if allele A^*1 at locus A is found to be significantly more frequent in people affected by a specific complex disease D than would be expected (from the individual population frequencies of A^*1 and the D gene), we would say that allele A^*1 is positively associated with disease, a disease-susceptibility allele. Conversely, if significantly less frequent in affected individuals, A^*1 would be a disease-resistance allele (negatively associated with disease).

Unlike linkage, the association of alleles in a population is not intrinsically genetic. There can be several ways in which the association can be explained, not all of which are genetic. Four possibilities are listed below:

- *Direct causation.* Simply by having allele A^*1 , a person is more susceptible to disease D . Somehow A^*1 confers an increased risk in the *population* of developing the disease (but A^*1 may not be necessary or sufficient for someone to develop the disease).
- *Linkage disequilibrium.* Allele A^*1 is not directly involved in pathogenesis but is nevertheless positively associated with disease. That can happen if allele A^*1 is located very close on the chromosome to a true susceptibility factor locus B where, for example, allele B^*2 is a high risk allele. Affected individuals would tend to have haplotype A^*1 - B^*2 ; although both A^*1 and B^*2 would be positively associated with disease, allele B^*2 is the allele contributing to pathogenesis. We explain linkage disequilibrium more fully below.
- *Epistasis.* People who have disease D plus a high-risk allele A^*1 may be more likely to survive and have children if they also have allele M^*1 at *modifier* locus M . If so, M^*1 might also appear to be associated with disease. The modifier locus might make a product that interacts with other gene products in a pathogenetic pathway for disease D .
- *Population stratification.* A population happens to have various genetically distinct subpopulations and both disease D and allele A^*1 might just happen to be common in one subpopulation, whereupon allele A^*1 appears to be associated with disease. Eric Lander and Nicholas Schork gave the light-hearted example of association of

the *HLA*A1* allele and the “trait” of being able to eat with chopsticks (*HLA*A1* is more frequent in Chinese than in Caucasians).

Candidate gene association analyses, case-control studies, and odds ratios

Association studies have a long history, beginning long before the molecular genetics revolution, at a time when certain protein polymorphisms were commonly used, including alleles of the ABO and other blood group systems, and especially HLA polymorphisms. Because of their extremely high polymorphism, numerous alleles could be identified at each classical HLA gene locus using panels of antisera (that is, *serological* typing was performed instead of DNA typing).

In those early days, there was no possibility of carrying out genome-wide analyses to hunt for markers that showed association with a specific disease. That was so because unlike genetic linkage, which works over long ranges in DNA molecules, genetic association works over very short distances only. Genome-wide linkage analyses require just a few hundred markers distributed across the genome, but genomewide association analyses typically need many hundreds of thousands of markers to find a marker allele that is both associated with—and very tightly linked to—the disease-susceptibility allele.

Instead, in the absence of abundant DNA polymorphisms, there was simply the possibility of candidate gene studies: testing individual protein polymorphisms in turn to see if any showed significant evidence of association with a specific disease. Because of their key importance in T-cell function and cell-mediated immunity, HLA alleles were suspected to show disease associations and were comprehensively investigated using candidate gene association studies.

To investigate any disease associations, **case-control studies** are carried out in which genetic variants are typed in affected individuals (cases) and in controls from the population under study. In the smaller studies, such as the early HLA association studies, controls were selected to be individuals *known* to be unaffected, but in large-scale association analyses it is more convenient to use general population-based controls, for whom the disease status is simply unknown (with the proviso that the disease of interest is not too common in the population under study).

Different methods can be used to measure the disease risk for each tested genetic variant. One popular method is the **odds ratio**; that is, the odds of being affected when possessing a specific genetic variant divided by the odds of being affected when lacking the genetic variant (see [Table 8.6](#) for a worked example).

TABLE 8.6

A WORKED EXAMPLE OF THE ODDS RATIO IN CASE-CONTROL STUDIES

HLA-Cw6 status	Number of cases (with psoriasis)	Number of unaffected controls	Odds of being affected		Odds ratio
Present	900	330	900/330	→	(900/330) ÷ (100/670) =
Absent	100	670	100/670		(900/330) × (670/100) = 18.27

The odds ratio is the odds of being affected when possessing a specific genetic variant divided by the odds of being affected when lacking the genetic variant. In this entirely hypothetical example, we imagine a case-control study of psoriasis in which 1000 affected individuals (cases) and 1000 unaffected controls have been typed for the HLA-Cw6 marker, giving the calculation in the final column.

As described in [Box 8.3](#) and in [Section 8.3](#), certain HLA variants have been identified to be the largest genetic contributors to a wide variety of important autoimmune disorders, and the small HLA region, just over 3 Mb long, has by some distance, the highest density of significant disease associations in our genome.

BOX 8.3 HLA ASSOCIATIONS WITH AUTOIMMUNE DISORDERS

The human major histocompatibility complex (MHC; also called the HLA complex) extends over 3.3 Mb at 6p21.3. It contains many genes that function in the immune system, notably HLA genes. Some of the HLA genes are *classical* HLA genes that make highly polymorphic cell surface proteins involved in cell-mediated immune responses ([Box 4.3](#) on page 105–6 gives HLA nomenclature and a simplified HLA gene map). Classical HLA genes work in cell-mediated immunity to signal the presence of cells infected by viruses (or other intracellular pathogens) to suitably discriminating T cells, thereby initiating an immune response to kill the infected cells.

All proteins within our cells (whether of normal host origin or from intracellular pathogens such as viruses) undergo turnover, whereby the proteins are degraded to peptides within the proteasome. The resulting peptides are bound by newly synthesized HLA proteins and are then transported to the cell surface so that the HLA–peptide complex can be recognized by a specific T-cell receptor on the surface of T cells ([Figure 1A](#)).

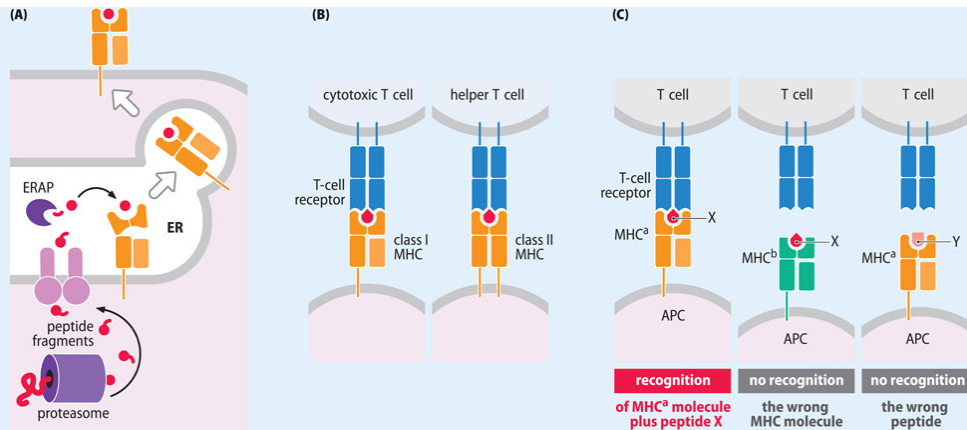


Figure 1 MHC-peptide binding and MHC restriction in antigen presentation. (A) Class I MHC proteins (class I HLA proteins in human cells) serve to bind peptides and display them on the cell surface. The peptides are produced by the degradation of any protein synthesized within the cell (a host cell protein or one made by an intracellular pathogen). Peptide fragments are produced within the proteasome and transported into the endoplasmic reticulum. Here they are snipped by an endoplasmic reticulum aminopeptidase (ERAP) to the proper size needed for loading on to a partly unfolded class I HLA protein. Once the peptide has been bound, the HLA protein completes its folding and is transported to the plasma membrane with the bound peptide displayed on the outside. (B) Receptors on cytotoxic T cells bind class I MHC–peptide complexes; those on helper T cells bind class II MHC–peptide complexes. (C) **MHC restriction.** Each T cell has a cell-specific receptor whose function is dependent on co-recognition of two molecules on the surface of an antigen presenting cell (APC): the combination of a *specific* peptide bound by a *specific* MHC protein. The T cell illustrated here is specific for a particular peptide X bound to a specific MHC allele (an imaginary allele that we designate here as MHC^a as shown on the left). If instead the APC had peptide Y bound to MHC^a, or peptide X bound to MHC^b, it would not be recognized by the same T cell (Adapted from Murphy K [2011] *Janeway’s Immunobiology*, 8th ed. Garland Science.)

Immune tolerance ensures that self-peptides (originating from normal host proteins) do not normally trigger an immune response. At an early stage in thymus development, T cells with receptors that recognize self-peptides bound to HLA are eliminated; thereafter T cells are normally focused on nonself (“foreign”) peptides (such as those from pathogens). Different T cells in a person contain different T-cell receptors to maximize the chance that a nonself peptide presented by an HLA protein can be recognized. When that happens, T cells are activated to mount an immune response (see [Figure 1B,C](#)).

Viruses readily mutate in an attempt to avoid triggering immune responses, and the number of potential foreign peptides is huge. This explains why T-cell receptors

are genetically programmed, like antibodies, to be extraordinarily diverse (detailed in Section 4.5). HLA proteins vary in their ability to present specific peptides for recognition and so they, too, are selected to be highly polymorphic.

In autoimmune disorders there is a breakdown in the ability to distinguish self from nonself. As a result, cells in the body can be attacked by *autoantibodies* and by autoreactive T cells that inappropriately recognize certain host antigens (autoantigens). In diseases such as type 1 diabetes, rheumatoid arthritis, and multiple sclerosis, activated T cells kill certain populations of host cells (such as insulin-producing pancreatic beta cells in type 1 diabetes). In autoreactive T-cell responses, host peptides (autoantigens) are presented by HLA proteins that may differ in their ability to bind the autoantigen. As a result, specific HLA antigens are associated with disease.

At the classical HLA loci, large numbers of alleles can be typed (previously, as serological polymorphisms by using panels of antisera; more recently as DNA variants). HLA–disease association studies involve typing HLA gene variants in affected individuals and controls and calculating the frequencies of a specific antigen or DNA variant in the two groups. This allows calculation of the odds of a disease occurring in individuals with or without a particular genetic variant, and calculation of odds ratios ([Table 1](#)).

TABLE 1

EXAMPLES OF ^{xs}HLA DISEASE ASSOCIATIONS IN THE NORWEGIAN POPULATION

Disorder	Class of HLA antigen	HLA antigen frequency		Odds ratio [*]
		Affecteds	Controls	
Ankylosing spondylitis	HLA-B27	>0.95	0.09	69.1
Celiac disease	HLA-DQ2and-DQ8	0.95	0.28	15.4
Multiple sclerosis	HLA-DQ6	0.59	0.26	4.1
Narcolepsy	HLA-DQ6	>0.95	0.33	129.8
Psoriasis	HLA-Cw6	0.87	0.33	13.3
Rheumatoid arthritis	HLA-DR4	0.81	0.33	3.8

All of the associations shown here indicate disease risk, except for the negative association of HLA-DQ6 and diabetes (with an odds ratio<1).That is, HLA-DQ6 is a protective allele: carriers have less risk of type 1 diabetes than the general population. (HLA antigen frequency courtesy of ErikThorsby.)

^{*}

Table 8.6 shows how odds ratios are calculated.

Disorder	Class of HLA antigen	HLA antigen frequency		Odds ratio *
		Affecteds	Controls	
Type 1 diabetes	HLA-DQ8and-DQ2	0.81	0.23	9.0
	HLA-DQ6	<0.1	0.33	0.22

All of the associations shown here indicate disease risk, except for the negative association of HLA-DQ6 and diabetes (with an odds ratio<1).That is, HLA-DQ6 is a protective allele: carriers have less risk of type 1 diabetes than the general population. (HLA antigen frequency courtesy of ErikThorsby.)

*
—

Table 8.6 shows how odds ratios are calculated.

From [Table 1](#) it is clear that possession of certain HLA antigens confers a substantially increased risk for certain disorders, and the odds ratios can be very impressive. If, for example, you carry an HLA-B27 antigen, you have a much-increased risk of developing ankylosing spondylitis (a form of inflammatory arthritis affecting the joints of the lower back). But HLA-B27 is merely a *susceptibility* factor: although the odds ratio approaches 70, only 1–5 % of individuals with HLA-B27 develop ankylosing spondylitis.

Linkage disequilibrium as the basis of allelic associations

Associations between genetic variants and disease can be caused by different factors, both genetic and nongenetic. In the former case, population substructure and history are important. As previously considered in [Section 5.4](#), human populations within countries, regions, and cities are often stratified into different groups (organized along ethnic, cultural, and religious lines) whose members preferentially mate within the group rather than with members of another group. As a result of population *stratification*, different subgroups within a broad population often have significantly different frequencies of a genetic variant, and this can confound genetic analyses. To minimize problems arising from population stratification, association studies need controls with the same type of population ancestry as those with the trait being studied.

Genetic variants associated with disease might be directly involved in pathogenesis, or they may be tightly linked to a disease-susceptibility allele. In the latter case the haplotype containing the genetic variant and the disease susceptibility allele has a higher frequency than would be predicted from the individual frequencies of the genetic variant and susceptibility allele. This is an example of **linkage disequilibrium**, the nonrandom association of alleles at two or more loci.

As a concept, linkage disequilibrium describes *any* nonrandom association of alleles at different loci; in practice, the alleles are at very closely linked loci. For example, in populations originating in northern Europe linkage disequilibrium is often evident for alleles at closely neighboring HLA genes, such as the *HLA-A* and *HLA-B* loci which are separated by 1.3 Mb of DNA at 6p22. Thus, in the population of Denmark the frequencies of *HLA-A1* and *HLA-B8* are 0.311 and 0.237, respectively, but the frequency of the *HLA-A1–HLA-B8* haplotype is 0.191, more than 2.5 times the expected value of 0.074 ($= 0.311 \times 0.237$).

Linkage disequilibrium might occur if a particular combination of alleles at neighboring loci were positively selected because they worked together to confer some advantage. However, linkage disequilibrium may often simply reflect reduced recombination between loci. This can happen in areas of the genome where there are low recombination rates. We now know, for example, that the HLA complex, the human major histocompatibility complex, is a region of low recombination (with 0.49 cM per Mb of DNA, compared to a genome wide average of 0.92 cM/Mb).

When a new DNA variant emerges by mutation it will show very tight linkage disequilibrium with alleles at very closely linked loci. The linkage disequilibrium will be gradually eroded by recombination, but that will take a very long time for any locus that is physically very close to the locus with the new mutation.

Sharing of ancestral chromosome segments

Association studies depend on linkage disequilibrium, which in turn reflects shared chromosome segments in large numbers of people because of a very distant common ancestor. Throughout this book we talk about families—groups of people who share large parts of their genomes because of common recent descent. We speak about people being *related* to each other, but we are all related if we go far enough back in history.

What we mean by “related” is having a *known* common ancestor (usually one that can be identified within the previous four generations). And when we say that two persons are unrelated, we generally mean that they do not have any great-grandparent in common, and that they are unaware of any more distant common ancestor. So-called unrelated people do, however, share small common chromosome segments that they have inherited from more distant common ancestors. If the common ancestor lived a long time ago, each shared segment will be quite small but will be shared by a large number of descendants (see [Figure 8.13A](#) for the principle).

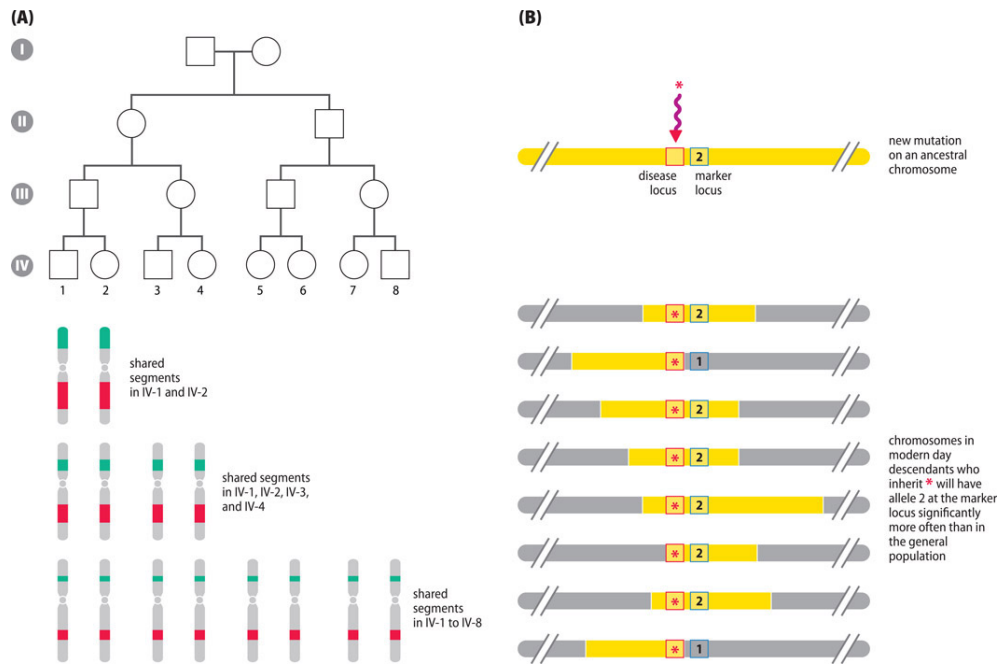


Figure 8.13 Shared ancestral chromosome segments and linkage disequilibrium in the immediate vicinity of an ancestral mutation. (A) The more distant a common ancestor is, the smaller each shared chromosome segment will be, but the larger will be the number of people who share it. In this highly idealized representation, sharing of two chromosome segments on a single chromosome extends to all eight individuals in generation IV (with common great-grandparents). The extent of the shared region is greatest in sibs, but progressively decreases the further the individuals are separated from a common ancestor. (B) Linkage disequilibrium around an ancestral mutation that confers disease susceptibility. Upper panel: imagine a newly emergent mutation (red asterisk) appeared on a chromosome that had a minor (infrequent) allele 2 at a very closely linked SNP locus where allele 1 is the major allele. Lower panel: after passing down through multiple generations, meiotic homologous recombination will ensure that most of the original chromosome (yellow) will have been replaced by segments from other copies of the chromosome (gray). Descendants who inherited the part of the ancestral chromosome with the disease-susceptibility variant have an increased chance of having allele 2 at the very closely linked marker locus. Affected individuals will therefore have a significantly higher frequency of allele 2 at the marker locus than the general population, or than control unaffected individuals. (Adapted from [Ardlie KG et al. \[2002\]](#) *Nat Rev Genet* 3:299–309; PMID 11967554. With permission from Macmillan Publishers Ltd.)

Shared ancestral chromosome segments can explain linkage disequilibrium. Shared segments contain loci that have not been separated by recombination, and so there is a nonrandom association of alleles at linked loci within such segments. By chance, an ancestral chromosome segment might contain an allele that confers susceptibility to a complex disease. In that case, people living now who suffer from the same disease would tend to share that chromosome segment ([Figure 8.13B](#)).

Usually, the susceptibility allele is neither necessary nor sufficient to cause the complex disease; not everyone with the disease will have that allele, and not everyone with the

susceptibility allele will have the disease. But, overall, people with the disease are more likely than unaffected people to have that ancestral chromosomal segment. This is the underlying principle that makes disease association studies possible.

Linkage disequilibrium decreases very rapidly with distance between alleles. If genomewide association studies are to be carried out successfully, a marker

map with a very high density is therefore needed to cover the genome. On the basis that ~1 nucleotide in 300 is polymorphic, a total maximum of ~10 million single nucleotide polymorphism (SNP) loci are potentially available and the International HapMap Consortium and follow-up studies have mapped and genotyped several millions of SNP loci in different human populations, providing an excellent resource for genomewide association studies. GWA studies have however saved costs by using SNP chips (microarrays) that have just a subset of the total SNP markers (often just 500 000 or a million SNP markers). As we explain later, when we describe how GWA studies are conducted, it has been possible to use statistical methods to infer genotypes at untested SNP loci (ones not represented on the SNP chips used in GWA studies).

The HapMap data show that our nuclear genome is a mosaic of small blocks of sequence, **haplotype blocks**, in each of which there is very limited genetic diversity. The blocks vary in size, but average ~5 kb, and at most chromosomal locations most genomes have just one out of only 3–5 different ancestral blocks. That does not, of course, mean that most of us are descended from 3–5 remote ancestors (at a neighboring block most genomes may again have one of only 3–5 ancestral blocks but they would be inherited from a different 3–5 remote ancestors, and so on; our remote ancestry is with *populations*, not individuals). [Box 8.4](#) provides more details and visual representations of haplotype blocks.

BOX 8.4 HAPLOTYPE BLOCKS AND THE INTERNATIONAL HAPMAP PROJECT

Initial attempts to define ancestral chromosome segments began with high-resolution mapping of haplotype structure in defined small genome regions in populations of European ancestry. The results suggested that our nuclear DNA might be composed of defined blocks of limited haplotype diversity (**haplo-type blocks**). [Figure 1A](#) illustrates an example—a haplotype block 84 kb long that spans most of the *RAD50* gene at 5q31. Eight common SNP loci were genotyped in this block, and two alleles at each of eight SNP loci means the potential for $2^8 = 256$ different haplotypes. Yet, within this block, almost every chromosome 5 that is tested has 1 of only 2 out of the 256 possible haplo-types—either the orange haplotype in [Figure 1A](#) (which we can represent by listing the nucleotides at the eight consecutive SNP loci as GGACAACC) or the green haplotype (AATTCGTG).

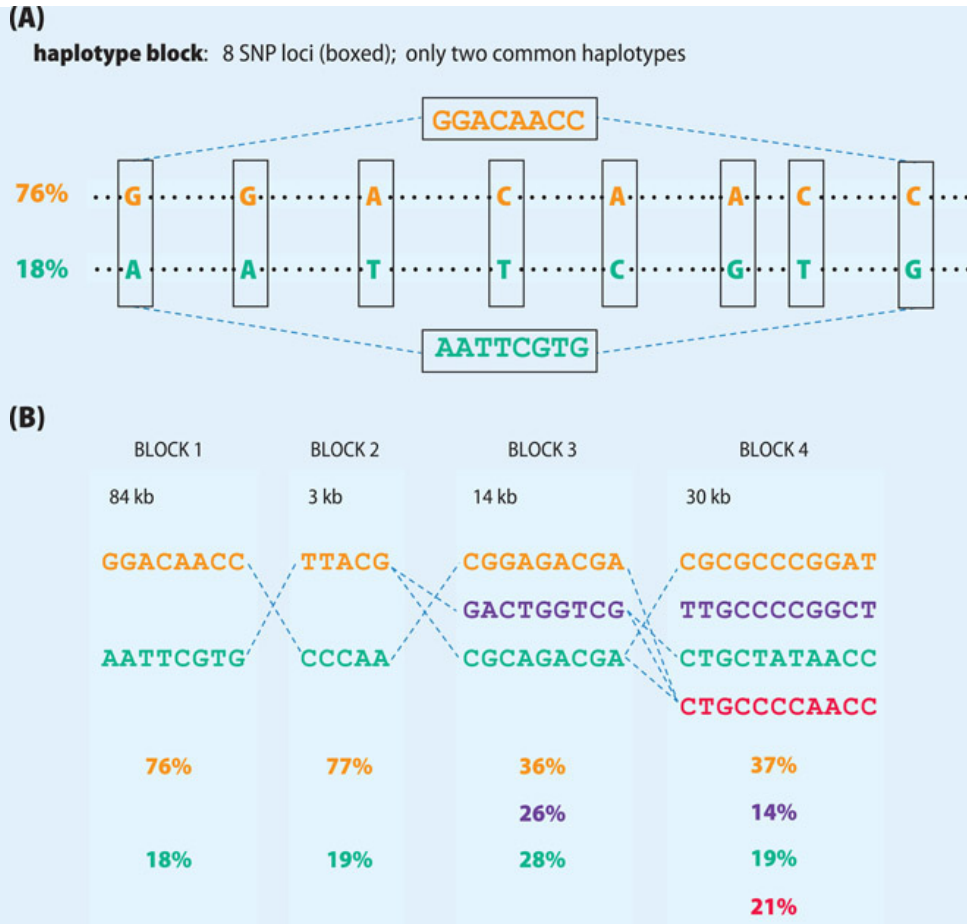


Figure 1 Haplotype blocks. (A) Genotyping at eight SNP loci (vertical boxes) spanning most of the *RAD50* gene at 5q31 reveals an 84 kb haplotype block. Just two haplotypes account for the vast majority (76 % and 18 %, respectively) of the chromosomes 5 from a sampled European population. (B) Adjacent haplotype blocks at 5q31. The 84 kb the block from panel (A) above, is represented as block 1 here. Neighboring blocks 2, 3, and 4 were genotyped at respectively five, nine, and eleven SNP loci and had between two and four haplotypes shown in different colors at population frequencies given at the bottom. Dashed blue lines signify locations where >2 % of all chromosomes 5 are seen to switch from one common haplotype to another. (Adapted from [Daly MJ et al. \[2001\]](#) *Nat Genet* 29:229–232 PMID 11586305. With permission from Macmillan Publishers Ltd.)

The low haplotype diversity is apparent in adjacent haplotype blocks. In [Figure 1B](#), block 1 (the same block that is shown in [Figure 1A](#)) and the neighboring block 2 are dominated by two haplotypes, and the next two blocks by three and four haplotypes, respectively. It suggests that the DNA in block 1 was contributed mostly by two ancestors, and that in blocks 2, 3, and 4 by a different set of two, three, and four ancestors, respectively.

The International HapMap project set out to make comprehensive maps of linkage disequilibrium in the human genome. The project began by genotyping common single

nucleotide polymorphisms (SNPs) in samples from four populations: the Yoruba from Nigeria (YRI); a white population from Utah, USA, descended from northern and western Europe (CEU); Han Chinese from Beijing (CHB); and Japanese from Tokyo (JPT). Haplotype maps were constructed by genotyping 3.1 million SNPs (or about one every kilobase).

The HapMap project confirmed that humans show rather limited genetic variation (by comparison, chimpanzees show very much more genetic diversity). At a fairly recent stage in population history, the human population was reduced to a very small number—perhaps just 10 000 or so individuals—that remained quite constant until comparatively recently. First agriculture, and then urbanization led to a very rapid massive expansion

in population size to the current eight billion individuals. As a result, about 90 % of the genetic variation in humans is found in all human populations.

Overall, about 85 % of our nuclear genome is a mosaic structure, composed of haplotype blocks. The average size of the haplotype blocks in the populations of European and Asiatic ancestry was 5.9 kb with an average of about 3.6 different haplotypes per block. In the Yoruban population there were an average of 5.1 different haplotypes per haplo-type block and the blocks averaged 4.8 kb in size (all human populations originated in Africa, and African populations have greater genetic diversity). Note, however, that the number, size, and identity of blocks depends on the statistical criteria used to define a block.

How genomewide association studies are carried out

Genomewide association (GWA) studies (or **GWAS**) began to really take off in the mid-2000s because of two technological developments. First, the International HapMap Project delivered hundreds of thousands and then millions of mapped SNP (single nucleotide polymorphism) loci. Secondly, by the mid-2000s the extension of microarray technology allowed the automated geno-typing of huge numbers of SNPs across the genome. We described the principles of microarray technology previously in [Figure 3.9](#) at the end of [Section 3.3](#). In the case of whole genome SNP microarrays, the microarrays carry oligonucleotides specific for each allele at many hundreds of thousands of SNP loci across the genome, plus controls.

GWA projects were designed to identify *common* variants (it was assumed that common complex diseases are predominantly caused by common variants). The bulk of GWA studies have therefore focused on case-control studies in which panels of affected individuals and matched controls are genotyped at hundreds of thousands of common SNPs (where the

minor allele usually has a frequency of at least 0.05). SNPs are then identified in which allele frequencies are significantly different in cases than in controls ([Figure 8.14](#)).

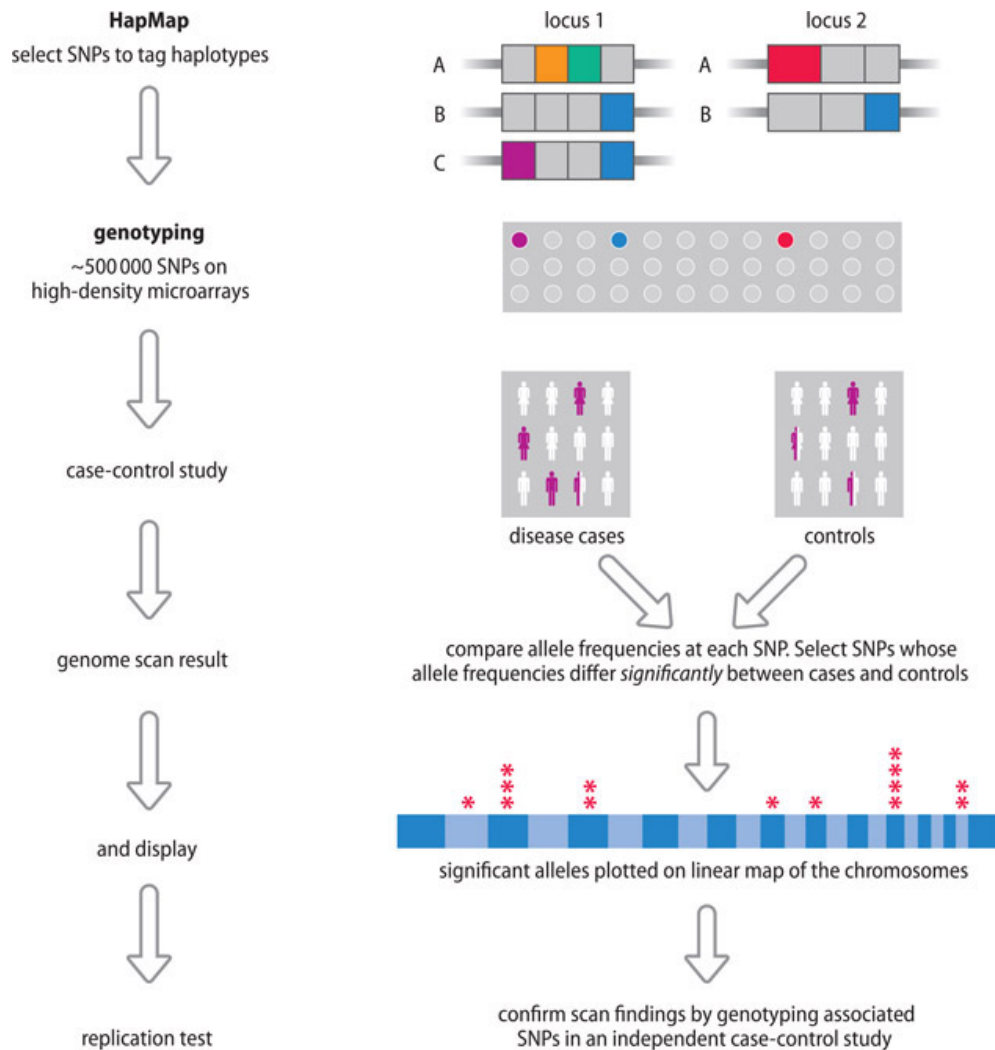


Figure 8.14 Carrying out a genome-wide association scan using SNP chips (microarrays). Using HapMap data (which map linkage disequilibrium across the human genome), representative single nucleotide polymorphisms (SNPs) are selected that will differentiate (or tag) the common haplotypes at each locus. In this example, three common haplotypes (A, B, and C) at locus 1 are tagged by four SNPs (with strong color if present, or gray if absent). But just two SNPs (purple and blue) are sufficient to discriminate between the three haplotypes. Similarly, the two haplotypes at locus 2 can be distinguished by either the red (chosen here) or the blue SNP. The tag SNPs are then genotyped in disease cases and controls using microarrays, and the allele frequencies for each SNP are compared in the two groups. SNPs associated with disease at an appropriate statistical threshold are genotyped in a second independent sample of cases and controls to establish which of the associations from the primary scan are robust. (Adapted from Mathew CG [2008] *Nat Rev Genet* 9:9–14; PMID 17968351. With permission from Macmillan Publishers Ltd.)

Statistical thresholds and data visualization

SNP microarray hybridization typically involves many hundreds of thousands of parallel DNA hybridizations, one for each of the fixed oligonucleotides. Because such huge numbers of hybridization tests are being carried out, stringent statistical significance thresholds are required to assess the significance of individual hybridization results. In order to set a more stringent genome-wide significance threshold, the standard P value of 0.05 is divided by the number of tests carried out. If the microarray hybridization involves one million different hybridization assays, for example, a stringent P value would then be $0.05/1\,000\,000 = 5 \times 10^{-8}$. One consequence of having such a stringent cut-off is that true weak positives might not be recorded. We consider the significance of that below.

The genotype test statistics are calculated for each variant and referenced against statistics expected under the null hypothesis of no disease association. The data can be visualized in different types of plot, notably Manhattan plots as shown in [Figure 8.15](#).

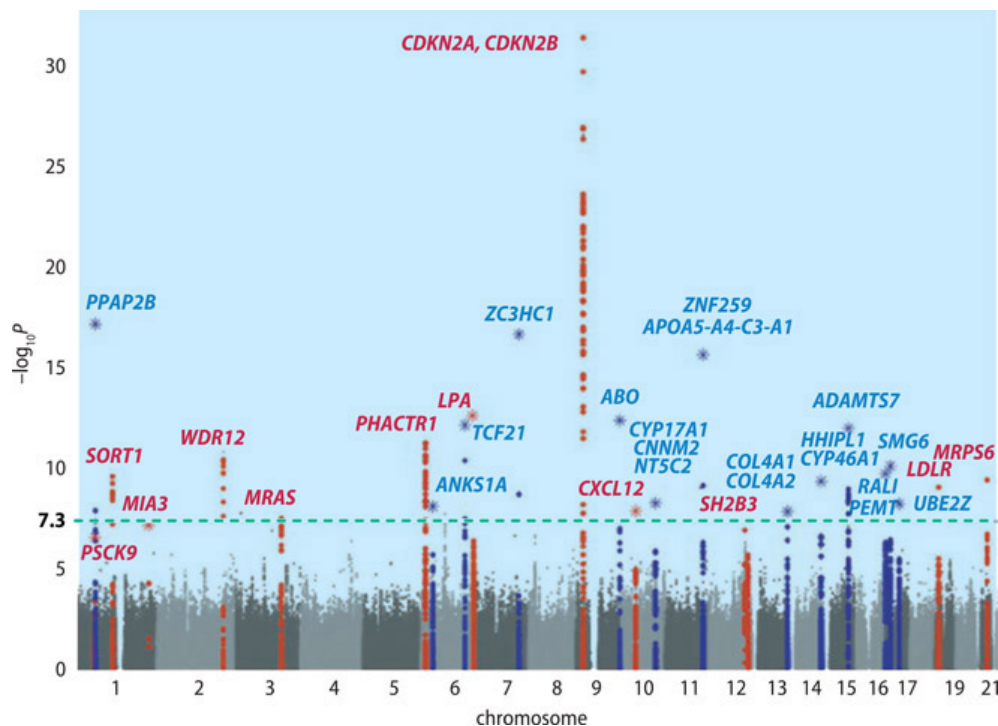


Figure 8.15. Visualizing genomewide association (GWA) data. The most common representation is a genome-wide Manhattan plot (think of skyscrapers). It displays GWA signals according to their genomic positions (horizontal scale) and statistical significance (vertical scale; the negative $\log_{10}P$ scale helps reveal signals of particular interest). This plot is from a large study of coronary artery disease; newly discovered disease-susceptibility loci are in blue, and previously discovered ones in red. The dashed green horizontal line at position 7.30 on the vertical scale indicates the threshold of statistical significance (corresponding to $P = 5 \times 10^{-8}$ in this case). The most significant associations were with previously recorded SNPs in the immediate vicinity of the closely neighboring *CDKN2A* and *CDKN2B* genes at 9p21. (From Schunkert H et al. [2011])

The need for phasing and imputation

The output from standard SNP microarray GWA studies is in the form of geno-types. However, as explained above, population associations are explained by shared ancestral chromosome segments. Accordingly, disease risk is thought to be defined by *haplotypes*, not genotypes. In modern GWA studies the procedure known as **phasing** is employed to arrange genotypes of neighboring loci into haplotypes. Interested readers can find more detail in the [Browning and Browning \(2011\)](#) reference under Further Reading. There are two immediate benefits of phasing. First, it provides haplotype-disease associations, which can be expected to be more accurate than the simple allele-disease associations obtained when analysing the raw genotype data. Secondly, phasing is essential for being able to carry out an important process widely used in GWA studies: genotype imputation.

In statistics, **imputation** is any process designed simply to estimate missing data. SNP chips used in GWA studies usually allow genotyping of hundreds of thousands or sometimes a million SNP loci. That is a rather small fraction of the SNP loci in the human genome (around one nucleotide is 300 is polymorphic, giving ~10 million or so SNP loci across the genome). That is, there is a lot of missing data that could potentially be mined.

Genotype imputation seeks to estimate the identity of the alleles at many of the untested SNP loci, by using reference genotype data and by taking advantage of linkage disequilibrium. Extensive information is available from reference human haplotypes across the genome obtained by various projects where extensive genotyping has been carried out. The initial reference human haplo-types came initially from the HapMap project, then subsequently from other sources: the 1000 Genomes, UK10K or TopMed projects, and then the Haplotype Reference Consortium (HRC), the most widely used imputation reference panel.

By looking for ancestrally shared regions of chromosome between a GWAS sample and individuals in the reference panel, alleles can be inferred with a very high degree of probability because haplotype sharing can extend over significant regions. When a typical sample of European ancestry is compared to haplo-types in the reference panels, for example, shared stretches of >100 kb in length are often identified. It is not easy to represent this visually, but [Figure 8.16](#) gives the general idea.

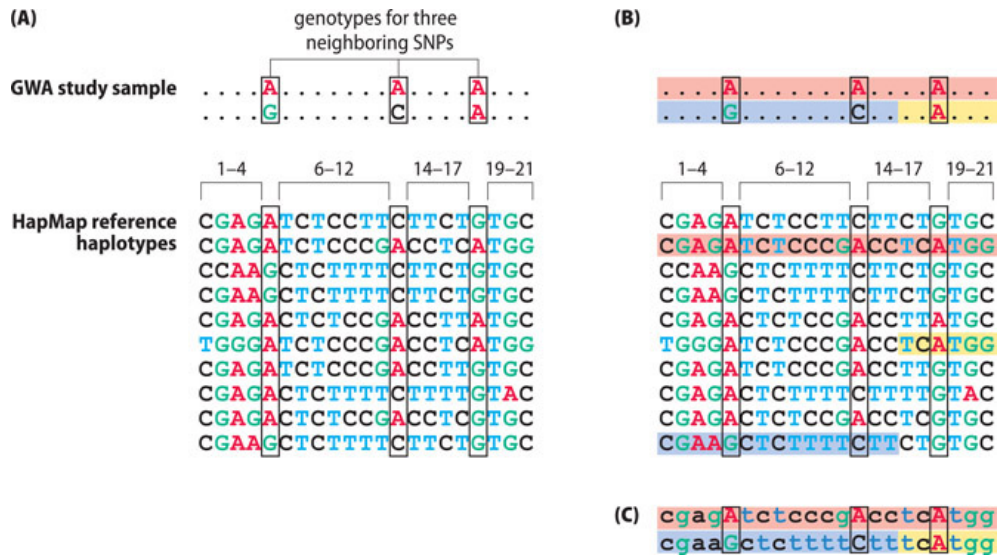


Figure 8.16 Genotype imputation in GWA studies. (A) The observed data. Genotypes obtained in the imagined GWA study here are given for three SNP loci that happen to be represented in the SNP microarray out of a total of 21 SNP loci in the immediate region (the dots represent the missing data, that is, the unknown alleles at the untested 18 SNP loci). For the same region various haplotypes are available from the HapMap where alleles at all 21 SNP loci have been identified. (B) Identifying regions of chromosome shared between a study sample and individuals in the reference panel (sharing is indicated by color shading). (C) The genotypes of the original three SNP loci (shown in upper case letters in the black open boxes), have been supplemented by the haplotype sharing information from (B) so as to reconstruct haplotypes for all 21 loci, even though only 3 of the 21 SNP loci were tested in the GWA study.

Genotype imputation is helpful in finding new disease-associated loci (simply by imputing the additional SNP loci, you get extra statistical power for the study). It can also be useful in fine mapping. A signal indicating association may be obtained with one of the tested SNP markers. But there will be other untested SNP markers nearby. If additional markers in this region are imputed a marker may be found that is more highly correlated with the disease; it would then become a priority for being included in replica studies. But by far the most important contribution made by genotype imputing to GWA studies is to enable large *meta-analyses*, as described below.

Dealing with confounding sample structure

Elements of sample structure, such as population subgroups, family relatedness and/or cryptic relatedness, are important confounders in GWA studies. (A *confounder* is a statistical term for a variable factor that influences both an independent and a dependent variable, such as, respectively genotype and disease incidence.) For example, when there are subgroups in the population with ancestry differences, one subgroup of the population may accidentally

be overrepresented in the study sample and another subgroup may be underrepresented, causing spurious (false-positive) disease associations.

Family relatedness can also confound studies. Ideally, individuals in a population under study would not be (recently) related, but that assumption may often not be true. Various *linear mixed models* perform well in dealing with confounding sample structure in GWA studies. Interested readers can find detailed explanations of the approaches used in recent reviews such as at PMID 24473328 and PMID 25033443.

The importance of very large well-designed studies and meta-analyses

Initial promising GWA subchromosomal locations need to be confirmed. To do this, candidate SNPs of high statistical significance are genotyped in an independent replication panel. In addition to low P values, extra confidence is obtained when the same location is replicated in independent studies on different populations.

Initial GWA studies were of comparatively small scale and more suited to mapping the comparatively few susceptibility factors of quite large effect. To map the more numerous susceptibility factors with more modest effect, large numbers of cases were needed. The template for successful large GWA studies was established by the Wellcome Trust Case Control Consortium (WTCCC), a consortium of British researchers who maximized their chances by pooling their individual collections of cases and controls in well-designed studies. In a landmark paper in 2007, the WTCCC reported considerable success in mapping susceptibility to seven complex diseases with 2000 cases for each disease and a common set of 3000 controls. There followed an explosion in the number of similar GWA studies and very considerable success, as many initial findings have been replicated and confirmed.

Human SNP chips made by different companies often show little overlap in the SNP markers chosen to be represented on the chips. Direct pooling of the data from research studies that use different types of SNP chip is, therefore, extremely problematic. But this is where imputation is so valuable: by imputing the same sets of alleles from many HapMap reference SNP loci, a common set of genotyping data become available for research studies allowing legitimate larger-scale pooling of the data. Such high-powered **meta-analyses** have been the foundation of the important successes of GWA studies in recent times.

Moving from candidate subchromosomal region to identify causal genetic variants in complex disease can be challenging

An SNP that shows a significant disease association is expected to be closely linked to the “causal variant”, the genetic variant that plays a role in causing the disease (by altering how a gene is expressed). However, moving from an associated SNP to a genetically linked causal variant may be extremely difficult for a variety of reasons, as listed below.

- GWAS variants often fall within regions of high linkage disequilibrium. As a result, many non-causal variants are inherited together with the causal variant. Because they, too, are on the same shared small chromosome segment, all of them will also be associated with the disease. Picking the causal variant normally means having to sift through multiple variants, and the candidate region may sometimes be quite large (when the local region shows relatively low recombination).
- Unlike in monogenic disorders, there is the problem that the causal variant is often not causative! It will be absent in a proportion of people with the disease and will be found in many normal people—the “causal” variant is merely a susceptibility factor. Take, for example, the hypothetical association of allele *A*1* with disease *D*. Imagine that *A*1* has a frequency of 0.457 in 2000 affected cases and a frequency of 0.361 in 2000 controls. That might seem a small difference, but because of the large number of samples genotyped it would be highly significant.
- Most GWAS variants, and even variants in strong linkage disequilibrium with the lead variant, are located in noncoding DNA regions, making it difficult to infer their importance. A noncoding variant might be expected to affect some noncoding regulatory DNA sequence (which might be poorly understood or previously uninvestigated), causing a change in expression of a nearby gene that might be limited to specific tissues or cell types. Occasionally, causative variants in coding DNA are found, but like the variants affecting regulatory elements, they might be expected to have a mild effect (for a causal variant to be common in the population it must be a comparatively ancient mutation: it cannot have been exposed to strong purifying (negative) selection otherwise it would have been eliminated by now). A causal variant in coding DNA is also likely to be subtle (such as a mild missense mutation that might not be readily recognized). There are exceptions: easily identified frameshifting and nonsense mutations may sometimes be implicated that would be expected to result in retention of most of the normal protein sequence (for an example, see the Crohn’s disease-associated *NOD2* 3020insC variant described in the legend to [Figure 8.11A](#)).

Towards identifying causal variants and disease-susceptibility loci

As mentioned above, a disease-associated SNP will have assorted neighboring variants mapping to the same haplotype block; they, too, will be disease-associated. But because the borders of haplotype blocks are not precisely defined (the linkage disequilibrium is always significantly <100 %), there may be differences in the degree of association of the different variants within a block. Within each block are a limited number of haplotypes; variants that happen to be present in multiple haplotypes will be much more strongly associated than

those on a single haplotype. Ideally, what we would hope to find is a peak area where a few variants show especially strong association, against a general background of association for the variants within a critical region (which can encompass adjacent haplotype blocks). Readers who may be interested in the statistical fine-scale mapping methods to home in on causal variants can find a relevant review at PMID 29844615. As the search draws closer to identifying the causal variant and the locus of the disease susceptibility factor, additional approaches can be used.

- *Candidate genes.* If a previously studied gene in the immediate neighbourhood appears to be related to previously identified disease-susceptibility loci at other genomic locations, or is involved in similar biological pathways, it immediately becomes a candidate susceptibility factor locus. It would then be a priority for intensive studies, as listed below.
- *Expression studies.* In principle, genes in the immediate vicinity, or a candidate gene if identified, can be studied to determine if their expression patterns appear to be significantly affected when comparing the presence or absence of the causal variant in laborious wet-lab experiments.
- *Sequencing studies.* Possible coding DNA variants can be screened with programs that assess the likely effect of amino acid substitution on the structure and function of the predicted protein, using programs such as PolyPhen-2 (<http://genetics.bwh.harvard.edu/pph2/>), SIFT (<https://sift.bii.a-star.edu.sg/>), and PROVEAN (<http://provean.jcvi.org/index.php>).

The limitations of GWA studies and the issue of missing heritability

Early linkage studies identified subchromosomal locations for some important genetic variants underlying complex disease, but the overall returns from later studies were minimal. Genomewide association (GWA) studies have had much greater success, and meta-analyses, in which the data from individual large disease association studies are pooled, provide even more power to detect significant associations, and have been especially valuable. Thousands of significant disease associations have been obtained, very many of which have been replicated.

Despite initial high hopes, the common disease variants identified by GWA studies have generally been of very weak effect—often with an odds ratio of 1.2 or less. Exceptions include some novel factors strongly predisposing to age-related macular degeneration (a leading cause of vision loss in older adults due to progressive deterioration of a central region of the retina). But many of the variants with high odds ratios were identified in the pre-GWAS era (such as *APP*, the amyloid precursor protein gene, in Alzheimer disease, the

common *NOD2* alleles in Crohn's disease, and especially HLA alleles that remain, by some distance, the strongest known genetic variants in autoimmune disorders).

The “missing heritability” problem

Once the effect size of an individual GWAS variant is known, its contribution to the heritability can be calculated. By doing this type of calculation for each GWAS variant associated with a specific disease, an aggregate estimate can be obtained for the combined contributions of all of the GWAS variants. This bottom-up approach, however, has frequently given much lower heritability estimates than those from top-down family studies (such as the twin studies described above). That then posed some important questions. Are the GWA studies missing something? And how can we account for the “missing heritability”?

Many different explanations have been put forward to explain the missing heritability problem. However, current thinking identifies two major reasons, resulting from self-imposed constraints in the statistical analysis of GWA studies and in the experimental GWA study design—see below.

- *The statistical significance cut-off for association.* The statistical cutoffs were chosen on the assumption that most association signals represent noise, and that only a few can be real (this assumption was based on what used to be a widespread oligogenic view of complex disease; as explained below, truly polygenic contributions may be the reality for individual complex diseases).
- *The filtering out of low frequency alleles.* The experimental design of GWA studies also requires a cut-off in terms of permitted allele frequency: only common SNP alleles (segregating with frequencies >5 % in the population) have been eligible as standard GWAS alleles. Rarer alleles, which collectively might have made an important contribution to the phenotype variation, were excluded for technical reasons (because of the reduced probability of linkage with causal variants and low statistical power).

The limitations of GWA studies

The motivation for GWA studies was, largely, for two reasons: to improve the prediction of disease risk (identifying individuals as being at higher risk for a specific disease should hopefully enable preventative measures and/or better targeting of clinical resources); and to provide a systematic approach to identify genes involved in pathogenesis, offering more possibilities for developing drugs and more effective therapies.

Early GWA studies were underpowered and were largely limited to identifying common variants of quite large effect, but thereafter increased sizes of individual GWA studies and then combinatorial meta-analyses were able to extend the scope of GWA studies very significantly. GWA studies are now widely viewed to have been technically very successful, delivering many thousands of unique, robust associations between common variants and common diseases. Nevertheless, they have been widely believed to have underperformed for a variety of reasons—see [Table 8.7](#).

TABLE 8.7

FIVE REASONS WHY GWA STUDIES HAVE NOT BEEN AS SUCCESSFUL AS HOPED

Reason	Explanation
Common variants almost always have very small effects	The genetic contribution to disease risk is often made up of an extremely large number of variants with very small effects, often with very low odds ratios (the great majority would need astronomically large study sizes to be definitely implicated).
Missing heritability	The common SNP variants assayed in GWA studies can explain only a part of the heritability. Rare variants and copy number variants also make important contributions (see text).
Limits on resolution	GWA studies are made cheaper by testing a limited percentage of SNP markers and relying on <i>imputation</i> to infer, from linkage disequilibrium, the genotypes at most SNP loci. The cost savings are made at the expense of resolution (GWA studies rely on implicating SNP-causal variant haplotypes rather than genotyping all SNP variants).
Causal variants are very difficult to identify	The great majority of common causal variants are in comparatively unstudied noncoding DNA. Time-consuming experimental approaches and/or very large-scale targeted sequencing may be needed to implicate variants and so far, few causal variants have been identified.
Interconnected regulatory networks	Many genuine disease associations may be due to genes that only subtly impact genes in core pathways contributing to pathogenesis.

The value of carrying out future GWA studies of ever greater scale to hunt down variants with very small effects is questionable. Instead, the priority must be to investigate further the

thousands of disease associations already obtained. As described above, the work involved in moving from a single disease-associated variant to identify the linked causal variant unambiguously can still be significant.

As described in [Section 8.3](#) GWA studies have been very important in helping us understand the underlying pathology of many common genetic diseases. As yet, however, at the beginning of the 2020s they have not yet had much effect on public health, and have had little clinical utility. GWA-derived polygenic risk scores seemed to have some promise, and we explain the background to them in the final subsection of [Section 8.2](#).

Alternative genome-wide studies and the role of rare variants and copy number variants in complex disease

Standard SNP chips used in GWA studies are not suited for identifying rare variants or copy number variants (which can be important in some disorders). Genomewide DNA sequencing can be readily used to identify rare variants that have one or a small number of nucleotides changed, but is not so suited to identifying many CNV loci (Note: the term *copy number variant* is poorly defined, but in the context of genomewide studies it is commonly used to mean copy number changes in sequences of intermediate length—from 0.1kb to a few megabases—where variation is due to simple deletions or duplications rather than replication slippage). For such large CNVs, microarray-based methods are used such as the CytoScan HD platform (which has 1.9 million copy number markers; interested readers can find a review at PMID 30223503). As described below, rare variants and CNVs have been found to be important in some complex diseases.

The importance of rare variants in complex disease

The proportion of phenotypic variance captured by all common SNPs, the SNP-based heritability, is substantially less than estimates of pedigree heritability. Even when using SNP genotypes imputed from a fully sequenced reference panel to gain additional additive variance, a significant gap remains between SNP-based and pedigree-based heritability estimates. The hypothesis that causal variants are rare, and consequently not well tagged (or imputed) by common SNPs, has recently been tested.

After carrying out whole genome sequencing (WGS) to estimate heritability of height on a large sample of individuals from the Trans-Omics for Precision Medicine (TOPMed) program, Pierrick Wachstein, Peter Visscher and colleagues estimated the WGS heritability for height to be significantly greater than SNP-based heritability estimates but approaching pedigree heritability estimates. The implications of the [Wachstein et al. \(2022\)](#) paper, listed in Further Reading, is that rare variants, particularly those in regions of low linkage disequilibrium, are a major source of the missing heritability in complex traits and disease.

More support for the importance of rare variants in complex disease has come from recent disease studies, such as the study of schizophrenia described by [Singh et al. \(2022\)](#), as listed under Further Reading. After whole exome sequencing of a large panel of schizophrenia cases and controls, ultra-rare coding variants were implicated in ten genes, and genes prioritized from studies investigating common variants were found to be enriched in rare variant risk.

Copy number variants associated with complex diseases

The poor initial returns from early (underpowered) GWA studies prompted alternative studies of copy number variants. Some CNVs are quite common (with a frequency of more than 1 %) and have been described as *copy number polymorphisms (CNPs)*. Although the overall contribution of copy number variation to complex disease susceptibility may not be so very high. CNPs have been found to be associated with a variety of different disorders, often by changing the copy number of genes in a clustered multigene family (see [Table 8.8](#) for examples).

TABLE 8.8

EXAMPLES OF COPY NUMBER POLYMORPHISMS (CNPs) ASSOCIATED WITH COMPLEX DISEASES

CNP Allele	Disease
Deletion upstream of the <i>IRGM</i> gene (involved in the innate immune response)	Crohn’s disease
Low copy number allele for an intragenic CNV within the <i>LPA</i> gene encoding lipoprotein Lp(a); shown in Figure 2.13B .	coronary artery disease
High copy number allele that spans the β -defensin multigene family and so provides extra β -defensin genes (which make antimicrobial peptides that provide resistance to microbial colonization of epithelial cells)	psoriasis
Single copy of the complement <i>C4</i> gene (instead of the normal two complement <i>C4</i> gene copies)	systemic lupus erythematosus
Low-copy-number <i>FCGR3A</i> alleles (notably deletions). The <i>FCGR3A</i> gene makes the Fc portion of immunoglobulin G and is involved in removing antigen-antibody complexes from the circulation and in other	

Data from [Girirajan S et al. \(2011\)](#). *Annu Rev Genet* 45:203-226; PMID 21854229.

CNPallele	Disease
antibody-dependent responses	

Data from [Girirajan S et al. \(2011\)](#). *Annu Rev Genet* 45:203-226; PMID 21854229.

Rarer copy number variants have been particularly implicated in neuropsychiatric disease, such as in autism and schizophrenia. High-resolution karyo-typing has shown that about 5 % of individuals with autism spectrum disorder have cytogenetically visible chromosome rearrangements. Global screens also suggest that there is a greater load of subcytogenetic common CNPs and rare CNVs in individuals with autism than in controls. Duplications, as well as deletions, contribute to disease. Many of the CNVs are found to occur *de novo*; others are transmitted, sometimes from an unaffected parent.

Inherited CNVs are scarcely more frequent in autism spectrum disorder than in controls, but *de novo* CNVs in affected individuals are generally larger than in controls, and typically about three to six times more frequent. Many CNVs associated with autism spectrum disorder contain multiple genes ([Table 8.9](#)), even if the pathogenesis might be due to altered expression of a single gene in many cases. Schizophrenia-associated CNVs cover some of the same regions found to be associated with autism, such as 1q21.2 (deletion), 22q11.2 (deletion), and 16p11.2 (duplication), plus large deletions of variable size at the *NRXN1* locus at 2p16.3.

TABLE 8.9

EXAMPLES OF LOCI THAT FREQUENTLY UNDERGO COPY NUMBER VARIATION IN AUTISM SPECTRUM DISORDER (ASD)

Locus size	Genes	Location	Pathogenic allele	Frequency in ASD (%)
0.7 Mb	30 genes	16.11.2	deletion and duplication	0.8
~1Mb	(<i>PTCHD1</i> and <i>PTCHD1AS</i>)	Xp22.1	deletion; mostly affecting upstream <i>PTCHD1AS</i> antisense noncoding RNA	0.5
Variable	<i>NRXN1</i>	2p16.3	mostly deletion	0.4

*

The same region is deleted in Williams-Beuren syndrome. (Data from Devlin B & Scherer SW [2012] *Curr Opin Genet Dev* 22:229-237; PMID 22463983.)

Locus size	Genes	Location	Pathogenic allele	Frequency in ASD (%)
1.4 Mb [*]	22 genes	7q 11.2	duplication	0.2
2.5 Mb	56 genes	22q11.2	deletion and duplication	0.2
1.5 Mb	14 genes	1q21.1	duplication	0.2
Variable	<i>SHANK2</i>	11q13.3	deletion	0.1

*

The same region is deleted in Williams-Beuren syndrome. (Data from Devlin B & Scherer SW [2012] *Curr Opin Genet Dev* 22:229-237; PMID 22463983.)

The *de novo* CNVs cannot contribute to heritability, and although the CNVs in autism spectrum disorder and schizophrenia are quite often of large effect, they account for only a small proportion of the observed genetic variance.

The assessment and prediction of risk for common genetic diseases and the development of polygenic risk scores

GWA studies have disappointed by offering poor predictive capacity. Take type 2 diabetes, for example. A study published in 2011 by de Miguel-Yanes et al. (PMID 20889853) reported that the predictive power of genetic data added very little to the predictive power of clinical data. The predictive power was measured using a standard statistical measure, the AUC (the receiver operator area under the curve statistic—see [Box 8.5](#)). The clinical indicators measured were: age, sex, family history, body mass index, blood pressure, blood glucose, HDL cholesterol, and triglycerides, and the AUC from the combined clinical indicators was 0.903. When genotypes from 40 known genetic susceptibility factors were added to the clinical indicators, the combined AUC statistic increased to just 0.906. Many previous studies had reported similar results.

BOX 8.5 ASSESSMENT AND PREDICTION OF DISEASE RISK

We describe genetic testing in detail in [Chapter 11](#). For now, note that two important parameters of any genetic test are its sensitivity (the proportion of all people who have the condition who are identified by the test) and its specificity (the proportion of all people who do not have the condition in whom the test result correctly predicts absence of the condition).

Identified genetic variants for complex disease susceptibility generally show rather low odds ratios. If genetic testing is ever to have high predictive accuracy in complex disease, a battery of tests would be needed. To measure the prediction accuracy of such testing, receiver-operating characteristic (ROC) curves are used. Here, the test sensitivity is plotted against $1 - \text{specificity}$ (the value of the specificity subtracted from 1.0). The area under the curve (AUC) is a measure of how well the test can distinguish between the tested people who have the condition and those who do not. AUC values range from 0.5 (providing no discrimination between those with the condition and those without it) to 1.0 (perfect discrimination). As shown in [Figure 1](#), simulations show that AUC values can increase as more genetic susceptibility factors are included.

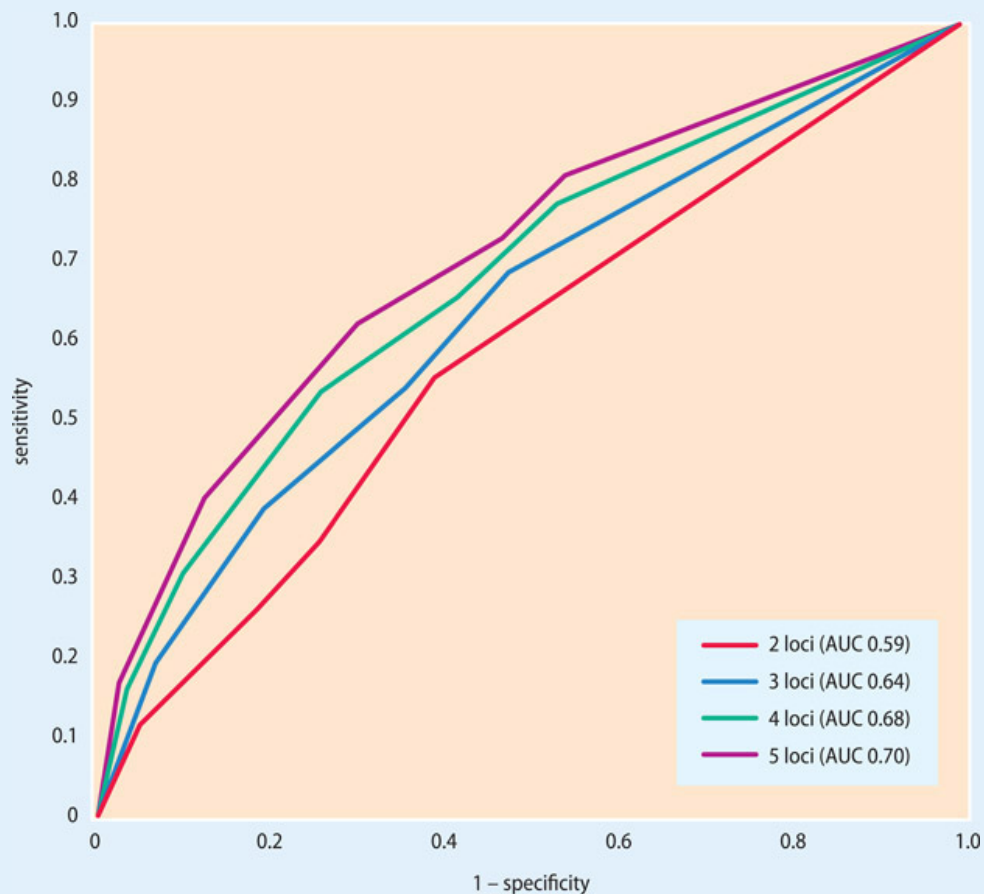


Figure 1 Predictive accuracy of testing for multiple genetic susceptibility factors in complex disease. A receiver-operating characteristic (ROC) curve plots sensitivity against $(1 - \text{specificity})$ for a test. The figure shows ROC curve simulations for testing with two, three, four, or five independent disease susceptibility factors; in this case susceptibility factors 1 to 5 are imagined to have relative disease risks of, respectively, 1.5, 2.0, 2.5, 3.0, and 3.5. (A relative risk of 2.0, for example, would mean that a person with the susceptibility factor has twice the risk of developing the disease compared with a person without it.) Testing for multiple susceptibility factors can lead to an increase in the area under the curve (AUC); the greater the AUC value, the more discriminating the test. (From Janssens AC et al. [2004] *Am J Hum Genet* 74:585–

Note that a very high AUC predictor may be of little practical use when the disease is quite rare. HLA-B27 is very strongly associated with ankylosing spondylitis, a rare type of chronic arthritis that affects parts of the spine. Despite the very impressive odds ratio of close to 70, and a test sensitivity and specificity each of 99 %, the disease risk conferred by typing positive for HLA-B27 is low (in different populations only about 1–5 % of individuals with HLA-B27 will develop the disease).

For many common complex diseases, even multiple variants identified by GWA studies fail to endow the genetic tests with any great predictive value (most SNP variants have odds ratios of less than 1.3). Current prediction of individual disease risk is not accurate because, for most diseases, only a small proportion of genetic variation in risk between people can be explained by known genetic variants. Type 1 diabetes is at the upper end of the scale: about 70 % or more of familial (genetic) risk can be accounted for by a combination of the major histocompatibility complex (the dominant contributor) and more than 50 additional GWA risk loci. The predictive model has an AUC of close to 0.9, but that is still some distance from what would be desired.

Even if we were to know—and be able to test for—every single genetic risk factor for a disease, the resulting whole genome genetic test would have only partial predictive success, because complex disease is caused by a combination of genetic and environmental factors. Depending on the heritability of a complex disease, the accuracy of genome-wide genetic prediction would have an upper limit of 60–90 % (assuming that we could identify every single genetic variant that affects risk and were able to estimate their effects without error). To obtain truly accurate testing in complex disease, environmental factors need to be taken into account.

Polygenic risk scores

The paradigm of establishing risk of a complex genetic disorder by testing individuals at loci known to affect risk was first challenged in 2009 by David Evans and Shaun Purcell and their collaborators. The question asked was: why confine ourselves to testing *known* risk factors when we have the technology from GWA studies to genotype individuals at huge numbers of loci across the genome. So began the new concept of using GWA-derived *genome-wide* genotypes to construct what was initially called polygenic scores (PGS) but came to be called **polygenic risk scores** (PRS). A typical procedure comprises the three steps shown in [Figure 8.17A](#).

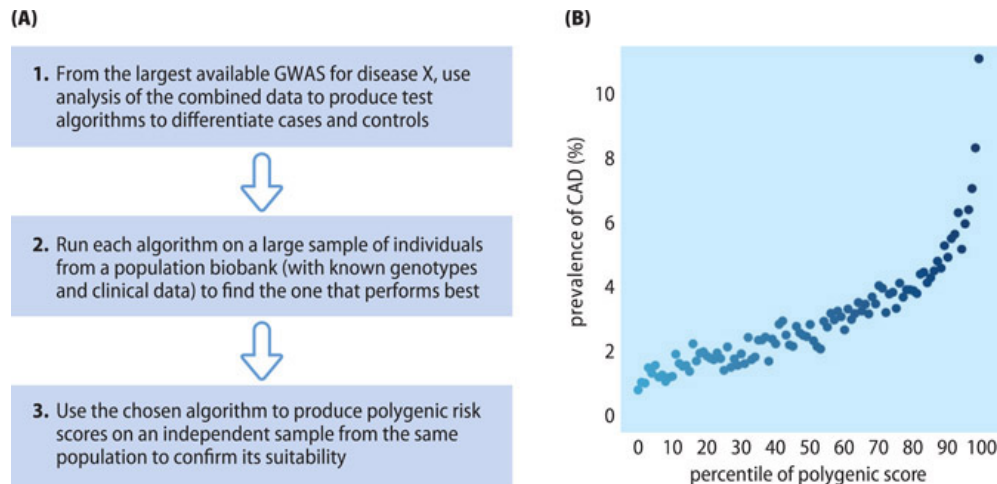


Figure 8.17 Polygenic risk scores: study design and an example. (A) A typical protocol. In step 1 the “combined data” means data on both cases and controls. In step 2, the best-performing algorithm is the one giving the highest polygenic risk scores (B) Distribution of polygenic risk scores for coronary artery disease (CAD) reported by [Khera et al \(2018\)](#) *Nat Genet* 50:1219–1224; PMID 30104762 with permission from Nature Publishing Group.

Initial efforts were hampered by the small sizes of GWA studies, limited computational methods for predicting genome-wide polygenic risk scores, and a lack of large datasets needed to validate and test the scoring system. However, large datasets from more recent GWA meta-analyses have allowed much greater precision in estimating the impact of individual variants on disease risk. And recently developed large population biobanks, such as the UK Biobank (which has medical and genetic data on 500 000 volunteers), have provided very large datasets for validating and testing the algorithms.

Enthusiasm for polygenic risk scores initially soared after an influential study by Khera et al. in [2018](#). They reported the development of useful algorithm predictors for each of coronary artery disease, atrial fibrillation, type 2 diabetes, inflammatory bowel disease and breast cancer. In the case of coronary artery disease (CAD), for example, a GWA meta-analysis involving 184 305 CAD cases and controls was the starting point, and 31 algorithms were developed as polygenic predictors of disease risk using the GWA data. The individual predictors were run on a sample of data from 120 281 participants in the UK Biobank and examined to see how well they could detect those participants that had been diagnosed with CAD. The best predictor had an AUC of 0.81 (referenced against a total of over 6.6 million SNP variants). When then used to compute polygenic risk scores for a separate second group of 288 978 UK Biobank participants, the best predictor performed equally well (AUC of 0.81). The testing with this predictor found that 8 % of the population had a genetic predisposition that conferred a threefold or greater risk for CAD, and the prevalence of CAD rose sharply in the highest polygenic score percentiles (see [Figure 8.17B](#)).

More recently, however, the usefulness and clinical utility of polygenic risk scores has been questioned. We come back to consider this point when we cover genetic testing in [Chapter 11](#).

8.3 ASPECTS OF THE GENETIC ARCHITECTURE OF COMPLEX DISEASE AND THE CONTRIBUTIONS OF ENVIRONMENTAL AND EPIGENETIC FACTORS

At the outset of the third decade of the twenty-first century, the information from studies on most common genetic diseases has yet to make a big impact on clinical practice. The medical specialty that has benefited most is oncology: genomic techniques—notably genome-wide sequencing of tumor—have long been deployed in cancer studies. We consider cancers separately in [Chapter 10](#).

For common genetic diseases, GWA studies had been launched in the hope of delivering two major benefits: greater understanding of the molecular basis of disease; and the prospect of developing new drugs and treatments. Many of these diseases have been expected to be collections of different but related diseases; knowing the major genetic determinants might permit disease *stratification* into disease subtypes to allow more targeted treatments and more efficient clinical management. Cancer genetics has led the way here, and we consider in [Chapter 10](#) how genetic investigations are stratifying cancers into multiple different subtypes.

Novel treatments for a complex disease may also be designed after identifying a *protective factor*, a genetic variant that is negatively correlated with disease. Finally, as new genetic susceptibility factors are identified, there is the prospect of novel *biomarkers*, that is, biological molecules that can be objectively measured and evaluated as indicators of different stages of the disease process; they can be of help in assessing the efficacy of drugs or other new treatments.

As detailed above, the path to understanding the molecular pathologies of complex diseases has not, however, been smooth. Many common genetic diseases are truly polygenic—small contributions can be made by each of a host of different genes that might be expressed slightly differently from normal. We now appreciate that lying between the rare, highly penetrant variants underlying Mendelian disease and the common variants of weak effect, is a much smaller group than initially expected, of low-frequency variants with intermediate effects ([Figure 8.18](#)).

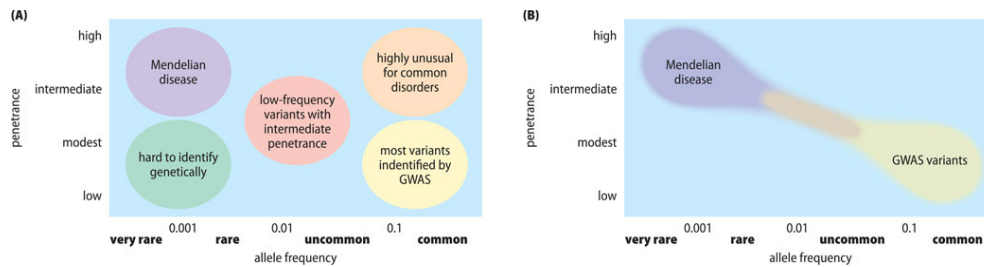


Figure 8.18 Changing views of the roles of genetic factors in determining phenotypes. (A) The hope in the early days of GWA studies. (B) A view after 10 years of GWA studies. (Part A reproduced from McCarthy MI et al. (2008) *Nat Rev Genet* 9:356–369; PMID 18398418, with permission from Nature Publishing Group.)

Because GWA studies rely on linkage disequilibrium to commonly occurring SNP variants, they cannot identify rare variants that may contribute to a complex disease. Rare variants of strong effect may be uncovered in Mendelian subsets, and rare variants of weaker effects might be uncovered through genome-wide sequencing, especially coding DNA variants, or by using micro-array chips designed to identify copy number variants. In this section we focus primarily on how genetic studies have illuminated the molecular basis of common genetic diseases. The genetic studies have involved not just studies of the nuclear genome and SNPs, but also studies of mtDNA variants and of copy number variants.

The genetic architecture of common genetic diseases is a large subject and so we choose to provide select examples to illustrate various principles as follows:

- an outstanding success story: revealing major disease pathways in inflammatory bowel disease (profiled in [Clinical Box 10](#))
- how genetic variants underlying Mendelian subsets relate to those in sporadic forms of the same disease: the examples of Alzheimer and Parkinson disease
- the importance of immune system pathways in complex disease
- unexpected linkages between pathogenetic pathways in different diseases and hybrid roles for individual variants as risk factors for some diseases and protective factors for others.

We finish the chapter by taking a look at how non-genetic factors contribute to complex diseases, and examining the role of environmental factors in complex disease and how epigenetic chromatin modifications might be involved.

CLINICAL BOX 10 ILLUMINATING THE PATHOGENESIS OF INFLAMMATORY BOWEL DISEASE (IBD) USING GENOMICS

Inflammatory bowel diseases are characterized by a chronic relapsing intestinal inflammation. Two major subtypes have been distinguished: Crohn’s disease, which can occur anywhere along the gastrointestinal tract and affects the entire bowel wall; and

ulcerative colitis, which is restricted to the epithelial lining of the colon and rectum. Disease results from abnormal immune responses to commensal organisms, the intestinal microbiota. Heritability is high: estimates from pooled twin studies are 0.75 for Crohn's disease and 0.67 for ulcerative colitis.

Before GWA studies were carried out, almost nothing was known about the genes involved in pathogenetic pathways leading to IBD. One of the very few clues came out of linkage analyses that ultimately led to identifying the *NOD2* gene as a novel susceptibility factor for Crohn's disease (described in [Figure 8.11](#) and associated text). But right from the outset, GWA studies quickly identified many disease associations and by 2017 they had delivered over 200 risk loci for IBD. The associated genes work in a variety of biological processes (see [Figure 1](#)), and mostly participate in biological pathways shared by the two disease subtypes.

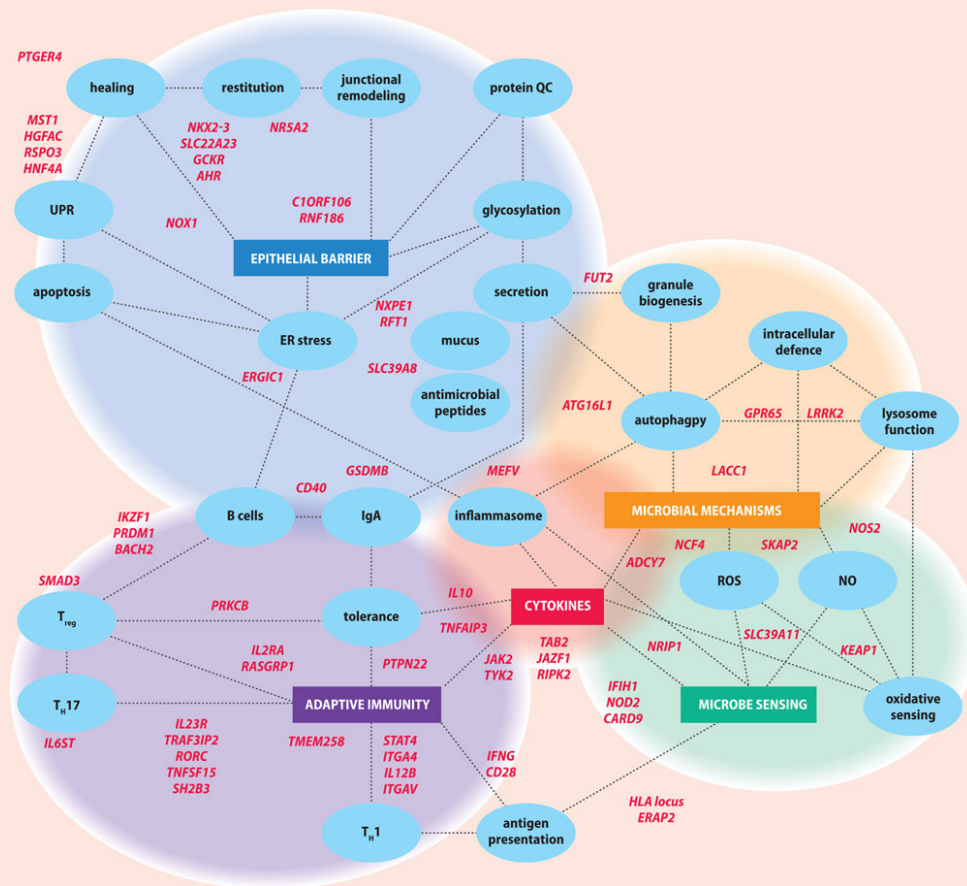


Figure 1 IBD genes and pathways controlling mucosal immunity. IBD risk genes regulate a complex network of interconnected functional pathways. IBD genes (red text) have been implicated in key biological functions (gray circles) that are controlled by interconnected molecular pathways (coloured rectangles). Lines connecting nodes reflect overlapping molecular regulation by common genes. Several IBD risk genes regulate several distinct biological functions depending on their cell type-specific activities. (Reproduced from Graham DB & Xavier RJ [2020] *Nature* 578:527–539; PMID 32103191 with permission from Nature

The GWA findings caused a substantial rethink about the pathogenesis of IBD. The importance of some pathways came as a surprise. The GWA risk variants implicated, for example, as many as five genes with a role in autophagy in Crohn's disease (a lysosomal degradation pathway that naturally disposes of worn-out intracellular organelles and very large protein aggregates). The autophagy machinery is now known to interact with many different stress response pathways in cells, including those involved in controlling immune responses and inflammation.

Another striking—and unexpected—finding was the important role of interleukin-23 (IL-23) pathways in both subtypes of IBD. Tissue injury in these conditions had once been thought to be primarily mediated by classical helper T-cell populations. However, GWA studies have clearly implicated IL-23 and activation of Th17 (a recently discovered subpopulation of helper T cells), resulting in IL-17 production and chronic inflammation. These findings prompted clinical trials using monoclonal antibodies against IL-23, one of which, ustekinumab, has recently been approved for the treatment of both Crohn's and ulcerative colitis.

Common neurodegenerative disease: from monogenic to polygenic disease

Mendelian subsets of common genetic disease are often infrequent, being rare or absent from schizophrenia, bipolar affective disorder, asthma, and stroke. For some complex diseases, however, they are quite common, accounting for ~5 % of all cases of breast cancer, colorectal cancer, and also prostate cancer (where some families show a shared origin with breast and ovarian cancer). In between are disorders where Mendelian subsets are infrequent but significant, accounting for ~1 % of all cases, as for other cancers and coronary heart disease (familial hypercholesterolemia).

Here, we consider the two most common neurodegenerative diseases, Alzheimer disease and Parkinson disease, and the relationship between Mendelian subsets and sporadic disease. In Alzheimer disease, Mendelian subsets are autosomal dominant and rare, accounting for <1 % of all cases, but they are more common in Parkinson disease, accounting for ~5 % of all cases, and comprise both autosomal dominant and autosomal recessive forms.

Linkage analyses provided the first genes to be implicated in these two diseases. Standard linkage analyses were employed for autosomal dominant pedigrees; for autosomal recessive cases autozygosity (homozygosity) mapping was initially used, but thereafter candidate gene approaches or exome sequencing approaches were employed. A list of gene loci implicated by subsequently identifying disease-specific mutations is given in [Table 8.10](#).

TABLE 8.10**GENES UNDERLYING ALZHEIMER (A.D.) AND PARKINSON DISEASE/PARKINSONISM (PARK) IN MENDELIAN SUBSETS**

Gene loci (protein product)	Familial disease*	AD/AR	Onset**	OMIM
<i>APP</i> (amyloid precursor protein);	A.D. type 1	AD	Early	104300
<i>PSEN1</i> (presenilin1)	A.D. type 3	AD	Early	607822
<i>PSEN2</i> (presenilin2)	A.D. type 4	AD	Early	606889
<i>SNCA</i> (synuclein alpha)	PARK1 & PARK4***	AD	Early/Juvenile	168601/605543
<i>PRKN</i> (parkin)	PARK2	AR	Juvenile	600116
<i>PINK1</i> (PTEN-induced kinase 1)	PARK6,	AR	Early	605909
<i>PARK7</i> (DJ1)	PARK7	AR	Early	606324
<i>LRRK2</i> (leucine rich repeat kinase 2)	PARK8	AD	Late	607060
<i>HTRA2</i> (HtrA serine peptidase)	PARK13	AD	Late	610297
<i>PLA2G6</i> (phospholipase A2 group VI)	PARK14****	AR	Juvenile	612953
<i>FBXO7</i> (F-box protein 7)	PARK15*****	AD	Early	260300
<i>VPS35</i> (vacuolar protein sorting 35)	PARK17	AD	Late	614203
<i>EIF4G1</i> (Euk. translation initiation factor4?1)	PARK18	AD	Late	614251
<i>DNAJC6</i> (DnaJ Hsp40 family member C6)	PARK19	AR	Juvenile	615528

Gene loci (protein product)	Familial disease*	AD/AR	Onset**	OMIM
<i>SYNJ1</i> (synaptojanin 1)	PARK20	AR	Early	615530

AD, autosomal dominant. AR, autosomal recessive. OMIM, the Online Mendelian Inheritance in Man database at <https://www.omim.org>

*

Familial disease simply means two or more affected individuals in a family.

**

Juvenile onset, average age <40 yrs; early onset, average age <50 yrs; late onset, average age similar to, or slightly less than that for sporadic patients.

The *SNCA* gene is heterozygously triplicated in PARK4, but has missense mutations in PARK1.

Also known as adult-onset dystonia Parkinsonism.

Also known as Parkinsonian-pyramidal syndrome.

Naturally, heritability is high in Mendelian subsets of both diseases (the causal variants are highly penetrant). In the common polygenic disorders, however, there is a significant difference: common Alzheimer disease is estimated to have a quite high heritability of 0.6 to 0.8, but common Parkinson disease has a low heritability score, around 0.35 to 0.4, and environmental factors are thought to play important roles in Parkinson disease.

The connection between monogenic subsets and sporadic disease

In addition to the gene loci implicated from Mendelian subsets listed in [Table 8.10](#), GWA studies have recently had some successes in Alzheimer and Parkinson disease. The study reported by Schwartzenuber et al. in 2021 (PMID 33589840) describes a GWA meta-analysis of Alzheimer disease that identified 37 risk loci, and a large meta-analysis of Parkinson disease reported 90 significant GWA signals that, however, are almost all located outside coding DNA and explain at most only about one-third of the heritable risk.

The pathology in the monogenic subsets generally mirrors that of sporadic Alzheimer and Parkinson disease, suggesting similar pathways are involved in pathogenesis, even if the monogenic diseases can be more severe, with earlier onset being very frequent. But there may be some phenotype variation as in Parkinson disease types 14 and 15 ([Table 8.10](#)).

A series of questions present themselves. Why should some gene loci be implicated in monogenic disease? And can individual gene loci be involved in both monogenetic and sporadic disease? A gene implicated in a monogenic subset must be displaying a rare variant of strong effect, one that has high penetrance. That type of variant cannot be associated with the common sporadic forms of disease, but it does not mean that other variants of the same gene locus are not involved in common disease.

An example of a gene locus involved in both a monogenic subset of Parkinson disease, and also in a common sporadic form, is the *LRRK2* gene. Autosomal dominant Parkinson disease type 8 is due to certain pathogenic missense mutations in *LRRK2*, such as the p.R1441C substitution and p.G2019S substitution, both reported to increase the kinase activity of the LRRK2 protein. (The G2019S mutation is much the more common of the two mutations and is not quite so penetrant.) Common *LRRK2* variants, however, can also act as susceptibility factors for sporadic Parkinson disease, notably those producing the p.R1682P and p.G2385R protein variants. Each of them doubles disease risk (each being found in 3 % to 4 % of healthy individuals but in 6 % to 8 % of individuals with Parkinson disease).

In the final subsection we consider where genes causing autosomal dominant Alzheimer disease act within the known Alzheimer pathogenetic pathways, but first we take a look at the special case of *APOE* in Alzheimer disease.

***APOE*-ε4, the most significant risk variant in Alzheimer disease**

The human *APOE* gene is the major susceptibility locus for Alzheimer disease. The common *APOE*-e4 allele is such a potent risk factor that in the late 1980s an affecteds-only nonparametric linkage analysis to look for genes underlying late-onset Alzheimer disease was able to map the disease in some late-onset pedigrees to 19q13, the same location as the previously mapped *APOE* gene. *APOE* makes apolipoprotein E, a key component of lipoprotein complexes that direct the transport and delivery of lipids from one tissue cell type to another. It is produced primarily in the liver, and then in the brain where it has long been known to be a component of the senile plaques characteristic of Alzheimer disease. Differences were found between apoE isoforms in binding to amyloid b (Aβ) *in vitro*, and subsequent candidate gene association analyses confirmed the importance of the *APOE*-e4 allele as a major risk factor in Alzheimer disease.

Apolipoprotein E is a member of the vertebrate family of apolipoproteins, but humans are unique in having a functionally polymorphic apolipoprotein E. As a result of allelic variation, three human apoE protein variants can be produced. The variants show differences

at two amino acid positions, and the three different variants confer quite different risk of Alzheimer disease ([Figure 8.19](#)).



Figure 8.19 Allelic variation at the human APOE locus confers significant differences in Alzheimer disease risk. The three human APOE alleles, APOE* ϵ 2, APOE* ϵ 3 and APOE* ϵ 4, make proteins—apoE2, apoE3, and apoE4, respectively—that vary at just two out of the 299 amino acid positions. Compared with people homozygous for APOE* ϵ 3, the most common allele, people with one copy of the APOE* ϵ 4 allele have a roughly threefold greater risk of Alzheimer disease, and people with two APOE* ϵ 4 copies have a roughly fifteenfold increased risk. APOE* ϵ 2, by contrast, is a protective allele, conferring a reduced risk compared to APOE* ϵ 3.

The APOE* ϵ 4 allele also predisposes to cardiovascular disease as well as to Alzheimer disease (and so will have an effect on reproductive rates). That begs the question of why this allele is so common. Evolutionary studies indicate that APOE* ϵ 4 is the ancestral allele (the monomorphic chimp and gorilla apoE proteins both have arginine residues at positions 112 and 158). APOE* ϵ 4 is thought to have been selectively advantageous to early humans who had a low-calorie, low-fat diet. Over time, however, it has increasingly been replaced by the APOE* ϵ 3 allele, which offers the advantage of decreased cholesterol metabolism (reducing the risk of cardiovascular disease).

Rare variants and common susceptibility factors in Alzheimer disease

Rare variants of large effect underlie monogenic forms of Alzheimer disease, but common susceptibility factors are key in sporadic cases. The early-onset and late-onset forms have the same post-mortem brain pathology—abundant extra-cellular plaques, largely composed of amyloid- β (A β) peptides of slightly different sizes, and intracellular neurofibrillary tangles mostly made of tau protein. Based on the similar brain pathologies, it had long been supposed that the rare large-effect variants and common susceptibility factors work in common pathways. Molecular studies have confirmed that, as described below.

Amyloid- β (A β) is now known to be a central focus of the disease pathways. A β peptides are formed by cleavage of the 770-residue transmembrane amyloid-b precursor protein (APP), a neuronal receptor involved in different neuronal functions (including neuronal adhesion and the formation and growth of axons). A β peptides are known to be metal

chelators, binding to metal ions such as copper, zinc, and iron, and reducing them; they also seem to have antimicrobial function. The A β peptides are thought to be the causative agent in Alzheimer disease, partly on the basis of the pathology and on the observation that A β is prone to aggregation in the same way as prions (as detailed in Clinical Box 8 on page 229–30), and partly on genetic analyses.

Standard linkage studies of autosomal dominant early-onset Alzheimer disease have identified three causative genes: the *APP* gene, which produces APP, and *PSEN1* and *PSEN2*, which are both involved in processing APP to make A β . The APP processing reaction requires sequential cleavage by two endoproteinases: first, a b-secretase (also called BACE1) cuts off most of the large N-terminal extracellular portion of APP; then a multisubunit g-secretase cleaves the trans-membrane segment. The catalytic subunit of g-secretase is a presenilin protein, either presenilin-1 or presenilin-2 (encoded by *PSEN1* and *PSEN2*, respectively).

Cleavage by g-secretase occurs at alternative single locations to generate a series of A β isoforms of different lengths (Figure 8.20A). The A β_{42} isoform (42 residues long) is thought to be the greatest contributor to pathogenesis (it is more prone to forming amyloid aggregates) but is not normally produced in large quantities, unlike the predominant A β_{40} isoform.

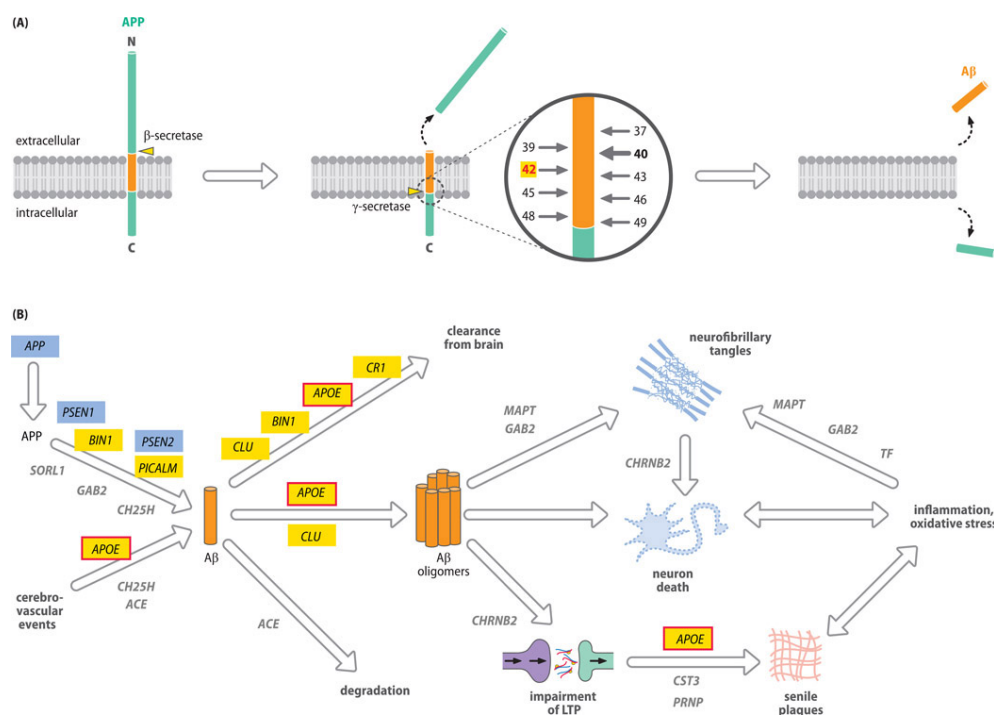


Figure 8.20 Biological pathways in Alzheimer pathogenesis. (A) Production of amyloid- β . The 770-amino acid amyloid precursor protein APP is first cleaved by b-secretase, releasing most of the large extracellular region. Subsequently, the membrane-bound g-secretase cleaves at position 714 or 715, initially generating an amyloid- β (A β) peptide 48 or 49 bp long (orange rectangle). It can then go on to trim three nucleotides at a time, generating a set of isoforms of different lengths, from 37 to 49 amino acids long. Of these, A β_{40} is the

most frequent isoform, but A β ₄₂ is especially prone to aggregation. (B) Early-onset Alzheimer disease genes and common late-onset Alzheimer susceptibility factors belong to common pathways. Gene symbols highlighted in blue at the top left are early-onset disease genes in which causal variants are highly penetrant. Some GWA loci are shown (highlighted in yellow); they are common disease susceptibility factors, mostly with generally modest or weak effects. However, *APOE* (shown with a distinguishing red border) has comparatively strong effects and plays key roles in multiple pathways. Other GWA loci shown here are: *BINI*, bridging integrator 1; *CLU*, clusterin; *CR1*, complement component 3b/4b receptor 1; *PICALM*, phosphatidylinositol binding clathrin assembly protein; and *CD33*. Gene symbols given in pale gray were implicated by functional studies. *SORL1* has also been implicated by DNA sequencing in some cases of early-onset Alzheimer. LTP, long-term potentiation. (Adapted from Bertram L & Tanzi RE [2008] *Nat Rev Neurosci* 9:768–778; PMID 22482448. With permission from Macmillan Publishers Ltd.)

A β metabolism involves a balance between the production from APP and its removal, either by enzymatic degradation (proteolysis) or by receptor-mediated transport out of the brain via the blood-brain barrier (clearance). Pathogenesis results from an increase in the amount of A β or the amount of A β ₄₂ relative to

A β ₄₀, and soluble A β oligomers may have a primary contribution (in addition to affecting synaptic transmission—by impairing long-term potentiation). They may exert some of their effects by regulating the production and phosphorylation of tau protein to induce the formation of neurofibrillary tangles that can cause neurons to die ([Figure 8.20B](#)). A β oligomers can further aggregate into fibrils that end up in extracellular senile plaques that can provoke inflammation responses that can further contribute to pathogenesis.

GWA studies have identified additional variants, some of which have been well replicated and are considered established susceptibility factors. Like *APOE*, they have been implicated in pathways involving A β , but principally in the production of A β and its clearance from the brain (see [Figure 8.20B](#)). However,

several of the genes have a role in inflammation (*CR1* and *CLU*) or the innate immune response (*CD33*; not shown). None of the newly implicated variants in late-onset susceptibility have strong effects (typically odds ratios of 1.15 or 1.10). But as the biological pathways in disease are mapped, new targets may become available for drug therapy.

The importance of immune system pathways in common genetic disease

We have long been aware of the importance of immune system pathways in a subset of complex disease: autoimmune diseases. Variants in the HLA complex at 6p21.3 are the strongest genetic risk factors for diseases such as rheumatoid arthritis, type 1 diabetes, systemic lupus erythematosus, multiple sclerosis, celiac disease, myasthenia gravis, Graves' disease, psoriasis, and so on. Earlier in this section, the importance of immune system pathways in the pathogenesis of inflammatory bowel disease could be seen in the bottom half of [Figure 1](#) in Clinical Box 10 above. Chronic low-grade inflammation is also present in

type 2 diabetes, a metabolic disorder (as well as in the autoimmune type 1 diabetes). Activated innate immune cells accumulate in metabolic tissues with the release of inflammatory mediators.

Although the eye and the brain have been considered as immune-privileged organs—ones where foreign tissue grafts can survive for extended time periods while similar grafts placed at most other sites in the body are acutely rejected—GWA studies have revealed the importance of certain immune system pathways, notably innate immunity, in common genetic disorders of the eye and brain. Take the most common cause of blindness in developed countries as an example: age-related macular degeneration. In this condition, the macula, a small patch of the retina responsible for central vision, is affected. A very early GWA study, using a small number of cases and controls only, identified the *CFH* (complement factor H) gene as a major risk factor in this disorder—a case of a common variant with an unusually strong effect. Follow-up GWA studies also identified three other complement genes as prominent susceptibility factors—*C2* and *CFB* (neighboring genes in the class III HLA region, encoding complement C2 and complement factor B, respectively)—and the *C3* gene at 19q13.

Immune cells actively contribute to homeostatic processes in the nervous system, and include microglia, the brain's primary resident immune cells, mast cells (in the parenchyma) and even small populations of T and B cells in the developing brain. It may not be altogether surprising, then, that immunity pathways are important in common genetic disorders affecting the brain. As an example of neurodegenerative disease, consider Alzheimer disease. As mentioned above, GWA studies have implicated several genes with a role in inflammation—such as the *CR1* (complement C3b/C4b receptor 1) and *CLU* (clusterin) genes (see [Figure 8.20](#))—or in the innate immune response, such as *CD33*. The accumulation of cerebral b-amyloid plaques is thought to result from imbalanced production and removal of amyloid-b peptides, arising from innate immune cells losing the ability to restrict their accumulation.

The immune system plays an important role in neurodevelopment, regulating neuronal proliferation, synapse formation, and plasticity, as well as removing apoptotic neurons. As an example of a neurodevelopmental disorder, consider autism spectrum disorder (ASD) where immune dysfunction in ASD has been repeatedly described, with symptoms including neuroinflammation, increased T cell responses, autoantibodies and enhanced innate NK cell and monocyte immune responses. Not unexpectedly, therefore, innate immune response genes are dysregulated in autism (as revealed by transcriptome analysis—interested readers can find an example at PMID 25494366).

Finally, taking schizophrenia as an example of a psychiatric disorder, GWA studies have identified more than 30 susceptibility factors known to have an immune system function or to be expressed in T or B lymphocytes (PMID 29701842). The strongest genetic risk factor is genetic variation within the major histocompatibility complex. Classical class I or class II HLA genes are not involved but, yet again, a complement gene is: an increased number of

complement *C4* genes in the class III region of the HLA complex is a very strong disease risk factor in schizophrenia. Because of evolutionarily recent tandem duplication of an ~30 kb segment of DNA, there are two slightly different complement *C4* genes (see [Figure 8.21A](#)).

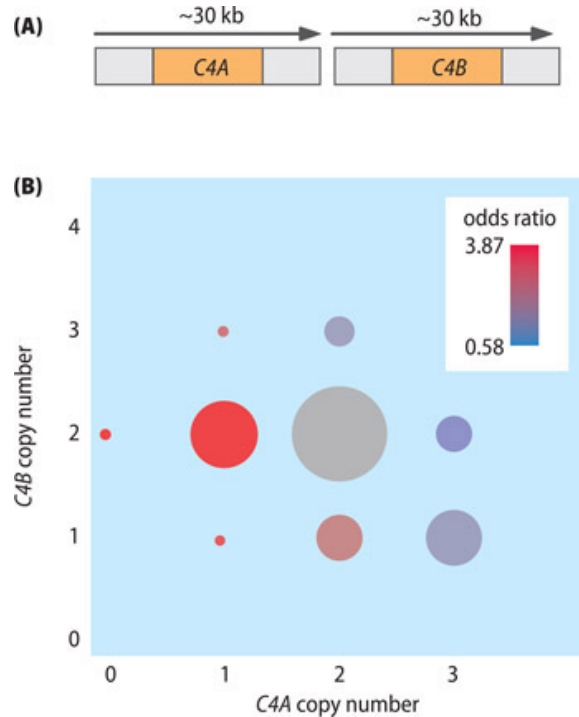


Figure 8.21 Susceptibility to, and protection against, systemic lupus erythematosus (SLE) correlates with complement *C4* gene copy number. (A) The standard (= most common) haplotype has a tandem duplication of ~30 kb with two slightly different *C4* genes, *C4A* and *C4B*. (B) Decreased *C4* copy number is a risk factor for SLE, notably when there is a single *C4A* allele (high odds ratio); increased *C4* copy number, notably three *C4A* copies protects against SLE (low odds ratio). The area of each circle is proportional to the number of individuals with that number of *C4A* and *C4B* genes. Panel B reproduced with permission from Kamitaki N et al. (2020) *Nature* 582: 577–581; PMID 32499649

In addition to the most common haplotype (which has two *C4* genes), unequal crossover can result in haplotypes with one, three, or sometimes four *C4* genes, and excess *C4* genes predispose to schizophrenia. Complement *C4* proteins are found in neuronal synapses, dendrites, axons, and cell bodies, and mouse studies suggest a role for complement *C4* in synaptic pruning. Producing an excess of complement *C4* is thought to cause the reduction in the number of synapses seen in the brains of schizophrenia individuals. A strong risk factor for schizophrenia, increased complement *C4* copy number is simultaneously a protective factor for some other diseases, as described in the next section.

The importance of protective factors and how a susceptibility factor for one complex disease may be a protective factor for another disease

Unravelling the molecular pathology of complex reveals connections between molecular components and biological pathways in different diseases. For example, the common R620W variant of the PTPN22 protein is known to modify disease risk in several autoimmune disorders. But what is now emerging are unexpected links between rather different diseases. According to the extent to which they share GWA variant profiles, heatmaps can be generated to compare the genetic profiles of different diseases—interested readers can find an example in [Figure 1](#) of PMID 20041220.

As well as susceptibility factors, which confer increased risk of disease, genetic investigations are identifying a series of protective factors that reduce disease risk. In most cases identified protective alleles are genetic variants that result from point mutation. They often produce a changed protein that works in a different way, but some are nonsense mutations or splice variants—see [Table 8.11](#) for examples.

TABLE 8.11

EXAMPLES OF PROTECTIVE VARIANTS OR ALLELES THAT REDUCE THE RISK FOR A COMMON DISEASE

Protective variant	Disease	Comments
POINT MUTATIONS		
<i>APOE*ε2</i>	Alzheimer disease	common allele (has cysteines at positions 112 and 158—see Figure 8.19)
<i>APP*A673T</i>		inhibits cleavage of APP, reducing production of amyloid-β
Blood group O	Coronary heart disease	AB blood group is a significant risk factor
<i>CARD9*IVS11 + 1G>C</i>	Crohn’s disease	a rare splice variant, but highly protective (odds ratio 0.29)
<i>PTPN22*R620W</i>		common allele; simultaneously a strong risk factor for both type 1 diabetes and rheumatoid arthritis

Non-HLA genes: *APOE*, apolipoprotein E; *APP*, amyloid protein precursor; *CARD9*, caspase recruitment domain family, member 9; *CCR5*, chemokine (C-C motif) receptor 5; *PCSK9*, proprotein convertase subtilisin/kexin type 9; *PTPN22*, protein tyrosine phosphatase, non-receptor type 22.

Protective variant	Disease	Comments
<i>HLA-DRB1*1301</i>	Rheumatoid arthritis	Affords protection in the 70% of affected cases who have anti-citrullinated protein antibodies
<i>PCSK9*C679X</i>	Coronary artery disease	<i>PCSK9</i> is important in cholesterol homeostasis, and inactivation reduces lipid levels; the C679X allele has a frequency of 1.8% in the US black population
COPY NUMBER CHANGES		
Complement <i>C4A</i> gene increased copy number	SLE (lupus), Sjogren syndrome	see Figure 8.21 for SLE; simultaneously a risk factor for schizophrenia
Complement <i>C4A</i> gene decreased copy number	schizophrenia	see text; simultaneously a risk factor for SLE and Sjogren syndrome

Non-HLA genes: *APOE*, apolipoprotein E; *APP*, amyloid protein precursor; *CARD9*, caspase recruitment domain family, member 9; *CCR5*, chemokine (C-C motif) receptor 5; *PCSK9*, proprotein convertase subtilisin/kexin type 9; *PTPN22*, protein tyrosine phosphatase, non-receptor type 22.

The lower part of [Table 8.11](#) lists examples of DNA variants that act as protective factors through gene dosage, that is alteration in the copy number of a gene. Increasing the number of complement C4 genes, especially the *C4A* gene, provides protection against lupus; conversely, a reduced number of complement C4 genes is a risk factor for lupus—see [Figure 8.21](#). The reverse is true for schizophrenia, as noted above.

A strong risk factor for a common genetic disease may also be a protective factor for an infectious disease. One such example is the common *FUT2* non-secretor

allele, a nonsense mutation in the gene that makes a(1,2)-fucosyltransferase. This enzyme completes the synthesis of H antigens, precursors of the ABO histo-blood group antigens found on cells in body fluids and on the surface of the intestinal mucosa. Homozygotes for the non-secretor allele fail to present ABO antigens in secretions and in the intestinal mucosa and have an increased risk of Crohn's disease and type 1 diabetes (most probably because of alterations to the diverse microorganisms resident in the gut). But the same individuals are strongly resistant to some strains of norovirus, the most common cause of non-bacterial gastroenteritis. A balance between acting as a risk factor for one common condition and acting as a protective factor for a different condition may explain why certain

genetic variants associated with risk of a common genetic disease have reached relatively high frequencies.

Gene–environment interactions in complex disease

Environmental factors are clearly important in cancers and infectious disease where some associations, such as *Helicobacter pylori* infection and ulcer formation, have only recently become evident. But they are increasingly being recognized to be important in complex diseases outside those two categories (as well as in some monogenic disorders); see [Table 8.12](#). That should not be surprising because monozygotic (identical) twins are often seen to be discordant for complex diseases (as shown previously in [Table 8.4](#)). Identical twins arise from the same zygote and might be expected to have identical DNA profiles. If one twin develops a complex disease, such as Crohn’s disease, but the other lives a long healthy life, factors other than DNA can be expected to be important (although post-zygotic mutations could also have a role in some cases).

TABLE 8.12

EXAMPLES OF DIFFERENT TYPES OF ENVIRONMENTAL FACTORS THAT CONTRIBUTE TO COMMON NON-INFECTIOUS AND NON-CANCEROUS DISEASE

Environmental	Examples
Teratogens and abnormal metabolite levels in uterine environment	low folic acid increases the risk of neural tube defects such as spina bifida
Unbalanced diets	over-consumption and excess of fatty foods can predispose to type 2 diabetes
Smoking	increased risk for disorders such as coronary artery disease, Crohn’s disease*, and aging-related macular degeneration
Commensal microorganisms	gut microbiota in inflammatory bowel disease, type 2 diabetes

*

But not ulcerative colitis-if anything, smoking protects against ulcerative colitis.

Striking evidence for the importance of environmental factors comes from increased risk for a specific disease that often befalls migrants who have moved from a community with a low general risk of that disease to join a society in which the disease is much more prevalent. In addition, as populations across the globe change their eating habits and lifestyles, there is a relentless rise in the frequency of conditions such as obesity and type 2 diabetes.

Gene–environment interactions are also important in the sense that they can make it difficult to detect a genetic (or environmental) effect if they are not identified and controlled for. That can lead to inconsistent disease associations when populations are variably exposed to certain environmental factors that modify the effect of a given genetic variant ([Figure 8.22](#)). Understanding gene–environment interactions can therefore allow us to develop protective strategies for complex diseases: by seeking to minimize exposure to an environmental factor, the harmful effect of a genetic susceptibility factor can be minimized.

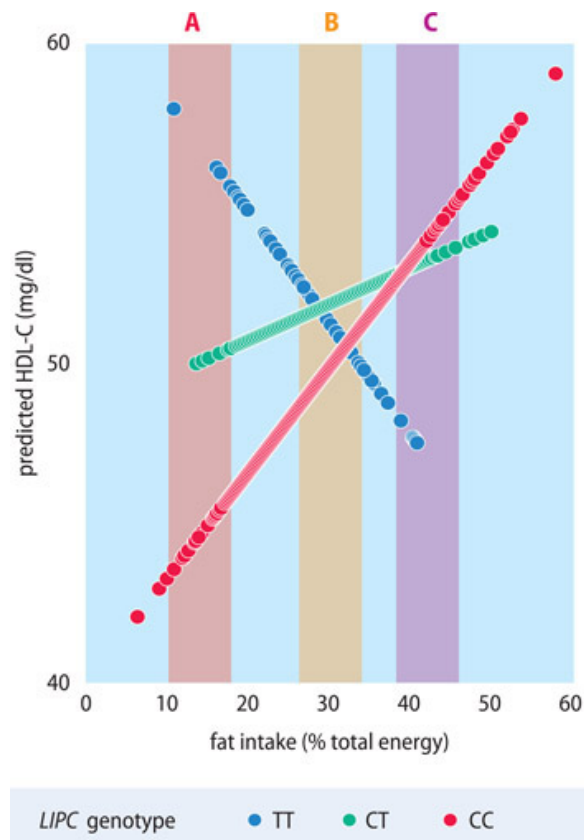


Figure 8.22 The importance of gene–environment interactions—an example. According to different total levels of dietary fat intake, variations in predicted values of high-density lipoprotein cholesterol (HDL-C) are shown for three genotypes at the –514(C/T) polymorphism (rs1800588) in the *LIPC* hepatic lipase locus. Low fat intake (band A) combined with the TT genotype (homozygous for the T allele) results in the highest HDL-C level. For a moderate fat intake (band B), there is no relationship between genotype and HDL-C level. For a high fat intake (band C), the TT genotype has the lowest HDL-C level. Gene–environment interactions are

therefore important in identifying genetic and environmental determinants of medically relevant phenotypes such as HDL-C levels; depending on the dietary fat intake, one could variously conclude that the TT genotype produces high HDL-C levels (band A) or low HDL-C levels (band C), or that it is not associated with HDL-C levels at all (band B). (Adapted from [Manolio TA et al. \[2006\]](#) *Nat Rev Genet* 7:812–820; PMID 16983377.

With permission from Macmillan Publishers Ltd.)

A plethora of “environmental factors”

The term *environment* has a multi-layered meaning in this context—it effectively includes any component that alters disease risk without originating from DNA variation in our cells. There are external physical environments. Right from the earliest stages and throughout life, we are exposed to a range of diverse radiation sources, and also to infectious agents that can influence susceptibility to non-infectious diseases. In the womb, we are exposed to a uterine environment and will be variously affected by what our mother consumes during pregnancy. After birth and throughout life, we ingest, or have surface contact with, a huge range of additional foreign molecules. The molecules that we ingest intentionally (in food and drink, stimulants, and so on) may be considered, in part, lifestyle choices. They, too, along with the amount of physical and mental exercise that we experience and the degree of stress to which we are subjected, are important in disease susceptibility.

Then there is our internal microbiome, the diverse range of microorganisms (*microbiota*) that constitute part of us. Mostly composed of bacteria, our personal microbiomes have 10 times more cells than we have and are mostly located within the gut (the gut microbiome in an average person has possibly 5000 different bacterial species. In addition to being beneficial to us (see above), our microbiomes have a major influence on susceptibility to disease—see above and the review by [Virgin & Todd \(2011\)](#) under Further Reading. Finally, there is the environment inside our body cells and how it links with the extracellular environment. Two important, and interlinked, components here are mtDNA variation ([Box 8.6](#)) and chromatin modifications, including DNA methylation.

BOX 8.6 MITOCHONDRIAL DNA HAPLOGROUPS AND MITOCHONDRIAL DNA VARIATION IN COMMON DISEASE

Because mitochondria are the power sources of our cells, the performance of cells (notably brain and muscle cells, which have high energy requirements) is hugely dependent on mitochondrial efficiency. In an environment where food (and therefore calories) is plentiful, mitochondria efficiently generate energy to keep cells in optimal condition; severely restricting calorie intake impairs mitochondrial and cell efficiency.

Mitochondria have an important influence on how nuclear genes are expressed, because they make the ATP and acetyl coenzyme A needed for diverse cell signaling pathways and

for phosphorylating and acetylating histones in chromatin. They are also the principal generators of reactive oxygen species that damage our cells. Ageing is a major risk factor for most common diseases, and accumulating oxidative damage through a lifetime is a principal contributor to increasing cellular inefficiency. The genetic control of mitochondrial function is mostly specified by nuclear genes, but mitochondrial DNA (mtDNA) is much more susceptible to mutation than is nuclear DNA.

EVOLUTION OF mtDNA HAPLOGROUPS

Because mtDNA is strictly maternally inherited, mtDNA undergoes negligible recombination at the population level, and so SNPs in mtDNA form branches of an evolving phylogenetic tree. The major subdivisions of the world mtDNA phylogeny occurred more than 10 000 years ago and are called mtDNA haplogroups, which developed as humans migrated into new geographic regions, leading to region-specific haplogroup variation ([Figure 1](#)).

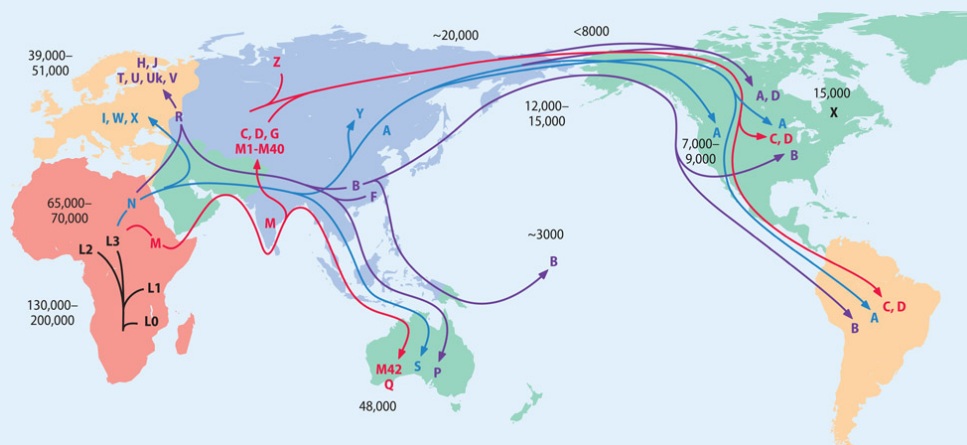


Figure 1 Evolution of mtDNA haplogroups. The estimated mutation rate is 2.2–2.9 % per million years. Time estimates are in years before the present. (From the MITOMAP database at <http://www.mitomap.org>)

More than 95 % of Europeans belong to one of 10 major haplogroups, namely H, J, T, U, K (a subgroup of U), M, I, V, W, and X; several of these are associated with complex human traits (see below for examples). Each haplogroup defines a group or “clade” of related mt DNAs containing specific sequence variants within the population. Mitochondrial DNA haplogroups influence the assembly and stability of the mitochondrial respiratory chain, the synthesis of respiratory chain proteins, and the propensity to develop intracellular

oxygen free radicals, which are implicated in the pathophysiology of several common human diseases.

Although contentious, some evidence suggests that the distribution of human mtDNA haplogroups has been influenced by environmental pressures, including climate. (Nuclear-mitochondrial DNA coevolution has been implicated in climatic evolution in other species.)

MITOCHONDRIAL DNA AND COMMON DISEASE

A variety of rare mitochondrial disorders is known to be due to variants of large effect in single genes in mtDNA, or rare multigenic mtDNA deletions and duplications. Common mtDNA variants with weak effects, notably single nucleotide variants, are known to alter the penetrance of these rare disorders. More recently, diverse association studies have shown that common mtDNA variants also influence susceptibility to a wide range of complex diseases, including neurodegenerative, psychiatric, cardiovascular, and many other diseases—the supplementary table in Gomez-Duran A et al. (2010) *Hum Mol Genet* 19:3343–3353 (PMID 20566709) gives a list of examples.

Certain mtDNA haplogroups have been shown to be associated with disease. The most common mtDNA haplogroup H, found in about 40 % of Europeans, is associated with a more than two-fold increased change in surviving severe infection (sepsis), but subgroups of this haplogroup are emerging as a risk factor for late-onset degenerative diseases, including those affecting the nervous system. This may suggest that infectious disease has shaped mtDNA evolution in Europe over a relatively short period, increasing the frequency of mtDNA haplogroup H and thereby predisposing modern humans to late-onset common disease.

Despite the frequent implication of mtDNA in different common complex diseases, the precise haplogroup associations are not always consistent. This is partly due to the limited cohort size in some studies, restricting the statistical power. Another issue is the different frequency of mtDNA haplogroups in different ethnic groups. A further confounder is occurrence of the same base substitution on different branches of the phylogenetic tree as a result of recurrent mutation, called mtDNA *homoplasy* (but not to be confused with homoplasmy), which accounts for up to 20 % of genetic variation in Europeans. The frequency of a subhaplotype containing a functional homoplasy can vary in different populations, and the distribution of subhaplogroups also varies in different populations. As a result, the major haplogroups can be associated with the disease in some populations but not in others.

The study of gene–environment (GxE) interactions has traditionally involved case-control studies of candidate genes, but the advent of GWA studies has prompted hypothesis-free genome-wide studies. They require large sample sizes, however—a GxE GWA study needs about four times as many samples as a standard GWA study to detect a main effect of the

same magnitude. Various GxE GWA studies have been launched, such as a scan to identify genes conferring susceptibility to air pollution in childhood asthma.

Prospective cohort studies

Case-control studies are the most widely used method of investigating the genetic and environmental basis of complex disease. Cases and controls are typically investigated *retrospectively* (that is, the disease cases have already occurred, and subjects need to be quizzed about previous events such as exposure to environmental factors). As a result, the studies are open to all kinds of bias, for example in the selection of subjects to be studied.

Prospective cohort studies have the big advantage of removing much of the bias by studying individuals over a long timeframe that commences *before* the onset of disease. They involve periodic assessment of subjects, including recording detailed information on them and collecting samples for future laboratory tests. Studies such as these do not select affected individuals, and so need to be very large to ensure that eventually there will be statistically significant numbers of affected individuals.

A leading example of a prospective cohort study is the UK Biobank project. From 2007 to 2010 it recruited 503 000 British people aged between 40 and 69 years and will go on to follow them with periodic testing over a period of 30 years ([Table 8.13](#)). By comparing those who remain healthy with those who develop disease within the 30-year timeframe of the study, researchers hope to gain important information on the genetic determinants of a range of common late-onset diseases (including cancers, heart diseases, stroke, diabetes, arthritis, osteoporosis, eye disorders, depression, and forms of dementia). The study will also help measure the extent to which individual diseases have genetic and environmental causes.

TABLE 8.13

COMPONENTS OF THE UK BIOBANK PROSPECTIVE COHORT STUDY

Baseline questionnaire	Baseline physical measurements	Follow and future measures
Sociodemographic	blood pressure	stored blood, urine, saliva repeat baseline assessment (20000 participants)
Family history	weight, body impedance	
Environmental	waist and hip circumference	

From Manolio TA et al. (2012) *Am J Epidemiol* 175:859-866; PMID 22411865. With permission from Oxford University Press.

Baseline questionnaire	Baseline physical measurements	Follow and future measures
Lifestyle	seated and standing heights	access national health records: death, cancer, hospitalizations, primary care
Cognitive function	grip strength	
Food frequency	bone density	
Internet-administered 24-hour dietary questionnaire	mailed triaxial accelerometers enhanced phenotyping (last 100000-150000 participants recruited): hearing, vascular reactivity, visual acuity, refractive error, intraocular pressure, corneal biomechanics, optical coherence tomography, fitness assessment	

From Manolio TA et al. (2012) *Am J Epidemiol* 175:859-866; PMID 22411865. With permission from Oxford University Press.

Epigenetics in complex disease and aging: significance and experimental approaches

How do environmental factors work to have an impact on complex disease? Somehow they must affect how our genes are expressed. They can do that at the DNA level by changing the DNA sequence in our cells (recall the environmental mutagens that we considered in [Section 4.1](#)).

Alternatively, if they are infectious agents they can introduce some novel genes or proteins that change how our cells work. Yet another way—and one that is now seen to be very common—is to change the epigenetic settings (the *epi-genome*) in our cells. We described before how epigenetic effects regulate gene expression ([Section 6.2](#)) and how they are important in some monogenic disorders ([Section 6.3](#)). And in [Chapter 10](#) we illustrate how epigenetic effects are very important in cancer. Here, we focus on epigenetic effects in other complex diseases.

Unlike the genome, which is very stable, the **epigenome**—effectively the chromatin states across all chromosomes (determined primarily by patterns of cytosine methylation, histone modification, and the positions of nucleosomes)—is comparatively fluid. In response to certain environmental cues (signals), the epigenome can be significantly altered, and that can result in important changes in gene expression.

As described in the second and third subsections below, epigenetic changes occur throughout the life of an individual. They are thought to be important in aging—a frequent

risk factor in complex disease—and they can explain, at least in part, why identical twins develop to become different (different post-zygotic mutations arising in the identical twins can also be expected to play a part).

As shown in [Figure 6.16](#) on page 158, early development is a period of rapid changes in the epigenomes of cell, with a global resetting of methylation marks. And at this stage epigenomes can be particularly sensitive to environmental factors. As described below, a popular theory holds that chronic adult diseases originate in early life, and perturbation of epigenetic settings by environmental factors is an attractive explanation. We consider ways in which that can happen below.

Experimental investigations

Because epigenomes are highly variable between different types of cell, analyzing epigenomes is potentially more complicated than genome analysis. As a result, the investigation of epigenetic factors in complex disease has lagged behind genome analysis. In the past few years, however, great strides have been taken in defining epigenetic settings in cells.

Analysis of global patterns of DNA methylation (the “methylome”) is comparatively advanced—the positions of 5-methylcytosines have been mapped across the genome to single nucleotide resolution in some cell types. Investigators can now carry out large-scale DNA methylation scans across the genome by using microarrays such as Illumina’s Infinium HumanMethylation450 BeadChip (it scans 485 000 cytosine methylation sites distributed across virtually all protein-coding genes with an average of 17 CpG sites per gene region, including CpG sites in the promoter, untranslated sequences, first exon, and elsewhere in the gene body).

Genome-wide investigations to identify the extent of epigenetic contributions to complex disease have been launched for common neurological and autoimmune disorders and some other disorders.

Epigenetic changes during aging

Aging, a very important risk factor for complex disease, is marked by progressive inefficiency in cell, tissue, and organ function. There are inherent limits on the efficiency of cellular processes, including endogenous errors in DNA replication, in DNA repair, and in the regulation of gene expression. As we age, therefore, both genetic and epigenetic changes accumulate progressively, and changes in the genomes and epigenomes of somatic stem cells (that are involved in maintaining tissue homeostasis) may be fundamental in the aging process.

The epigenetic changes that accumulate in our cells can be secondary to genetic changes (mutations in DNA sequences that regulate epigenetic mechanisms) or to inherent errors in the epigenetic regulation machinery. However, they can quite often be induced by environmental factors or as a result of stochastic (chance) factors. Both cytosine methylation and histone modification patterns change with aging. In the former case, for example, there is a progressive loss of cytosine methylation across the genome during aging, but against this general pattern of global hypomethylation (which includes very many methylation sites outside gene regions), hypermethylation occurs at promoters of certain genes.

Epigenetic changes in monozygotic twins

Epigenetic changes may constitute a major reason why monozygotic (identical) twins, who are initially extremely similar in appearance and behavior, go on to develop significant differences in various aspects of the healthy phenotype. And there is quite frequent discordance between identical twins for a variety of complex genetic disorders.

Identical twins derive from a single zygote (the embryo splits at a very early stage in embryonic development), and so they initially have identical genetic profiles (but may accumulate different post-zygotic mutations). Epigenetic differences between identical twins are initially minimal but can begin to occur even in prenatal development. Stochastic factors may be involved, such as different X-chromosome methylation patterns in female identical twins arising from the random choice of which parental X chromosome to inactivate.

Environmental factors can also play a part in prenatal development, because the *in utero* environments can be different. The vast majority of identical twins are located in separate amniotic membranes, and in diamniotic twins there is an increased risk of congenital heart disease that usually affects just one twin. Differences in exposure to postnatal environmental factors most probably contribute to the significant epigenetic differences observed in older monozygotic twins (in both cytosine methylation and histone modification patterns).

The developmental origins of adult health and disease

Pioneering epidemiological studies have established that low birth weight confers an increased risk of developing different common adult diseases, including various cardiovascular diseases, hypertension, and stroke. When significantly fewer nutrients are provided to the fetus during pregnancy, reprogramming in early development seems to cause the fetus to develop a “thrifty phenotype” with a low metabolic rate and reduced pancreatic beta cell mass and islet function. The *thrifty phenotype* is thought to be an adaptation that maximizes the chance of surviving in an adverse environment where calorie intake is restricted, but the altered metabolism is not well adapted to a later life where food is

plentiful, increasing the risk of metabolic syndrome (with strong risk determinants for type 2 diabetes, obesity, and hypertension).

The effects of the Dutch *Hongerwinter*, a wartime famine that took place in western parts of the Netherlands for six months in 1944/1945, provide support for the “thrifty phenotype” hypothesis. Women who endured semi-starvation conditions during mid to late gestation gave birth to underweight babies who were then exposed in later life to normal levels of calorie intake, with an increased incidence of common metabolic and cardiovascular diseases. Individuals born to mothers who experienced starvation conditions during early gestation only, so that they were of average weight at birth, had even higher rates of obesity than those who suffered sharply reduced nutrition in mid to late gestation, and they also had an increased risk of schizophrenia. By implication, early gestation appears to be a particularly critical time in which environmental factors can have an influence.

Because other nutritional cues during infancy and childhood were also found to be associated with adverse effects in later life, the “thrifty phenotype” hypothesis has broadened into a more general theory that proposes that a wide range of environmental conditions during embryonic development and early life determine susceptibility to different adult diseases.

Environmental factors seem to have an impact on development so as to increase the risk of disease in later life, but how do they work? The comparative plasticity of epigenomes makes them likely targets of environmental factors, and this is supported by data from experimental models and human studies (notably environmentally induced changes in DNA methylation patterns). Because epigenetic processes, such as DNA methylation and histone modifications, rely on metabolic factors, a differential availability of dietary components can be expected to influence epigenetic mechanisms. For example, methylation of cytosines and histones uses S-adenosylmethionine as a methyl donor, and dietary factors, notably folate (vitamin B9), are known to have a key role in the pathway that produces S-adenosylmethionine. As described in [Section 10.3](#), cancer studies have also shown a direct link between inflammation (which is often triggered by environmental factors) and epigenetic modification causing altered gene expression.

Transgenerational epigenetic effects

Epigenetic effects are clearly transmitted through mitosis so that chromatin states are heritable through cell generations. For example, when a liver cell divides it gives rise to two liver cells with the same type of epigenome (genome-wide pattern of chromatin states) as the parent cell. But can epigenetic effects be transmitted through meiosis? Might a pattern of increased disease risk deriving from environmentally induced epigenetic modifications be passed on to children so that they, too, have increased disease risk?

Transgenerational epigenetic inheritance is common in plants but rare in animals. In the nematode worm *Caenorhabditis elegans*, experimental manipulation of specific chromatin modifiers in parents can result in an extended lifespan up to the third generation. Suggestive evidence for human transgenerational epigenetic effects comes from certain studies in northern Europe, such as a Swedish study that seems to link the availability of food supply during the early life of the *paternal* grandparents and longevity of the grandchildren, including associations with cardiovascular disease and diabetes. However, a defined mechanism for transgenerational inheritance in humans and animal models is currently missing. Interested readers should consult recent reviews such as those by [Grossnicklaus et al. \(2013\)](#) and [Cavalli and Heard \(2019\)](#), listed under Further Reading.

SUMMARY

- Identifying genes for Mendelian disorders used to rely on first finding a subchromosomal location for the disease gene, usually by genome-wide linkage analyses, but sometimes by looking for disease-associated chromosome breakpoints. Genes would be sought in the target region, and promising candidates tested for evidence of disease-associated mutations.
- Genetic linkage investigates whether alleles at two or more loci co-segregate in families. Two loci located on different chromosomes, or far apart on the same chromosome, are unlinked; alleles at the loci will have a 50 % chance of being inherited together at meiosis. Alleles at linked loci (lying close together on a single chromosome) will often be co-inherited as a haplotype.
- For Mendelian disorders, parametric linkage analyses would be used, ones where the mode of inheritance, disease gene frequency, and penetrance of disease genotypes were inputted into the program. Genome-wide linkage analyses to map a Mendelian disorder require several hundred polymorphic markers from across the genome. If a marker locus is physically close to the disease locus, a marker allele will tend to co-segregate with disease during meiosis.
- The modern alternative for finding Mendelian genes involves whole exome sequencing (sequencing the exons plus immediately flanking intron sequence of protein-coding genes plus miRNA genes).
- A polygenic trait, such as adult height or blood pressure, shows a continuous range of values and a normal (bell-shaped) distribution within a population.

The genetic susceptibility is due to alleles at many loci, each of weak effect.

- In complex (multifactorial) diseases no single gene locus dominates to the same extent as in monogenic conditions. As well as being polygenic, various nongenetic (environmental) factors have important roles in disease.
- DNA variants that cause a Mendelian disorder are rare variants of strong effect. By contrast, many DNA variants involved in pathogenesis of a complex disease are common (occur at high frequency) and are of weak effect. They are susceptibility factors, being found in unaffected individuals but at lower frequencies than in affected people.
- In complex disease, the disease susceptibility needs to cross some high threshold value for a person to be affected. Affected people will have high-risk alleles at multiple susceptibility loci, and so their first-degree relatives will be at higher risk than the general population. Depending on environmental factors, the liability threshold value can change and often shows sex differences.
- The variance of a phenotype is the square of the standard deviation; the heritability is the proportion of the variance that is due to genetic factors.
- In diseases with a strong genetic component, siblings of an affected person have a much higher risk of disease than the general population, and monozygotic twins are much more likely to show concordance in disease status than are dizygotic twins.
- Heritability is not a fixed property—for any disease it varies between populations, and can vary within the same population when environmental factors change.
- Except in the case of rare Mendelian subsets, linkage analyses used in complex disease are nonparametric. They test affected relatives only, typically affected sibs, and look for chromosomal segments that they share more often than expected by chance.
- Association studies are superior to linkage analyses in finding susceptibility factors for complex diseases. They test affected individuals (cases) and unrelated controls from the same population to seek *statistical* associations between individual variants and the disease.
- Association studies aim to identify common alleles (on short chromosome segments) that have significantly different frequencies in cases and controls.

The associations may arise because many people share a short chromosome segment inherited from a distant common ancestor who carried a susceptibility factor.

- The international HapMap project has defined ancestral chromosome segments in various human populations. The shared ancestral chromosome segments are usually very small (often just a few kilobases).
- Candidate gene association studies test for association between a disease and alleles of a specified gene of interest (often because of a suspected role in the disease).
- Association works over very short ranges only on the DNA, and so in genomewide association (GWA), densely spaced markers are needed (usually with at least 500 000 SNP markers across the genome).
- DNA variants may confer increased disease risk (susceptibility factors) or reduced risk (protective factors). A risk factor for one complex disease may sometimes be a protective factor for a different disease.
- Association studies may reveal haplotype blocks harboring a disease susceptibility variant but identifying the pathogenic variant can often be difficult because of linkage disequilibrium, the nonrandom association between all the variants in the block.
- Common disease susceptibility alleles are of ancient origin and are mildly deleterious (such as regulatory sequence mutations and weak missense mutations). They avoid being eliminated by purifying selection by having little effect on reproductive rates, or by simultaneously conferring some selective advantage now, or in the past.
- GWA studies have successfully identified thousands of disease-associated SNP markers; because almost all are of weak effect, they have limited use in predicting disease risk.
- Copy number variants (CNVs) do not generally make a large contribution to genetic susceptibility to disease but do occur at increased frequencies in some disorders.
- GWA studies have been of great value in elucidating the biological pathways in complex diseases, with prospects for identifying new drug targets and treatments.

- Nongenetic factors are clearly very important in complex diseases, but standard case-control studies are limited in their ability to detect gene–environment interactions. Prospective cohort studies are more suited to that task; they study individuals over a long timeframe that commences before the onset of disease.
- Environmental factors work at different levels to influence disease susceptibility. One important way is to alter the epigenetic settings of cells, resulting in altered gene expression. Altered epigenetic settings in early life are thought to alter the risk of various adult diseases such as diabetes and cardiovascular diseases.

QUESTIONS

Questions can be downloaded by visiting the following link, under Support Materials: www.routledge.com/9780367490812.

FURTHER READING

General genetic mapping and meiotic recombination frequency

Altshuler D, Daly MJ & Lander ES (2008) Genetic mapping in human disease. *Science* 322:881–888; PMID 18988837.

[Cheung VG](#) (2007) Polymorphic variation in human meiotic recombination. *Am J Hum Genet* 80:526–530; PMID 17273974.

[Coop G](#) (2008) High-resolution mapping of crossovers reveals extensive variation in fine-scale recombination patterns among humans. *Science* 319:1395–1398; PMID 18239090.

Ott J (1999) *Analysis of Human Genetic Linkage*, 3rd ed. Johns Hopkins University Press. [Authoritative, detailed account.]

Gene identification in Mendelian disorders

[Bamshad MJ](#) (2011) Exome sequencing as a tool for Mendelian disease gene discovery. *Nat Rev Genet* 12:745–755; PMID 21946919.

Puliti A (2007) Teaching molecular genetics: chapter 4—positional cloning of genetic disorders. *Pediatr Nephrol* 22:2023–2029; PMID 17661092.

Heritability and heritability studies

- Lichtenstein P (2009) Common genetic determinants of schizophrenia and bipolar disorder in Swedish families: a population-based study. *Lancet* 373:234–239; PMID 19150704.
- Visscher PM (2008) Heritability in the genomics era—concepts and misconceptions. *Nat Rev Genet* 9:255–266; PMID 18319743.
- Wells JCK & Stock JT (2011) Re-examining heritability: genetics, life-history and plasticity. *Trends Genet* 10:421–428; PMID 21757369.

Quantitative traits and the liability/threshold model

- Falconer DS (1965) The inheritance of liability to certain diseases estimated from the incidence among relatives. *Ann Hum Genet* 29:51–76; doi [10.1111/j.1469-1809.1965.tb00500.x](https://doi.org/10.1111/j.1469-1809.1965.tb00500.x). [The original formulation of the liability threshold model to explain dichotomous traits.]
- Lango Allen H (2010) Hundreds of variants clustered at genomic loci and biological pathways affect human height. *Nature* 467:832–838; PMID 20881960.

Linkage analysis in complex disease

- [Altmüller J](#) (2001) Genomewide scans of complex human diseases: true linkage is hard to find. *Am J Hum Genet* 69:936–950; PMID 11565063.
- Hugot JP (1996) Mapping of a susceptibility locus for Crohn's disease on chromosome 16. *Nature* 379:821–823; PMID 8587604.
- [Risch N & Merikangas K](#) (1996) The future of genetic studies of complex human diseases. *Science* 273:1516–1517; PMID 8801636.
- Weeks DE (2004) Age-related maculopathy: a genomewide scan with continued evidence of susceptibility loci within the 1q31, 10q26, and 17q25 regions. *Am J Hum Genet* 75:174–189; PMID 15168325.

Linkage disequilibrium, haplotype blocks, and the HapMap Project

- [Ardlie KG](#) (2002) Patterns of linkage disequilibrium in the human genome. *Nat Rev Genet* 3:299–309; PMID 11967554.
- [Daly MJ](#) (2001) High-resolution haplotype structure in the human genome. *Nat Genet* 29:229–232; PMID 11586305.
- The International HapMap Consortium (2007) A second generation human haplotype map of over 3.1 million SNPs. *Nature* 449:851–862; PMID 17943122.

Slatkin M (2008) Linkage disequilibrium—understanding the evolutionary past and mapping the medical future. *Nat Rev Genet* 9:477–485; PMID 18427557.

General reviews on genetic association and GWA studies

Lewis CM & Knight J (2012) Introduction to genetic association studies. *Cold Spring Harbor Protoc* 3:297–306; PMID 22383645.

Manolio TA (2013) Bringing genomewide association findings into clinical use. *Nat Rev Genet* 14:549–558; PMID 23835440.

Tam V (2019) Benefits and limitations of genomewide association studies. *Nat Rev Genet* 20:467–484; PMID 31068683.

Visscher PM (2017) 10 years of GWAS discovery: biology, function and translation. *Am J Hum Genet* 101:5–22; PMID 28686856.

Development and technical aspects of GWA studies

[Browning SR & Browning BL](#) (2011) Haplotype phasing: existing methods and new developments. *Nat Rev Genet* 12:703–714; PMID 21921926.

Caliskan M (2021) A catalog of GWAS fine-mapping efforts in autoimmune disease. *Am J Hum Genet* 108:549–563; PMID 33798443.

Cooper GM & Shendure J (2011) Needles in stacks of needles: finding disease-causal variants in a wealth of genomic data. *Nat Rev Genet* 12:628–640; PMID 21850043.

GWAS Catalog. *The NHGRI-EBI Catalog of human genomewide association studies*. Available at www.ebi.ac.uk/gwas/ (a general background description is also available at PMID 30445434).

Marigorta UM (2018) Replicability and Prediction: Lessons and Challenges from GWAS. *Trends Genet* 34:504–517; PMID 29716745.

Wellcome Trust Case Control Consortium (2007) Genomewide association study of 14 000 cases of seven common diseases and 3000 shared controls. *Nature* 447:661–678; PMID 17554300.

Missing heritability, evolution of common variants, and rare variants and copy number variants in complex disease

Manolio TA (2009) Finding the missing heritability of complex diseases. *Nature* 461:747–753; PMID 19812666.

Nakagome S (2012) Crohn's disease risk alleles on the NOD2 locus have been maintained by natural selection on standing variation. *Mol Biol Evol* 29:1569–1585; PMID 22319155.

Raychaudhuri S (2011) Mapping rare and common causal alleles for complex human diseases. *Cell* 147:57–69; PMID 21962507.

[Girirajan S](#) (2011) Human copy number variation and complex disease. *Annu Rev Genet* 45:203–226; PMID 21854229.

Malhotra D & Sebat J (2012) CNVs: harbingers of a rare variant revolution in psychiatric genetics. *Cell* 148:1223–1241; PMID 22424231.

[Singh T](#) (2022) Rare coding variants in ten genes confer substantial risk for schizophrenia. *Nature* 604:509–516; PMID 35396579.

[Wainschtein P](#) (2022) Assessing the contribution of rare variants to complex trait heritability from whole-genome sequence data. *Nature Genet* 54:263–273; PMID 35256806.

Genetic risk prediction in complex disease

Choi SW (2020) Tutorial: a guide to performing polygenic risk score analyses. *Nat Protocol* 15:2759–2772; PMID 32709988.

[Khera AV](#) (2018) Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations. *Nat Genet* 50:1219–1224; PMID 30104762.

Visscher PM & Gibson G (2013) What if we had whole-genome sequence data for millions of individuals? *Genome Med* 5:80; PMID 24050736.

Visscher PM (2021) Discovery and implications of polygenicity of common disease. *Science* 373:1468–1473; PMID 34554790.

Wray NR (2010) The genetic interpretation of area under the ROC curve in genomic profiling. *PLOS Genet* 6:e1000864; PMID 20195508.

Genetic architecture and biological pathways in complex disease

Bertram L & Tanzi RE (2012) The genetics of Alzheimer's disease. *Prog Mol Biol Transl Sci* 107:79–100; PMID 22482448.

Khor B, Gardet A & Xavier RJ (2011) Genetics and pathogenesis of inflammatory bowel disease. *Nature* 474:307–317; PMID 21677747.

King RA, Rotter JI & Motulsky AG (eds) (2002) *The Genetic Basis of Common Disease*, 2nd ed. Oxford University Press.

Sullivan PF (2012) Genetic architectures of psychiatric disorders: the emerging picture and its implications. *Nat Rev Genet* 13:537–551; PMID 22777127.

9

Genetic approaches to treating disease

DOI: [10.1201/9781003044406-9](https://doi.org/10.1201/9781003044406-9)

CONTENTS

[9.1 AN OVERVIEW OF TREATING GENETIC DISEASE AND OF GENETIC TREATMENT OF DISEASE](#)

[9.2 GENETIC INPUTS INTO TREATING DISEASE WITH SMALL MOLECULE DRUGS AND THERAPEUTIC PROTEINS](#)

[9.3 PRINCIPLES OF GENE AND CELL THERAPY](#)

[9.4 GENE THERAPY FOR INHERITED DISORDERS: PRACTICE AND FUTURE DIRECTIONS](#)

[SUMMARY](#)

[QUESTIONS](#)

[FURTHER READING](#)

Treatment of genetic disease and genetic treatment of disease are two separate matters. The cause of a disease (whether mostly genetic or mostly environmental) and its treatability are quite unconnected. Standard medical treatments to alleviate disease symptoms—hearing aids or cochlear implants for treating profound deafness, for example—are just as applicable if the disease is mostly genetic or mostly environmental. In this chapter the primary focus is on how genetic technologies are being applied to treat disease, but we begin by taking a broader look at different treatment strategies for genetic disorders.

For the great majority of genetic conditions, even for single-gene disorders, existing treatments are lacking or unsatisfactory. [Figure 9.1](#) represents a snapshot taken in 1999 when the treatability of 372 genetic diseases was assessed by Charles Scriver and Eileen Treacy. The situation has improved since then, but we still have a long way to go.

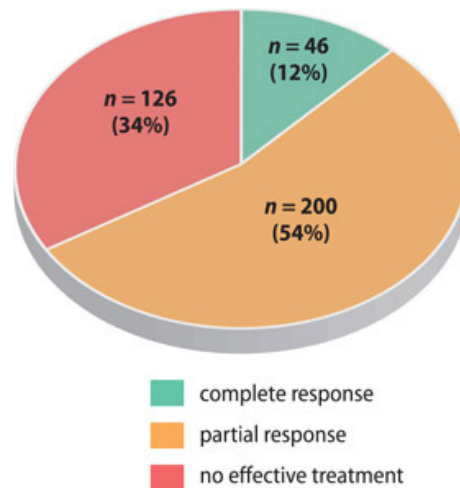


Figure 9.1 Treatment of genetic disorders has often not been very effective. The data here record the response to treatment for 372 single-gene disorders in which the underlying gene or its biochemical function were known and representative treatment data were available by 1999. (Data from Scriver CR & Treacy EP [1999] *Mol Genet Metab* 68:93–102; PMID 10527662.)

Causative genes for many single-gene disorders have been identified comparatively recently, and it may take many years of research to identify how the underlying genes function normally in cells and tissues. Armed with that knowledge, we might hope to devise better treatments in the future, including those that target the *cause* of disease, rather than just dealing with the symptoms. Towards that end, a variety of biological drugs/treatments (sometimes collectively known as *biologics*) have been devised including the use of therapeutic proteins and RNAs, plus gene therapies that offer the opportunity for highly effective treatment for certain monogenic disorders; for other monogenic disorders there may be difficult technical obstacles to devising effective biological therapies—we provide examples below. But despite their promise, biological drugs/treatments can be very expensive, and as genetic causes of disease are understood more precisely new conventional drug therapies may be devised—we give examples below.

We cover treatments for cancers in [Chapter 10](#). For other complex diseases, reasonably satisfactory treatments may exist, such as in the case of diabetes; for many others the treatments are less than satisfactory, or ineffective. By definition, complex diseases are complex at the genetic level: only a decade ago we knew very few of the underlying genetic factors, but recent studies have since revealed many of the contributing genetic factors. In some cases, genetic studies will be able to divide individual complex diseases into subtypes (*disease stratification*), allowing different treatments to be tailored to suit different disease subtypes. The emerging information will place us in a better position to develop novel, more effective treatments.

Environmental factors are clearly very important in complex diseases and have been notably well documented in many cancers. Some environmental factors are also well recognized in some noncancer conditions. Cigarette smoking is a powerful factor in age-

related macular degeneration and emphysema, for example, and the importance of a healthy diet and regular exercise is well recognized in conditions such as type 2 diabetes. Considerable work needs to be done to extend our knowledge of contributory environmental factors. That will provide opportunities for effective interventions, because exposure to an environmental factor can often be modified.

In this chapter we are primarily concerned with molecular approaches to treating disease. In [Section 9.1](#) we give an overview. First, we look at how treatments can be classified into different categories. We take a broad view of the different levels at which disease can be treated, and explore the different genetic technology inputs that can be applied. In [Section 9.2](#) we cover genetic inputs to treating disease with synthetic hydrocarbon-based chemical drugs (“small molecule drugs”) and therapeutic proteins, including genetically altered antibodies, and we explore aspects of pharmacogenetics that deal with the different responses of patients to small molecule drugs and aspects of drug metabolism. Variation in how we respond to chemical drugs is very important: it leads to hundreds of thousands of fatalities per year.

In [Section 9.3](#) we cover the principles and general methodology of different therapeutic methods involving the genetic modification of a patient’s cells (gene therapy). In this section we also describe related stem cell therapy methods. All the methods described above need to be tested in animal disease models before clinical trials are carried out, and we briefly deal with different approaches to disease modeling in the closing section of [Section 9.3](#).

Finally, in [Section 9.4](#) we describe how gene therapy has been applied in clinical trials, assess the progress made, and consider future prospects, including the use of therapeutic RNAs. Ethical issues are considered later, in [Chapter 11](#).

9.1 AN OVERVIEW OF TREATING GENETIC DISEASE AND OF GENETIC TREATMENT OF DISEASE

In this introductory section we first look at broad categories of treating genetic disease and illustrate the diversity of treatments with examples from inborn errors of metabolism. We then consider the different levels at which molecular-based disease treatments can be applied.

Three different broad approaches to treating genetic disorders

Two types of treatment can be used, according to whether pathogenesis is due to some defined genetic deficiency or some positively harmful effect (rather than a *lack* of some important gene product or metabolite). A third type of treatment seeks to reduce susceptibility to disease by understanding the pathway involved ([Figure 9.2](#)). We expand on these themes in the sections below, taking into account both current practice and experimental therapies.

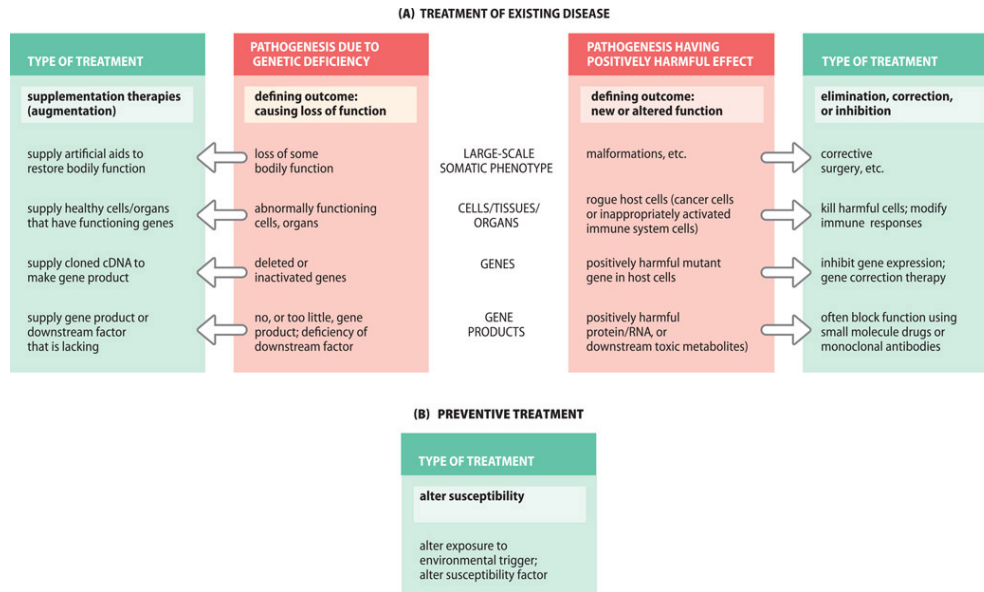


Figure 9.2 Different major treatment strategies for genetic disorders. Note that some strategies are experimental. (A) Different types of treatment for an existing genetic condition according to whether the disease is caused by a genetic deficiency or some other genetic effect, and according to the level at which treatment is applied. Supplementation (augmentation) therapies seek to compensate for a genetic deficiency in various ways: by supplying purified functional gene product directly (protein supplementation); by supplying a purified downstream factor that is lacking; or by indirectly supplying either cloned DNA or healthy cells (either from a donor, or genetically modified cells from the patient) to make the missing gene product. Therapies for conditions where the pathogenesis has a positively harmful effect, may work at different levels. At the cell/tissue level, the object is to deal with rogue cells that behave abnormally to cause disease (notably cancer cells, or immune system cells that attack host cells in autoimmune and inflammatory diseases). At the gene level, the object is to prevent the harmful effects of a gain-of-function mutation (or a gene from a pathogenic microorganism), and at the gene product level, the aims may be to eliminate or reduce the production of elevated toxic metabolites, as in some inborn errors of metabolism. (B) Disease prevention strategies include altering exposure to environmental triggers, such as through extreme dietary modifications in some inborn errors of metabolism, and the use of drugs such as statins to alter disease susceptibility.

Supplementation therapy for genetic deficiencies

In some genetic disorders the problem is the loss of some normal function. In principle, these disorders might be treated by *supplementation (augmentation) therapy*: something is provided to the patient that *supplements* a severely depleted, or missing, factor, thereby overcoming the deficiency and restoring function. Different types of supplement can be provided to restore function at different levels. At the level of the somatic phenotype, treatment can be conventional—providing cochlear implants or hearing aids to treat hereditary deafness, for example.

At the molecular level, the phenotype can be restored by providing a purified gene product that is lacking—a missing enzyme, say, in many inborn errors of metabolism. Alternatively, when the gene product works in a biological pathway required to synthesize some important downstream factor, such as a lipid hormone, it might be a lack of the downstream factor that is treated (by providing purified lipid hormone, in this case). At a higher molecular level, the aim of some types of supplementation (augmentation) therapy has been to transfer a cloned cDNA into the tissues of a patient where it can be expressed to make a missing protein.

At the cellular and organ levels, healthy cells and organs can be transplanted into a patient to make a product that the patient lacks, or to compensate for deficiency of a particular cell type. That can involve transplanting cells from a donor, as in bone marrow transplantation or organ transplantation. More recently, some cellular gene therapies have been used very successfully; here, the cells of the patient are genetically modified so that they can now express the desired gene product. Novel stem cell therapies seek to treat disease by supplying cells of a particular type that are lacking.

Applicability of molecular supplementation therapy

Recessive disorders (where both alleles have lost their function) are more suited to molecular supplementation therapy than are dominant disorders. Affected individuals quite often cannot make any functional copies of some normal gene product. Even a modest efficiency in delivery (of healthy cells, genes, or proteins) to an affected individual can often allow effective treatment, and recently there have been substantial breakthroughs. As illustrated below, however, supplementation gene therapy is currently not practical for some recessive disorders—it can often be difficult to get efficient delivery and production of the desired molecules.

In dominant disorders due to haploinsufficiency, disease occurs even when one allele is normal and present in all diploid cells; efficient delivery and production of the missing gene product would be essential. In addition, the underlying gene is dosage sensitive (described in [Box 7.3](#)). Very precise supplementation therapy—a very difficult prospect—would be needed, and is currently unavailable. Supplementation therapy can also be applied to certain complex diseases, for example by treating diabetes using purified insulin, or by transplantation of pancreatic islet cells.

Treatment for disorders producing positively harmful effects

A second, different approach to treatment is needed for diseases in which the pathogenesis involves a positively harmful effect, rather than a deficiency. Here, supplementation therapy cannot be used: something has gone wrong that cannot be corrected by simply administering

some normal gene, normal gene product, or normal cells to the patient. Different methods are needed (see [Figure 9.2B](#)).

The harmful effect might be treatable at the somatic phenotype level, as in the case of some developmental malformations: corrective surgery is highly effective, for example, in treating various complex disorders such as congenital heart defects, cleft lip and palate, and pyloric stenosis.

At the molecular level, treatments can be conducted at different stages. In many inborn errors of metabolism the problem is elevated levels of harmful metabolites that can be tackled in different ways, as described in the next major section. A more general problem is presented by actively harmful gene products from a mutant gene. Examples include mutant prion proteins and b-amyloid, which are liable to form protein aggregates that can kill cells (described in Clinical Box 8 on page 229), and also harmful proteins or RNAs formed after the unstable expansion of short oligonucleotide repeats (detailed in [Section 7.2](#)). Dangerous mutant gene products may be combatted by using a small molecule drug or therapeutic monoclonal antibody to bind selectively to the mutant molecule and inhibit its activity.

In some cases, the therapeutic strategies are used to selectively inhibit the expression of a harmful gene at the mRNA level, as described in detail in [Section 9.3](#). In addition, at the gene level, experimental corrective gene therapy has the potential to repair damaged genes, replacing a pathogenic mutation by a normal sequence, as described below.

At the cellular level, the problem may manifest itself as harmful cells. Some mutations can induce cells to behave abnormally, proliferating excessively to cause cancers that have been treated by long-standing methods (surgical excision, radiation, and chemotherapy) and more recently by targeted chemical and biological drugs (we cover these plus the prospects of cancer gene therapies in [Chapter 10](#)). In some genetic disorders, the problem is excessive immune responses in which certain immune system cells inappropriately attack host cells (in autoimmune disorders such as rheumatoid arthritis, and in inflammatory diseases such as Crohn's disease). Here, there is the potential to employ therapies that down-regulate immune responses, but in some cancer gene therapy trials the exact opposite approach has been taken (upregulating immune responses in an attempt to kill cancer cells).

Treatment by altering disease susceptibility

A third way of treating disease seeks to reduce susceptibility to disease in some way, for certain monogenic disorders and also some complex diseases. In some inborn errors of metabolism, blockage of one step in a metabolic pathway can drive alternative pathways, causing a buildup of toxic metabolites. A solution here may be to reduce disease susceptibility by removing an environmental trigger. And in some diseases, key susceptibility factors can be manipulated to reduce the chance of disease recurrence, or the effects of a progressive disease.

Very different treatment options for different inborn errors of metabolism

Inborn errors of metabolism, founded on the work of Archibald Garrod on alkaptonuria in the early 1900s, were the first genetic disorders to be understood at the biochemical level. Since then, we have developed a detailed understanding of the molecular pathology for many of these disorders.

Early optimism that disease treatments would follow once we knew all the major details of pathogenesis has suffered quite a setback: effective treatment is unavailable for many of these disorders. But there has been a steady improvement. In a longitudinal study of 65 inborn errors of metabolism published in 2008, significant improvement was recorded: 31 disorders showed no significant response to treatment in 1983 but only 17 in 2008; and the number of conditions fully responding to treatment jumped from 8 in 1993 to 20 in 2008. As we describe later, there have been some important and encouraging successes using various therapies in the last dozen years.

In this section, we use inborn errors of metabolism to illustrate the different ways in which treatment can be offered (each of the three general strategies shown in [Figure 9.2](#) has been successfully applied), and why effective treatment will be difficult for some disorders.

Two broad phenotype classes

For the simplest cases, picture a metabolic pathway composed of sequential steps, each catalyzed by an individual enzyme. Now imagine that loss-of-function mutations have inactivated the gene encoding one of the enzymes. The resulting absence of that enzyme will lead to a lack of downstream product plus a buildup of substrate proximal to (before) the blocked step.

Sometimes, and often in biosynthetic pathways, the most noticeable effect is the lack of end product. In these cases, supplementation therapy can compensate for a lack of a gene product, or of some other downstream molecule whose production depends on the gene product.

In other cases, as previously described for phenylketonuria in Clinical Box 9 on page 234, the buildup of precursors proximal to the blocked step drives alternative pathways, producing abnormal concentrations of some metabolites that have harmful effects. While a disorder such as this is caused by a recessive loss of function at the disease locus, the disease phenotype results from what is at the cellular and physiological levels, a gain of function: a buildup of positively harmful metabolites that requires different treatment strategies. As we show below, the disease phenotype for some disorders may have components that are treatable by supplementation therapy, and others that are due to positively harmful effects.

Supplementation therapy

Here the missing product is provided to overcome the deficiency. It might be the gene product itself (protein supplementation) or a critically important downstream factor that it regulates, which may not be a protein. For example, recessive congenital hypothyroidism (OMIM 275200) is due to a deficiency in thyroid hormone (a *lipid* hormone whose production is largely dependent on a peptide hormone, thyroid-stimulating hormone). Affected individuals have a mutant thyroid-stimulating hormone receptor that fails to respond to thyroid-stimulating hormone, but they are effectively treated with purified thyroid hormone. Mutations in *CYP21A2* cause 21-hydroxylase deficiency by disturbing the production of steroid hormones. The effects include greatly reduced levels of two types of lipid hormones: steroid hormones of the mineralocorticoid class (such as aldosterone) and glucocorticoid class (such as cortisol); the deficiencies in these hormones can be treated by relevant steroid supplementation therapies ([Figure 9.3A](#)).

prenatal period (which suppresses production of fetal adrenal steroid hormones). (B) Tyrosine catabolism in type I tyrosinemia. Genetic deficiency of fumarylacetoacetate hydrolase (FAH) produces elevated levels of fumarylacetoacetate (which may induce liver damage), and of succinylacetone, an inhibitor of δ -aminolevulinate dehydratase (ALAD). The resulting buildup of δ -aminolevulinate may precipitate neurological crises. The drug nitisinone inhibits a proximal enzyme, *p*-hydroxyphenylpyruvate dioxygenase (*p*-HPPD), causing a change in metabolite levels (purple arrows), including a compensatory reduction in fumarylacetoacetate levels. NTBC, 2-(2-nitro-4-trifluoromethylbenzoyl)-1,3-cyclohexanedione.

Supplementation therapy can also involve transplanting cells or organs to supplement a genetic deficiency. Bone marrow transplantation, effectively a way of transplanting hematopoietic stem cells, has frequently been used to treat disorders of blood cells (or other cells originating from hematopoietic stem cells). Liver transplantation has been used for many serious inborn errors of metabolism (many metabolic enzymes are synthesized by the liver). Because of the possibility of graft rejection, organ transplantation is a serious matter; it is indicated when the disorder is expected to progress to organ failure—the treatment is intended to save lives, not just cure patients.

A more recent, and quite different, treatment uses a form of gene therapy: the gene product is obtained after the cells of a patient have been genetically modified to contain and express a functional copy of the relevant gene (as detailed in [Sections 9.3](#) and [9.4](#)). And stem cell therapy is designed to replace cells of a particular type that have been lost through disease (or injury).

Treating or preventing harmful effects of elevated metabolites

When abnormally elevated levels of metabolites cause disease, different treatments can be devised. One way is to use drugs to cause a compensatory change in metabolite levels. Take type 1 tyrosinemia (PMID 20301688). This disorder results from a deficiency of fumarylacetoacetate hydrolase, which catalyzes the terminal step in the tyrosine degradative pathway ([Figure 9.3B](#)). The resulting buildup of precursors leads to liver and renal tubule dysfunction, and untreated children may have repeated neurologic crises. Oral administration of nitisinone (also called NTBC) provides effective treatment—it inhibits a proximal (upstream) enzyme step, causing a compensatory reduction in fumarylacetoacetate hydrolase levels ([Figure 9.3B](#)).

Normal levels of elevated metabolites can be restored by the removal of excess amounts from the body. Phlebotomy is a possibility if the excess metabolite is present in blood ([Table 9.1](#)). Alternatively, some indirect means can be used to force the metabolite into another metabolic pathway to decrease the levels to normal values—see the example of dealing with excess ammonia in urea cycle disorders in [Figure 9.4](#).

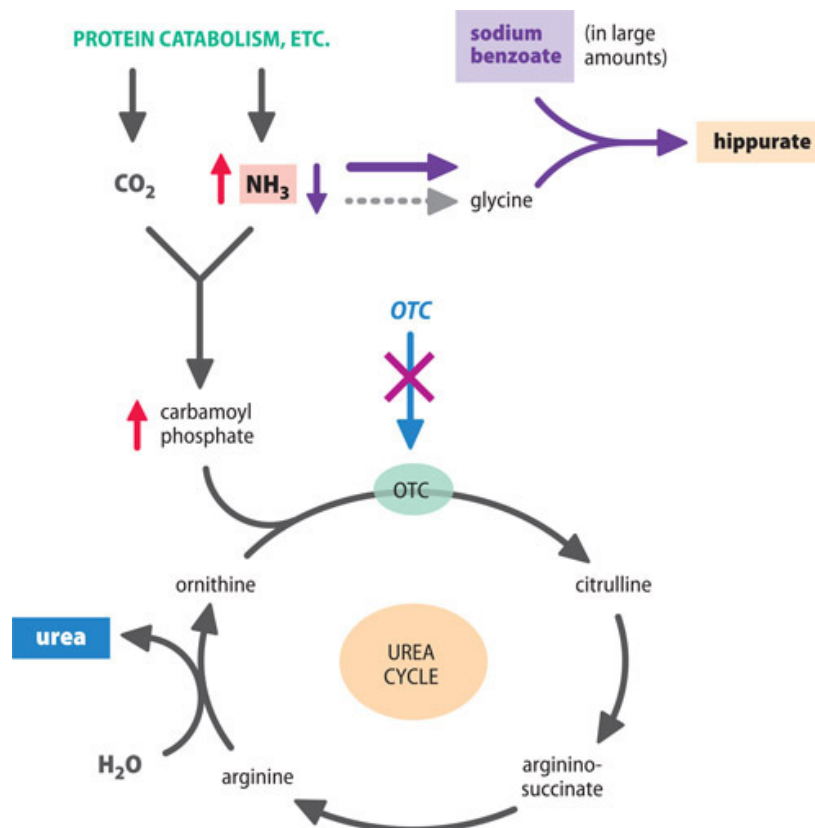


Figure 9.4 Reducing elevated metabolite levels by shunting the metabolite into an alternative metabolic pathway. The urea cycle normally serves to convert ammonia (NH₃), which is neurotoxic, to nontoxic urea. But in urea cycle disorders, ammonia cannot be converted to urea and builds up. In ornithine transcarbamylase (OTC) deficiency (OMIM 311250), the metabolic block causes an increase in levels of the proximal metabolites, carbamoyl phosphate and ammonia (vertical red arrows). The therapy here involves treating a patient with large amounts of sodium benzoate and takes advantage of a normally minor pathway in which some ammonia is naturally converted into small amounts of glycine. Benzoate ions conjugate with glycine to form hippurate, which is excreted in urine. By removing glycine, the treatment drives the production of replacement glycine from ammonia (thick purple horizontal arrow), thereby reducing ammonia levels (vertical purple arrow).

TABLE 9.1 EXAMPLES OF DIFFERENT TYPES OF TREATMENT OF INBORN ERRORS OF METABOLISM

Treatment	Type of action	Examples and comments
Supplementation (augmentation) therapy	protein supplementation	enzyme replacement therapies for many inborn errors of metabolism; provision of blood clotting factors in hemophilias (see Section 9.2 for details)
	hormone replacement	thyroid hormone for infants with congenital hypothyroidism; growth hormone for growth hormone deficiency (see also Figure 9.3A)

Treatment	Type of action	Examples and comments
Counteracting harmful effects of abnormally elevated	bone marrow transplantation	useful for disorders affecting blood cells and some other immune system cells (as illustrated in Figure 9.20), such as mucopolysaccharidosis type 1 (Hurler syndrome, OMIM 607014)
	organ transplantation	liver transplantation has been used successfully for various inborn errors of metabolism, including α_1 -antitrypsin deficiency and urea cycle disorders
	gene supplementation	successfully used for different types of severe combined immunodeficiency, hemophilia (Sections 9.3 and 9.4)
	manipulated excretion of metabolite	periodic phlebotomy (blood removal) is a very effective treatment for directly removing excess iron in the iron overload condition hemochromatosis (OMIM 235200)
metabolites Prevention (avoiding or reducing susceptibility)	shunting of elevated metabolite into side metabolic pathway	for urea cycle disorders the buildup of toxic ammonia can be alleviated by sodium benzoate treatment, driving excess ammonia to be metabolized in a side pathway (Figure 9.4)
	inhibition of a proximal step in a pathway leading to harmful metabolites	babies with type 1 tyrosinemia (OMIM 276700) are unable to metabolize tyrosine effectively and suffer liver damage from toxic intermediates. The drug NTBC inhibits a proximal enzyme to prevent buildup of toxic intermediates (Figure 9.3B)
	substrate restriction (diet modified to severely reduce or eliminate intake of substrate for a deficient enzyme)	reduced intake of phenylalanine in phenylketonuria. Elimination of galactose in galactosemia (OMIM 230400); affected individuals completely lack galactose-1-phosphate uridylyltransferase. Galactose is a component of the lactose in milk but is inessential, and milk is completely withdrawn from the diet
	reduction of a susceptibility factor	in familial hypercholesterolemia, <i>LDLR</i> gene mutations result in low levels of the low-density lipoprotein receptor; the resulting

Treatment	Type of action	Examples and comments
		elevated plasma low-density lipoprotein cholesterol levels predispose to cardiovascular disease but can be effectively lowered by statins, drugs that inhibit a proximal enzyme, HMGCoA reductase, in the cholesterol biosynthesis pathway

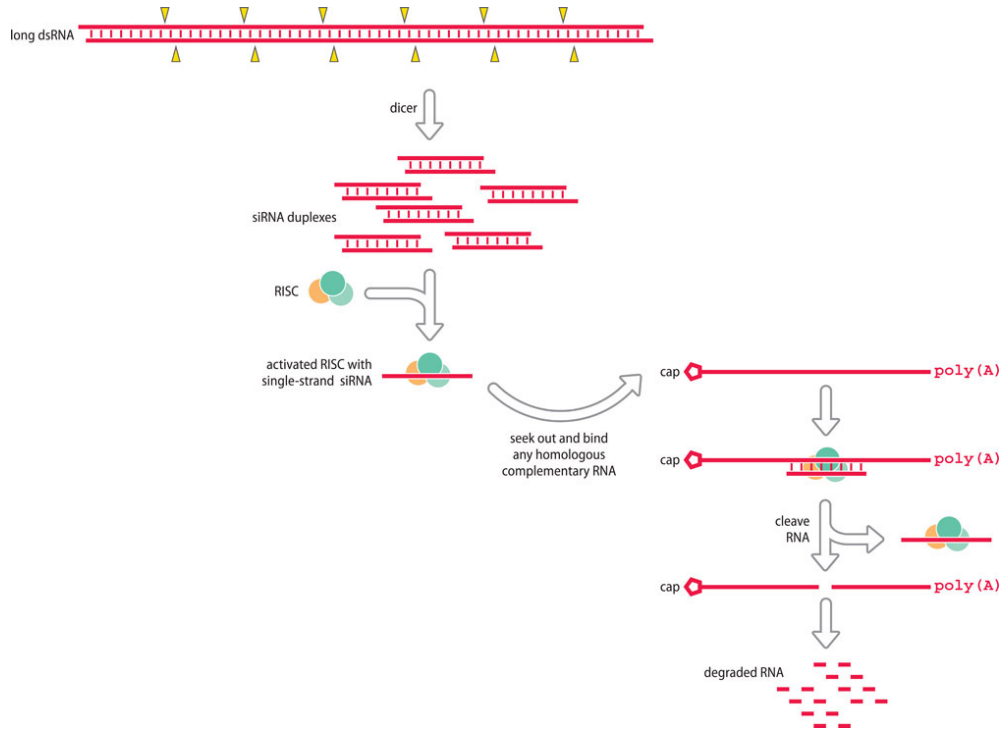


Figure 6.13 RNA interference (RNAi). Long double-stranded (ds) RNA is an unusual structure in our cells, but may signal the presence of some invading viruses or excess transposon activity. To defend cells against these threats, Dicer, a cytoplasmic ribonuclease, cleaves dsRNA asymmetrically on the two strands at positions 21 nucleotides apart (yellow triangles). The resulting short interfering RNA (siRNA) consists of a duplex of two 21-nucleotide sequences with overhangs of two nucleotides at the 3' ends. The siRNA duplexes are bound by RNA-induced silencing complexes (RISC) that degrade one of the siRNA strands to give an activated RISC complex with a single siRNA strand (called the *guide RNA*). The RISC-siRNA complex is now activated and will bind (by RNA–RNA base pairing) to any RNA sequence that is complementary in sequence to the guide RNA (such as a specific viral mRNA sequence). The cleaved RNA fragments, lacking a protective cap or poly(A) sequence, are vulnerable to attack by cellular exonucleases and are rapidly degraded. Note that introducing long dsRNA into mammalian cell culture results in the indiscriminate destruction of mRNA, so siRNAs need to be short, often 21 nucleotides long.

A different approach, *disease prevention*, seeks to *prevent* the buildup of toxic levels of metabolite. In some cases, *substrate restriction* is used: the diet is modified to severely

reduce or eliminate the intake of a substrate of the deficient enzyme. That can work very successfully when the blocked enzyme is at the start of a pathway that metabolizes a dietary component. This approach prevents the harmful effects of toxic metabolites that build up in phenylketonuria (as described in Clinical Box 9 on page 234). Even in this case, the treatment requires lifelong compliance with a rather restricted, and difficult, diet. Prevention is sometimes possible at the prenatal level, as in the case of treating virilized female fetuses in 21-hydroxylase deficiency ([Figure 9.3A](#)).

Mixed success in treatment

Treatment of some inborn errors of metabolism is very effective, as with phenylketonuria. However, the treatment can be sub-optimal in some cases, and very difficult or essentially nonexistent in others for various reasons. If a disorder is congenital and harmful effects have occurred during development, treatment options may be limited. In some cases, potential therapy can be frustrated by delivery problems. In Tay–Sachs syndrome (PMID 20301397), for example, deficiency in hexosaminidase A leads to an inexorable buildup of a sphingolipid GM2 ganglioside to toxic levels, causing damage to brain cells.

Genetic treatment of disease may be conducted at many different levels

Any disease, whether it has a genetic cause or not, is potentially treatable using a range of different procedures that apply genetic manipulations or genetic knowledge in some way ([Figure 9.5](#)).

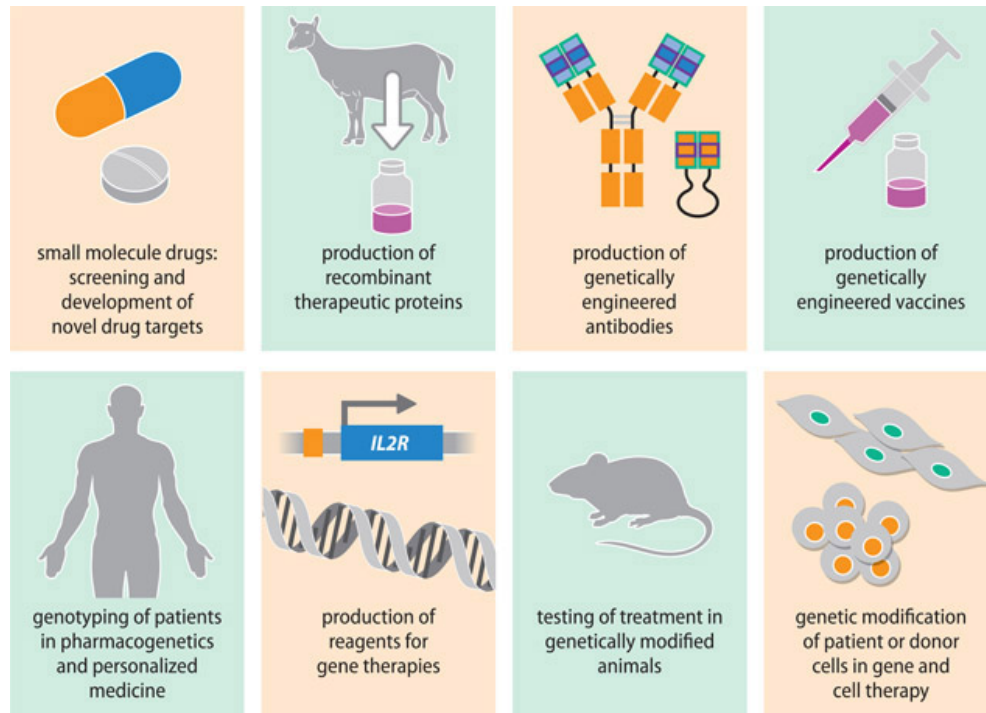


Figure 9.5 Some of the many different ways in which genetic technologies are used in the treatment of disease.

Sometimes genetic techniques form part of a treatment regime that also involves conventional small molecule drugs or vaccines.

Pharmacogenetics is concerned with how the actions of drugs and the reactions to them vary according to variation in the patient’s genes. Genotyping of individuals might then be used to predict patterns of favorable and adverse responses to specific drug treatments. Such genotyping may become routine as massively parallel DNA sequencing permits extensive screening of genes in vast numbers of people.

New targets for drug development are being identified using a knowledge of genetics and cell biology. Genetic techniques can also be used directly in producing drugs and vaccines for treating disease. Another active area concerns treating disease with therapeutic proteins that are produced or modified by genetic engineering. Genes are cloned and expressed in suitable cultured cells or organisms to make large amounts of a specific protein that is then purified (*recombinant proteins*), including hormones, blood factors, and enzymes, and especially genetically engineered antibodies.

Gene therapies are the ultimate genetic application in treating disease; they rely on genetically modifying the cells of a patient. Delivering therapeutic constructs into the stem cells of a patient is particularly valuable when the disease primarily affects short-lived cells, such as blood cells. Animal models are especially important resources for testing new therapies before they are used in clinical trials. As described in [Section 9.3](#), the vast majority

of animal models of disease have been generated by the genetic manipulation of rodents, notably mice.

9.2 GENETIC INPUTS INTO TREATING DISEASE WITH SMALL MOLECULE DRUGS AND THERAPEUTIC PROTEINS

Chemical treatments for disease are developed by the pharmaceutical industry. Previously they relied almost exclusively on hydrocarbon-based small molecule drugs synthesized by standard chemical reactions. More recently they have been joined by a new class, biological drugs (*biologics*), including therapeutic proteins that have been prepared by genetic engineering. We give a brief description of the two classes below. Then we cover the important question of how genetic variation between people can result in very wide differences in how we respond to therapeutic drugs, and consider aspects of drug metabolism. We finish with examples of therapeutic uses of the two principal drug classes.

Small molecule drugs

The conventional drug discovery process has involved screening huge numbers of small synthetic molecules with hydrocarbon backbones for evidence that they can reduce pathogenic effects. A drug such as this typically works by binding to a specific target protein with a key role in the pathogenesis—often a receptor, ion channel, or enzyme. The drug is able to bind to the protein by fitting into some cleft, groove, or pocket at a key position in the protein structure. Binding of the drug to the target protein may often prevent the protein from interacting with other molecules (and so block its function), or it may change the function of the protein in some way.

The drug screening process normally begins with preclinical assays in cell culture and in animal models to see whether a candidate drug has some encouraging properties. Promising candidates may be used in clinical trials, when their potential usefulness is monitored in different ways (**Figure 9.6**). To bring a drug to market is both costly (about US \$1 billion) and time-consuming (often 12 years or more). Sometimes, however, a drug previously developed to treat one type of disease can be used to treat other diseases (*drug repurposing*); that is valuable because the drug has already been through lengthy and expensive clinical trials to assess its safety profile.

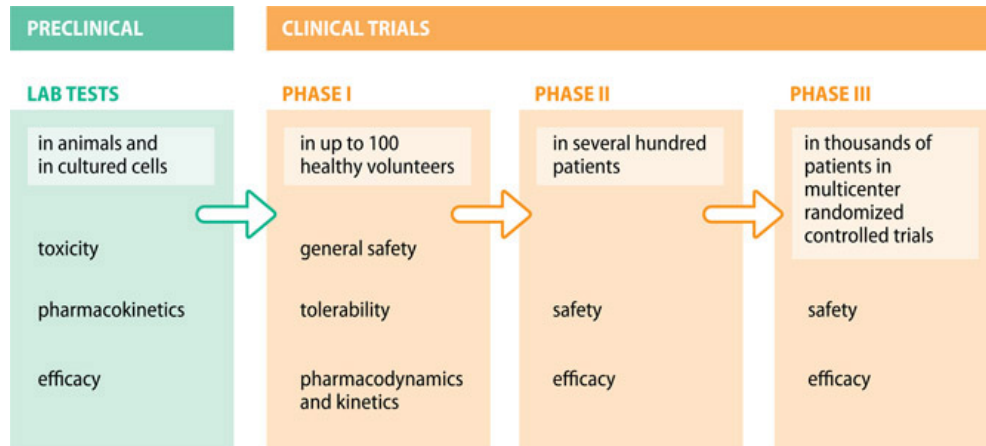


Figure 9.6 Major stages in drug development. The lists indicate the principal parameters tested at each stage. Pharmacokinetic testing assesses the absorption, activation, metabolism, and excretion of drugs. Pharmacodynamics monitors what the drug does to the body. Successive stages toward regulatory approval are increasingly expensive. To avoid unnecessary expenditure, any effects of genetic variation among patients that might influence marketability need to be identified as early as possible in the process.

Small molecule drugs have been with us for some time. Two important questions are: how effective are they, and how safe are they? Although the therapeutic value of many small molecule drugs on the market is questionable, many others have undoubtedly been of great service. But individual drugs affect different people in different ways. As explained below, many of our drug-handling enzymes are polymorphic; genetic variation between individuals has an important influence on both the efficacy and the safety of drugs.

All drugs currently on the market act through only a few hundred target molecules (they were first developed when information about possible targets was scarce), and the declining number of new drug applications and approvals over the past few years has reflected a crisis in drug target identification and validation.

New approaches

Recently, both genomics and genetic engineering have been making an impact on drug development. Genomic advances offer a broader perspective on how genetic variation affects drug metabolism, and the ways in which people respond to drugs. They are also providing additional potential drug targets, as described below.

Genetic engineering has been applied to allow the production of large quantities of different biopharmaceuticals (or *biologics*), including therapeutic “recombinant” proteins and genetically engineered monoclonal antibodies. Many of these are currently licensed to treat various disorders.

An overview of how genetic differences affect the metabolism and performance of small molecule drugs

Small molecule drugs have been with us for some time, and accumulated data tell us how effective they are and how safe they are. Taking the first point, drugs vary widely in effectiveness. Even when a drug receives regulatory approval, it is rarely effective in 100 % of the patients to whom it is prescribed. Some people might need higher or lower doses to achieve the same therapeutic effect. For others, a drug might simply have no therapeutic effect at all.

Most of us have benefited from effective antibiotics and painkillers, and drugs such as statins and beta-blockers (for reducing the risk of heart disease) have been of very considerable value. Certain other drugs, especially those used to treat psychiatric disease, are much less effective. That can mean wasted time and money, and extra suffering for patients. Then there is the enormous problem concerning adverse reactions to drugs, as described below.

Some of these differential effects are due to environmental causes: a person's ability to absorb or metabolize a drug may be changed by illness or lifestyle. Sometimes adverse effects occur as a result of interactions between different combinations of drugs taken by a sick person. But many differences are due to genetic variation between people, and genetic variation in our drug-handling enzymes is often pronounced.

Natural selection fosters genetic variation in the genes encoding our drug-handling enzymes because a major role of these enzymes has been to deal with unusual exogenous chemicals (*xenobiotics*) in our diet and environment. Like immune system molecules that recognize foreign antigens, drug-handling enzymes need to deal with potentially dangerous invaders some of which are subject to genetic control (such as ingested plant or fungal metabolites that might be toxic). The genetic variation in drug-handling enzymes that affects how the body deals with drugs is simply a small part of how we respond to the extraordinary range of chemicals encountered in our diet and environment.

Pharmacogenetics is the study of the roles of specific genes in these effects (or *pharmacogenomics*, when taking a genome-wide perspective of the role of genetic variation). Genetic factors affect both drug metabolism and effect, as studied by:

- **Pharmacokinetics**—the study of what the body does with, and to, the drug (encompassing drug absorption, activation, metabolism, and excretion).
- **Pharmacodynamics**—the study of the target response: how the person is affected by the drug.

Different stages at which genetic variation influences drug metabolism

Drug transport and metabolism within the body involve a series of stages. The pharmacological effect of the drug will be measured by its effect on cells within the desired *target tissue* or organ, but drug delivery is mostly nonspecific. Usually cells throughout the body are exposed to the drug so that the drug can exert a beneficial effect on what often might be a small group of desired target cells. The circulation is the delivery route. That might mean direct intravenous injection, but most drugs are given orally or through intramuscular injection. After oral ingestion, for example, drugs are transported from intestinal epithelial cells into the portal vein and from there to the liver, then through hepatic veins to the heart for general distribution in the bloodstream.

Only a small proportion of a given drug dose will be responsible for producing a specific pharmacological effect—most of the drug will be broken down by metabolizing enzymes (principally within the liver) or excreted unchanged. Although some drug transport is passive, many of the different drug-handling events require dedicated proteins whose function is to transport specific drugs into or out of cells. Within the target tissue are receptor proteins and other proteins in a biological pathway that the drug interacts with to produce the pharmacological effect.

Genetic factors are implicated in the variation between individuals in drug metabolism at each of the different levels: drug absorption (differences in the ability to transport a drug into the bloodstream); drug activation (differences in the ability of liver enzymes to convert a prodrug into the active drug); target response (differences in how the targeted process or pathway responds to a given local concentration of the drug); and catabolism and excretion (differences in the rate at which the drug is catabolized and disposed of).

Phase I and phase II reactions in drug metabolism

Drug metabolism is a defense mechanism: it facilitates excretion of the parent drug and its metabolites, and so limits their ability to accumulate within the body and cause dose-dependent toxicity. It is mostly carried out in the liver (which has multiple enzymes that are responsible for detoxifying drugs and assisting in their excretion), but significant drug metabolism also occurs in some other sites, such as the intestines and kidneys. Small molecule drugs are based on hydrocarbon backbones and so are lipophilic, but drug metabolism allows them to be converted into hydrophilic forms that are easier to excrete from the body.

Phase I reactions are usually carried out by monooxygenases; these work by adding an oxygen atom from molecular oxygen to produce a more polar substance ([Figure 9.7](#)). Often a hydroxyl group is introduced, or a bulky alkyl group bound to a nitrogen, sulfur, or oxygen atom is replaced by a hydrogen atom. The drug derivative is typically left with a more reactive group, a molecular “handle” that makes it easier for a secondary reaction to be carried out (see below). Sometimes, a phase I reaction also results in *drug activation*,

converting an inactive form of a drug, a **prodrug**, into an active drug. For example, the painkiller codeine (methyl-morphine) is a prodrug that is converted by a phase I enzyme in the liver into an active form, morphine.

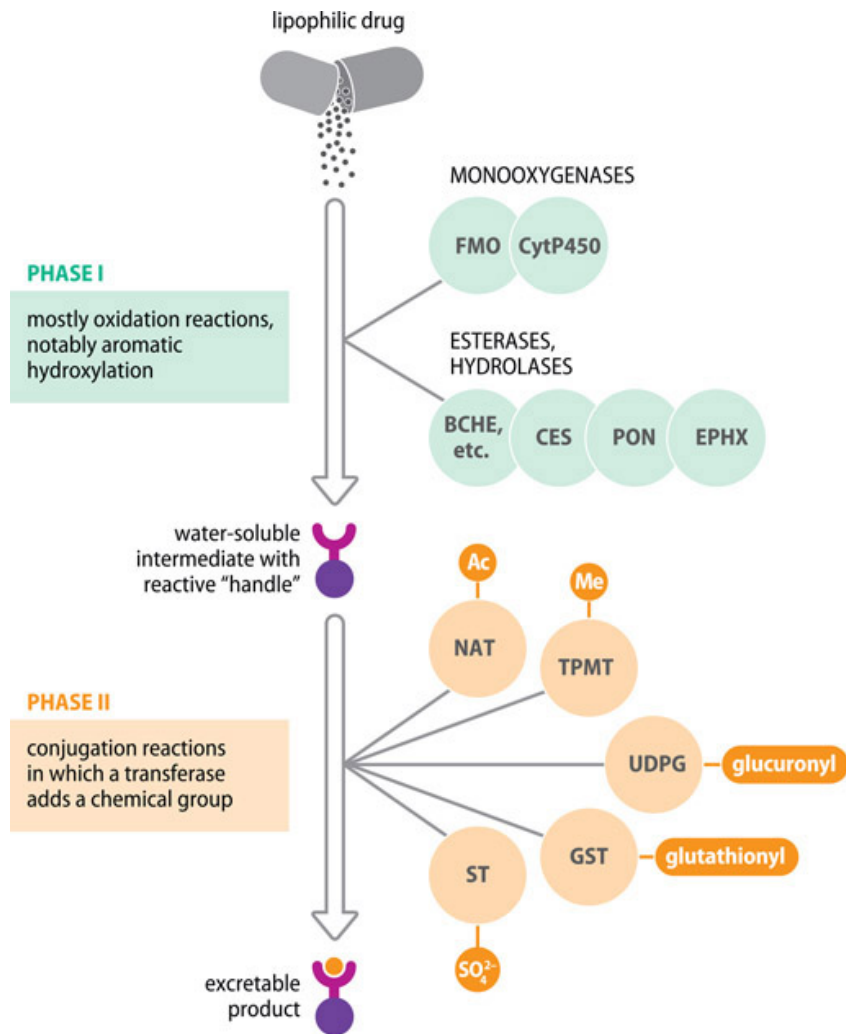


Figure 9.7 Two major stages in drug metabolism. Phase I drug reactions typically result in a more polar drug derivative with a reactive group, a molecular “handle” that makes it easier to accept a chemical group donated by a phase II enzyme. Phase I enzymes are often monooxygenases, notably cytochrome P450 enzymes (CytP450), but also include various other enzymes, notably esterases and other hydrolases. In phase II drug metabolism one of a variety of different transferase enzymes adds a chemical group that facilitates excretion. Note that the sequence shown here occurs commonly, but phase II reactions can sometimes occur without a previous phase I reaction. BCHE, butyrylcholinesterases; CES, carboxyesterases; EPHX, epoxide hydrolases; FMO, flavin-containing monooxygenases; GST, glutathione S-transferase; NAT, *N*-acetyltransferase; PON, paraoxonases; TPMT, thiopurine methyltransferase; UDPG, UDP glucuronosyltransferases; ST, sulfotransferases.

Phase II reactions are conjugation reactions, catalyzed by transferases that add one of a variety of chemical groups. As illustrated in [Figure 9.7](#), phase II reactions commonly occur after phase I reactions have introduced a molecular handle for attaching the secondary chemical group. A hydroxyl group attached during phase I, for example, provides a convenient site for an acetyl group or a sugar (glucuronyl) group to be attached by a phase II enzyme, detoxifying the drug and assisting in its excretion.

Phenotype differences arising from genetic variation in drug metabolism

All drugs work optimally within a certain **therapeutic window**, a range of concentrations within which the therapeutic benefit is optimal without posing any great risks to health. If the concentration is below this range, the therapeutic benefit might be insufficient (drug underdose); if above this range, there is an increasing risk of toxicity (drug overdose).

When drugs are detoxified by metabolism, the concentration of active drug falls; repeated drug doses are required to maintain drug concentrations within the safe therapeutic range. The speed at which a drug is metabolized has consequences for both *drug efficacy*—the degree to which the drug gives therapeutic benefit—and safety. Individuals who eliminate or inactivate drugs comparatively slowly (“slow metabolizers”) will have a longer or stronger response to a given concentration of the drug than fast metabolizers will. They can be at risk of a drug overdose if given the usual dose ([Figure 9.8](#)). They may also be more at risk of adverse reactions if breakdown products from the drug are toxic. Ultrafast metabolizers might gain little therapeutic benefit from a drug (see [Figure 9.8](#)).

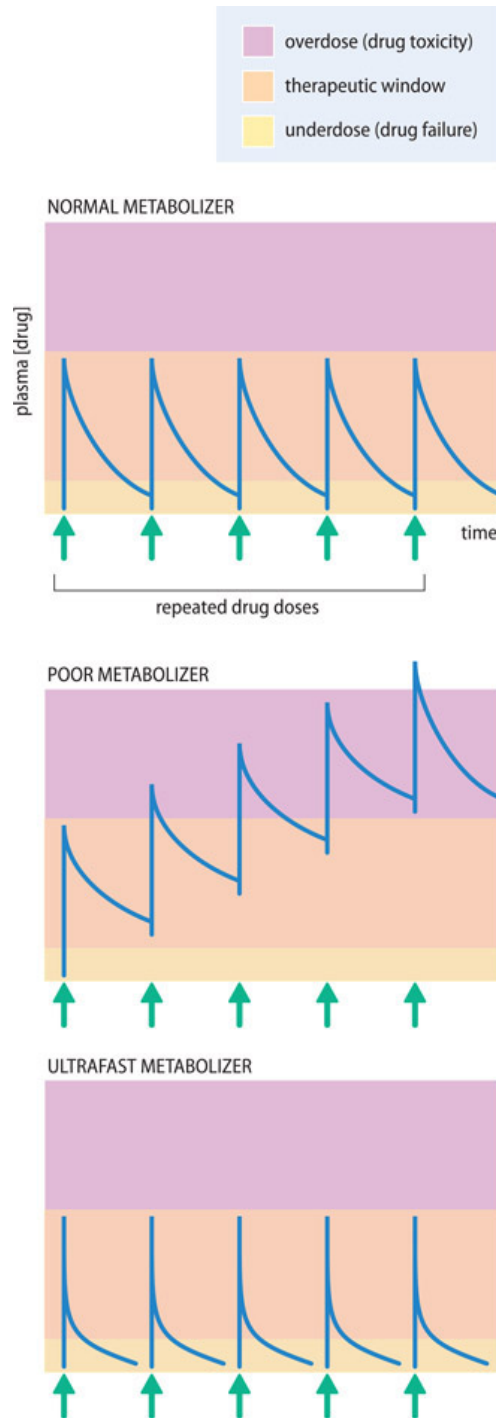


Figure 9.8 The effect of different drug metabolizing rates on plasma drug concentration. The *therapeutic window* is the range of plasma drug concentrations that are of therapeutic benefit without causing extra safety risks due to drug toxicity. Normal metabolizers are expected to benefit from having drug concentrations in the therapeutic window for long periods. If given the normal drug concentration, poor metabolizers can suffer from an overdose: the failure to metabolize the drug quickly means that the drug concentration progressively increases to very high, unsafe levels after repeated doses. Ultrafast metabolizers might gain little therapeutic

benefit, because the drug is rapidly cleared from the plasma after each drug dose. See [Figure 9.9](#) for specific examples of different classes of drug metabolism.

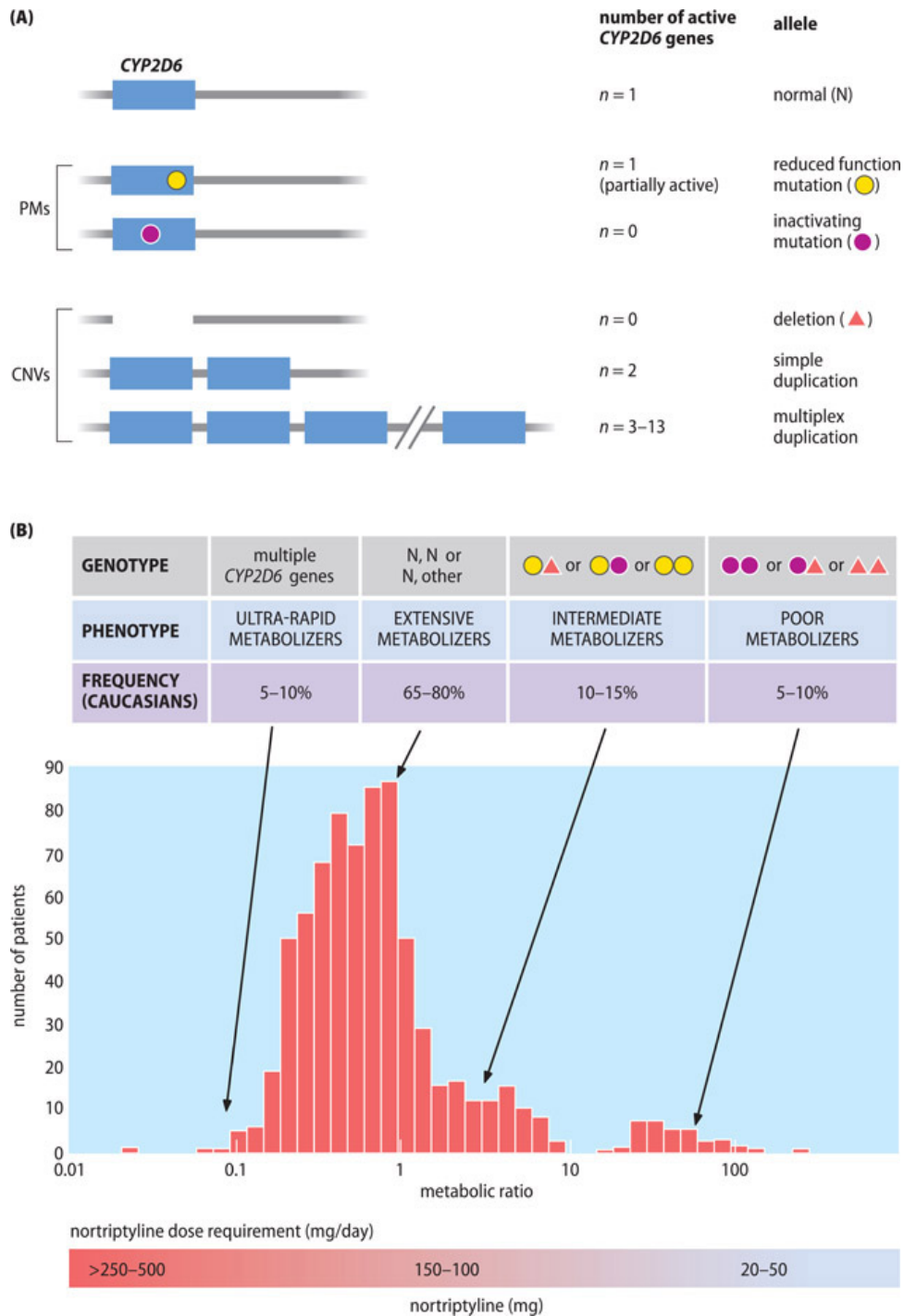


Figure 9.9 *CYP2D6* allele classes and correlation between genotypes and drug-metabolizing abilities. (A) *CYP2D6* allele classes. Variation includes both common point mutations (PMs) and copy number variants (CNVs). Multiple *CYP2D6* genes are quite common in some populations, most probably as a result of natural selection (much like the gene amplification giving rise to multiple α -amylase genes in some populations, as

described in [Figure 4.8](#)). (B) *CYP2D6* genotypes and drug-metabolizing ability. The histogram in the middle panel shows a range of *CYP2D6*'s drug-metabolizing ability (metabolic ratio) in a study group. To assay *CYP2D6* activity, a urinary metabolic ratio is used: after a standard drug dose, the urinary concentration of the substrate drug (debrisoquine in this case) is measured and divided by that of its metabolic product. High ratios show poor conversion as a result of low enzyme activity. The upper panel shows the four classes of metabolizer, their frequencies (in Caucasian populations) and how phenotype relates to genotype. Low metabolizers are at risk of drug overdose and should be given lower drug doses. As an example, the lower panel shows recommended doses of the *CYP2D6*-metabolized antidepressant nortriptyline, arranged on a sliding scale according to the metabolic ratio shown above. (Adapted from Meyer UA [2004] *Nature Rev Genet* 5:669–676; PMID 15372089. With permission from Macmillan Publishers Ltd.)

Individual drugs can be metabolized by multiple different enzymes encoded by different genes; as well as genetic variation in different genes, environmental factors contribute significantly to how a drug is metabolized. As a result, drug response phenotypes are often multifactorial. Sometimes, however, a specific drug might be metabolized by just one enzyme; genetic variation in the gene making that enzyme can have a predominant contribution to the phenotype—we provide examples in the next sections.

Genetic variation in cytochrome P450 enzymes in phase I drug metabolism

The great majority of phase I drug reactions are carried out by monooxygenases, notably heme-containing enzymes in the cytochrome P450 superfamily (they have in common a spectral absorption peak at 450 nm). Cytochrome P450 enzymes catalyze different types of reaction. In addition to specific drugs, their substrates can be endogenous chemicals, notably certain steroids, and xenobiotics in our food and in the environment.

We have >110 different cytochrome P450 genes, classified into 18 families and multiple subfamilies. Thus, for example, the *CYP2C19* gene gets its name from cytochrome P450 family 2, subfamily C, polypeptide 19. Six cytochrome P450 enzymes catalyze 90 % of the phase I reactions on commonly used drugs. *CYP3A4* is involved in metabolizing ~40 % of all drugs; *CYP2D6* is another prolific drug-handler. Individual drugs are also often substrates for more than one P450 enzyme. The antidepressant amitriptyline, for example, can be metabolized by each of *CYP1A2*, *CYP2C19*, and *CYP2D6*.

Specific cytochrome P450 enzymes can be induced or inhibited by certain drugs. That can result in unexpected interactions between these drugs and those that are substrates of the enzyme. Comprehensive drug interaction lists are maintained in the Drug Interactions Flockhart Table (available at <https://drug-interactions.medicines.uoi.edu/MainTable.aspx>).

The wide and overlapping specificities of cytochrome P450 enzymes mean that it can often be difficult to correlate how a person metabolizes a specific drug with the activity of any one P450 enzyme. Nevertheless, some drugs are metabolized by just one P450 enzyme, and DNA variation in just a single gene can result in wide differences between people in how they

metabolize the drug. Often the variation is due to simple mutations that change key amino acids or that inactivate gene expression, but occasionally excess gene activity occurs when there are multiple copies of the same cytochrome P450 gene.

Genetic variation in CYP2D6 and its consequences

The *CYP2D6* gene has more than 100 genetic variants, with a continuum of enzyme activity; it is a prime example of how different types of genetic variation in a single enzyme can have a marked effect on the metabolism of certain drugs. As a result of severe inactivating mutations or deletions in both *CYP2D6* alleles, rare individuals have a very low activity of the enzyme. When treated with certain drugs normally metabolized by this enzyme alone, they fail to metabolize and excrete the drug (with high plasma levels for the drugs and low levels of expected catabolic products in urine samples).

Depending on their CYP2D6 activity, people vary in their ability to metabolize drugs for which this enzyme has the predominant role). Four classes of metabolizers are recognized: poor metabolizers (who lack normal alleles, and are comparatively frequent in Caucasian populations); intermediate metabolizers; extensive metabolizers (with one or two active *CYP2D6* alleles); and ultrafast metabolizers (having multiple *CYP2D6* genes as a result of gene amplification) ([Figure 9.9](#)).

People with very low CYP2D6 activity can show unusually marked sensitivity to certain drugs; they are also at risk of drug overdose if prescribed with normal doses of certain beta-blockers and tricyclic antidepressants. Because CYP2D6 is also the enzyme that converts codeine to morphine, people with very low CYP2D6 activity also get minimal painkilling benefit from codeine. Ultrafast metabolizers may also get little benefit from drugs principally metabolized by CYP2D6 (because they metabolize and detoxify the drugs so quickly).

Genetic variation in other cytochrome P450 enzymes

CYP3A4 activity in the liver shows extensive variability between individuals, but unlike for CYP2D6 there are very few coding sequence variants. And unlike CYP2D6, CYP3A4 is inducible; regulatory mutations are thought to be significant contributors to the variability.

CYP3A4 is highly related to CYP3A5 (and to CYP3A7, which is normally expressed only at fetal stages); the drug-metabolizing activities of CYP3A4 and CYP3A5 strongly overlap, complicating matters. However, CYP3A5 is less biologically active than CYP3A4: in Caucasian populations *CYP3A5* null alleles predominate and only 10 % of the alleles make an active enzyme.

CYP2C9 deficiency is important in metabolizing the anticoagulant warfarin (as described below), and also results in an exaggerated response to tolbutamide, a hypoglycemic agent used in treating type 2 diabetes. CYP2C19 deficiency shows marked differences between

different ethnic groups, as revealed by the frequency of poor metabolizers ([Table 9.2](#)); poor metabolizers require a lower dose of certain drugs such as clopidogrel, an antiplatelet agent that is used to inhibit blood clotting in some diseases such as coronary artery disease.

TABLE 9.2 SIGNIFICANT POPULATION DIFFERENCES IN THE FREQUENCY OF POOR METABOLIZERS OF CYP2D6 AND CYP2C19

Population origin	Frequency (%) of CYP2D6 poor metabolizers	Frequency (%) of CYP2C19 poor metabolizers
Amerindian	0	2.0
Caucasian	7.2	2.9
East and South East Asian*	0.5	15.7
Middle East and North African	1.5	2.0
Polynesian	1.0	13.6
Indian subcontinent (Sri Lankan)	0	17.6
Sub-Saharan African	3.4	4.0

* Excluding the Indian subcontinent. (Data from Burroughs VJ, Maxey RW & Levy RA [2002] *J Natl Med Assoc* 94(10 Suppl): 1–26; PMID 12401060.)

Genetic variation in enzymes that work in phase II drug metabolism

Aromatic *N*-acetylation is a frequent type of phase II drug metabolism. Two types of *N*-acetyltransferase, NAT1 and NAT2, deal with different sets of drugs. Whereas NAT1 is comparatively invariant, NAT2 is polymorphic in a wide range of human populations, with rapid acetylators (who eliminate drugs rapidly) and slow acetylators (who have low NAT2 levels).

Variation in acetylating abilities of NAT2 can be clearly seen as a bimodal distribution using the drug isoniazid, which is used to treat tuberculosis ([Figure 9.10](#)). The proportion of slow acetylators is highly variable between different ethnic groups but can be very high in some populations, notably in Caucasian populations ([Table 9.3](#)).

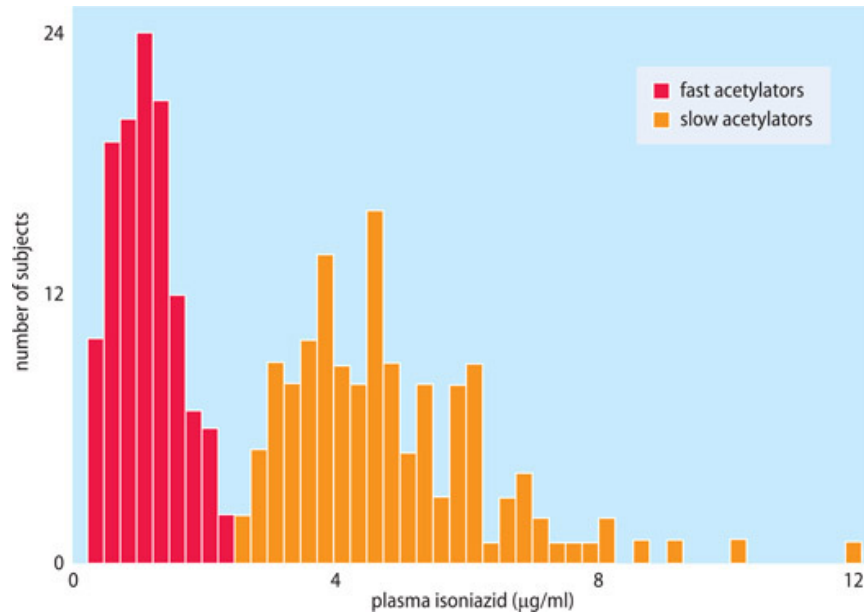


Figure 9.10 A bimodal distribution of plasma levels of isoniazid as a result of genetic polymorphism in the *NAT2* (*N*-acetyltransferase 2) gene. Plasma concentrations were measured in 267 normal subjects 6 hours after an oral dose of isoniazid. Fast acetylators removed the drug rapidly. The number of slow acetylators (presumed to be homozygotes or compound heterozygotes for severe inactivating mutations) was almost the same as the number of fast acetylators. This suggests that only about 30 % of *NAT2* alleles in the study group were producing active enzyme. (Adapted from Price Evans DA, Manley KA & McKusick VA [1960] *Br Med J* ii:485–491. With permission from BMJ Publishing Group Ltd.)

TABLE 9.3 SIGNIFICANT POPULATION DIFFERENCES IN THE FREQUENCY OF *NAT2* SLOW ACETYLATORS

Population origin	Frequency (%) of <i>NAT2</i> slow acetylators
Caucasian	58
Chinese	22
Eskimo	6
Japanese	10
Sub-Saharan African	51

Data from Wood AJ & Zhou HH [1991] *Clin Pharmacokinet* 20:350–373; PMID 1879095.)

The slow acetylators take longer to eliminate drugs (and other xenobiotics) and so often show enhanced sensitivity to drugs metabolized by *NAT2*; they also appear to be more susceptible to certain cancers, notably bladder cancer. Nevertheless, natural selection appears to have driven an increase in frequency of the slow-acetylator phenotype in some populations. A possible explanation is that certain chemicals in well-cooked meat are converted by *NAT2* into carcinogens; individuals with slow-acetylator phenotypes would be

comparatively protected in populations that have had a long tradition of eating well-cooked meat.

Variation in other phase II enzymes is also significant. Polymorphism in thiopurine methyltransferase is a particular clinical concern when using certain immunosuppressant drugs such as 6-mercaptopurine, which is commonly used to treat childhood leukemia. (About 1 in 300 children do not express thiopurine methyltransferase; for them 6-mercaptopurine is toxic.) The glutathione S-transferase (GST) superfamily includes some enzymes such as *GSTM1* and *GSTT1* that are encoded by genes susceptible to gene deletion via unequal crossover. As a result, inactive alleles are very common (about 50 % of the *GSTM1* alleles in people of northern European ancestry are gene deletions, for example). People with consequently low levels of these enzymes find it difficult to cope with high doses of drugs that are processed by them.

The UDP glucuronosyltransferase superfamily includes polymorphic enzymes involved in metabolizing different substrates, including drugs used in cancer chemotherapy. For example, the prodrug irinotecan is converted into an active anti-tumor form in the liver (it inhibits DNA topoisomerase, an enzyme needed for DNA replication) and is normally processed by the enzyme *UGT1A1*. The common *UGT1A1**28 promoter polymorphism results in reduced production of this enzyme, and is frequent in many populations (1 in 3 people are homozygotes in sub-Saharan Africa, and 1 in 5 in the Indian subcontinent). The *28/*28 homozygotes have a much higher risk of serious bone marrow and gastrointestinal toxicity.

Altered drug responses resulting from genetic variation in drug targets

In the sections above we have considered genetic variation in enzymes performing phase I and phase II drug metabolism, and how that variation affects drug pharmacokinetics (including how fast a drug is metabolized and excreted). In this section we consider how genetic variation affects the pharmacodynamics of drugs.

The efficacy of a drug is partly determined by genetic variation in *drug targets*, the molecules that the drug must interact with in the cells of the target tissue. Drug targets will typically include receptors, signaling molecules, and other molecular components of biological pathways that the drug interacts with to have its pharmacogenetic effect.

Genes encoding drug receptors quite often show polymorphisms or variants that lead to clinically significant altered responses to drugs. Examples of clinically significant genetic variation in drug receptors include variants of beta-adrenergic receptors, cell surface receptors that have central roles in the sympathetic nervous system. Two of these receptors, *ADRB1* and *ADRB2*, are widely used as drug targets in therapeutic approaches for various common and important diseases including asthma, hypertension, and heart failure. Genetic variation in both *ADRB1* and *ADRB2* has been linked to altered responses to drugs. Other examples include variation in the H2RA serotonin receptor and the RYR1 ryanodine receptor ([Table 9.4](#)).

TABLE 9.4 EXAMPLES OF HOW GENETIC VARIATION IN DRUG TARGETS CAUSE ALTERED RESPONSES TO THERAPEUTIC DRUGS

Drug target	Function	Polymorphism or variant	Example of drug treatment	Effect of polymorphism or variant
ACE	angiotensin-converting enzyme	Alu repeat indel polymorphism in intron of <i>ACE</i> gene	use of ACE inhibitors—captopril and enalapril—to treat heart failure	drugs are more effective in Alu ⁻ /Alu ⁻ homozygotes
ADRB1	β ₁ adrenergic receptor	common R389G polymorphism	beta-blockers, such as bucindolol, for reducing heart disease risk	reduced cardiovascular response to drugs
ADRB2	β ₂ adrenergic receptor	common R16G polymorphism	albuterol for treating asthma	homozygotes are much less likely to respond to treatment*
RYR1	ryanodine receptor	different mutations	inhalation anesthetics	potentially fatal (see Clinical Box 11)

* Note: the R16G polymorphism has significant effects on short-acting agonists but little effect on long-acting agonists, which now constitute the more widely used treatment.

Some therapeutic drugs are designed to specifically inhibit key enzymes that have pivotal roles in biological pathways underlying common diseases. Examples include statins, which were designed to inhibit HMG CoA reductase (lowering cholesterol levels, and so reducing blood pressure and the risk of cardiovascular disease); warfarin, an inhibitor of a key enzyme in the maturation of several blood-clotting factors (see the next section); and inhibitors of angiotensin-converting enzyme (ACE). The last enzyme is also an important regulator of blood pressure (and several other functions), and an insertion/deletion polymorphism due to the variable presence of an Alu repeat in intron 15 of the *ACE* gene is associated with variation in ACE activity. People who are homozygous for the Alu deletion allele have about twice the enzyme activity of those who have two Alu insertion alleles; this difference is thought to be an important contributor to variable responses to ACE inhibitors (see [Table 9.4](#)).

We give a summary of serious adverse drug reactions in [Clinical Box 11](#). In most cases, the genetic variation associated with the effect is primarily confined to a single locus. However, for many drugs genetic variation at multiple loci might be important. One common example where we know some of the details concerns treatment with the anticoagulant warfarin.

CLINICAL BOX 11 WHEN PRESCRIBED DRUGS CAN BE DANGEROUS AND SOMETIMES DEADLY, DEPENDING ON A PATIENT'S GENOTYPE

Therapeutic drugs and other drugs administered in medical procedures (such as anesthetics) can produce extreme responses in some people. Adverse drug reactions are very common, being responsible for a significant proportion of all hospital admissions (nearly 7 % in a UK study) and can result in disability or permanent damage, birth defects, and an extraordinary number of fatalities (about 100 000 deaths each year in the US alone).

Adverse drug reactions have various causes. Type A reactions are relatively common and dose-dependent; they are predictable from the drug pharmacology and are usually mild. Type B reactions are idiosyncratic reactions that are not related simply to drug dose; they are rare but can often be severe ([Table 1](#)). Genetic variants are important in both types of reaction.

TABLE 1 SOME CLASSES OF SEVERE TYPE B ADVERSE DRUG REACTIONS

Drug-induced injury or toxicity	Associated drugs (examples)	Comments and examples
Apnea (respiratory paralysis)	suxamethonium (succinyl choline)	this drug works as a fast-acting muscle relaxant and is used before surgery. Normally the effects of the drug wear off quite quickly when the drug is metabolized by the enzyme butylcholinesterase. Low metabolizers are at risk of apnea—they remain paralyzed and unable to breathe after surgery because they cannot regain their muscle function quickly enough and may require extended ventilation
Prolongation of cardiac QT interval	thioridazine, clarithromycin, terfenadine	induced by many different drugs. Associated with polymorphic ventricular tachycardia, or torsades de pointes (see Figure 1), which can be fatal
Hematologic toxicity	6-mercaptopurine, azathioprine	thiopurine S-methyltransferase (TPMT) inactivates these immunosuppressant drugs by adding a methyl group. In people with two low-activity TPMT alleles, the drugs are metabolized slowly; if normal doses are given, the drugs accumulate and can result in life-threatening bone marrow toxicity
Hemorrhage	warfarin	see text

Drug-induced injury or toxicity	Associated drugs (examples)	Comments and examples
Hypersensitivity reactions	abacavir, carbamazepine, allopurinol	inappropriate immune reactions to otherwise nontoxic drugs can have broad manifestations. When treated with the anti-HIV drug abacavir, about 5 % of patients demonstrate skin, gastrointestinal, and respiratory hypersensitivity reactions that can sometimes be fatal. Treatment with the anticonvulsant carbamazepine or allopurinol (used in treating gout) can induce cutaneous adverse drug reactions, including toxic epidermal necrolysis
Liver injury	flucloxacillin, isoniazid, allopurinol	individuals with certain HLA antigens are at increased risk of induced liver disease for some drugs, such as <i>HLA-B*5701</i> in the case of flucloxacillin
Muscle toxicity	halothane, isoflurane, various statins	statins and several other drugs are associated with usually mild myopathies. Sometimes, however, more severe cases show rhabdomyolysis (breakdown of muscle tissue) that can result in death. In response to inhalation anesthetics (halothane, isoflurane), individuals who have inactivating mutations in the ryanodine receptor gene develop life-threatening rhabdomyolysis and an extreme rise in temperature, a form of malignant hyperthermia (OMIM 145600)

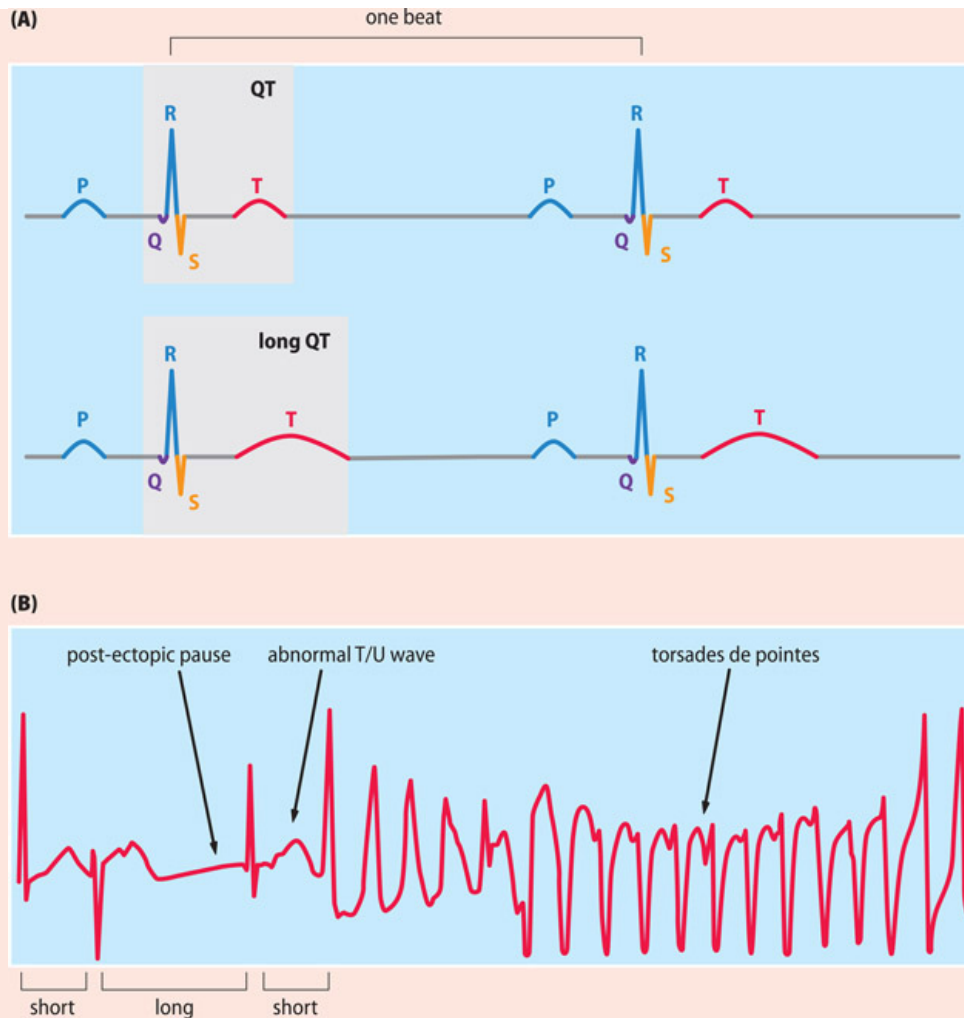


Figure 1 Drug-induced prolongation of the cardiac QT interval and torsades de pointes. (A) Cardiac depolarization–repolarization cycle. Specific repeated features are labeled from P to T. The *QT interval*, the shaded interval that spans the onset of the QRS complex until the end of the T wave, represents the time taken for one complete cycle of ventricular depolarization and repolarization. Certain drugs can prolong the QT interval (*long QT*), and this can sometimes induce a rapid beating of the heart, which often manifests itself as torsades de pointes (TdP). (B) Cardiac rhythm profile in a patient with drug-induced TdP. Notice the short–long–short initiating ventricular cycle, pause-dependent long QT interval, and abnormal TU wave preceding the development of TdP. This type of ventricular arrhythmia can self-terminate, but it can also degenerate into potentially fatal arrhythmias such as ventricular fibrillation. (B, Adapted from Yap YG & Camm AJ [2003] *Heart* 89:1363–1372; PMID 14594906. With permission from BMJ Publishing Group Ltd.)

When genotypes at multiple loci in patients are important in drug treatment: the example of warfarin

Warfarin is prescribed for patients at risk of developing clots within blood vessels (thrombosis), including clotting that can block arteries (embolism). Delivering the optimal warfarin dosage is clinically very important because there is a narrow therapeutic window: if the administered warfarin level is too low, the patient remains at risk of thrombosis and embolism; if it is too high, there is a risk of life-threatening hemorrhage. The final warfarin dose is critical, but because of genetic variation the optimal dose varies enormously between individuals.

Chemically, warfarin is a mixture of two isomers. The (*S*)-warfarin isomer is three to five times as potent as the (*R*)-isomer, has a shorter half-life, and is metabolized predominantly by CYP2C9. Two common polymorphisms in CYP2C9 result in a large decrease in enzyme activity, down to 12 % of wild-type activity for CYP2C9*1, and to just 5 % of normal enzyme activity for CYP2C9*3. Patients with one or two of these common alleles are at increased risk of hemorrhage (presumably because drug metabolism takes longer). It soon became clear, however, that these polymorphisms could explain only a part of the genetic variation in the final warfarin dose.

In 2004 the drug target of warfarin was found to be vitamin K epoxide reductase complex subunit 1 (VKORC1), an enzyme that converts oxidized vitamin K to its reduced form; afterward, association studies showed that genetic variation in VKORC1 was associated with variation in the final warfarin dose. Vitamin K is an indispensable cofactor for the enzyme that converts inactive clotting proteins to give four of our blood clotting factors (factors II, VII, IX, and X). By inhibiting vitamin K epoxide reductase, warfarin inhibits the recycling of vitamin K; the consequent decreased supply of vitamin K inhibits the formation of these four clotting factors.

Subsequently a genomewide association study also implicated a common V433M polymorphism in CYP4F2, a cytochrome P450 enzyme that works as a vitamin K oxidase ([Figure 9.11](#)). Variations in VKORC1 and CYP2C9 remain the largest genetic determinants, accounting for about 40 % of the variation in the final warfarin dose. However, other factors, such as aging and the simultaneous administration of other medicines, also have a very significant influence on the dose of warfarin that is prescribed.

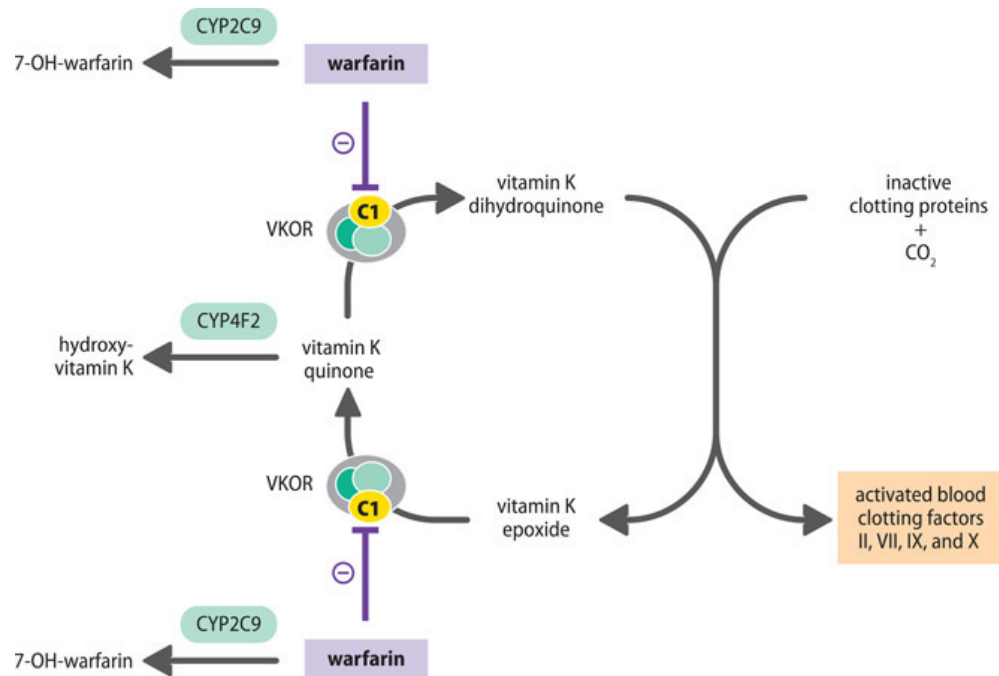


Figure 9.11 Roles of warfarin in anticoagulation and genetic variants affecting final warfarin drug dose.

Warfarin is a therapeutic anticoagulant prescribed for people at risk of thrombosis and embolism. It works by inhibiting VKORC1, the C1 subunit of the vitamin K epoxide reductase complex (VKOR). VKOR is a precursor of vitamin K dihydroquinone, which activates four blood clotting factors: factors II, VII, IX, and X. When VKORC1 is inhibited, the supply of vitamin K is decreased, resulting in a reduced supply of activated clotting factors II, VII, IX, and X. In addition to VKORC1, genetic variation in at least two cytochrome P450 enzymes is known to be associated with variation in the final warfarin dose needed: CYP2C9 converts warfarin to an inactive form, 7-hydroxywarfarin, and CYP4F2 metabolizes vitamin K quinone.

Translating genetic advances: from identifying novel disease genes to therapeutic small molecule drugs

The quest to define the molecular basis of genetic disorders has identified many previously unknown genes, often after labor-intensive studies. Time-consuming studies are then needed to work out how the genes work normally, and in disease states (as mimicked using cultured cells or animal models). Thereafter, therapies can be developed to tackle the *cause* of the disease rather than just the symptoms.

As we will see later in this chapter, gene therapy is becoming an attractive option for treating the cause of several genetic diseases (notably recessive monogenic disorders, as described later in this chapter). An alternative possibility, however, is to develop therapeutic small molecule drugs that possess aromatic hydrocarbon backbones (just like the familiar aspirin, codeine, and so on that we have long been used to).

The pharmaceutical industry is the environment for developing therapeutic small molecule drugs. Assays are devised in which the fault caused by the mutant gene is replicated in cultured cells (or sometimes even in very simple model organisms in the case of highly

conserved genes), and then high throughput screening is carried out by exposing suitable mutant cells to tens of thousands of different synthetic small molecules in parallel. The aim is to identify individual small molecules that somehow overcome the adverse effects caused by pathogenic mutation; they then become candidate therapeutic drugs for further detailed testing. We provide three examples below to illustrate how identifying novel genes for monogenic disorders has led to developing novel therapeutic small molecule drugs.

The first example is an early success story in which genetic advances led to a promising drug target that was screened to identify a new class of very valuable drugs. The second illustrates treating a common recessive disorder which, however, is not so amenable to gene therapy, and the need for stratifying mutations according to their functional effects in cells. In the third example, working out what the disease normally does and discovering that the defect involved a specific molecular pathway allowed the application of a previously identified drug that works in the same pathway and has largely replaced a surgical treatment method.

Familial hypercholesterolemia: new and valuable drugs

With a frequency of more than 0.2 % in most populations, familial hypercholesterolemia (OMIM 143890), an autosomal dominant disorder, is the most common single-gene disorder. Affected persons have extremely high cholesterol levels, irrespective of diet. Most cases are due to mutations in the low-density lipoprotein receptor gene, *LDLR*. Heterozygotes typically develop coronary artery disease in the fourth or fifth decade; rare homozygotes are much more severely affected, and most suffer a heart attack before the age of 20 years.

LDLR imports cholesterol-containing low-density lipoprotein into liver cells, where it represses cholesterol synthesis as part of a homeostatic mechanism. An *LDLR* loss-of-function mutation results in less *LDLR* being made, resulting in an increase in endogenous cholesterol synthesis. Hydroxymethylglutaryl (HMG) CoA reductase was known to be the rate-limiting enzyme in the endogenous cholesterol biosynthesis pathway, and so represented a very promising drug target.

Screening for hydrocarbon-based small molecules that inhibit HMG CoA reductase identified statins, a class of drugs effective in lowering cholesterol. Drugs such as these have been hugely important, being widely prescribed to reduce the general risk of heart disease, not just to treat familial hypercholesterolemia.

Cystic fibrosis: not an easy prospect for therapy

In 1989 a previously unstudied gene was found to be the locus for cystic fibrosis. Named the cystic fibrosis transmembrane regulator gene (*CFTR*), it was subsequently shown to function as an anion channel that allows chloride and bicarbonate ions to pass through the plasma

membrane of cells. A great deal was discovered about aspects of CFTR biology but even twenty years after gene discovery, slow progress had been made in understanding the pathogenesis, and new therapies were lacking. In the last decade, however, significant progress has been made.

As we explain in [Section 9.4](#), cystic fibrosis is an example of a recessive disorder that is difficult to treat by gene therapy, prompting alternative efforts to develop therapeutic small molecule drugs. Different drugs may be applicable, according to their effect at different functional levels (see [Figure 9.12](#) for one scheme that classifies *CFTR* mutations into six categories according to how they affect normal gene function).

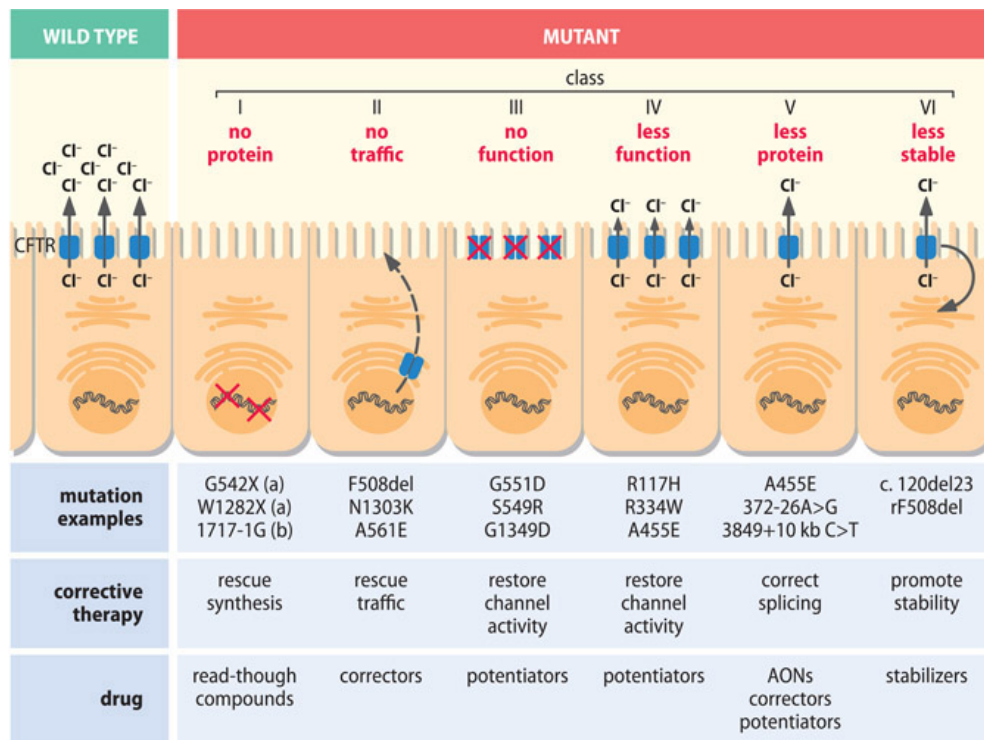


Figure 9.12 Functional classification of *CFTR* mutations causing cystic fibrosis and major targeted *CFTR* modulator therapies and drug type nomenclature. For mutation classes 1 and 2, the *CFTR* protein isn't produced or doesn't get to the apical membrane because of problems with protein folding or trafficking within the cell. For mutation classes III-VI, the *CFTR* protein does get incorporated into the apical membrane, but the ion channel doesn't open (III), doesn't conduct ions properly (IV), is present in low amounts (V), or shows instability (VI). See text for drug classes and progress using them. AONs, antisense oligonucleotides; WT, wild type. Adapted with permission from Amaral MD (2015) *J Intl Med* 277:155–166; PMID 25266997.

The bulk of mutations leading to failure to make a *CFTR* protein result from premature termination codons (nonsense mutations and mutations causing translational frameshifting). For patients with these mutations a generic small molecule drug may be considered that can suppress the effect of premature termination codons irrespective of the type of disease. We consider such “read-through compounds” later in this section.

To obtain small molecule drugs that specifically target mutant CFTR protein, high throughput screening is required. In principle, the small molecule drugs can work by binding to and modulating the effect of mutant CFTR proteins in different ways as listed below.

- *Correctors* assist mutant CFTR to adopt a suitable conformation, enabling them to move from the endoplasmic reticulum to the plasma membrane.
- *Potentiators* facilitate ion channel gating (opening the channel) and conduction (increasing the flow of ions through an open channel).
- *Stabilizers* help keep the CFTR protein anchored to the plasma membrane.
- *Amplifiers* increase the amount of CFTR protein made.

The first breakthrough came when ivacaftor, a potentiator drug marketed by Vertex Pharmaceuticals, was shown to be effective in treating cystic fibrosis patients with the Gly551Asp (G551D) mutation. In the cells of these patients the mutant CFTR protein causes the ion channel to fail to open, but ivacaftor helps to reopen it. After promising clinical results, ivacaftor received regulatory approval in the US in 2012. Although it was subsequently found to be useful for treating patients with some other minor *CFTR* mutations, however, ivacaftor treatment is appropriate for just 5 % of cystic fibrosis patients.

The predominant cystic fibrosis mutation, accounting for around 70 % of pathogenic *CFTR* mutations in many Caucasian populations, produces the Phe508del (F508del) CFTR mutant protein. The problem here is that the mutant CFTR not only misfolds and gets trapped in the endoplasmic reticulum (where it is then targeted for destruction), but any of the mutant protein that does manage to get to the plasma membrane is functionally inactive. In 2019, however, a couple of phase 3 multicenter clinical trials demonstrated that a combination of two folding correctors (elexacaftor and tezacaftor) and the ivacaftor potentiator were both efficacious and safe in treating cystic fibrosis patients with the Phe508del mutant. The combined drug, called Trikafta, is an effective therapy for patients possessing a *CFTR* allele making the mutant F508 del. variant, representing 90% of cystic fibrosis patients in the USA.

Tuberous sclerosis: from a biological pathway to new treatments

Tuberous sclerosis complex is an autosomal dominant disorder in which benign (noncancerous) tumors develop in many organs and can disrupt how they function (tumors of the central nervous system and kidneys are the leading causes of the morbidity and mortality). Additional abnormalities of cell migration and function in the brain lead to seizures, autism, and learning difficulties. The disorder can be caused by mutations in the *TSC1* or *TSC2* genes that encode components of the TSC1–TSC2 protein complex. Until these novel genes were identified, nothing was known about the molecular pathogenesis of the disorder, and the only treatment options were the often problematic surgical removal of tumors.

The TSC1–TSC2 complex was then found to be part of the mTORC1 growth signaling pathway. When the TSC1–TSC2 complex is disrupted by mutations in either *TSC1* or *TSC2*, mTORC1 signaling is constitutively active; downstream targets become activated by phosphorylation, driving protein synthesis and cell growth ([Figure 9.13](#)). A principal subunit of the mTORC1 complex is the mTOR (mammalian target of rapamycin) protein. Rapamycin, also called sirolimus, is an antifungal antibiotic first isolated from a strain of *Streptomyces* in the 1970s. Both it and a closely related drug, everolimus, are powerful mTOR inhibitors. They had initial applications as immunosuppressants to prevent organ transplant rejection and then as anti-proliferative agents for treating certain cancers.

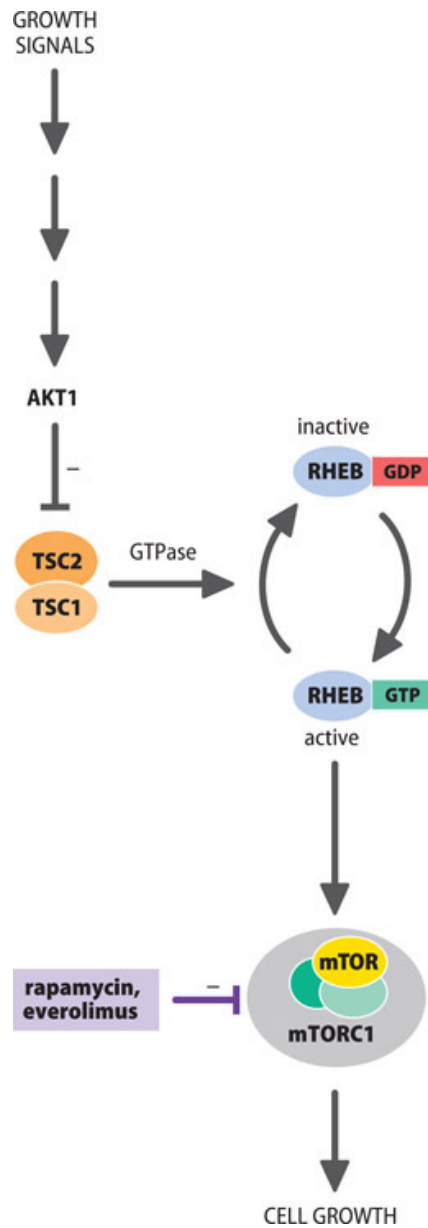


Figure 9.13 Therapeutic targeting of mTORC1 signaling in tuberous sclerosis complex. In tuberous sclerosis complex, the problem is that mutations in either *TSC1* or *TSC2* disrupt the TSC1–TSC2 protein

complex. Normally, TSC2 acts as a GTPase and stimulates the formation of an inactive form of the RHEB regulator of the mTORC1 complex (but can be countermanded when growth signals repress TSC1–TSC2). But in tuberous sclerosis complex, the disruption of the TSC1–TSC2 complex causes the RHEB regulator to be activated so that mTORC1 signaling is constitutively active, and growth is no longer regulated as normal. Rapamycin (also called sirolimus) and everolimus, an *O*-(2-hydroxyethyl) derivative of rapamycin, are effective inhibitors of mTOR, a major subunit of the mTORC1 complex, and work to suppress cell growth. AKT1 is also known as protein kinase B.

The tuberous sclerosis complex was discovered to arise from unregulated growth due to abnormal regulation of the mTORC1 pathway. Thereafter various clinical trials were carried out with mTOR inhibitors. Everolimus and related mTOR inhibitors were found to be effective and safe drugs for treating various aspects of tuberous sclerosis, including angiomyolipomas (which are very common in adult Marfan syndrome patients and can lead to renal failure and a need for dialysis) and also subependymal giant cell astrocytomas (SEGAs). Surgical removal used to be the treatment option for managing SEGAs but the surgery can be particularly difficult, and has now largely been replaced by treatment with mTOR inhibitors.

Translating genomic advances and developing generic drugs as a way of overcoming the problem of too few drug targets

Most small molecule drugs work by binding to specific protein targets, blocking their interactions with other molecules. However, only a rather small percentage of protein targets are susceptible to drugs. In many cases, small molecule drugs cannot block interactions between two types of protein because the interacting surfaces of the proteins are too smooth (small molecule drugs are most effective when they can sneak into clefts and pockets within proteins). Some types of protein are easier targets: more than 50 % of drug targets belong to one of four types of protein (class I G-protein-coupled receptors, nuclear receptors, ligand-gated ion channels, and voltage-gated ion channels); protein kinases are another favorite target.

A survey published in 2006 estimated that there were just 324 protein targets for all approved drugs, but by a decade later that number had more than doubled, reaching a total of 667 human protein targets (PMID 27910877). Recently acquired massive data sets generated through genomics, high-throughput transcriptomics, proteomics, and bioinformatics have helped.

Another potential solution to obstacles in identifying novel drug targets is to develop generic drugs not focused on a specific gene product. Because they might have quite widespread applicability, generic drugs might also have the merit of reducing costs (which can be exceptionally high for “orphan drugs” used to treat rare diseases).

A prominent class of generic drugs is made up of “read-through” compounds, small molecule drugs that can suppress stop codons so that translation continues (translational

“readthrough”). The potential applications are huge: nonsense mutations are responsible for causing anywhere from 5–70 % of individual cases for most inherited diseases. A recently identified drug, Ataluren (or PTC124), seemed promising: it appeared to cause readthrough of premature nonsense mutations (notably UGA), without affecting the recognition of normal stop codons, and with little evidence of toxicity. However, there have been concerns that PTC 124 activity might be due to off-target effects (PMID 23824517), and its efficacy can be limited (when treated with Ataluren, patients with cystic fibrosis due to *CFTR* nonsense mutations showed just a modest improvement in lung function).

Developing biological drugs: therapeutic proteins produced by genetic engineering

In recent years a new drug class, biological drugs (often called **biologics**), has been developed. The great majority are therapeutic recombinant proteins, that is, genetically engineered proteins administered to patients. They include genetically engineered antibodies (the subject of the section following this one) and other genetically engineered proteins (described in this section).

Certain genetic disorders resulting from deficiency of a specific protein hormone or blood protein can be treated by administering an external supply of the missing protein. To ensure greater stability and activity, the proteins are often PEGylated, that is, conjugated with PEG [poly(ethylene glycol)]. The increased size of the protein–PEG complex means reduced renal clearance, so that the protein spends more time in the circulation. Pegylation can also make the protein less immunogenic.

Therapeutic proteins were often previously extracted from animal or human sources, but there have been safety issues. A safer alternative is to use therapeutic “recombinant” proteins made by cloning the desired human gene and expressing it to make protein, usually within mammalian cells, such as human fibroblasts or the Chinese hamster ovary cell line. (Mammalian cells are often needed because many proteins undergo post-translational modifications, such as glycosylation, that show species differences in the pattern of modification.) Recombinant human insulin was first marketed in 1982; [Table 9.5](#) gives several subsequent examples.

TABLE 9.5 EXAMPLES OF THERAPEUTIC RECOMBINANT PROTEINS

Recombinant protein	For treatment of
Insulin	diabetes
Growth hormone	growth hormone deficiency
Blood clotting factors VIII and IX	hemophilia
Interferon α	hairy cell leukemia; chronic hepatitis

For genetically engineered therapeutic antibodies, see [Table 9.6](#).

Recombinant protein	For treatment of
Interferon β	multiple sclerosis
Interferon γ	infections in patients with chronic granulomatous disease
Tissue plasminogen activator	thrombotic disorders
Leptin	obesity
Erythropoietin	anemia

For genetically engineered therapeutic antibodies, see [Table 9.6](#).

TABLE 9.6 EXAMPLES OF LICENSED THERAPEUTIC MONOCLONAL ANTIBODIES (mAbs) FOR TREATING COMMON GENETIC DISEASE

Disease category	Target	mAb generic name (trade name)	Disease treated
Autoimmune or immunological	IgE	omalizumab (Xolair)	asthma
	Integrin α_4	natalizumab (Tysabri)	multiple sclerosis; Crohn's disease
	TNF α	certolizumab pegol (Cimzia)	Crohn's disease; rheumatoid arthritis
adalimumab (Humira)*			
Cancer	EGFR	panitumumab (Vectibix)*	EGFR-expressing metastatic colorectal cancer
	HER2	trastuzumab (Herceptin)	HER2-positive metastatic breast cancer
	VEGF	bevacizumab (Avastin)	colorectal, breast, renal, and NSCL cancer; age-related macular degeneration
Other diseases	PCSK9	alirocumab (Praluent)	hypercholesterolemia that doesn't respond well to statin treatment; atherosclerosis
		evolocumab (Repatha)*	

CD11a, white blood cell antigen; IL2R, interleukin type 2 receptor; IgE, immunoglobulin E; TNF α , tumor necrosis factor α ; EGFR, epidermal growth factor receptor; HER2, human epidermal growth factor receptor 2; NSCL cancer, non-small-cell lung cancer; VEGF, vascular endothelial growth factor; RSV, respiratory syncytial virus

* Fully human antibodies; all others are humanized antibodies (see [Figure 9.14](#) for an explanation of different monoclonal antibody classes).

Disease category	Target	mAb generic name (trade name)	Disease treated
	VEGF	ranibizumab (Lucentis)	“wet” age-related macular degeneration

CD11a, white blood cell antigen; IL2R, interleukin type 2 receptor; IgE, immunoglobulin E; TNF α , tumor necrosis factor α ; EGFR, epidermal growth factor receptor; HER2, human epidermal growth factor receptor 2; NSCL cancer, non-small-cell lung cancer; VEGF, vascular endothelial growth factor; RSV, respiratory syncytial virus

* Fully human antibodies; all others are humanized antibodies (see [Figure 9.14](#) for an explanation of different monoclonal antibody classes).

Some human proteins are required in very high therapeutic doses, beyond the production capabilities of cultured cell lines. Transgenic animals such as transgenic sheep or goats are an alternative source; here, the desired protein is secreted in the animal’s milk, aiding purification. In 2009, Atryn became the first therapeutic protein produced by a transgenic animal to be approved by the FDA. Atryn is an antithrombin expressed in the milk of goats, and was designed to be used in therapy to prevent blood clotting.

Genetically engineered therapeutic antibodies with improved therapeutic potential

One class of recombinant protein has notably been put to therapeutic use: genetically engineered antibodies. As detailed in [Section 4.4](#), each of us has a huge repertoire of different antibodies that act as a defense system against innumerable foreign antigens. Antibody molecules function as adaptors: they have binding sites for foreign antigens at the variable end, and binding sites for effector molecules at the constant end. Binding of an antibody may be sufficient to neutralize some toxins and viruses; more usually, the bound antibody triggers the complement system and cell-mediated killing.

Artificially produced therapeutic antibodies are designed to be mono-specific (specific for a single antigen). Traditional monoclonal antibodies (mAbs) are secreted by *hybridomas*, immortalized cells produced by fusing antibody-producing B lymphocytes from an immunized mouse or rat with cells from an immortal mouse B-lymphocyte tumor. Hybridomas are propagated as individual clones, each of which can provide a permanent and stable source of a single mAb.

The therapeutic potential of mAbs produced like this is, unfortunately, limited. Rodent mAbs, raised against human pathogens for example, have a short half-life in human serum, often causing the recipient to make anti-rodent antibodies. And only some of the different classes can trigger human effector functions.

Genetically engineered antibodies

To make rodent monoclonal more stable in humans, some or all of the rodent protein sequence is replaced by the human equivalent. That happens by genetic engineering at the DNA level: coding DNA sequences encoding part or all of the rodent antibody are replaced by the equivalent human sequences, and the altered coding DNA is expressed to make the desired antibody. The first such attempts were designed to generate a chimeric V/C antibody containing the original rodent variable chains but human constant regions ([Figure 9.14](#)).

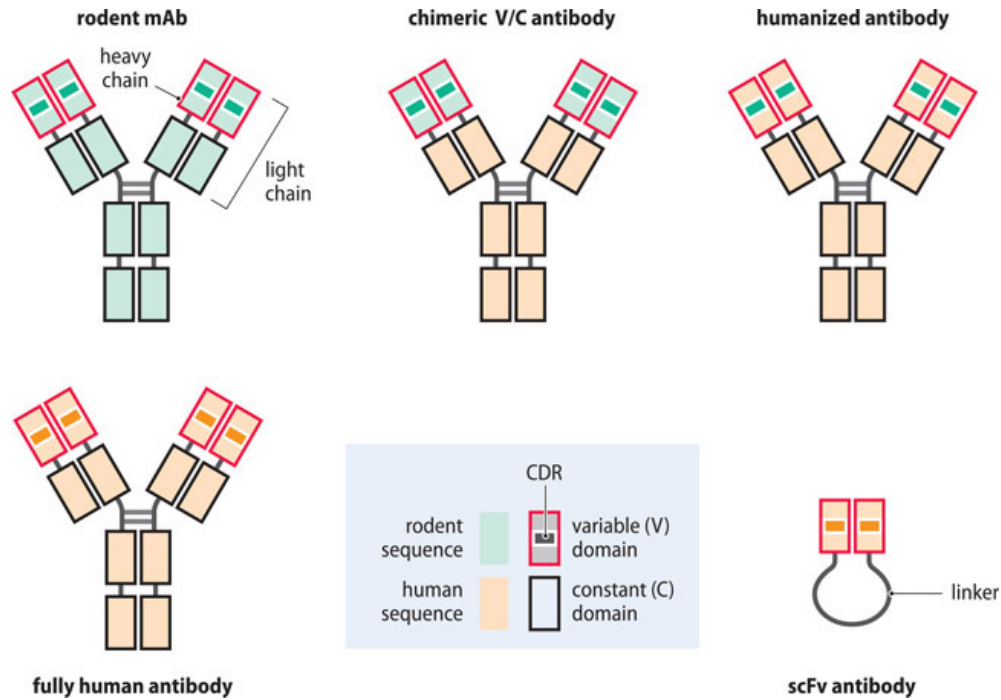


Figure 9.14 Using genetic engineering to make improved therapeutic antibodies. Classical antibodies consist of heavy and light chains with variable (V) and constant (C) domains. Rodent monoclonal antibodies (mAbs) are monospecific antibodies synthesized by hybridomas. Chimeric V/C antibodies are engineered to have human constant domains joined to rodent variable domain sequences (containing the critically important hypervariable complementarity-determining region, CDR). In humanized antibodies all the sequence is human except the hypervariable CDR. Fully human antibodies are obtained by different routes (see text). Single-chain antibodies have also been made, with two variable domains only, connected by a linker peptide. These single-chain variable fragment (scFv) antibodies are particularly well suited to working within the reducing environment of cells; they can serve as *intrabodies* (intracellular antibodies) by binding to specific antigens within cells. Depending on the length of the linker, they bind their target as monomers, dimers, or trimers. Multimers bind their target more strongly than monomers.

Subsequently, *humanized antibodies* were constructed in which all the rodent sequence was replaced by human sequence, except for the complementarity-determining regions (CDRs), the hypervariable sequences of the antigen-binding site (see [Figure 9.14](#)). More recently, it has been possible to prepare *fully human antibodies* by different routes. For example, mice have been genetically manipulated to delete their immunoglobulin loci and

replace them with an artificial chromosome containing the entire human heavy-chain and g light-chain loci so that they can make fully human antibodies only.

From inauspicious beginnings in the 1980s, mAbs have become the most successful biotech drugs ever, being used to treat a variety of common genetic diseases (see [Table 9.6](#)). The market for mAbs is the fastest-growing component of the pharmaceutical industry, and of the therapeutic mAbs currently in use, the eight bestsellers are expected to generate together an annual income of more than US \$170 billion by the end of 2021. Many additional mAb products are in the pipeline. Of the FDA-approved mAbs, most are partly or fully human and the majority are aimed at treating autoimmune or immunological disease or cancers. In the latter case, the latest antibodies are being developed as antibody–drug conjugates so that the antibodies deliver powerful toxins to kill cancer cells.

Intracellular antibodies (intrabodies)

Simplified antibodies containing just the variable sequences important in antigen recognition can function inside cells and bind to specific intracellular antigens. They are produced by genetic engineering: the appropriate genetic construct is made, then transfected into suitable cells to produce the desired intrabody. They therefore complement standard-size antibodies (which bind to epitopes on cell surfaces), and their therapeutic potential is being tested. There are two significant categories, as listed below.

- *Single-chain antibodies.* Engineered to have a one variable chain, single-chain variable fragment (scFv) antibodies have almost all the binding specificity of a mAb (see [Figure 9.14](#)). They can be made on a large scale in bacterial, yeast or even plant cells. Unlike multichain standard antibodies, scFv antibodies are stable in the reducing environment within cells, and well suited to acting as intrabodies. They are designed to bind to specific target molecules within cells. As required, they can be directed towards specific subcellular compartments.
- *Nanobodies (single-domain antibody fragments).* The starting point was the discovery that camelids (camels, llamas and so on) have fully functional antibodies that lack light chains, and their heavy chain-only antibodies have a single variable domain. Thereafter, cloned, isolated single variable domains were found to have full antigen-binding capacity and to be very stable compared to normal antibodies.

Intrabodies can carry effector molecules that perform specific functions when antigen binding occurs. However, for many therapeutic purposes they are designed simply to block specific protein–protein associations within cells. As such, they complement conventional drugs. Protein–protein interactions usually occur across large, flat surfaces and are often unsuitable targets for small molecule drugs (that normally operate by fitting snugly into clefts on the surface of macromolecules). Potentially promising therapeutic target proteins for

intrabodies include mutant proteins that tend to misfold in a way that causes neurons to die, as in various neurodegenerative diseases including Alzheimer, Huntington, and prion diseases.

9.3 PRINCIPLES OF GENE AND CELL THERAPY

Gene therapy involves the direct genetic modification of cells to achieve a therapeutic goal. The genetic modification can involve the insertion of DNA, RNA, or oligonucleotides. Gene therapy can be classified into two types, according to whether somatic cells or germline cells are genetically modified. Somatic cell gene therapy seeks to modify specific cells or tissues of the patient in a way that is confined to that patient. Germline gene therapy would produce a permanent modification that can be transmitted to descendants; this could be achieved by modifying the DNA of a gamete, zygote, or early embryo.

Germline gene therapy that involves modifying nuclear DNA is widely banned in humans for ethical reasons, but, as described in [Section 9.4](#), replacement of mutant mtDNA by normal mtDNA in oocytes or zygotes is licensed in some countries to prevent transmission of certain mitochondrial DNA disorders. Ethical issues relating to gene therapy are discussed in [Section 11.5](#).

Gene therapy has had a checkered history. Tremendous initial excitement—and quite a bit of hype—was followed by a fallow period of disappointing results and safety concerns (with unexpected deaths of patients arising from unforeseen deficiencies in the treatment methods). More recently, there has been a greater appreciation of safety risks, and significant successes.

In this section and [Section 9.4](#), we are mostly concerned with gene therapy for inherited disorders, which has focused predominantly on recessive Mendelian disorders. Cancer gene therapy and other approaches to treating cancer are described in [Chapter 10](#).

The first real successes of gene therapy were not achieved until the early 2000s and involved treating very rare cases of severe combined immunodeficiencies. They took advantage of previous experience of bone marrow transplantation, which is effectively a crude type of hematopoietic stem cell therapy.

As we describe later, *stem cells* are cells that have the property of being able to renew themselves and also being able to give rise to more specialized cells. For many types of gene therapy, it is important to maximize gene transfer into appropriate stem cells in the patient. Cell therapies based on the genetic modification of stem cells are also fundamental in *regenerative medicine*, where the object is to treat disease by replacing cells or tissues that have been lost through disease or injury.

In this section we consider the principles underlying gene and cell therapy. In [Section 9.4](#) we deal with the progress made, and discuss future prospects.

Two broad strategies in somatic gene therapy

The cells targeted in somatic cell gene therapy are normally those directly involved in the pathogenic process, but in many cancer gene therapy trials the object has been to genetically modify *normal* cells in a patient to provoke specific immune responses and killing of harmful cells.

Using molecular genetic approaches to treat disease might involve many different strategies. But at the level of the diseased cells there are two basic strategies: disease cells are simply genetically modified in some way so as to alleviate disease, or they are selectively killed. Within each of the two main strategies are different substrategies, as described below.

Modifying disease cells (Figure 9.15A). According to the molecular pathology, different strategies are used. If the problem is loss of function, a simple solution (in theory) is to add functioning copies of the relevant gene. In genetic disorders in which the pathogenesis results from a gain of function, there is some harmful or toxic gene product within cells. The approach then might be to selectively inhibit the expression of the harmful gene product without affecting the expression of any normal genes. This can often be done by selectively blocking transcription of the harmful mutant gene or by targeting transcripts of the gene so that they are destroyed (*gene silencing*). Yet other approaches seek to repair a genetic lesion by some type of *genome editing* or find a way of minimizing its effect. We detail the approaches below.

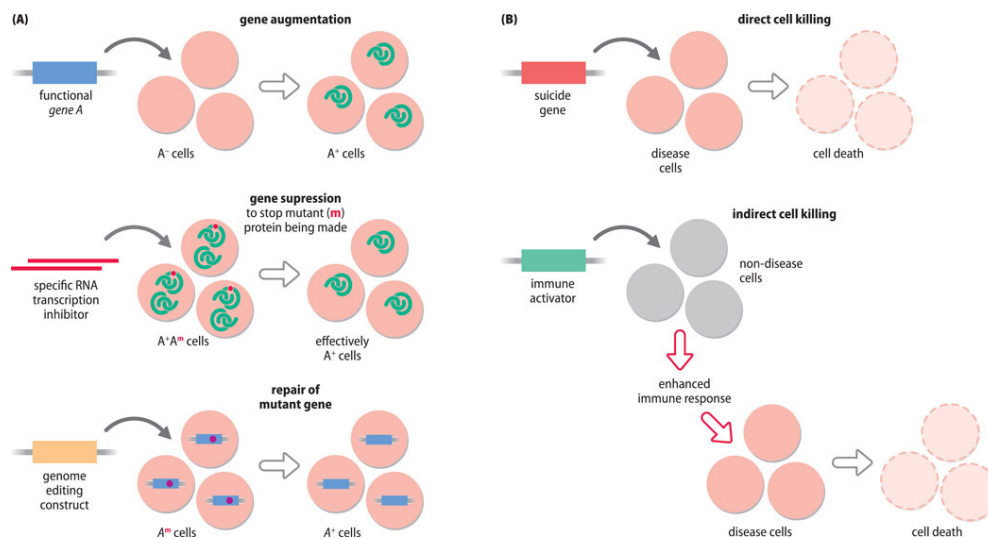


Figure 9.15 Different general types of gene therapy strategy. (A) Therapies aimed at modifying disease cells. *Gene augmentation* therapy (also called gene supplementation or gene addition therapy) can be applied to loss-of-function disorders but is currently limited to treating recessive disorders (in which the disease results from a lack, or an almost complete lack, of some gene product). The object is simply to transfer a cloned working gene copy into the cells of the patient in order to make some gene product that is lacking. *Gene suppression therapy* can be applied to disorders that result from positively harmful gene products. If the disease is caused by a gain-of-function mutation that produces a harmful mutant gene product, A^m, in addition to the normal gene product, A⁺, one might try to specifically inhibit the expression of the mutant allele without inhibiting expression of the normal allele. The same approach can be applied to treating autoimmune and

infectious diseases. *Genome editing* can be used to repair a DNA lesion, converting the sequence of the mutant allele, A^m , to a normal allele sequence, A . (B) Therapies aimed at killing harmful cells. The overwhelming application has been in cancer gene therapy trials, either seeking to kill cancer cells directly (by inserting and expressing cloned genes that give rise to some cytotoxic product and cell death) or indirectly (by transferring genes into non-disease cells, such as immune system cells, to provoke an immune response directed at tumors).

Killing disease cells ([Figure 9.15B](#)). This approach has frequently been used in cancer gene therapy trials. Traditional cancer treatments have often relied on killing disease cells by using blunt instruments, such as high-energy radiation and harmful chemicals that selectively kill dividing cells. Gene therapy approaches can kill harmful cells either directly or by modifying immune system cells to enhance immune responses that can kill the harmful cells.

The delivery problem: designing optimal and safe strategies for getting genetic constructs into the cells of patients

In gene therapy, a therapeutic *genetic construct* of some type—often a cloned gene, but sometimes RNA or oligonucleotides—is transferred into the cells of a patient. (A nucleic acid molecule introduced in this way is often referred to as a **transgene**.) Depending on the disease, the type of cells to be targeted can be very different, and different strategies are needed. And, depending on the target cells, some disorders are easier to treat in principle than others.

Consider access to the desired target cells. Some cells and tissues—notably blood, skin, muscle, and eyes—are very accessible; others, such as brain cells, are not easily accessed. Then there is the question of overcoming various barriers that impede the transfer and expression of genetic constructs. Strong immune responses constitute important barriers. And, as we will see, mechanical barriers can also be important.

For target cells, another significant factor is the extent to which the cells divide. Regular cell division is required to replenish short-lived cells, such as blood and skin cells, unlike in the case of long-lived cells, such as terminally differentiated muscle cells. The distinction is important. For nondividing cells the key parameters would simply be the efficiency of transfer of the therapeutic construct into the cells of the patient and the degree to which the introduced construct was able to function in the expected way. However, for dividing cells we also need to take into account what happens to the descendant cells.

Even if we were to achieve significant success in getting the desired genetic construct into short-lived cells, the cells that have taken up the genetic construct are going to die and will be replaced by new cells. A small minority of the cells, **stem cells**, divide to continuously replenish cells lost through aging, illness, or injury (see [Box 9.1](#) for a brief overview of stem cells). To ensure that copies of the therapeutic construct keep getting into newly dividing cells, therefore, it would be best to target the relevant stem cells if possible, and get the therapeutic construct integrated into chromosomes (so that it gets replicated, allowing copies to be passed to both daughter cells at cell division).

BOX 9.1 AN OVERVIEW OF STEM CELLS AND ARTIFICIAL EPIGENETIC REPROGRAMMING OF CELLS

Stem cells have two essential properties: they can self-renew and they can also give rise to more **differentiated** (more specialized) cells. As well as undergoing normal (symmetric) cell divisions, stem cells can undergo asymmetric cell division to give two *different* daughter cells. One daughter cell is identical to the parent stem cell, allowing self-renewal; the other daughter cell is more specialized and can undergo further rounds of differentiation to give terminally differentiated cells (**Figure 1**).

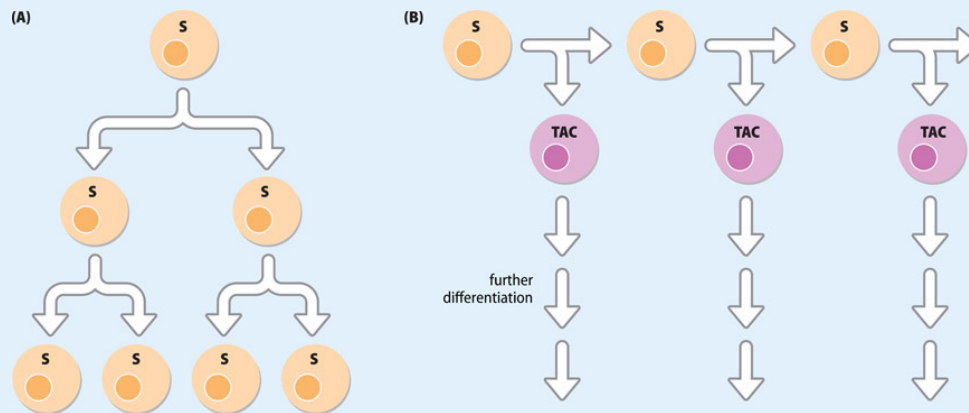


Figure 1 Symmetric and asymmetric stem cell divisions. (A) Populations of stem cells (S) can expand quickly by symmetrical cell division during growth and when there is a rapid need for new stem cells (to replace cells lost through disease or injury). (B) Stem cells give rise to more differentiated cells by asymmetric cell divisions—the stem cell produces two different daughter cells. One daughter cell is a stem cell identical to the parent cell. The other is a stem cell derivative, sometimes called a **transit amplifying cell** (TAC), that is more differentiated than its sister or parent cell and can subsequently undergo additional differentiation steps to form a terminally differentiated cell.

Different classes of stem cell are used in experimental investigations. Some are cultured *somatic* stem cells derived from naturally occurring somatic stem cells in the body. They can give rise to a limited number of differentiated cell types. The other major class are artificially created *pluripotent* stem cells that have the capacity to give rise to all of the different cell types in the body. Two major types of cultured pluripotent stem cells are described below.

- **Embryonic stem cell (ESC) lines** are artificially derived from naturally pluripotent cells of the very early embryo (which, however, are not stem cells, being *transient* cells that will change into more specialized cells during development).
- **Induced pluripotent stem cells (iPSCs)** are obtained by artificially changing the normal epigenetic settings of easily obtained cells such as skin cells; this is a type of artificial epigenetic reprogramming.

SOMATIC STEM CELLS

These cells—also called adult stem cells or tissue stem cells—occur naturally in the body, and help replace cells with naturally short life spans (notably in the blood, skin, intestines, and testis) or help replenish cells lost in disease or injury. (Our powers of tissue regeneration are, however, rather limited.)

Most somatic stem cells give rise to a limited set of differentiated cells. Some, such as spermatogonial stem cells, are unipotent, giving rise to a single type of differentiated cell. Others are multipotent, being able to give rise to several different classes of differentiated cells. For example, hematopoietic stem cells can give rise to all types of blood cell as well as to certain tissue cell types ([Figure 9.17](#)). Cultured somatic stem cell lines have been used for studying differentiation, and purified populations of genetically modified somatic stem cells, notably hematopoietic stem cells, have been used in gene therapy.

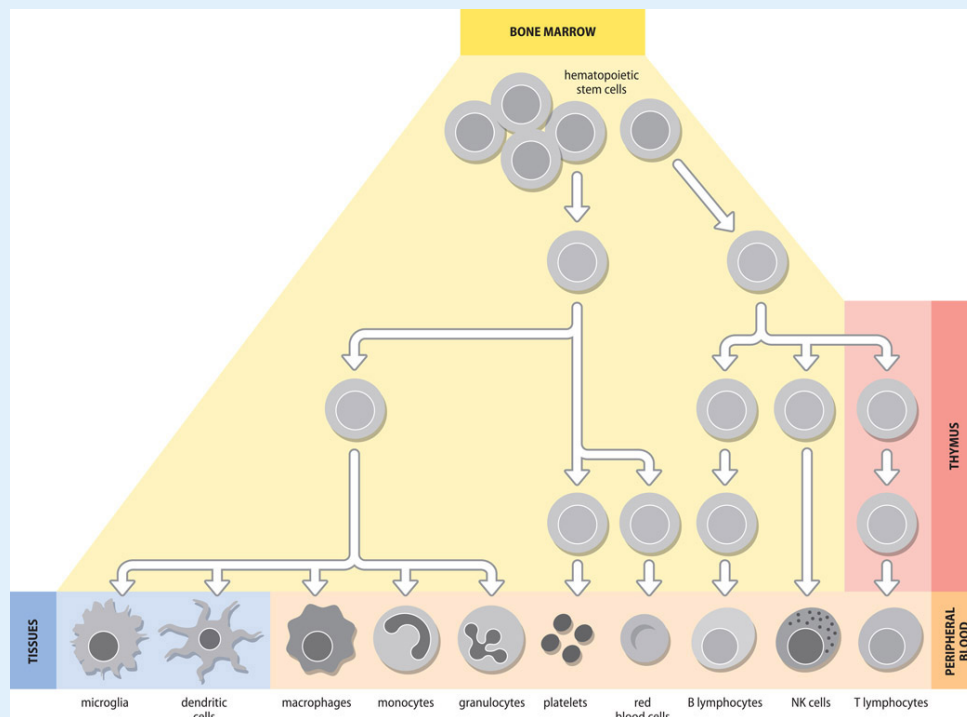


Figure 9.17 All blood cells and some tissue immune system cells originate from hematopoietic stem cells. All differentiated blood cells have limited life spans, and there is a continuous cycle of cell death and cell replacement. The replacement blood cells are derived from hematopoietic stem cells that are particularly concentrated in the bone marrow. Hematopoietic stem cells also give rise to some tissue cells, including tissue macrophages (such as microglia, the resident macrophages of the brain and spinal cord) and dendritic cells (a class of immune system cells that work in antigen presentation in varied tissues). NK cells, natural killer cells.

EMBRYONIC STEM CELL LINES

The mammalian zygote and cells descending from it through the first few cleavage divisions) are entirely unspecialized and are said to be *totipotent*—they can give rise to every type of cell in both the embryo and in the extra-embryonic membranes. Subsequently, the *blastocyst* forms as a hollow ball of cells with two quite distinct layers: an outer layer of cells known as the *trophoblast* (which will give rise to the extraembryonic membranes such as the chorion and the amnion), and a group of inner cells, the *inner cell mass*, located at one end of the blastocyst ([Figure 2](#)).

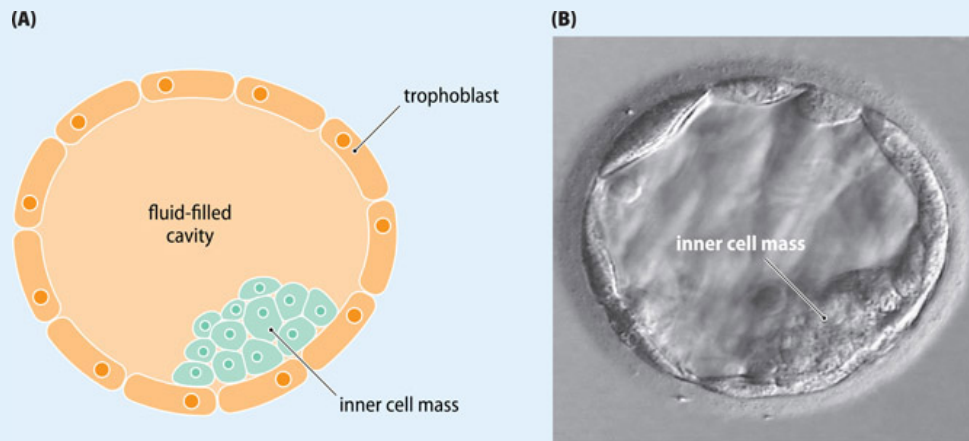


Figure 2 Blastocyst structure and the inner cell mass. (A) The blastocyst is a hollow, fluid-filled ball of cells with two distinct cell populations: an outer cell layer (the trophoblast) and an inner group of cells located at one end (the inner cell mass). (B) A 6-day-old human blastocyst containing about 100 cells, showing the location of the inner cell mass. (B, Courtesy of M. Herbert, Newcastle University, UK.)

Cells from the transient inner cell mass are pluripotent and can be cultured to make a pluripotent embryonic stem cell (ESC) line. ESCs can then be experimentally induced to make desired types of differentiated cell (including derivatives of the germ cell layers—ectoderm, mesoderm, and endoderm—and also germ cells). They have been vitally important for making animal (mostly, mouse) models of disease, as described in [Box 9.2](#). Human ESC lines are produced from cells derived from surplus embryos in assisted reproduction (*in vitro* fertilization; IVF) clinics. They may have promise in cell therapy if immune responses in recipients are minimized in some way, but because human ESC lines are derived from human embryos, the creation of new ESC lines remains controversial.

INDUCED PLURIPOTENT STEM CELLS AND CELL REPROGRAMMING

For decades, cell differentiation in mammals was thought to be irreversible. Then a cloned sheep called Dolly proved that terminally differentiated mammalian cells could be reprogrammed to become unspecialized cells resembling the pluripotent cells of the early embryo. Cloning mammals is, however, extremely arduous and technically difficult.

Alternative, comparatively simple, methods can be used to re-set the epigenetic marks of a cell. For example, by providing certain key transcription factors, or by inducing the cells

to make them, terminally differentiated mammalian cells may be induced to *dedifferentiate* (to give less specialized cells), or to *transdifferentiate* (to give specialized cells of a different type)

Like ESCs, induced pluripotent stem cells (iPSCs) can be directed to differentiate into more specialized cells ([Figure 3](#)). Because iPSCs may retain some characteristics of their progenitor cells, they are less robust than ESCs.

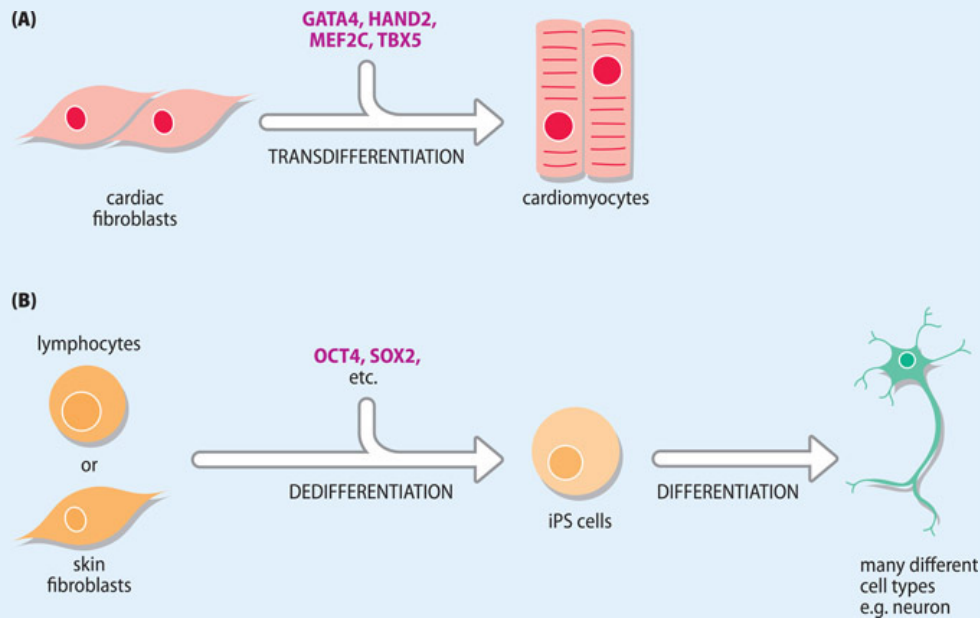


Figure 3 Cell reprogramming. (A) An example of direct reprogramming. Here, certain cardiac transcription factors (in pink) induce the transdifferentiation of fibroblasts to myocytes. (B) Reprogramming to give induced pluripotent stem cells (iPSCs) with the use of transcription factors, such as OCT4 and SOX2, that are important in embryonic development and pluripotency. By providing suitable transcription factors, the resulting iPSCs can be induced to differentiate to give a desired cell type.

From a medical perspective, iPSCs offer two exciting applications: human cellular models of disease and genetically modified cells for therapeutic purposes. Animal disease models have been very valuable, enabling the use of invasive studies to infer the molecular basis of human disease, and providing frontline testing of new therapies. But they are only *models*; they quite often show important differences from humans. Accessible skin cells from a patient can now be reprogrammed to become iPSCs that can then be directed to differentiate into cells relevant to the disease process (such as normally inaccessible neurons for a neurodegenerative disorder). Genetically impaired disease cell lines are miniature disease models, useful for drug screening (testing for toxicity, efficacy, and so on) and for studying the molecular basis of disease in *human* cells.

In providing genetically modified cells for therapeutic purposes, iPSCs have the advantage that they can be made from the cells of a patient, genetically modified, and then returned to the patient *without provoking an immune response*. Successful artificially

induced reprogramming of human cells may transform the prospects of using dedifferentiated human cells therapeutically. We describe this aspect in more detail in [section 9.4](#).

Efficiency and safety aspects

In any gene delivery system used in gene therapy, two key parameters are fundamental: efficiency and safety. Most gene therapy methods rely on transferring genes into the cells of a patient and expressing them to make some product. For the gene delivery method to be effective, it is important to maximize transfection efficiencies for the optimal target cells and to get long-lasting high-level expression of the therapeutic genes.

For disorders where target cells are short-lived, targeting the relevant stem cells should maximize the chances of success. However, stem cells occur *in vivo* at very low frequencies. For blood disorders, happily, it is possible to obtain preparations of bone marrow cells or peripheral blood lymphocytes from patients, grow the cells in culture, and enrich for hematopoietic stem cells. The purified cells can be genetically modified in culture to overcome a genetic defect and then returned to the patient, a type of *ex vivo* gene therapy, as described in the next section.

As we describe below, viral vectors are commonly used to get therapeutic gene constructs into cells at high efficiency, and they often allow high-level expression of the therapeutic transgenes. Some viral vectors are deliberately used because they are adept at getting DNA inserted into chromosomes, which is important when targeting tissues in which cells are short-lived. But the features that make the gene therapy process efficient come with significant safety risks.

One important risk concerns the integration of some therapeutic recombinant viruses into chromosomes—there has often been little control over where they will insert into the genomic DNA of patient cells. They might insert by accident into an endogenous gene and block its function, but the greatest danger is the accidental activation of an oncogene, causing tumor formation.

An additional important risk is that the patient might mount a strong immune or inflammatory response to high levels of what might appear to be foreign molecules. Components of viral vectors might pose such risks, but even if a perfectly normal therapeutic human gene were inserted and expressed to give a desired protein that the patient completely lacked (through constitutional homozygous gene deletion, for example), an immune response might occur if the protein had never been produced by the patient. We enlarge on these issues in [Section 9.4](#).

Different ways of delivering therapeutic genetic constructs, and the advantages of *ex vivo* gene therapy

In gene therapy, a genetic construct is inserted into the cells of a patient by using either viral delivery systems or nonviral methods. Viral vector systems are generally much more efficient than nonviral methods, but they pose greater safety risks.

Using viruses to transfer DNA into human or other animal cells (a process called **transduction**) might be expected to be efficient: over long evolutionary timescales, viruses have mastered the process of infecting cells, and getting their genes to be expressed, often after inserting their genomes into host cell DNA. Depending on the type of virus, a virus may have a DNA or RNA genome that is single-stranded or double-stranded. To be useful for ferrying genes into cells, a virus vector is used, a modified double-stranded DNA copy of the viral DNA or RNA genome (making it easy for a therapeutic DNA to be joined to it to form a recombinant DNA).

Viral vectors for use in gene therapy have been engineered to lack most, and quite often all, of the coding capacity of the original viral genome. The idea is that the recombinant DNA (virus vector plus therapeutic DNA) can nevertheless get packaged into a viral protein coat to make a recombinant virus that is still efficient at infecting cells. In some cases the recombinant viral DNA can integrate into the nuclear genome of a cell, permitting long-lasting therapeutic gene expression; but the integration of vectors poses safety risks. When using other types of virus vector, the recombinant viral DNA does not integrate into the host cell genome; instead, it remains as an extrachromosomal **episome** in cells.

The nonviral transfer of DNA, RNA, or oligonucleotides into human or other animal cells (**transfection**) is much less efficient than viral transduction; the overall amount of transgene expression is therefore more limited. The transfection procedures also do not result in appreciable integration of DNA into the genome of the cell. As a result, transfection has the advantage of greater safety in therapeutic applications, but with reduced efficiency. In addition, transfection methods do not have the same size constraints for the packaged nucleic acid that applies to virus vectors—they can be used to ferry very large nucleic acids.

***In vivo* and *ex vivo* gene therapy**

Some types of gene therapy procedure occur *in vivo*: the transfer of the therapeutic constructs is carried out *in situ* within the patient. Often the therapeutic construct is injected directly into an organ (such as muscle, eye, or brain). It may in some cases be introduced indirectly to target cells. For example, coding sequences of genes important in vision have been successfully delivered into the eyes of patients with hereditary loss of vision with quite good outcomes. We describe below how certain viruses are adept at infecting human cells of a particular type; that property has also been exploited to increase the efficiency of delivering therapeutic genes to the desired target cells.

Because there is no way of selecting and amplifying cells that have taken up (and, in some cases, expressed) the genetic construct, the success of *in vivo* gene therapy is crucially

dependent on the general efficiency of gene transfer and, where appropriate, expression in the correct tissue.

Ex vivo gene therapy means removing cells from a patient, culturing and genetically modifying them *in vitro*, and then returning suitably modified cells back to the patient (Figure 9.16). Because the cells of the patient are genetically modified in the laboratory, they have the enormous advantage that the cells can be analyzed at length to identify those in which the intended genetic modification has been successful. The correctly modified cells can then be amplified in culture and injected back into the patient.

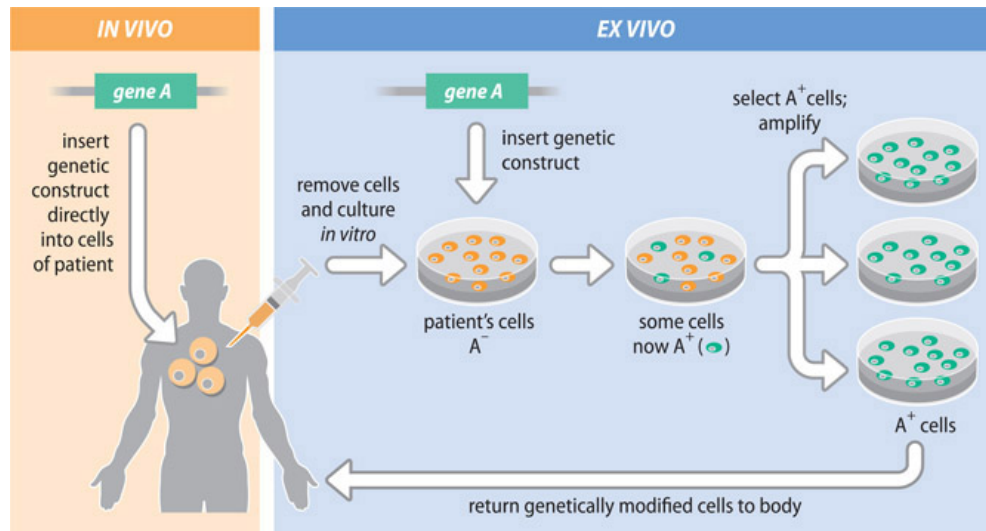


Figure 9.16 *Ex vivo* and *in vivo* gene therapy. In *ex vivo* gene therapy, cells are removed from the patient and genetically modified in some way in the laboratory (in this case we illustrate a gene supplementation procedure in which a therapeutic transgene, gene A, is expressed to make a gene product, A, that is lacking in the cells of the patient). The modified cells are selected, amplified in culture and returned to the patient. The procedure allows detailed checking of genetically modified cells to ensure that they have the correct genetic modification before they are returned to the patient. For many tissues, this is not possible, and the cells must be modified directly within the patient's body (*in vivo* gene therapy).

In practice, *ex vivo* gene therapy has been directed at certain disorders—mostly blood disorders but also some storage disorders—in which the genetically modified cells are bone marrow cells that have been taken from the patient and then treated in such a way so as to enrich for hematopoietic stem cells. As described later, this procedure has been at the core of a series of successful gene therapies.

Nonviral delivery of therapeutic genetic constructs

Interest in nonviral vector delivery systems has mostly been propelled by safety concerns over the use of viral vectors. Nonviral vector systems are certainly safer—they do not

integrate into chromosomes and they are not very immunogenic.

The therapeutic gene is typically carried in a plasmid vector, but transport of plasmid DNAs into the nucleus of nondividing cells is normally very inefficient (the plasmid DNA often cannot enter nuclear membrane pores). Various tricks can be used to help get the plasmids into the nucleus (such as compacting the DNA to a small enough size to pass through the nuclear pores). Because the transfected DNA cannot be stably integrated into the chromosomes of the host cell, nonviral methods of therapeutic gene delivery are more suited to delivery into tissues such as muscle, which do not regularly proliferate, and in which the injected DNA may continue to be expressed for several months.

Different delivery systems can be used. **Liposomes**, synthetic vesicles that form spontaneously when certain lipids are mixed in aqueous solution, have often been used to enclose the desired DNA construct. The lipid coating allows the DNA to survive *in vivo*, bind to cells, and be endocytosed into the cells. Another method uses compacted DNA nanoparticles. Because of its phosphate groups, DNA is a polyanion. Polycations bind strongly to DNA and so cause the DNA to be significantly compacted. Because of their much reduced size, compacted DNA nanoparticles are comparatively efficient at transferring genes to dividing and nondividing cells and have a plasmid capacity of at least 20 kb.

Currently, there is still some way to go for nonviral delivery systems. The efficiency of getting genetic constructs into cells using these methods remains low, as does the expression levels of transfected therapeutic genes. Interested readers can find a recent review at PMID 32580326.

Viral delivery of therapeutic gene constructs: relatively high efficiency but safety concerns

Before we detail how viruses can be used as vectors in gene therapy, let us first consider some properties of viruses. Viruses have a DNA or RNA genome packaged within a protein shell (known as a *capsid*), and in some viruses, called *enveloped viruses*, the protein capsid is in turn enclosed by a lipid bilayer containing viral proteins. Enveloped viruses enter cells either by fusing with the host plasma membrane to release their genome and capsid proteins into the cytosol, or by first binding to cell surface receptors, and then entering via receptor-mediated endocytosis, fusion-based transfer, or endocytosis-based transfer.

Some viruses infect a broad range of human cell types and are said to have a broad *tropism*. Other viruses have a narrow tropism: they bind to receptors expressed by only a few cell types. Herpes viruses, for example, are tropic for cells of the central nervous system. The natural tropism of viruses may be retained in vectors or genetically modified in some way, so as to target a particular tissue.

For an introduced transgene to be expressed, it needs to be ferried to the nucleus. Some viruses can gain access to the nucleus only after the nuclear envelope has dissolved during mitosis. They are limited to infecting dividing cells. Other viruses have devised ways to

transfer their genomes efficiently through nuclear membrane pores, so that both dividing and nondividing cells can be infected.

Some viruses are able to integrate their genome into the genome of the host cell. They include retroviruses, whose genome consists of a single RNA strand, such as lentiviruses and gammaretroviruses. They are able to convert their RNA into a single-stranded cDNA using a viral reverse transcriptase; the single DNA strand is then copied into a double-stranded DNA (replicative form) that can integrate into the host genome using viral enzymes. Other viruses, such as adenoviruses and adeno-associated viruses, do not integrate into the genome.

Virus vectors used in gene therapy

TABLE 9.7 FOUR MAJOR CLASSES OF VIRAL VECTORS THAT HAVE BEEN USED IN GENETHERAPY PROTOCOLS

	Virus class	Viral genome	Cloning capacity	Target cells (D or ND cells)	Transgene expression	Vector yield*; other comments
INTEGRATING	Gamma-retroviruses	ssRNA; ~8-10kb	7-8 kb	D cells only	long-lasting	moderate; risk of oncogene activation
	Lentiviruses (notably HIV)	ssRNA; ~9kb	Up to 8 kb	D and ND cells; tropism varies	long-lasting and high-level	high; risk of oncogene activation
NON-INTEGRATING	Adenoviruses	dsDNA; 38-39 kb	often 7.5 kb; up to 34 kb	Mostly ND cells	transient but high-level	high; immunogenicity can be a major problem
	Adeno-associated viruses (AAVs)	ssDNA; 5 kb	<4.5 kb	Mostly ND cells	high-level in medium/long-term (year)	high; small size capacity; immunogenicity less than for adenovirus

Abbreviations: ss, single-stranded; ds, double-stranded; D, dividing cells; ND, non-dividing cells.

* High vector yield, 10^{12} transducing units/ml; moderate vector yield, 10^{10} transducing units/ml.

Four major classes of virus have been used as vectors for gene therapy (see [Table 9.7](#) for a summary of their properties). The big advantage of using virus vectors is the efficiency in

getting transgenes into cells, which far exceeds that of nonviral transfection methods. Over evolutionary time scales, viruses have become adept at infecting cells, and in expressing their genes within cells. There can be significant safety concerns, however, when using viruses as compared with non-viral transfer methods, and we describe these later.

Vectors based on integrating viruses allow therapeutic genes to be inserted into chromosomes of cells and to be passed on to any descendant cells (an important advantage if the target cells are blood cells or other cell types that have a high rate of cell turnover). For both gammaretroviruses and lentiviruses, the vector is made by isolating viral replicative forms that consist of double-stranded DNA, and genetically modifying them in various ways.

Non-integrating vectors are traditionally based on DNA viruses, and they can be especially useful when the object is to get high-level expression in nondividing target cells, such as muscle. Vectors based on adenovirus have been popular because they can permit very high levels of gene expression, but here, too, there have been safety issues (which relate to their immunogenicity). Safer vectors based on adeno-associated virus (AAV) have subsequently been widely used. We detail the use of viral vectors in gene therapy and the issue of safety concerns in [Section 9.4](#).

The importance of disease models for testing potential therapies in humans

Cellular disease models can be very helpful in understanding the molecular basis of disease, and can be important in drug screening and drug toxicity assays. Recent advances in stem cell technology have allowed the production of a wide range of human cellular disease models. Readily accessible blood or skin cells from a patient can be genetically reprogrammed so that they are converted to some desired cell type that is principally involved in the pathology, such as normally inaccessible neurons. We cover the relevant technology—*induced pluripotent stem cells*—in [Section 9.4](#).

To test novel therapeutic approaches, a robust whole-animal model of disease is necessary. Some animal models of disease, such as the *mdx* mouse model of muscular dystrophy, originated by spontaneous mutation, but the vast majority are artificially generated by genetic manipulation. Primate models might be expected to be the most faithful disease models, but for decades the preferred disease models have been rodent models, notably mice. There are several reasons: rodents breed quickly and prolifically; they are reasonably closely related to humans, sharing 99 % of our genes; maintaining rodent colonies is not too expensive; and there are fewer ethical concerns than with primate models. An additional compelling reason is that for decades certain important genetic manipulation technologies have effectively been available in mice only.

The vast majority of rodent disease models have been created by genetically modifying the germ line, in which foreign DNA is typically engineered into the chromosomal DNA of germline cells. One way is to make a **transgenic animal** disease model by inserting a transgene (= any foreign DNA) into the zygote. This approach can be used in a wide range of different animals.

A second, powerful, technology relies on first genetically modifying the genome of intact embryonic stem (ES) cells in culture in a precise, pre-determined way. The modified ES cells are then transferred into the early embryo in order to produce an animal with genetically modified cells, including modified germline cells. Certain mouse ES cell lines have been particularly amenable to this genetic modification (which is why mouse disease models are so prevalent). The technology is so sophisticated that we can, in principle, make any desired change to the genome sequence of a mouse—even substituting a single nucleotide—at essentially any position we choose. See **Box 9.2** for the salient details.

BOX 9.2 TWO POPULAR WAYS OF MAKING MOUSE DISEASE MODELS

TRANSGENESIS THROUGH PRONUCLEAR MICROINJECTION

One important route for making transgenic mice (or other transgenic animals) is to inject a transgene into the zygote so that the exogenous DNA gets into the genome of the zygote. There is usually no control over where the transgene integrates. The resulting animal will have the transgene in all cells and can transmit it to future generations ([Figure 1](#)).

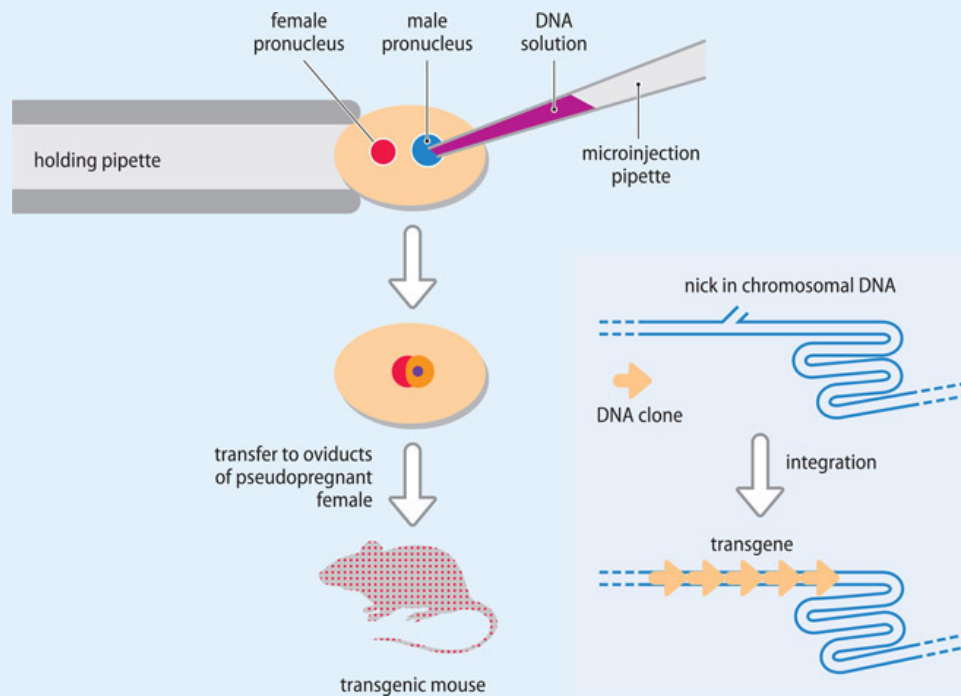


Figure 1 Construction of transgenic mice by pronuclear microinjection. A fine-pointed microinjection pipette is used to pierce first the oocyte and then the male pronucleus (which is bigger than the female pronucleus), delivering an aqueous solution of a desired DNA clone. The introduced DNA integrates at a *nick* (single-stranded DNA break) that has occurred randomly in the chromosomal DNA. The integrated transgene usually consists of multiple copies of the DNA clone. Surviving oocytes are reimplanted into the oviducts of

foster females. DNA analysis of tail biopsies from resulting newborn mice checks for the presence of the desired DNA sequence.

This method is often used for modeling dominantly inherited disease due to gain of function or overexpression. In the former case, for example, the trans-gene might often be a mutant human cDNA with an attached promoter sequence to drive expression of the mutant protein in the same cells as those in which it is expressed in humans. Larger transgenes are possible, too, and have sometimes included artificial human chromosomes.

MAKING PRECISE GENETIC MODIFICATIONS IN INTACT EMBRYONIC STEM CELLS

Another popular way of getting foreign DNA into the mouse germ line begins with cultured **embryonic stem (ES) cells** derived from pluripotent cells of the early mouse embryo. ES cells are immortal and can be used to give rise to all cells of the organism, including gametes. A selected mouse ES cell line is genetically modified in culture and then transferred into an isolated mouse blastocyst. The genetically modified blastocyst is implanted into a mouse, which is bred to obtain genetically modified mice.

Genetic modification of an ES cell line in culture has the big advantage that a very precise change—sometimes just a specific single nucleotide change—can be made to order within any individual gene or locus of interest in intact ES cells. This procedure, known as **genome editing**, requires double-strand DNA breaks to be made at the locus of interest, and two major methods have been used as listed below.

- **Gene targeting.** This is a type of homologous recombination. The specificity is provided by transfecting a *homologous* DNA sequence containing an artificially created desired genetic modification. The normal sequence at the locus of interest is replaced by the introduced sequence with the genetic modification using endogenous nucleases in the cell. An appropriate double crossover will suffice as shown in [Figure 2C](#).

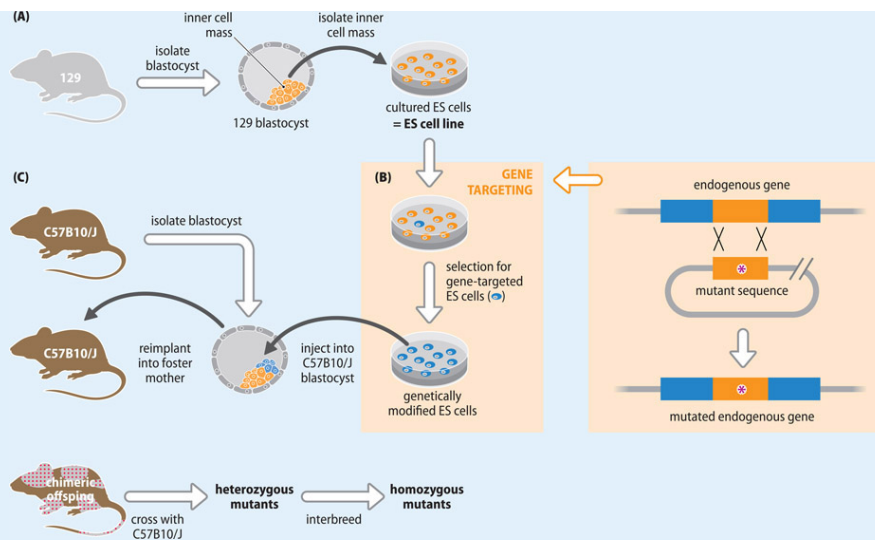


Figure 2 Gene targeting in embryonic stem cells to introduce mutations into the mouse germ

line. (A) Embryonic stem (ES) cell lines are made by excising blastocysts from the oviducts of a suitable mouse strain. Cells from the inner cell mass are cultured to eventually give an ES cell line. (B) An ES cell line can be genetically modified in culture by transfecting a linearized plasmid containing a DNA sequence (orange box) that is identical to part of the endogenous target gene, except for a genetically engineered desired mutation (magenta asterisk). Double recombination (X) allows the desired mutation to be introduced into the endogenous gene. (C) The genetically modified ES cells are injected into an isolated blastocyst from another mouse strain with a different coat color, and the blastocyst containing modified ES cells is implanted into a foster mother of the same strain. Subsequent development of the introduced blastocyst can generate chimeric offspring that can be readily identified because they have differently colored coat patches. Backcrossing of chimeras can produce heterozygous mutants (if the genetically modified ES cells have contributed to the germ line); subsequent interbreeding generates homozygous mutants.

- **CRISPR-Cas.** This new method is simple and can be conducted comparatively quickly. The specificity is provided by RNA sequences designed to bind to selected target sequences on opposing DNA strands at the desired locus. Cleavage of the DNA at the desired locus is carried out on both strands using engineered Cas nucleases introduced into the ES cells. We describe this method later in the text.

The techniques for germline modification have been widely used to make loss-of-function mutations to inactivate a gene (these mutations are known as **gene knockouts**, and mice containing them are called knockout mice). Homozygous loss-of-function mouse mutants are often used to model human recessive disorders, but the method also delivers heterozygous mutants that might show phenotypes too. If the homozygous condition is lethal, the mutation can be maintained in heterozygotes, and the mutant strain can be stored for decades by freezing cells in liquid nitrogen. Variant methods can be used to make

desired subchromosomal duplications and deletions, translocations, and so on (*chromosome engineering*).

Inevitably (because it is simpler to do so), most of the artificially created disease models are intended to replicate monogenic disorders. Some good disease models have been produced, and they have been very helpful in allowing us to gain insights into the molecular basis of human diseases, and in testing gene therapies and other new treatments.

When rodent disease models can be inadequate

Rodent models are generally extremely valuable, but are limited in some ways. Mice are small and are less well-suited than larger mammals to physiological analyses. Larger animal disease models including dog, pig, sheep and primate models have been constructed for some disorders.

Because of species differences, rodent models may quite often fail to replicate some aspects of the human phenotype that they were intended to mimic. In some cases, they are simply inadequate to the task. Disorders such as autism, schizophrenia, and Alzheimer disease cannot be fully replicated in mice (which lack the complex cognitive and social abilities of primates). Many neuroactive drugs have shown early promise in mice but failed in human trials.

As a result of these difficulties, and because of the recent emergence of a transformative technology, genome editing using the CRISPR-Cas system, there has been renewed interest in making primate disease models. Because it also offers interesting therapeutic potential we consider the CRISPR-Cas method of genome editing in the next section.

9.4 GENE THERAPY FOR INHERITED DISORDERS: PRACTICE AND FUTURE DIRECTIONS

Gene therapy has had a roller-coaster ride over three decades; periods of over-optimism would be followed by bouts of excessive pessimism in response to significant setbacks. The first undoubted successes were reported in the early 2000s, and the number of successful reports is beginning to increase significantly.

By 2021, the Gene Therapy Clinical Trials Worldwide Database (available at <http://www.abedia.com/wiley/>) had listed over 3000 such trials. Close to two-thirds of them have been aimed at treating cancers and have often been of limited clinical value—we consider the general difficulties against the broader background of cancer therapy in [Chapter 10](#). Here we focus on gene therapy for inherited disorders where the approach is to modify the disease cells genetically.

Other gene therapy trials have been split almost equally between monogenic disorders, complex cardiovascular diseases, infectious diseases, and other categories. However, only 3 % of the listed trials are phase III trials, in which the efficacy of the therapy is tested on a large scale. Despite the limited number of trials, monogenic diseases have always been high on the gene therapy agenda, and the first definitive successes have been in that area.

Multiple successes for *ex vivo* gene supplementation therapy targeted at hematopoietic stem cells

Successful *ex vivo* gene therapy trials have been carried out for various blood disorders and some storage disorders by targeting bone marrow cells or peripheral blood lymphocytes enriched for hematopoietic stem cells. Our blood cells are short-lived and need to be replaced by new cells derived from self-renewing hematopoietic stem cells. These cells, which are found mostly in the bone marrow (and to a smaller extent in peripheral blood), give rise to all of the many different types of blood cell and also to some tissue cells with immune functions ([Figure 9.17](#)).

For some disorders treated in this way, alternative treatments have sometimes been used, but they are either very expensive or very risky (see below). For some blood disorders, treatment with purified gene product (such as recombinant proteins) is an option, but it is extremely expensive. Bone marrow transplantation has occasionally been used.

In **allogeneic** bone marrow transplantation the donor is often a family relative, such as a sibling, but complete HLA matching of donor and patient is rare (even for siblings there is only a 1 in 4 chance) and sometimes transplantation is attempted using partial HLA matching between donor and recipient. That may result in a severe *graft-versus-host disease*, in which immune system cells originating from the donor bone marrow interpret the cells of the patient as being foreign and mount a strong immune response against them. As a result, the procedure carries a 10–15 % mortality risk that increases to 35 % if the recipient has previously received irradiation treatment (in an attempt to kill many of the original hematopoietic stem cells, so that the transplanted stem cells might expand to become the dominant type).

The advantage of *ex vivo* gene therapy here is two-fold. It is significantly less expensive than using purified proteins. Secondly, it is much less risky than bone marrow transplantation because the genetically modified transplanted cells derive originally from the patient (**autologous** transplantation).

Safety issues in gammaretroviral integration

The first gene therapy successes came in treating severe immunodeficiencies. In severe combined immunodeficiency (SCID) the functions of both B and T lymphocytes are

defective. Affected individuals have virtually no functioning immune system and are extremely vulnerable to infectious disease.

The most common form of SCID is X-linked; inactivating mutations in the *IL2RG* gene means a lack of the γ_c (common gamma) subunit for multiple inter-leukin receptors, including interleukin receptor 2. (Lymphocytes use interleukins as **cytokines** or chemical messengers that help in intercellular signaling, in this case between different types of lymphocyte and other immune system cells; lack of the γ_c cytokine receptor subunit has devastating effects on lymphocyte and immune system function.) Another common form of SCID is due to adenosine deaminase (ADA) deficiency; the resulting buildup of toxic purine metabolites kills T cells. B-cell function is also impaired because B cells are normally regulated by certain types of regulatory T cell.

The first SCID gene therapy trials involved *ex vivo* gammaretroviral transfer of *IL2RG* or *ADA* coding sequences into autologous patient cells. To aid the chances of success, bone marrow cells from the patient were further enriched for hematopoietic stem cells by selecting for cells expressing the CD34 surface antigen, a marker of hematopoietic stem cells ([Figure 9.18](#)). By 2008, 17 out of 20 patients with X-linked SCID and 11 out of 11 patients with ADA-deficient SCID had been successfully treated and retained a functional immune system (for more than nine years after treatment in the earliest patients).

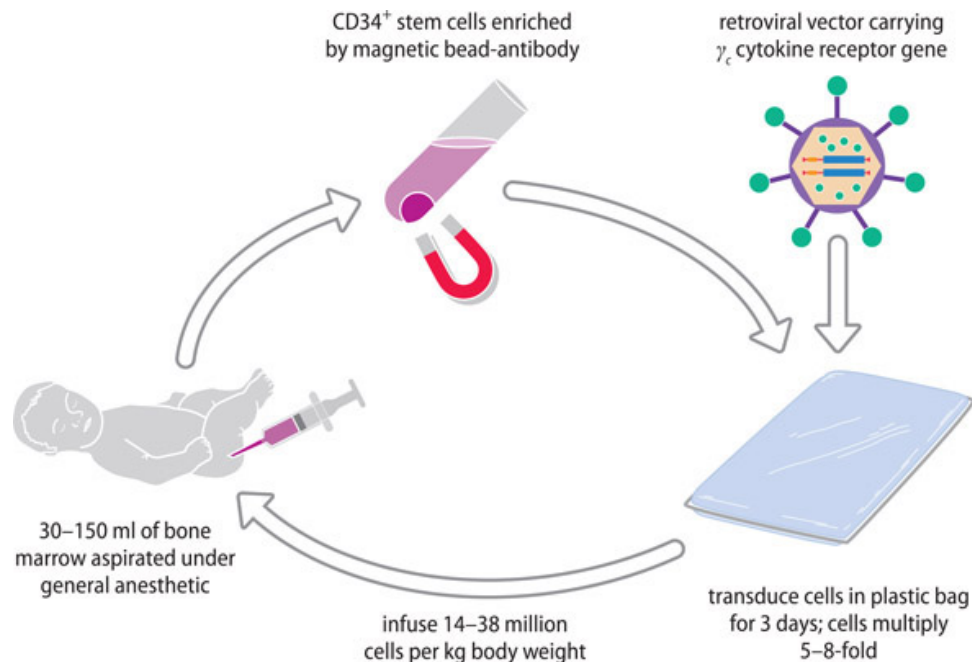


Figure 9.18 The first successful gene therapy: *ex vivo* gene therapy for X-linked severe combined immunodeficiency disease. Bone marrow cells were removed from the patient and antibody affinity was used to enrich for cells expressing the CD34 antigen, a marker of hematopoietic stem cells. To do this, bone marrow cells were mixed with paramagnetic beads coated with a CD34-specific monoclonal antibody; beads containing bound cells were removed with a magnet. The transduced stem cells were expanded in culture before being returned to the patient. Readers interested in the details can find them at PMID 10784449 and 11961146.

Although the use of integrating retrovirus vectors was beneficial in terms of efficiency, the chromosomal insertion of the transgenes was unsafe and led to the development of leukemia in several patients. The same kind of approach has been successfully applied to some other blood disorders. However, oncogene activation in some other cases also led to leukemia in patients or silencing of the inserted transgenes.

It has become clear that gammaretroviral vectors have a pronounced tendency to integrate close to transcriptional start sites, and the long terminal repeats carry very powerful promoter and enhancer sequences that can readily activate the expression of neighboring host cell genes. Retrovirus technology has been made safer by replacing the powerful virus promoter/enhancers by more moderate mammalian control sequences.

More recently, self-inactivating lentivirus vectors have become the first choice for integrating vectors in gene therapy. They have the advantage of long-lasting high-level expression but are much safer than retroviruses: abnormal activation of an endogenous gene is very rarely triggered when lentivirus vectors integrate into chromosomes, mostly because they do not have the same tendency as gammaretroviruses to insert close to transcriptional start sites.

***In vivo* gene therapy: approaches, barriers, and recent successes**

In vivo gene therapy involves the transfer (usually direct) of a genetic construct into post-mitotic disease cells at specific sites in the body (such as muscles, eyes, brain, liver, lung, heart, and joints). Because the intended target cells are nondividing cells, there is no need to insert genes into chromosomes, and so the viral vectors that are used are typically based on non-integrating DNA viruses.

Delivery using adenovirus and adeno-associated virus vectors

Early *in vivo* gene therapy trials often used adenovirus vectors to transfer therapeutic transgenes. These allow high-level expression, and some adenovirus vectors can accept inserts as large as 35 kb (much larger than the vast majority of full-length human cDNA sequences). However, harmful immune and inflammatory responses have sometimes resulted in fatalities. (Because the vectors are non-integrating, and the expression of introduced transgenes is often somewhat transient, short-term and repeated administration would be necessary for sustained expression, but that only exacerbates the immune response).

The safety problems with adenovirus vectors, has prompted a switch to using adeno-associated viruses (AAVs), which are nonpathogenic and are quite unrelated to adenoviruses (their name comes from their natural reliance on simultaneous infection by a helper virus, often an adenovirus). Their most important advantage is that they can permit the robust *in vivo* expression of transgenes in various tissues over several years while exhibiting little immunogenicity and little or no toxicity or inflammation. Multiple different serotypes of

AAV have been isolated, and some have a usefully narrow tropism, such as AAV8 (strongly tropic for the liver). There are two downsides. First, a maximum of just 4.5 kb of foreign DNA can be inserted into an AAV vector. Secondly, neutralizing antibodies may be a problem in some people after repeat exposure to the same AAV serovar.

Amenability of disorders to *in vivo* gene therapy

Different disorders may be more or less amenable to *in vivo* gene therapy, largely depending on the efficiency of transgene transfer and expression. That, in turn, partly depends on different types of barrier. Immunological barriers are particularly important when using recombinant virus vectors: as well as posing safety risks, immunological responses can result in transgene silencing (increased host cell cytokine signaling often attenuates the influence of viral promoters).

In addition to immunological barriers, mechanical barriers can also be a major obstacle. Take cystic fibrosis, a disorder that primarily affects the lungs. Gene delivery to the airways using aerosols might seem a very attractive option, given that lung epithelial cells interface directly with the environment. But a combination of immunological and mechanical barriers makes gene therapy a difficult proposition. Lung epithelial cells are locked together by intercellular *tight junctions*, and large numbers of macrophages are on patrol, readily intercepting and destroying viral vectors. And to top that, there is a natural layer of mucus on the epithelial surface that becomes thicker in individuals with cystic fibrosis, impeding gene transfer.

Some parts of the body are *immunologically privileged sites* in which immune responses to foreign antigen are much weaker than in most other parts of the body (as a result of blood-tissue barriers or a lack of lymphatics, for example). They include the brain and much of the eyes. Additional advantages of the eyes are their accessibility and also their compactness (compare the need for multiple injections at diverse skeletal muscle sites in disorders such as Duchenne muscular dystrophy).

The liver, too, is a quite accessible organ (via direct injection, injection into the hepatic portal vein, or even injection into a peripheral vein); because it has a primary role in biosynthesis, the liver has become a popular target for gene delivery. A wide range of metabolic disorders are caused by defective synthesis of proteins manufactured in the liver (such as blood clotting factors VIII and IX, which are deficient in hemophilia, and many enzymes in inborn errors of metabolism).

Two early examples of successful *in vivo* gene therapy

Hemophilia B (OMIM 306900) is an X-linked recessive disorder caused by a deficiency of blood clotting factor IX. The disorder can be treated by protein therapy (using clotting factor

concentrates), but at huge cost. Remarkably, in a study reported by Nathwani et al in the *New England Journal of Medicine* in 2011 (PMID 22149959) a single intravenous injection of a recombinant AAV construct with a factor IX coding sequence could successfully treat patients with hemophilia for more than a year, even though factor IX expression levels were about 10 % or less of the normal values.

In type 2 Leber congenital amaurosis (OMIM 204100), the principal clinical feature—profound loss of vision—usually presents at birth. In the type 2 form, the blindness results from inactivating mutations in both copies of the *RPE65* gene, causing severe retinal degeneration (*RPE65* encodes a retinal pigment epithelium enzyme). Different *in vivo* gene therapy trials have involved injecting a recombinant AAV construct containing a transgene with the *RPE65* coding sequence into the subretinal space, allowing the transduction of retinal pigment epithelial cells. The trials showed the procedure to be both safe and of considerable clinical benefit. In the largest clinical trial, all patients demonstrated increased pupillary response and increased visual field, and a majority of patients demonstrated improved visual acuity.

An overview of RNA and oligonucleotide therapeutics

Popular therapeutic applications for RNA and/or oligonucleotides are summarized in [Figure 9.19](#). All of them work by targeting RNA or oligonucleotide sequences to base pair with complementary RNA or DNA sequences at a disease gene locus in order to obtain some therapeutic benefit. They fill a gap that supplementation (augmentation) gene therapy cannot fill. Supplementation gene therapy has undoubtedly been successful in treating certain recessive monogenic disorders (those where the problem is a genetic deficiency). But it is not suited to treating diseases where the mutant gene makes a positively harmful product. To deal with disorders where pathogenesis results from some type of toxic RNA, or a mutant protein with a gain of function or a dominant-negative effect, RNA and oligonucleotide therapeutics offers two major possibilities:

- *gene suppression/silencing* by specific inhibition of, or induced cleavage of, transcripts of the target gene
- *gene repair* (the pathogenic mutation causing a harmful gene product to be produced is repaired by replacing the mutant sequence with a normal one).

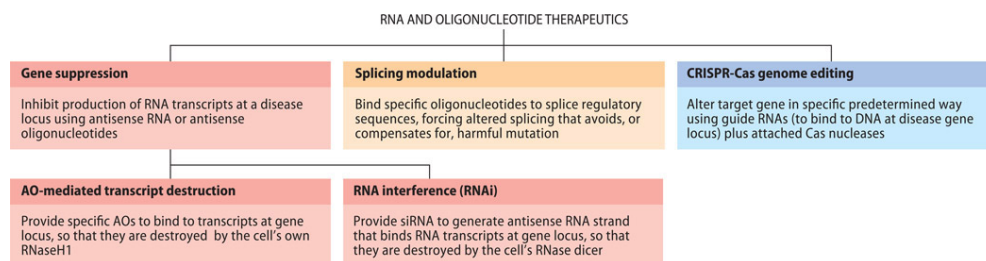


Figure 9.19 An overview of major strategies used in RNA and oligonucleotide therapeutics. All these approaches are focused on RNA transcripts (RNA-targeted therapeutics) except for CRISPR-Cas gene editing, which uses guide RNAs to target DNA sequences. The top right box summarizes the basic CRISPR-Cas method but newer variants called base editing and prime editing use additional enzymes, as described in the text. To make therapeutic oligonucleotides more robust and less susceptible to nuclease attack (when transfected into cells or tissues) their chemical structure is altered, often by having stable phosphorothioate bonds connecting nucleotides instead of phosphodiester bonds and with protective side chains at certain positions on the sugars. siRNA, short interfering RNA, is provided as a RNA duplex but gives rise to a desired antisense RNA within cells. AO, antisense oligonucleotide.

For gene suppression/gene silencing, RNA transcripts from the disease gene locus need to be tagged inside cells by a specific antisense oligonucleotide or RNA in such a way that the bound pathogenic RNA transcripts can be destroyed by a dedicated cellular ribonuclease. It would be optimal, of course, to use a mutant-allele specific antisense oligonucleotide or RNA, but some clinical trials have simply used gene-specific antisense RNA or oligonucleotides that bind to transcripts of both the mutant allele and normal allele in affected heterozygotes. The idea here is that sufficient reduction in expression of the mutant allele might be achieved to obtain therapeutic benefit, while because gene suppression is not 100 % efficient, there is sufficient expression from the normal allele. The two major approaches to gene suppression/gene silencing are listed below.

- *Antisense oligonucleotides (AO)*. The object is to induce ribonuclease RNaseH1 in the cells of an affected individual to selectively destroy transcripts at the disease gene locus. (The natural function of RNaseH1 is to destroy the RNA strands of DNA-RNA hybrids in the cell.) To do this an AO is designed to base pair specifically to transcripts from the disease gene locus, or mutant allele, then transfected into the cells of a patient. The AO must contain a significant number of deoxyribonucleotides so that the RNA-AO hybrid becomes a target for RNaseH1 cleavage of the bound RNA transcripts (often the AO is designed to have a central section of ~10 deoxyribonucleotides flanked by five ribonucleotides on each side).
- *Short interfering RNA (siRNA)*. The object is to exploit a natural innate defense mechanism, RNA interference (RNAi), to specifically destroy RNA transcripts from the disease gene locus or mutant allele. RNA interference is induced after transfecting into the cells of an affected individual a specific siRNA (a double-stranded RNA that will ultimately generate an antisense single-strand RNA) or a gene encoding a precursor of that specific siRNA. The antisense RNA will bind specifically to RNA transcripts from the disease gene locus or mutant allele and thereby tag them to be destroyed by the cells' dicer ribonuclease. In the section following this one we explain RNA interference in detail and show how it has been exploited for therapeutic purposes.

The transformative CRISPR-Cas genome editing method of genome editing (also called **gene editing**) can also be thought of as a type of RNA therapy, even although it works at the DNA level. A potentially powerful therapeutic application of CRISPR-Cas gene editing is in repairing a pathogenic point mutation, replacing the mutant sequence by a normal one. It is crucially dependent on transfecting genes into the desired cells in order to make *guide RNAs* (which are designed to base pair to specific sequences flanking a pathogenic point mutation in the gene of interest) and a Cas (Crispr-associated) nuclease. The object is usually to steer Cas nucleases to cut at pre-determined target sites in the vicinity of the pathogenic mutation within intact cells as a precursor to repairing the mutant gene. We outline this potentially highly promising procedure below.

Another application of oligonucleotide therapeutics, therapeutic splicing modulation, is necessarily limited in scope. The idea is that by designing suitable antisense oligonucleotides to bind to—and thereby blockade—a specific splice junction, an exon with a harmful mutation might be skipped, avoiding the harmful effect of that mutation. Of course, for most mutant genes, this type of exon skipping therapy cannot be applied: even if the induced exon skipping did not induce a translational frameshift, valuable sequence could be expected to be lost. Because of certain unusual characteristics, however, some examples of successful splicing modulation have been possible in treating Duchenne muscular dystrophy and spinal muscular atrophy (see [Clinical Box 12](#)).

CLINICAL BOX 12 SPLICE MODULATION THERAPY FOR DUCHENNE MUSCULAR DYSTROPHY AND SPINAL MUSCULAR ATROPHY

Splice modulation therapy has particularly been applied to treating neuromuscular diseases (see PMID 23631896, as exemplified by the two cases below).

EXON SKIPPING THERAPY FOR DUCHENNE MUSCULAR DYSTROPHY

Duchenne muscular dystrophy (DMD), a severe and progressive X-linked recessive muscular dystrophy, results from a deficiency of the dystrophin protein. Affected boys need to use wheelchairs by 12 years of age, develop additional cardiomyopathy after age 18 years, and often die before 30 years of age. The disorder is primarily due to intragenic deletions in the 2.4 Mb dystrophin gene.

Surprisingly, deletion of a large central portion of the dystrophin gene, up to 1 Mb, can result in the milder Becker muscular dystrophy (BMD), but deletion of a single nucleotide within an exon in that same 1 Mb central region can result in severe DMD. Two observations explain that apparent anomaly. First, the sequence of the central region of dystrophin is not so important: it acts simply as a flexible linker between the two functionally important parts, the N-terminal and C-terminal domains; large, in-frame

internal deletions may reduce dystrophin performance but some functional protein remains, resulting in a mild BMD phenotype. Secondly, internal frameshifting deletions are consistently associated with DMD.

Exon skipping therapy in DMD patients who have a central frameshifting deletion aims to restore the translational reading frame: the net effect should resemble in-frame deletions associated with milder BMD. [Figure 1](#) shows how the antisense oligonucleotide eteplirsen can induce skipping of exon 51 to restore the translation reading frame in patients with a deletion of exon 50. Skipping of exon 51 can also restore the translational reading frame for several other common internal deletions in the dystrophin gene. Therapeutic skipping of exon 53 has also been carried out using another AO, golodirsen—see [Table 9.8](#). For a review of splicing therapy in neuromuscular disease, see PMID 23631896

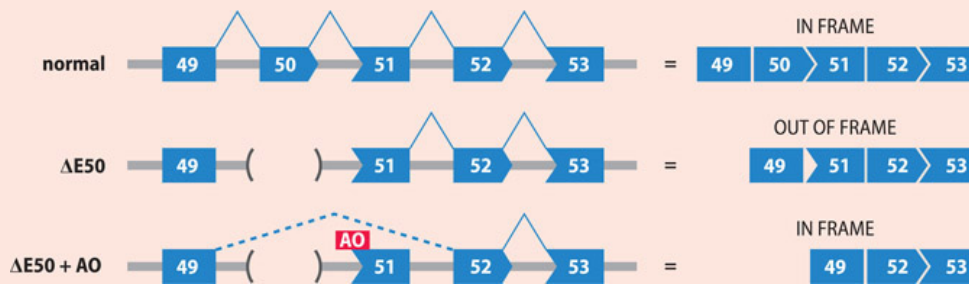


Figure 1 An example of exon skipping therapy for Duchenne muscular dystrophy (DMD). Deletion of the 109-nucleotide exon 50 ($\Delta E50$) in the dystrophin gene results in a shift in the translational reading frame for dystrophin mRNA, resulting in DMD. Local intramuscular injections with eteplirsen, a specific antisense oligonucleotide (AO) that can bind to and blockade splice regulatory sequences at the start of exon 51, causes skipping of exon 51 in $\Delta E50$ patients and splicing of exon 49 to exon 52 without a frameshift (the total number of missing nucleotides = 342). Significant clinical benefit is evident, as measured by improved walking statistics when compared with controls.

TABLE 9.8 THE FIRST WAVE OF MARKETED ANTISENSE OLIGONUCLEOTIDE (AO) AND RNA INTERFERENCE (RNAI) THERAPIES

Type of therapy	Name of therapeutic	Disorder treated	Target gene	PMID
AO gene suppression	inotersen	Hereditary transthyretin amyloidosis	<i>TTR</i> (transthyretin)	29972757, 29972750
AO splice modulation	eteplirsen	Duchenne muscular	<i>DMD</i> exon 51	29752304
	golodirsen	dystrophy	<i>DMD</i> exon 53	32139505
	nusinersen	Spinal muscular atrophy	<i>SMN2</i> exon 7	29443664
RNAi (siRNA)	patisiran	Hereditary transthyretin	<i>TTR</i> (transthyretin)	29972753, 29972750

Type of therapy	Name of therapeutic	Disorder treated	Target gene	PMID
		amyloidosis		
	givosiran	Acute hepatic porphyria	<i>ALAS1</i> (5-amino-levulinic acid synthase)	32521132
	lumasiran	Primary hyperoxaluria type 1	<i>GO</i> (glyoxylate oxidase)	33789010

EXON INCLUSION THERAPY FOR SPINAL MUSCULAR ATROPHY

Spinal muscular atrophy (SMA) is a degenerative motor neuron disorder that leads to muscle atrophy and respiratory failure. Individuals with the most severe form rarely survive beyond 2 years of age. The disease is due to defects in the *SMN1* gene which is part of a cluster of duplicated genes that arose by evolutionarily recent segmental duplication. *SMN2*, a paralog (gene duplicate) of *SMN1*, can produce a protein identical to the SMN1 protein. A single nucleotide change in a splice regulatory sequence, however, causes skipping of exon 7 in 90 % of the *SMN2* transcripts, producing an unstable protein; only 10 % of *SMN2* transcripts make the normal protein (see [Figure 2A](#)).

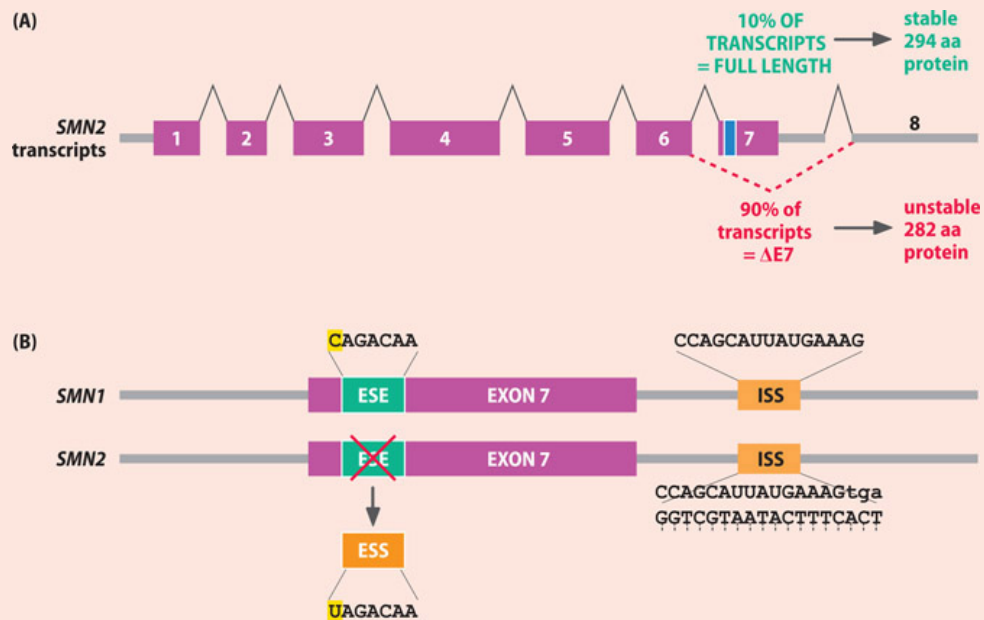


Figure 2 *SMN2* transcripts, exon 7 splice regulation and role of nusinersen. 90 % of *SMN2* transcripts lack exon 7 because a 7 bp exonic splice enhancer (ESE) sequence at the beginning of exon 7 in *SMN1* has been mutated; the resulting sequence UAGACAA now works more as a weak exonic splicing suppressor (ESS). Together with an intronic splice silencer (ISS) in intron 7, the effect is to strongly inhibit exon 7 inclusion in the absence of an exon 7 ESE. The antisense oligonucleotide nusinersen works by base pairing

with the intron 7 splice silencer sequence, thereby blockading it and promoting exon 7 inclusion in *SMN1* transcripts. Interested readers can find a recent review at PMID 29422644.

The nucleotide sequences of *SMN1* and *SMN2* mRNAs differ at just two nucleotide positions, but one of them, a C/U difference near the beginning of exon 7 is critically important. It falls within a critical exonic splice enhancer sequence, **CAGACAA** in *SMN1*, but the equivalent **UAGACAA** sequence in *SMN2* acts weakly in the opposite direction, as a splice suppressor.

The number of *SMN2* genes can vary as a result of unequal crossover and affected individuals with multiple *SMN2* genes are less severely affected. All individuals with spinal muscular atrophy type 4, the mildest form, have four to six *SMN2* gene copies. *SMN2* may be viewed as a poorly efficient back-up gene, but when there are no functional *SMN1* genes, the more backup *SMN2* genes the better. That prompted the idea of a novel therapy: making the back-up *SMN2* gene more effective by promoting exon 7 inclusion in *SMN2* transcripts using an antisense oligonucleotide, nusinersen (see [Figure 2B](#)).

RNA interference therapy

Different diseases are potentially amenable to treatment by taking advantage of RNA interference (RNAi), an innate defense mechanism that protects cells against invading viruses and over-active transposable elements. (A small percentage of our resident transposons are actively transposing; if that percentage were allowed to become too great, the genome could be overwhelmed by transposons inserting into essential genes.)

RNAi is triggered in cells by the presence of double-stranded RNA (which is not normally produced in our cells, except by invading viruses, and by the association of sense and antisense transcripts from highly repeated transposons). The double-stranded RNA is detected and cleaved in cells by a ribonuclease called dicer, producing fragments 21 bp long with recessed 5' ends, known as **short interfering RNA (siRNA)**. They are recognized by special protein complexes, RNA-induced silencing complexes (RISC), that initiate a pathway whereby *any* RNA transcripts containing the same nucleotide sequence as the siRNA are destroyed ([Figure 9.20](#)).

Using RNAi to silence a mutant allele

The pathway shown in [Figure 9.20](#) is concerned with *natural* gene silencing that destroys transcripts from the genes of invading viruses or from transposable elements. It can be artificially exploited to selectively inhibit the expression of a gene of interest within cells. To do that a genetic construct is transfected into cells to produce directly, or indirectly, a gene locus-specific or, preferably an allele-specific, double-stranded siRNA. RISC complexes can then be activated by the allele-specific siRNA to downregulate RNA transcription from, say,

a positively harmful gene in the cells of an affected individual. The bulky, highly charged double-stranded siRNA can be transfected into cells with the assistance of attached lipids; alternatively, a gene encoding short hairpin RNA, a siRNA precursor, is transfected into the cells ([Figure 9.21](#)).

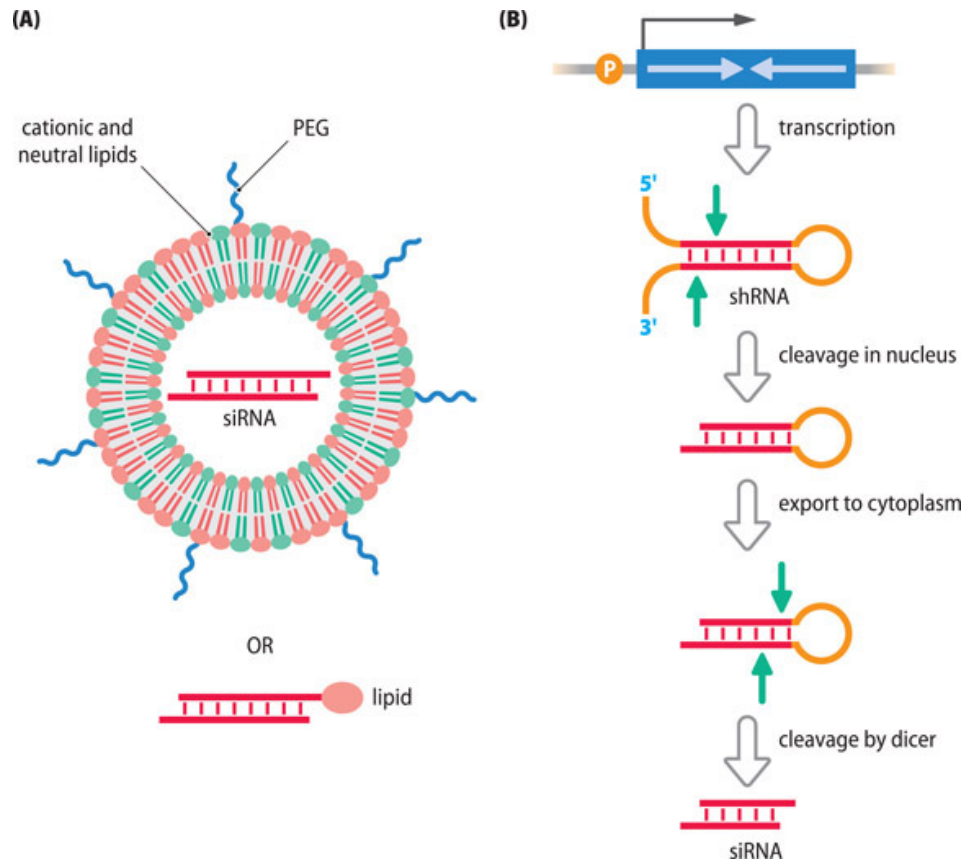


Figure 9.21 Two different types of siRNA delivery to cells. (A) Direct siRNA delivery. The interfering RNA needs to be short because transfecting long double-stranded RNA into mammalian cells results in *indiscriminate* destruction of mRNAs. Two short oligoribonucleotides can be chemically synthesized to form a siRNA duplex that will have two-nucleotide 3' overhangs like the natural siRNAs shown in [Figure 9.20](#). The RNA sequence can be chosen to be gene- and allele-specific (a unique sequence from an exon of the target gene that shows differences between mutant and normal alleles). Because siRNA is highly charged and comparatively large, it cannot easily cross plasma membranes; it has to be complexed with lipid-based carriers such as liposomes or conjugated with a lipid such as cholesterol. (B) Delivery of a gene encoding a siRNA precursor. Recombinant viruses are used to ferry an artificial gene construct into cells. The gene has inverted repeats (pale blue arrows) and is transcribed in the nucleus to make a single-stranded RNA with two long complementary sequences, allowing it to fold back to form a mostly double-stranded shorthairpin RNA (shRNA). The shRNA will be processed by the cell's RNAi machinery to yield a specific siRNA duplex in the cytoplasm.

RNAi therapy seeks to silence a specific gene by designing an artificial siRNA to target transcripts of that gene, causing their degradation. It has not been easy: complete gene silencing is difficult to obtain (but significant lowering of the amount of harmful gene product may produce clinical benefit), there is the risk of off-target effects, and efficient delivery to tissues and cells has been problematic. After a series of failures (mostly because of the delivery problem), a corner was turned in 2018: delivery of patisiran, a siRNA specific for the transthyretin gene *TTR*, provided clinical benefit for people with hereditary transthyretin amyloidosis, gaining regulatory approval. The normal transthyretin protein forms a tetramer, but mutant transthyretin leads to harmful amyloid deposits in different tissues. Affected individuals show slowly progressive peripheral sensorimotor and/or autonomic neuropathy as well as non-neuropathic changes of cardiomyopathy, nephropathy, vitreous opacities, and CNS amyloidosis.

Subsequently, Alnylam, the company responsible for producing transthyretin, developed an effective way of delivering siRNA to liver (by complexing the siRNA with N-acetylgalactosamine, targeting delivery to hepatocytes), and have been researching ways of efficiently targeting other tissues. **Table 9.8** provides a list of RNAi and antisense oligonucleotide drugs that have been approved for treating genetic disorders.

Future therapeutic prospects using CRISPR-Cas gene editing

Genome editing describes any method allowing precise genetic alteration to a pre-determined locus in intact cells. In standard genome editing a first requirement is to have some way of making DNA breaks specifically at the locus of interest. Thereafter, the breaks are exploited in some way in order to obtain a specific desired change to the DNA sequence at a locus of interest. The first such method—*gene targeting*—used endogenous endonucleases that work naturally in homologous recombination; the specificity comes from inserting a transgene carrying a DNA sequence homologous to the gene of interest plus a DNA marker sequence to select for recombination events in which the original sequence was replaced by a desired sequence. The method, detailed in [Box 9.2](#), is rather laborious and time-consuming.

In the newer genome editing methods, genetically engineered constructs are transfected into cells of an affected individual and expressed to make artificial *programmable* endonucleases, ones designed to cut the DNA at pre-determined target sites in the genome. The endonucleases must be transported to the desired target site by being bound to artificially designed RNA or protein **guide sequences** that are specifically designed to bind to the desired target sequences. (Effectively it is the guide sequence that is programmable, being designed to bind a specific sequence, usually 18–20 nucleotides long, in the genome). After a guide sequence has transported its attached nuclease to the correct target sequence, the nuclease cuts the bound DNA strand in the immediate vicinity of the binding site.

CRISPR-Cas: origins

Genome editing using protein guide sequences (such as zinc finger nucleases) is laborious. Happily, however, the CRISPR-Cas system, which uses RNA guide sequences, is fast, versatile and comparatively simple, and has been transformative. (The acronyms are: CRISPR—clustered, regularly interspersed short palindromic repeats; Cas—CRISPR-associated). Like restriction nucleases and RNAi, CRISPR-Cas genome editing was developed from a bacterial self-defense mechanism, in this case a form of adaptive immunity. Here two types of RNA play a critical role:

- guide RNAs, each having a distinctive guide sequence at the 5' end (originating from previously captured virus or plasmid sequences) plus a common 3' repeat sequence, the R sequence—see [Figure 9.22](#).

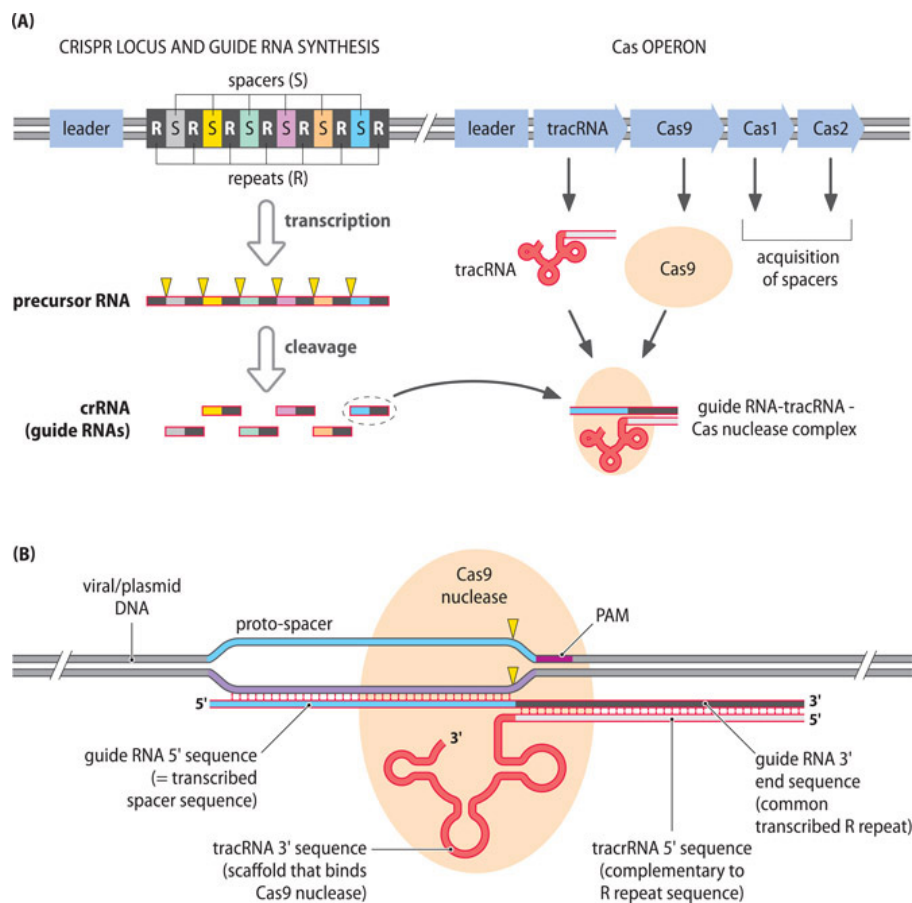


Figure 9.22 How CRISPR-Cas works to defend prokaryotes. (A) A prokaryotic CRISPR locus has multiple copies of an invariant host-cell repeat sequence (R) with interspersed DNA spacers (S), different sequences captured from “proto-spacer” sequences in the genomes of viruses or plasmids that have previously infected the prokaryotic cell. A CRISPR locus produces various short CRISPR *guide RNAs* (gRNAs), each with a distinctive transcribed spacer sequence at its 5' end (a *guide*

sequence) but a common (invariant) repeat sequence (R) at its 3' end. In this example from *S. pyrogenes*, the Cas operon makes a Cas9 (Crispr-associated) nuclease, and a tracrRNA (**trans**activator RNA). The latter works as an intermediate, forming a ternary complex by binding a guide RNA (using its 5' end sequence to base pair with the R sequence of a guide RNA), and its 3' end sequence to bind a Cas nuclease—see panel B for an expanded view. (B) Defence against recurring virus or plasmid invasion. The blue proto-spacer sequence on the invading DNA is identical to a spacer previously captured and stored in a CRISPR locus (the spacer in the dashed ring on the right of the CRISPR locus in panel A). The 5' end of an appropriate guide RNA (as part of a guide RNA-tracrRNA-Cas nuclease complex) can therefore bind to a complementary sequence on the invading DNA; the binding occurs just upstream of a short protospacer-associated motif (PAM) in the virus/plasmid DNA (which in the case of the Cas9 nuclease is 5' NGG 3', where N = A, C, G or T). Once the Cas9 nuclease has been brought close to the target viral/plasmid DNA it cuts the DNA on both strands (vertical yellow darts).

- a tracrRNA (**trans-activating** RNA) having at the 5' end a sequence complementary to the R sequence, and at the 3' end a sequence that can act as a scaffold to specifically bind a CRISPR-associated (Cas) nuclease—see [Figure 9.22](#).

Exploiting CRISPR-Cas for therapeutic genome editing

CRISPR-Cas genome editing began by exploiting the natural CRISPR-Cas system to make double-stranded DNA breaks in the genomes of complex cells. Synthetic guide RNAs would be designed to recognize target sequences 18–21 nucleotides in length. After a double-strand break would be made at the desired locus, cellular DNA repair mechanisms would be activated that could produce desired sequence changes at the gene of interest, or be manipulated to do so. Recall from [Section 4.2](#) that two DNA repair pathways are dedicated to repairing double-strand DNA breaks in our cells, and they can be exploited in genome editing as detailed below.

- Nonhomologous end joining (NHEJ) is commonly used by cells, operates throughout the cell cycle and prioritizes speed over accuracy. It rapidly joins the ends of a broken DNA but very small mistakes are made during repair (a nucleotide is often deleted or inserted, for example). When carrying out *ex vivo* gene therapy, cells subjected to genome editing can be sampled to see if, in some of them, NHEJ produces the desired sequence change. This type of DNA repair can be used to inactivate a functional sequence for a therapeutic purpose.
- Homology-dependent DNA repair. The homologous recombination DNA repair pathway is available to replicating cells after S phase. When used naturally by cells it can make accurate repairs to a double-strand DNA break using the unbroken sister chromatid as a DNA template. A type of homologous recombination can be used to repair a pathogenic mutation by engineering a double-strand break in a mutant allele

and simultaneously providing a transgene containing a sequence from a normal allele and flanking sequences with 100 % sequence homology. This procedure, known as *homology-directed repair (HDR)*, can repair the mutant allele, converting the sequence of a disease allele to that of a normal allele ([Figure 9.23](#)).

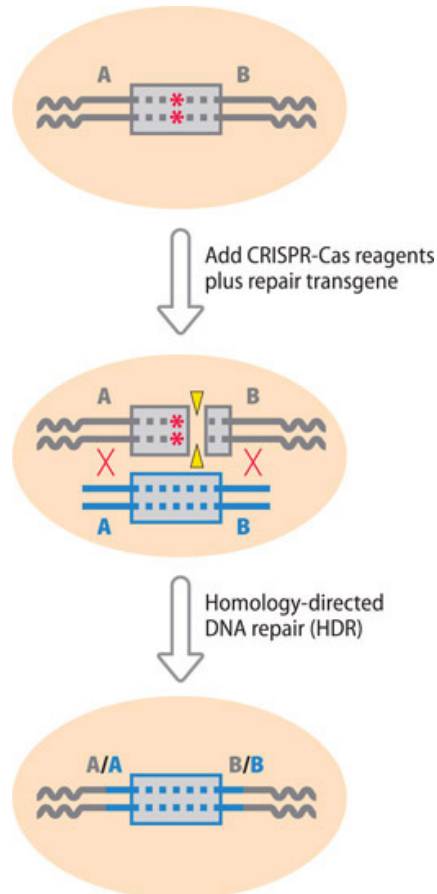


Figure 9.23 Homology-directed DNA repair A chromosomal gene with a pathogenic mutation (shown here as a red asterisk within an exon) can be repaired by using CRISPR-Cas to make a double-stranded break in the immediate vicinity of the mutation. A provided transgene (blue), with the normal DNA sequence for the exon and flanking intronic sequence A and B, can act as a template. Recombination (marked by red X) between the flanking sequences of chromosomal exon and the transgene can replace the mutant sequence by a normal sequence. Note that for point mutations, a short single-stranded oligonucleotide is often preferentially used as a template DNA.

A single hybrid RNA (sometimes called sgRNA) is commonly used in modern CRISPR-Cas genome editing, with a 5' guide sequence from a guide RNA joined by a linker sequence to the 3' scaffold sequence of tracrRNA. The target sequence must have a suitable protospacer motif (PAM) immediately downstream of it. In modern CRISPR-Cas genome editing, modified Cas nucleases, called *DNA nickases*, are often used that are designed to cut a *single* DNA strand, and two closely neighboring target sequences are often designed to be

bound, as shown in [Figure 9.24](#). That has two important consequences. First, it minimizes inappropriate binding of the guide sequence to *off-target* sequences (any sequence other than the desired target sequence). Secondly, the accuracy in making a desired change may be improved: errors are common in artificial homology-directed DNA repair when both DNA strands are broken.

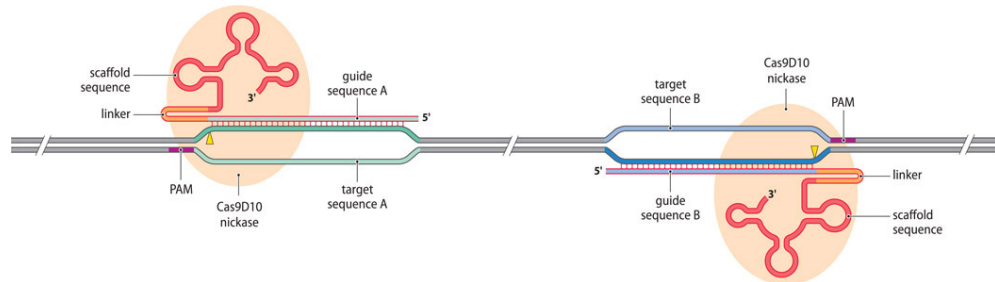


Figure 9.24 Making single-strand DNA breaks at neighboring target sequences using DNA nickases in CRISPR-Cas genome editing. The object is to reduce off-target effects by delivering a modified Cas nuclease to two closely located target sequences, A and B, at a locus of interest. The modified Cas nuclease is a DNA nickase, able to cut just one of the two DNA strands (at positions shown by the vertical yellow darts). The DNA nickases are delivered to their target sequence by a single type of RNA (containing the guide sequence from a guide RNA attached via a linker sequence to a tracrRNA scaffold sequence that binds the DNA nickase).

Subsequent variants of the standard CRISPR-Cas design include methods where Cas nucleases are fused to proteins with some type of enzyme activity. Two such methods are listed below.

- *Base editing.* A catalytically impaired Cas nuclease is fused to an enzyme that converts one base to another, without the need for cutting the DNA or for a DNA template. Initial work involves cytosine base-editors, such as the cytosine deaminase APOBEC (converts C to U; the U is subsequently converted to T after DNA replication or DNA repair) and adenine base-editors. This method may be superseded by the one directly below.
- *Prime editing.* A Cas nuclease is fused to a reverse transcriptase, and a special prime-editing guide RNA (pegRNA) is used. In addition to the usual guide sequence and Cas-binding scaffold sequence, the pegRNA has an additional “replace” sequence; one that acts as a template for inserting a short desired sequence. Because it offers precise genome editing, and can carry out any type of short sequence correction, this may become a very popular method.

At the time of writing (January 2022) it is still early days for therapeutic applications of genome editing, but the technology is advancing rapidly. Some clinical trials have used older-style technologies featuring protein guide sequences, but now the focus is very much on CRISPR-Cas genome editing.

As in the case of standard supplementation gene therapy, *ex vivo* gene editing therapies are primarily being used because of the huge advantage of being able to study autologous cells in culture and then selecting those that show the desired genetic modification before injecting the modified cells into a patient. Gene editing therapies may offer some advantages over gene supplementation therapies because of the size limitation of insert DNA that can be accommodated in gene therapy vectors. For certain genes, a cDNA may simply be too large to be accommodated in vectors, notably AAV vectors. But it too has some downsides— notably the potential problem for immune reactions when expressing the Cas nuclease.

Genome editing strategies that rely on NHEJ (nonhomologous end joining) DNA repair are inherently more efficient than HDR (homology-directed DNA repair), and are being pursued where the object is to inactivate a gene or a regulatory sequence (see [Table 9.9](#) for examples).

TABLE 9.9 SOME EARLY EXAMPLES OF USING GENOME ENGINEERING IN NON-CANCER CLINICAL TRIALS

Type of approach	Disorder treated and basis of method	Technology and reference
Disable a cell receptor to prevent virus infection	HIV-AIDS. Inactivate the gene making the CCR5 receptor on helperTcells (required for HIV infection) by the HIV virus. <i>Ex vivo</i> gene therapy using autologousCD34 ⁺ T cells.	ZFN, TALEN, CRISPR
Alter regulatory signals so as to reactivate a silenced gene to make a protein to supplement genetic deficiency of a closely related protein	Beta thalassemia/sickle cell disease. The idea is to restore gamma globin production (normal in fetal stages only) to make up for lack of normal beta globin in affected individuals. May involve targeting the cis-acting BC11A repressor of the gamma-globin gene or its target sequence.	ZFN, CRISPR (see Figure 9.3 of PMID 32775490)

ZFN (zinc finger nuclease) and TALEN (TALE nuclease) are older, cumbersome genome editing technologies that use protein guide sequences.

Therapeutic applications of stem cells and cell reprogramming

As detailed earlier, many of the successes in gene supplementation therapy, notably *ex vivo* gene therapies, have been dependent on cell transplantation and therefore also constitute a type of cell therapy. But stem cells also offer the prospect of **regenerative medicine**, a type of *cell supplementation therapy* in which stem cell cultures are manipulated so as to simply provide replacement cells for cells lost through disease (or injury). Take cultured human pluripotent stem cells, which can proliferate indefinitely and differentiate into all types of cells in the body. If efficiently directed down the correct differentiation pathway, they could in principle provide replacement cells to supplement a deficiency of some functioning cells in

a patient. Complex disorders arising from loss of a particular cell type, are possible targets for this type of therapy, including Type I diabetes and Parkinson disease (loss of pancreatic beta cells and dopaminergic neurons, respectively), as are some injuries (for example, to the spinal cord).

Sources of cells for cell therapy

Most cell therapies have used cultured human pluripotent stem cells. Of these, human embryonic stem cells (ESCs) were the first to be obtained (by culturing surplus human embryos from *in vitro* fertilization centers) and have been with us for more than 20 years. More recently, human somatic cells have been induced to dedifferentiate to produce **induced pluripotent stem cells (iPSCs)** – see [Box 9.1](#), which are more acceptable ethically than ESCs and are now being used in preference to ESCs for therapeutic applications. iPSC technology also has the advantage of permitting, in principle, *ex vivo* genetic modification to be extended to a wide range of genetic disorders, not just disorders of blood or of other cells originating from hematopoietic stem cells. (In this case, iPSCs derived from accessible blood or skin cells in an affected individual would first be genetically modified in culture and differentiated to give a desired cell type; then, the desired, genetically modified, differentiated cells would be returned to an appropriate location in the patient.)

Cells obtained after directed differentiation of ESCs or iPSCs have been used in clinical trials to treat various disorders, including various eye disorders and some complex diseases. But this is still a young field and although there is considerable promise, various obstacles need to be surmounted, as explained below. Another possible source of cells involves *transdifferentiation*, switching from one differentiated cell type to another, such as from fibroblast to neuron. The method is technically difficult, but one interesting possibility of a therapeutic application is to convert astrocytes into neurons *in vivo* to treat Parkinson disease. Astrocytes, a subtype of glial cells, are the most abundant cells in the central nervous system and can be converted into induced dopamine-releasing neurons by blocking expression of an astrocyte protein called PTB. Interested readers can find details of preliminary work using a mouse model of Parkinson disease at PMID 32581373.

Obstacles to overcome in cell therapy

Three major types of obstacle need to be surmounted in cell therapies, as shown in [Figure 9.25A](#). Of these immunogenicity is a major problem. Any transplant of cells runs some risk of immune rejection: cell surface antigens on transplanted cells, most notably variant HLA proteins, may be perceived as foreign. (Note, however that the degree of immune response is less in some sites, including the central nervous system and most of the the eye, which have *immune privilege*: foreign antigens may be tolerated without inducing an inflammatory immune response.) Cells originating from differentiation of ESCs, which derive from donors

of surplus embryos, and so are allogeneic, need to be transplanted into a patient where there is a high degree of matching for HLA antigens. Although transplant of *autologous* iPSC-derived cells back into the patient of origin is not normally expected to provoke immune reactions, the problem here is the huge expense involved in making autologous iPSCs from individual patients.

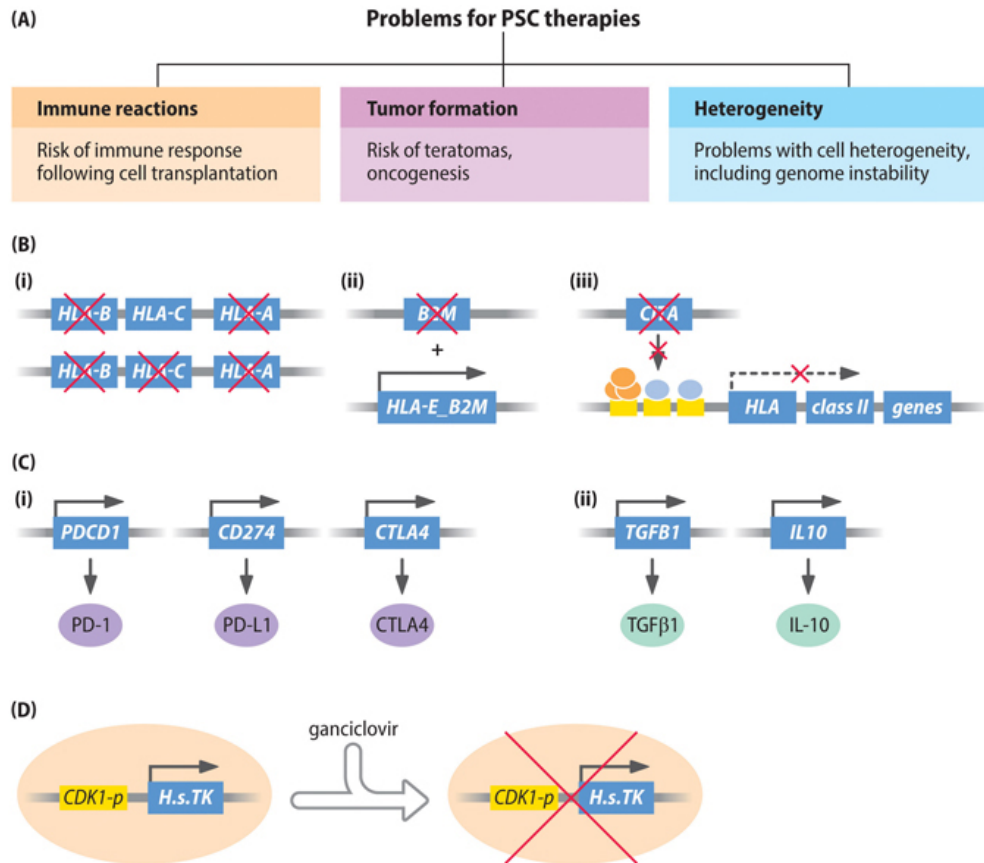


Figure 9.25 Overcoming problems for therapeutic application of pluripotent stem cells (PSCs). (A) The three main difficulties to overcome. (B) Suppressing HLA protein production by genome editing. Immunogenicity can be reduced by deleting five of the six polymorphic class I HLA genes leaving a single, mildly polymorphic HLA-C allele, as shown in B(i), or by deleting the *B2M* gene and supplying a fusion gene combining the non-polymorphic *HLA-E* gene with a *B2M* gene copy, as shown in B(ii). Class II HLA protein production can be switched off by deleting the *CIITA* gene that makes an activator protein required for class II HLA expression—see B(iii). (C) *Immune cloaking*. This can be done by stimulating (i) genes making proteins involved in checkpoint blockade; and (ii) genes making localized immune suppressants. (D) Protecting against teratoma formation. After introducing into cells a suicide gene, the *Herpes simplex* thymidine kinase (*TK*) gene, coupled to a *CDK1* gene promoter (always expressed in dividing cells), treatment with ganciclovir selectively kills dividing cells that express the *TK* gene.

To address the immunogenicity problem, banks of different iPSC lines have been set up in some countries from rare donors homozygous for frequently occurring HLA haplotypes.

According to the degree of HLA matching, an optimal iPSC line could be selected for use with a patient to reduce the chance of immune reaction. An alternative approach is to apply genome editing to iPSCs to make them “immunocompatible” in some way.

Another way of making cells immunocompatible is to suppress expression of polymorphic HLA proteins. The classical class I HLA genes, notably *HLA-A* and *HLA-B*, and to a lesser extent *HLA-C*, make highly polymorphic heavy chains that each combine with an invariant light chain, beta-2-microglobulin, produced by the *B2M* gene. If genome editing simply deletes the *B2M* gene, the genetically modified cells have no class I HLA antigen; although no longer detected by cytotoxic T cells, these “unnatural” cells are liable to be killed by natural killer cells. To counter natural killer cells, genome editing seeks to leave some residual modestly or poorly polymorphic HLA protein ([Figure 9.25Bi,ii](#)). Class II HLA expression can be suppressed by deleting the *CIITA* (Class II major histocompatibility complex transactivator) gene ([Figure 9.25Biii](#)).

A different way of reducing the immunogenicity of therapeutic cells is *immune cloaking*, which involves stimulating the expression of genes that cancer cells use to escape immune detection ([Figure 9.25C](#)). The ultimate aim is to generate “universal” hypoimmunogenic pluripotent stem cell lines to be used as “off-the-shelf” reagents; differentiated cells derived from them could be transplanted into any patient with minimal risk of immune reaction.

Another important obstacle in cell therapy is the possibility of tumorigenesis. Practical difficulties in accurately and efficiently directing iPSCs (or ESCs) to undergo differentiation steps towards a desired differentiated cell type could lead to incomplete differentiation. Residual pluripotent cells might then be transmitted to the patient that could form *teratomas*, tumors composed of heterogeneous cell types derived from the different embryonic germ layers. One way to safeguard against that is to genetically modify cells derived by iPSC differentiation so that they contain a *suicide gene* that is expressed only in dividing cells ([Figure 9.25D](#)).

A special case: preventing transmission of severe mitochondrial DNA disorders by mitochondrial replacement

Mutations in mitochondrial DNA (mtDNA) are a significant cause of human disease: pathogenic mutations are found in at least 1 in 200 of the population, and cause severe multisystem disease in approximately 1 in 10 000 of the population. Pathogenic mtDNA can be maternally inherited, but there are no effective treatments for mitochondrial DNA disorders.

In the clinical management of mtDNA disorders, the emphasis has therefore been on prevention. Preimplantation and prenatal diagnosis (as described in [Chapter 11](#)) are well established in clinical genetic practice as a way of selecting unaffected embryos. However, the results can be difficult to interpret for patients with *heteroplasmic* mtDNA mutations (with variable numbers of mutant and normal mtDNAs in each cell). In addition, an increasingly large group of diseases are recognized to be caused by *homoplasmic* mtDNA

mutations (all, or almost all, of the mtDNA molecules are mutant). Here, prevention by selecting an unaffected embryo is not an option—all the offspring would inherit the pathogenic mutation in the maternal egg, and this type of genetic defect can be associated with a very high disease recurrence risk.

Women who are carriers of serious mtDNA disorders caused by homoplasmic mutations or where the proportion of mutant mtDNA is close to 100 % therefore face the bleak prospect of having severely affected children in *each* pregnancy. The usual option of prenatal diagnosis to select healthy embryos cannot be achieved if every embryo will contain mutant mtDNA.

The transmission of homoplasmic mutations can, however, be avoided if the defective maternal mtDNA is replaced by mtDNA from an asymptomatic donor. That can be done by *in vitro* fertilization using either of two slightly different approaches, replacing mtDNA at the zygote level or the oocyte level (**Figure 9.26**). The resulting human embryos appear to be viable *in vitro*, and the degree of mutant DNA carryover is low or undetectable. The clinical application of this mitochondrial replacement method became legally permissible in the United Kingdom in 2015, and is now a nationally commissioned NHS services in England and Wales. It constitutes an exceptional example of germline gene transfer in humans, and we return to consider the associated ethical considerations in [Chapter 11](#).

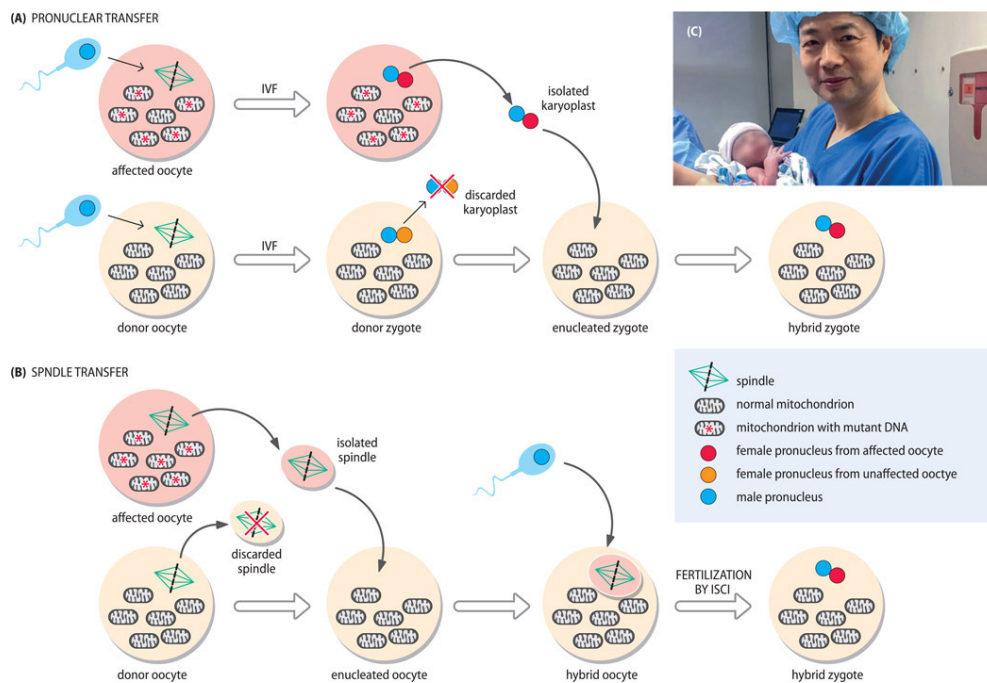


Figure 9.26 Mitochondrial replacement therapy to prevent transmission of severe mtDNA disease. A donor provides an enucleated oocyte with healthy mitochondria and normal mtDNA; the prospective parents provide the nuclear genome, either after or before *in vitro* fertilization (IVF). (A) Pronuclear transfer technique. An affected oocyte from the prospective mother (with many, sometimes all, mtDNAs having the pathogenic mutation) is fertilized by her partner’s sperm. The resulting normal *karyoplast* (combined male and female pronuclei) is isolated, then transferred into an enucleated donor zygote with normal mitochondria. The resulting

zygote has “foreign” but normal mtDNA. (B) Metaphase II spindle transfer technique. The metaphase II spindle is transferred from an oocyte with mutant mtDNA into a mitochondrial donor oocyte. The resulting hybrid oocyte has a nuclear genome from the prospective mother, but mtDNA from the donor. Fertilization by intracytoplasmic sperm injection (ICSI) produces a hybrid zygote. (C) In an attempt in 2016 to prevent transmission of Leigh syndrome, a severe neurological disorder, a hybrid human zygote produced by mitochondrial donation gave rise to a three-parent baby. The image shows the healthy baby boy. When tested some months later just 1 % of his mtDNA carried the harmful mutation. Holding him is Dr. John Zhang from the New Hope Fertility Center in New York. (A and B adapted from Craven L et al. [2011] *Hum Mol Genet* 20:R168–174; PMID 21852248, with permission from Oxford University Press; C reproduced courtesy of New Hope Fertility Center, New York.)

SUMMARY

- Treatment for inborn errors of metabolism sometimes involves supplementing a genetic deficiency, but often the treatment is directed at reducing the harmful effects of abnormally elevated metabolites.
- Drug development typically involves screening hydrocarbon-based small molecules for compounds that will bind to medically important protein targets. By binding to a protein, the drug affects its function in some way.
- Genetic variation means that different individuals can respond very differently to drugs; adverse reactions to drugs are very common and cause very many fatalities.
- The pharmacokinetics of a drug describes how it is absorbed, activated (in the case of a prodrug), metabolized, and excreted; pharmacodynamics describes the effect it has on the body.
- Phase I drug metabolism reactions are typically oxidative reactions carried out by monooxygenases; phase II reactions are conjugative reactions in which a transfer-ase enzyme adds a chemical group. The overall effect is to convert lipophilic hydrocarbon drugs into more polar forms that can be excreted more easily.
- In addition to dealing with artificial drugs, drug-metabolizing enzymes handle unusual exogenous chemicals (xenobiotics) in our diet and environment. They are often highly polymorphic because xenobiotics originating from other organisms are under genetic control and potentially harmful to us.
- The therapeutic window is the range of drug concentrations in which pharmaceutical benefit is achieved without safety risks.
- Poor drug metabolizers are at risk of a drug overdose (the drug does not get cleared quickly; repeated drug doses drive up the concentration). Others are

ultrafast metabolizers and may get little therapeutic benefit (the drug is cleared too rapidly).

- Six cytochrome P450 enzymes carry out 90 % of phase I drug metabolism. Each handles the metabolism of multiple drugs; conversely, some individual drugs may be metabolized by two or more cytochrome P450 enzymes.
- When a drug is metabolized principally by one enzyme, genetic variation in that enzyme can be mostly responsible for large differences between individuals in the ability to metabolize that drug. For some other drugs, such as warfarin, several different genetic factors determine how the drug is metabolized.
- Therapeutic “recombinant proteins” are made by expressing cloned human genes in cells to make a human protein that can be purified and used to treat a genetic deficiency of that protein.
- Therapeutic antibodies are usually designed to bind to harmful gene products to block their effects. Rodent monoclonal antibodies have limited lifetimes after injection into patients; genetic engineering allows the replacement of rodent sequences by human sequences to make more effective antibodies.
- Genetically engineered antibodies with a single variable polypeptide chain can work as intracellular antibodies by binding harmful proteins within cells.
- Gene therapy means inserting nucleic acids or oligo-nucleotides into the cells of a patient to counteract or alleviate disease.
- In gene supplementation therapy, diseased cells that are genetically deficient for some product are supplemented by transfecting a cloned gene to make the missing product inside the cells.
- Some therapies target RNA. In gene silencing, the expression of a positively harmful gene (such as a gene with a gain-of-function mutation or one expressed by a pathogen) is selectively repressed, usually by inhibiting the RNA. RNAs can sometimes also be induced to undergo alternative splicing to counteract disease.
- Stem cells are cells that can both renew themselves and give rise to more differentiated (more specialized) cells. Pluripotent embryonic stem cells are artificially cultured cells derived from the very early embryo that can be induced to give rise to virtually any differentiated cell. Somatic stem cells help to replace a limited set of short-lived cells.
- In cell reprogramming, the epigenetic settings of cells are artificially altered to induce changes in gene expression so that the cells acquire the characteristics of a different cell type. Differentiated cells can be induced to dedifferentiate to become unspecialized pluripotent stem cells or to form a different type of somatic cell (transdifferentiation).

- Virus vectors are more efficient but less safe than non viral vectors in transporting therapeutic genetic constructs into cells.
- Integrating virus vectors such as lentivirus vectors can allow a genetic construct to be inserted into the chromosomes of a cell. That is highly desirable when targeting short-lived cells that are replenished by stem cells; if a therapeutic transgene integrates into the stem cell, it will be transmitted by cell division.
- *Ex vivo* gene therapy involves removing cells from a patient, genetically modifying them in culture and returning the genetically modified autologous cells to the patient. It has been used to treat disorders by genetic modification of impure populations of hematopoietic stem cells that give rise to blood cells or some types of tissue immune system cell.
- In *in vivo* gene therapy, the cells of a patient are genetically modified *in situ*. Non-integrating virus vectors such as AAV virus are commonly used to transfect differentiated cells.
- Animal disease models are usually created by genetically modifying the germ line to mimic a human phenotype. They are important in permitting a detailed understanding of molecular pathology, and to provide a front-line system for testing new therapies.
- Therapeutic antisense oligonucleotides are designed to base pair with RNA transcripts so as to inhibit the expression of harmful gene products of the target gene or to modulate RNA splicing.
- RNA interference (RNAi) is a natural gene-silencing mechanism that evolved as a cellular defense against virus attack or excessive transposon activity. Therapeutic short interfering RNAs can be designed to inhibit expression of a harmful gene or allele after delivery into the cells of a patient.
- Genome editing involves making a precise genetic modification at a pre-determined location in the genome of intact cells.
- Standard CRISPR-Cas genome editing uses artificial hybrid RNAs having a variable guide sequence (programmed to base pair to a desired unique genomic site) and a scaffold sequence (binds to Cas endonuclease). The bound Cas nuclease is transported to the desired site, where it cuts the DNA to permit DNA repair and the desired genetic modification.
- Mitochondrial replacement is an *in vitro* fertilization method that avoids transmission of a severe mitochondrial DNA disorder to a child by moving the nuclear DNA from a maternal oocyte (either prior to, or after, fertilization) to an enucleated donor oocyte (before or after fertilization) which has normal healthy mitochondria, and mtDNA.

QUESTIONS

Questions can be downloaded by visiting the following link, under Support Materials: www.routledge.com/9780367490812.

FURTHER READING

General overviews

Dietz H (2010) New therapeutic approaches to mendelian disorders. *N Engl J Med* 363:852–863; PMID 20818846.

Pharmacogenetics and pharmacogenomics

Cytochrome P450 Drug Interaction Table. *Indiana University School of Medicine, Division of Clinical Pharmacology*. Available at <https://drug-interactions.medicine.iu.edu/Home.aspx>

Hockings JK (2020) Pharmacogenomics: an evolving clinical tool for precision medicine. *Cleveland Clinic J Med* 87:91–99; PMID 32015062.

Meyer UA, Zanger UM & Schwab M (2013) Omics and drug response. *Annu Rev Pharmacol Toxicol* 53:475–502; PMID 23140244.

Pharmacogenomics Knowledge Base (PharmGKB). <http://www.pharmgkb.org> [A knowledge resource offering clinical information including dosing guidelines and drug labels, potentially clinically actionable gene-drug associations and genotypephenotype relationships.]

Roden DM (2019). Genomic medicine 2. Pharmacogenomics. *Lancet* 394:521–532; PMID 31395440.

Wang L, McLeod HL & Weinshilboum RM (2011) Genomics and drug response. *N Engl J Med* 364:1144–1153; PMID 21428770.

Small molecule drug therapy for genetic disorders

Curatolo P & Moavero R (2012) mTOR inhibitors in tuberous sclerosis complex. *Curr Neuroparmacol* 10:404–415; PMID 23730262.

De Boeck K & Davies JC (2017) Where are we with transformational therapies for patients with cystic fibrosis? *Curr Opin Pharmacol* 34: 70–75; PMID 28992608.

Franz DN (2013) Efficacy and safety of everolimus for subependymal giant cell astrocytomas associated with tuberous sclerosis complex (EXIST-1): a multicenter, randomized, placebo-controlled phase 3 trial. *Lancet* 381:125–132; PMID 23158522.

Heilbron K (2021) Advancing drug discovery through the power of the human genome. *J Pathol* 254:418–429; PMID 33748968.

Santos R (2017) A comprehensive map of molecular drug targets. *Nat Rev Drug Discov* 16:19–34; PMID 27910877.

The Therapeutic Targets Database at <http://db.idrblab.net/ttd/>

Therapeutic antibodies and proteins

Cardinale A & Biocca S (2008) The potential of intracellular antibodies for therapeutic targeting of protein-misfolding diseases. *Trends Mol Med* 14:373–380; PMID 18693139.

Dimitrov DS (2012) Therapeutic proteins. *Methods Mol Biol* 899:1–26; PMID 22735943.

Hunter P (2019) The prospects for recombinant proteins from transgenic animals. *EMBO Reports* 20:e48757; PMID 30397525.

Kim SJ (2005) Antibody engineering for the development of therapeutic antibodies. *Mol Cells* 20:17–29; PMID 16258237.

Lu R-M (2020) Development of therapeutic antibodies for the treatment of diseases. *J Biomed Sci* 27:1 PMID 31894001.

Animal disease models for testing therapies

Shen H (2013) Precision gene editing paves way for transgenic monkeys. *Nature* 503:14–15; PMID 24201259.

Strachan T & Read AP (2019) *Human Molecular Genetics*, 5th ed. Garland Science. [Chapter 21 gives a detailed account of the technologies involved in making animal models and the extent to which the phenotypes faithfully replicate that of human disorders they were intended to mimic.]

Gene therapy: general

Anguela XM & High KA (2019) Entering the modern era of gene therapy. *Annu Rev Med* 70:273–288; PMID 30477394.

Dunbar CE (2018) Gene therapy comes of age. *Science* 359(6372):eaan4672; PMID 29326244.

Gene Therapy [Net.com](http://www.genetherapynet.com/). Available at <http://www.genetherapynet.com/> [Covers basic science, clinical trials and includes databases, publication and so on.]

Ginn SL (2018) Gene therapy clinical trials worldwide to 2017—an update. *J Gene Med* e3015; PMID 29575374.

High KA & Roncarolo MG (2019) Gene therapy. *N Eng J Med* 381:455–464; PMID 31365802.

Ma C-C (2020) The approved gene therapy drugs worldwide from 1998 to 2019. *Biotechnol Adv* 40:107502; PMID: 31887345

Clinical trials databases

ClinicalTrials.gov. www.clinicaltrials.gov [A comprehensive US government site.]
Gene Therapy Clinical Trials Worldwide. <https://a873679.fmphost.com/fmi/webd/GTCT>
[Provided by the Journal of Gene Medicine and published by Wiley.]

RNA and oligonucleotide therapeutics

Bennett CF (2019) Therapeutic antisense oligonucleotides are coming of age. *Annu Rev Med* 70:307–321; PMID 30691367.

Crooke ST (2018) RNA-targeted therapeutics. *Cell Metab*. 27:714–733; PMID 29617640.

Dowdy SF (2017) Overcoming cellular barriers for RNA therapeutics. *Nature Biotechnol* 35:222–229; PMID 28244992.

Hu B (2020) Therapeutic siRNA: state of the art. *Signal Transduct Target Ther* 5:101; PMID 32561705

Lieberman J (2018) Tapping the RNA world for therapeutics. *Nature Struct Mol Biol* 25:357–364; PMID 29662218.

Kuijper EC (2020) Opportunities and challenges for antisense oligonucleotide therapies. *J Inher Metab Dis* (in press); PMID 32391605.

Yu A-M (2020) RNA drugs and RNA targets for small molecules: principles, progress, and challenges. *Pharmacol Rev* 72:862-898; PMID 32929000.

CRISPR-Cas and therapeutic genome editing

Doudna JA (2020) The promise and challenge of therapeutic genome editing. *Nature* 578:229–236; PMID 32051598.

Doudna JA & Charpentier E (2014) The new frontier of genome engineering with CRISPR-Cas9. *Science* 346(6213): 1258096; PMID 25430774.

van Haasteren J (2020) The delivery challenge—fulfilling the promise of therapeutic genome editing. *Nature Biotechnol* 38:845–855; PMID 32601435.

Stem cells and cell therapy

Cossu G (2018) *Lancet* Commission: stem cells and regenerative medicine. *Lancet* 391:883–910; PMID 28987452.

Kimbrel EA & Lanza R (2020) Next-generation stem cells—ushering in a new era of cell-based therapies. *Nature Rev Drug Discov* 19:463–479; PMID 32612263.

Lanza R (2019) Engineering universal cells that evade immune detection. *Nature Rev Immunol* 19:723–733; PMID 31417198.

Yamanaka S (2020) Pluripotent stem cell-based cell therapy—promise and challenges. *Cell Stem Cell* 27:523–531; PMID 33007237.

Treatment for mitochondrial DNA disorders

Chinnery PF (2020) Mitochondrial replacement in the clinic. *N Eng J Med* 382:1855–1857; PMID 32374967.

Russell OM (2020) Mitochondrial diseases: hope for the future. *Cell* 181:168–188; PMID 32220313.

10

Cancer genetics and genomics

DOI: [10.1201/9781003044406-10](https://doi.org/10.1201/9781003044406-10)

CONTENTS

[10.1 FUNDAMENTAL CHARACTERISTICS AND EVOLUTION OF
CANCER](#)

[10.2 ONCOGENES AND TUMOR SUPPRESSOR GENES](#)

[10.3 GENOMIC INSTABILITY AND EPIGENETIC DYSREGULATION IN
CANCER](#)

[10.4 NEW INSIGHTS FROM GENOME-WIDE STUDIES OF CANCERS](#)

[10.5 GENETIC INROADS INTO CANCER THERAPY](#)

[SUMMARY](#)

[QUESTIONS](#)

[FURTHER READING](#)

Why should cancer merit a separate chapter in this book and not, say, neurology or cardiology? Well, the molecular pathogenesis in cancers is, for the most part, quite different from that of other genetic disorders: somatic mutations and epigenetic dysregulation are extremely common, and natural selection operates at the level of the cell as well as at the level of the organism. In addition, a number of specialized genetic mechanisms — kataegis, chromothripsis, chromoplexy and so on—are observed only in cancer cells.

Tumorigenesis involves an extraordinary and bewildering degree of changes to both the genome and the epigenome. Not only is there genetic heterogeneity between tumors, but also within individual tumors. Despite the heterogeneity of the very many different diseases that we call cancer, the phenotype—uncontrolled cell proliferation—is comparatively simple and more amenable to genetic analysis than some other common classes of disease, such as psychiatric disorders.

In [Section 10.1](#) we give an overview of the primary distinguishing biological capabilities of cancer cells, outline the broad multi-stage evolution of cancers and describe how intratumor heterogeneity evolves. [Section 10.2](#) is mostly devoted to considering the principles underlying two fundamental classes of genes in cancer development: oncogenes and tumor suppressor genes. As cancers evolve, genomic instability and epigenetic dysregulation become increasingly prominent; we consider selected aspects in [Section 10.3](#).

Genome-wide molecular profiling studies—notably genome-wide sequencing—are transforming our understanding of cancer, and in [Section 10.4](#) we take a look at new insights emerging from the burgeoning cancer genome studies. Finally, in [Section 10.5](#) we consider the challenges and prospects in deriving clinical benefit from all the extraordinary information coming out of cancer genetics and cancer genome studies.

10.1 FUNDAMENTAL CHARACTERISTICS AND EVOLUTION OF CANCER

In order to appreciate why cancers are so different from other genetic disorders, it is important to understand how cancers evolve and the role of natural selection in this process. First, however, we provide a summary of the fundamental characteristics of cancers.

The defining features of unregulated cell growth and cancer

The term **cancer** is applied to a heterogeneous group of disorders whose common features are uncontrolled cell growth and cell spreading; abnormal cells are formed that can invade adjacent tissues and spread to other parts of the body through the blood and lymph systems (but see below for a second usage). *Carcinogenesis*, the general process of cancer formation, may result from aberrant functioning of a wide range of genetic control mechanisms, as detailed below.

Aberrant regulation of cell growth results in an abnormal increase in cell numbers; growths can result that appear normal or abnormal. A growth containing excessive numbers of cells that appear to be virtually the same as those in the normal tissue is said to be *hyperplastic*; a growth that has cytologically abnormal cells is said to be *dysplastic*.

Sometimes a growth formed by excessive cell proliferation is localized. That is, it shows no signs of invading neighboring tissue, and is described as a **benign tumor**. Benign tumors are self-limiting: they grow slowly and can often be surgically removed with low risk of recurrence. They often do not present much danger. Sometimes, however, they grow quite large over time, and simply by expanding, they can press on neighboring structures in a way that can cause disease. For example, in tuberous sclerosis complex (caused by mutations in *TSC1* or *TSC2*, genes that work in the mTOR growth signaling pathway), benign tumors usually form in multiple different organs. By growing to a large size, they can sometimes disrupt organ function.

In the more than 100 different diseases that we call cancers, the abnormal cells resulting from uncontrolled cell growth have an additional defining property: they can spread. In these diseases the tumors may initially be benign, but they often progress to become **malignant tumors** (which are also commonly called **cancers**).

Malignant tumors have two distinguishing features: they can invade neighboring tissues, and the cells can break away and enter the lymphatic system or bloodstream, to be carried to another location where they cross back into tissues to form secondary tumors ([Figure 10.1](#)). Spreading to more distant sites in the body is known as **metastasis**; [Figure 10.2](#) shows dissemination via the bloodstream—the cancer cells cross capillary walls and migrate through the extracellular matrix.

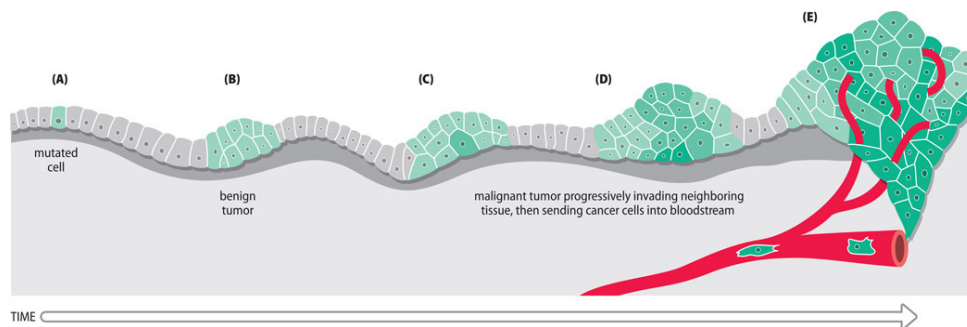


Figure 10.1 Progressive changes in the formation of malignant tumors. The initial mutated cell (A) can develop into a benign tumor (B) through the loss of some normal controls on cell division. Subsequent DNA changes and epigenetic changes can cause tumor cells to lose further normal controls to become a malignant tumor (C to E) that aggressively invades neighboring tissue. Cells from the malignant tumor can detach themselves and enter the bloodstream (as shown here) or the lymphatic system. In this way they are carried to remote sites in the body where they can exit the circulation and invade neighboring tissues to establish secondary tumors (*metastasis*—for detail of the mechanism see [Figure 10.2](#)). (From the website of the National Cancer Institute [<http://www.cancer.gov>].)

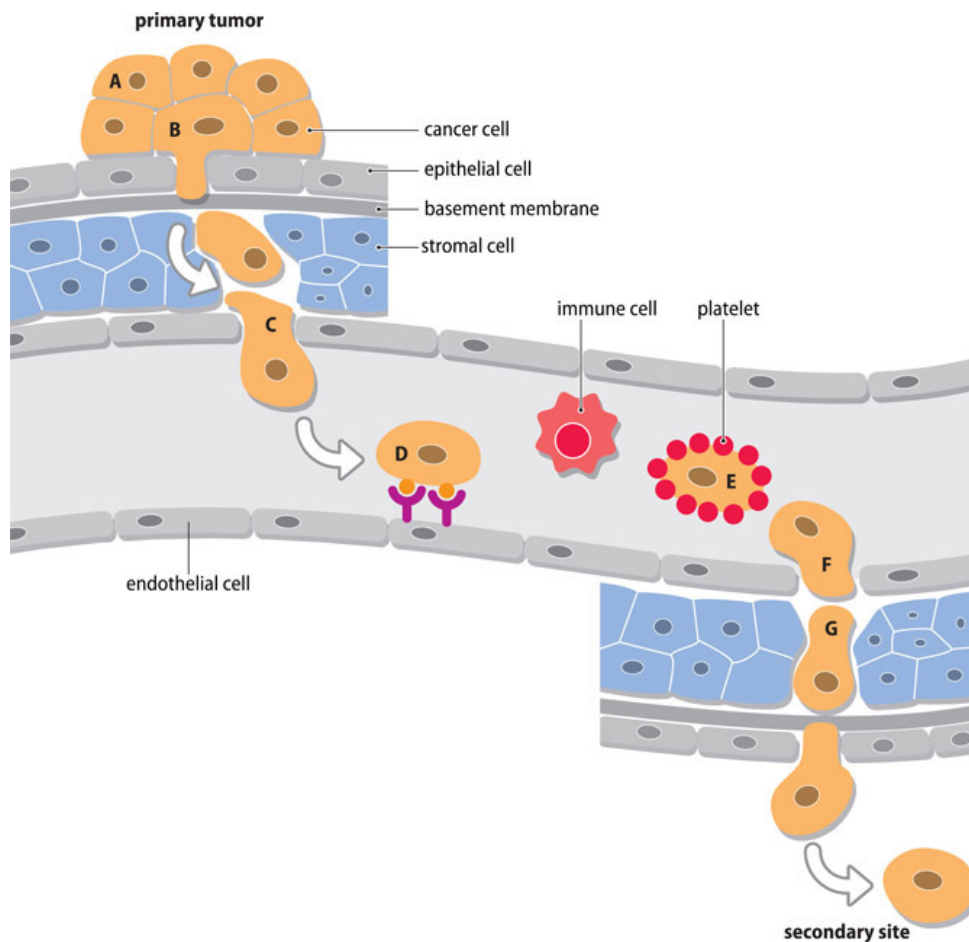


Figure 10.2 The multiple steps taken by metastatic cells to seed secondary tumors.

Metastatic cells must first break free from the primary tumor. To accomplish this, cancer cells (A) reduce adhesion to neighboring cells and (B) clear a path for migration into the vasculature-rich **stroma** (connective tissue plus blood vessels). Once at the vasculature, cells can freely enter the bloodstream if the vasculature is discontinuous, as in certain regions of the liver, bone marrow, and kidneys. *Intravasation* (C) is required if the vasculature is continuous; metastatic

cells either cause endothelial cells to retract (by releasing compounds such as vascular endothelial growth factor) or induce endothelial cell death (by releasing reactive oxygen species and factors including matrix metalloproteinases). In the bloodstream, cancer cell distribution is determined by blood flow and interactions between cancer cells and the secondary organs that they colonize: cells can get trapped in narrow capillary beds, such as those of the lung and liver, and can also express receptors that bind to metastasis-supporting sites (D) or to platelets (E), which protect the cancer cells from the immune system. Cancer cells can circulate for more than two hours, suggesting that they do not always become lodged in the first capillary beds that they reach. After reaching the secondary site, cancer cells can leave the bloodstream (F) by *extravasation* (inducing endothelial cell retraction or death). To proliferate in the secondary site, cancer cells co-opt the local environment by releasing proinflammatory compounds and proteinases that induce their neighbors to release growth factors (G). (Adapted from Schroeder A et al. [2012] *Nature Rev Cancer* 12:39–50; PMID 22193407. With permission from Macmillan Publishers Ltd.)

The tumors (also called *neoplasms*) in cancer can be broadly classified as solid or liquid. Solid tumors form discrete masses composed of epithelial or mesenchymal (stromal) cells. “Liquid tumors” are made up of neoplastic cells whose precursors are normally mobile blood cells; they include leukemias and also lymphomas (which, although generally forming solid masses in lymph nodes, are able to travel through the lymphatic system). According to the type of tissues or cells in which they arise, the tumors are classified into different categories ([Table 10.1](#)).

TABLE 10.1 MAJOR CATEGORIES OF TUMORS ACCORDING TO TISSUE OR CELLS OF ORIGIN

Tissue/cells of origin	Tumors
Epithelial tissue (single-layer or bilayer)	adenoma (benign); adenocarcinoma (malignant)
Epithelial tissue (multi-layer, as in skin and bladder)	papilloma (benign); squamous cell carcinoma (malignant, in skin); transitional cell carcinoma (malignant, in bladder)

Tissue/cells of origin	Tumors
Blood forming tissue (notably bone marrow)	lymphoma (of lymphocytes); leukemia (of leukocytes)
Stromal (mesenchymal) tissue	benign tumors have the simple -oma suffix; malignant tumors end in -sarcoma. Examples are: fibroma and fibrosarcoma (fibroblasts); osteoma and osteosarcoma (bone); chondroma and chondrosarcoma (cartilage); hemangioma and hemangiosarcoma (endothelial cells)
Glial cells	gliomas

Cancers form when cell division is somehow affected to cause uncontrolled cell proliferation. Changes at the DNA and chromatin levels are the primary contributors. A small minority of human cancers are associated with a specific virus, as described below. Mostly, however, human cancers form because of a series of mutations and epigenetic dysregulation in certain cancer-susceptibility genes.

Recall from [Section 4.1](#) that mutations frequently arise through errors in DNA replication and DNA repair; tissues that have actively dividing cells are therefore prone to forming tumors. Cells divide in development as an organism grows, and some childhood tumors can arise from mutations in cells of the embryo. Although the great majority of our cells are not actively dividing, an adult human has roughly one trillion (10^{12}) rapidly multiplying cells. There is a need to replace certain types of cell that have a high turnover, notably cells in the blood, skin, and the gastrointestinal tract. Thus, for example, each day about 4 % of the keratinocytes in our skin and a remarkable 15–20 % of the epithelial cells of the colon die and are replaced.

The short-lived cells that need to be replaced regularly are ones that interface, directly or indirectly, with the environment, and continuous turnover of these cells is a protective measure. Stem cells are the key cells responsible for manufacturing new body cells to replace the lost cells. As we explain below, considerable evidence suggests that cancers are often diseases of stem cells.

Why cancers are different from other diseases: the contest between natural selection operating at the level of the cell and the level of the organism

We are accustomed to thinking of how Darwinian natural selection works at the level of the organism: the key parameter is the reproductive success rates of individuals. *Selection pressure* is the effect of natural selection on allele frequencies. It ensures that a deleterious allele—one that reduces reproductive fitness—will normally be at a low frequency in the population (according to the penetrance, the frequency will be maintained by new mutation, or by transmission by unaffected carriers). Germline mutations make the key contribution to noncancerous genetic disorders; somatic mutations normally have minor roles.

Cancers are different. Yes, occasionally cancers can run in families, and germline mutations are clearly important in some cases. However, all cancers have multiple somatic mutations and the genetic contribution to cancers is dominated by somatic (post-zygotic) mutation. That happens because natural selection also operates at the level of *cells and cancers show abnormal cell proliferation*.

The balance between cell proliferation and cell death

The principal defining feature of cancer is uncontrolled growth in cell number. Growth occurs when the net balance between cell proliferation and cell death is positive. Cell proliferation is required for growth, but a complicated series of controls ensures that normally our cells do not divide in an uncontrolled fashion; sometimes there is a need for brakes to be applied, and cells are ordered to undergo cell cycle arrest. In the opposite direction is cell death, a natural way of removing inefficient cells, cells that are unwanted, and potentially dangerous cells. Like cell proliferation, cell death can be ordered to occur, and it too is highly regulated.

The mechanisms regulating cell proliferation and cell death involve sophisticated intercellular signaling. Some signaling pathways send instructions for certain cells to proliferate or undergo cell cycle arrest; other pathways induce the death of undesirable cells in some way (programmed cell death, or apoptosis). Classical cancer-susceptibility genes were identified as working to regulate cell

division or having direct roles in growth-signaling pathways. Additional cancer-susceptibility genes were found to have roles in apoptosis, but as well as these types of gene, there are many types of non-classical cancer-susceptibility gene that do not function directly in these areas. Instead, they have indirect roles, functioning in a variety of areas such as DNA repair/genome maintenance, cell metabolism and epigenetic regulation and so on. When such genes are faulty, the resulting increased mutation or epigenetic dysregulation can have consequences for genes directly regulating cell growth or apoptosis. See [Figure 10.3](#) for a summary.

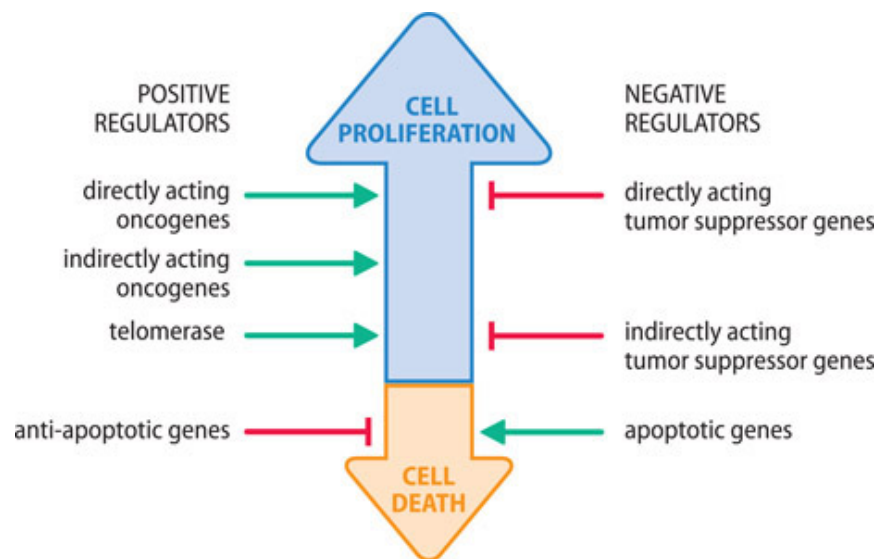


Figure 10.3 Major classes of cancer gene as positive or negative regulators of net cell proliferation (cell growth). Green arrows indicate stimulatory effects on cell proliferation or apoptosis; red T-bars indicate inhibitory effects. Some oncogenes are directly involved in promoting cell division and cell proliferation (by regulating the cell cycle or cell growth signaling pathways). Many other genes have similar effects, but act indirectly, such as oncogenes working in cell metabolism, telomerase and anti-apoptotic genes. Similarly, some tumor suppressor genes control cell proliferation directly; others work indirectly by, for example, regulating genome maintenance or epigenetic pathways. Apoptosis-promoting oncogenes negatively affect net cell proliferation. Note: some tumor suppressor genes, notably TP53, both suppress cell proliferation and promote apoptosis.

Throughout development up to the age of maturity there is an overarching priority for increased numbers of cells to sustain rapid growth of the organism—not in an unconstrained way, of course, but executed according to detailed

prescribed body plans and the requirements of intricate tissue architecture, and so on. But cells are also lost during development. In addition to short-lived cells, many cells are intentionally deleted during development as part of the natural process of sculpting our tissues and organs, and to ensure healthy immune and nervous systems. To distinguish self from nonself, immune system cells with receptors that bind to self-antigens must be deleted, and during nervous system development neurons with unproductive interneuron connections are deleted.

But when we reach adulthood, growth is restricted. Although most of our cells are non-dividing cells by then, a significant minority continue to divide to replace certain types of cell that turn over rapidly, such as skin, blood, and intestinal epithelial cells. Apoptosis is also used in adults to ensure the destruction of both damaged cells (arising as a result of natural wear and tear, or through injury) and potentially harmful cells (such as virus-infected cells).

Why we do not all succumb to cancer?

Cancer is initiated when cells develop capabilities to escape some normal controls limiting cell proliferation, or inducing apoptosis. It is no accident that the cells most likely to give rise to tumors are cells that already possess some elements of the required capabilities. Primarily, they are stem cells (which have either a high intrinsic proliferative capacity, or can be induced by inflammation, tissue damage, and so on to proliferate rapidly), and to a lesser extent populations of cells in the embryo or fetus that transiently undergo rapid proliferation before differentiating.

Cells can escape from these normal controls as a result of mutation in certain control genes. This is where natural selection at the cellular level is important: each successive mutation that disrupts normal controls on cellular proliferation or apoptosis confers an additional selective growth advantage on its descendants. The resulting expansion in mutant cells provides a greater target for successive cancer mutations. As a result, there is strong selection pressure on cells to evolve through a series of stages into tumor cells.

If there is such strong selection pressure on cells to evolve into tumor cells, why do we not all succumb to cancer? Certainly, if we were to live long enough, cancer would be an inevitable consequence of random mutations. However, an opposing force of natural selection works at the level of the organism (to keep us healthy and free from tumors—at least until we have produced and raised children). It

involves different mechanisms, not least immunosurveillance to detect and kill cancer cells, using cytotoxic T cells and natural killer (NK) cells. (Individuals whose immune systems are suppressed are more susceptible to cancer.)

There is, however, an imbalance between natural selection working at the level of the organism and at the level of cells. Luckily for us, natural selection operates over a much longer timescale than does the selection pressure in favor of tumor cell formation. Cancer cells can successfully proliferate and form tumors within an individual person, *but they do not leave progeny beyond the life of their human host*. That is, tumorigenesis processes must start afresh in a new individual. But at the level of the organism, natural selection continues down through generations. Individuals who have efficient cancer defense mechanisms are able to pass on good anti-cancer defense genes to their offspring, and anti-cancer defense systems continue to evolve from generation to generation.

Cancer cells acquire several distinguishing biological characteristics during their evolution

As described in the next section, the development of tumors occurs as a series of stages during which both genetic and epigenetic changes progressively accumulate in cells. During these stages the cells progressively acquire different biological capabilities that mark them out as cancer cells.

By definition, cancer cells show unregulated cell proliferation, and tumors develop by breaking away from normal control systems. They switch off various brakes that normally place limits on cell proliferation and genome instability, and counteract death (apoptosis) signals from neighboring cells. Cancer cells also lose the contact inhibition of normal cells that places limits on cell growth. Partly by overcoming these negative signals, they become masters of their own growth, and go on to acquire the characteristic ability to replicate indefinitely (**Box 10.1**).

To support continued cell proliferation, cancer cells re-adjust their metabolism. Thus, they show increased flux through the pentose phosphate pathway (PPP) and elevated rates of lipid biosynthesis, and they take up and use glucose at much higher rates than normal cells. (The last characteristic can be used by imaging systems to differentiate cancer cells from normal cells, so that the spread of cancer cells in the body can be visualized.)

Although apparently exposed to aerobic conditions, cancer cells nevertheless derive their energy from glycolysis, rather than from oxidative phosphorylation.

Glycolysis is normally used by cells in anaerobic conditions; the process involves converting glucose to first pyruvate, and then lactate, and energy production is inefficient (2 molecules of ATP generated per molecule of glucose, whereas under aerobic conditions, normal cells convert glucose to pyruvate and then pyruvate is catabolized in the tricarboxylic acid cycle, generating up to 36 molecules of ATP per molecule of glucose).

Why, under aerobic conditions, cancer cells normally use the much more inefficient glycolysis system of producing energy (the Warburg effect) remains poorly understood. The switch to glycolysis occurs early in oncogenesis and may be activated to drive cell survival. Interested readers can find a recent review at PMID 33347611.

BOX 10.1 TELOMERE SHORTENING AND SELECTION PRESSURE ON CANCER CELLS TO BECOME IMMORTAL BY ACTIVATING TELOMERASE EXPRESSION

Normal human cells can be grown in culture for limited periods. Fetal cells, for example, can divide 40–60 times in culture before reaching a state of *senescence* after which they cannot grow any further. Cancer cells, however, have unusual growth properties. In culture they do not exhibit contact inhibition, nor require adhesion to a solid substrate. Notably, they can replicate indefinitely, and so are immortal. (The HeLa cell line is the most outstanding human example; developed from a cervical cancer biopsy in the early 1950s, it has been extensively propagated to become the most intensively studied human cell line.)

THE END-REPLICATION PROBLEM

The above observations on the replicative behavior of cells relate to the end-replication problem: how can the extreme ends of linear chromosomes be replicated when new DNA strands grow in the 5' → 3' direction only? During DNA replication, new DNA synthesis is catalyzed by DNA polymerases that use an existing DNA strand as a template. As the replication fork advances, one new DNA strand is made in the same direction as the direction of travel for the replication fork, and can be synthesized continuously in the 5' → 3' direction. However, the other strand can be made only by synthesizing successive short

pieces of DNA (Okazaki fragments) in the opposite direction to that taken by the replication fork, and there is a problem with completing the synthesis at the very end (**Figure 1**). Because of the end-replication problem, using just DNA-dependent DNA polymerases means that a small amount of DNA will be lost from each telomere after every cell division.

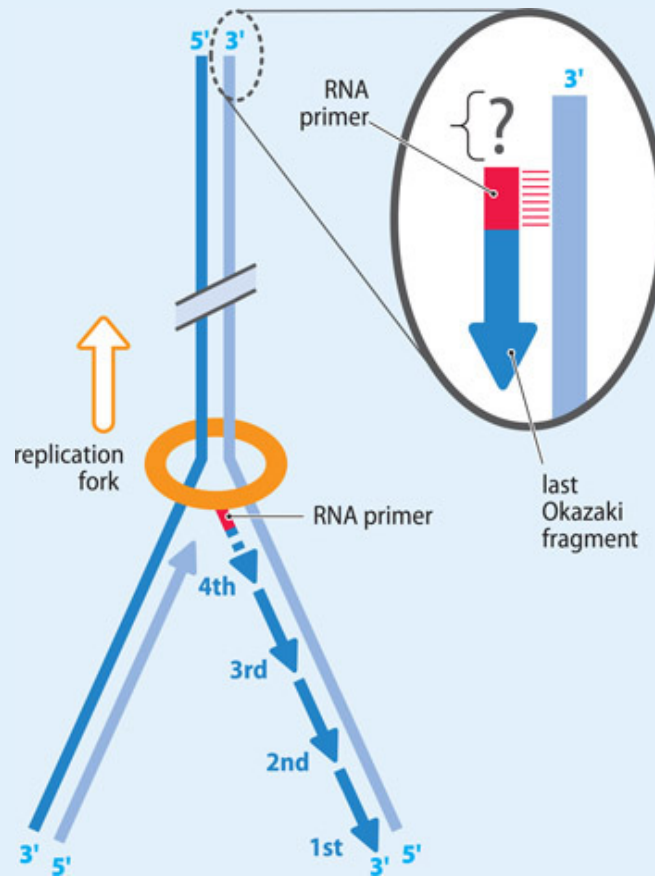


Figure 1 The problem with replicating the extreme ends of DNA in linear chromosomes.

In normal DNA replication by DNA-dependent DNA polymerases, an existing DNA strand is used as a template for making a complementary new DNA strand. Here, as the replication fork advances in the upward direction it can synthesize a continuous DNA strand upward in the 5' → 3' direction from one original DNA strand (colored deep blue) but for the pale blue original strand the 5' → 3' direction for DNA synthesis is in the opposite direction to the upward direction of the replication fork. The DNA must be synthesized in short pieces, called Okazaki fragments, starting from a position beyond the last fragment and moving backward toward it. (DNA-dependent polymerases use short RNA primers to initiate the synthesis of DNA, but the RNA primers are degraded, DNA synthesis fills in, and adjacent Okazaki fragments are ligated.) The question mark indicates a problem that is reached at the very end:

how is synthesis to be completed when there can be no DNA template beyond the 3' terminus?

The telomeres of our chromosomes have tandem TTAGGG repeats extending over several kilobases of DNA (see [Figure 1.9](#) on page 10), but because of the end-replication problem (plus oxidative damage and other end-processing events), the arrays of telomeric TTAGGG repeats normally shorten with each cell division (the number of telomere repeats lost varies between different cell types but is often in the range of 5–20 repeats). When a few telomeres become critically shortened, there is a growth arrest state, at which time DNA damage signaling and cellular senescence is triggered. In the absence of other changes, cells can remain in a senescent state for years.

TELOMERASE SOLVES THE END-REPLICATION PROBLEM

The end-replication problem can be solved—and telomeres restored to full-length—when cells express time DNA damage signaling and cellular senescence is triggered. In the absence of other changes, cells can remain in a senescent state for years. telomerase, an RNA-dependent DNA polymerase. Telomerase is a ribonucleoprotein consisting of a reverse transcriptase and a noncoding RNA (ncRNA). The ncRNA has a hexanucleotide sequence that is complementary in sequence to the telomere repeat; it serves as a template from which the reverse transcriptase can make tandem telomere repeats ([Figure 2](#)).

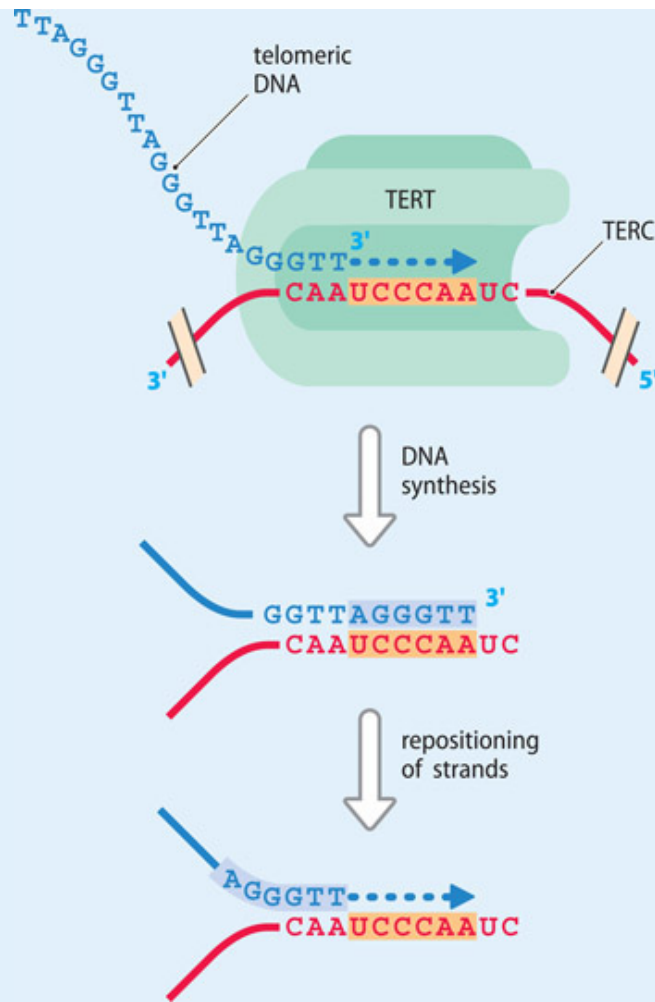


Figure 2 Telomerase uses a reverse transcriptase and a noncoding RNA template to make new telomere DNA repeats. The telomerase reverse transcriptase (TERT) is an RNA-dependent DNA polymerase: it uses an RNA template provided by the other subunit, TERC (telomerase RNA complex). Only a small part of the RNA is used as a template—the hexanucleotide that is shaded—and so the telomeric DNA is extended by one hexanucleotide repeat (blue shading). Repositioning of the telomeric DNA relative to the RNA template allows the synthesis of tandem complementary copies of the hexanucleotide sequence in the RNA template.

Telomerase is expressed during early human embryogenesis (and in embryonic stem cells). However, its expression is subsequently repressed in most somatic cells, but it is active in the male germ line, activated lymphocytes, and stem cells in certain regenerative tissues.

SELECTION PRESSURE ON CANCER CELLS TO ACHIEVE REPLICATIVE IMMORTALITY

The repression of telomerase and the resulting erosion of telomeres in our cells is thought to be yet another defense system to stop cancer from developing during our long lifetimes. Cancer cells require multiple successive mutations to become malignant. After one mutation has led to some growth advantage, about another 20–40 cell divisions might be required to achieve a cell population size that is sufficient for another spontaneous mutation to occur in a cell with the previous mutation. Premalignant cells would therefore often be expected to come up against the barrier of replicative senescence before they have sustained sufficient mutations to form malignant tumors.

Tumor cells are able to bypass replicative senescence by suppressing tumor suppressors, such as p53 and the RB1 retinoblastoma protein. However, after a few additional cell divisions past the point at which senescence normally occurs, the cells enter a *crisis* state. Now the telomeres can be so short that DNA repair mechanisms do not recognize them as legitimate ends of chromosomes; instead, they treat them as double-strand DNA breaks. As a result, chromosomes can undergo end-to-end fusions. The resulting chromosomes have two centromeres and may be pulled in opposite directions at mitosis. That causes further broken ends, new cycles of chromosome fusion and breakage, and an acceleration of genome instability.

Rare cells that escape this crisis stage are almost always able to do so by having reactivated expression of telomerase: the telomeres are stabilized and the cell becomes immortal. However, the telomerase produced is not present at excess (the telomeres in cancer cells with stem cell-like properties are generally of the same length or shorter than those in adjacent normal cells). Note: in a small number of cases the ALT (*alternative lengthening of telomeres*) pathway is deployed; here telomeres are lengthened instead via homologous recombination mediated by inactivating mutations in the *ATRX* and *DAXX* genes (which together make a protein complex that deposits histone variant H3.3 into the repetitive heterochromatin of telomeres to promote transcription).

During cancer progression, cancer cells also undergo epigenetic reprogramming so that they can become less differentiated. Solid cancers show a plastic

phenotype, with a differentiated tumor mass and also undifferentiated areas. The latter, notably marking regions that form an invasive front as the cancer spreads, allow flexibility to respond to different environments, and metastases can show striking re-differentiation.

To ensure their survival, cancer cells need to avoid being destroyed by immune system cells, and they develop appropriate counter-attacking measures. Not only that, but they also maximize their ability to survive by invading host tissue and co-opting normal cells to help them, and by sending out cells to form secondary tumors.

It is also common for cancers to gain access to the vascular system by inducing the sprouting of existing blood vessels, whereupon the tumors become linked to the existing vasculature (**angiogenesis**), as in the tumor shown in [Figure 10.1E](#). That then allows tumor cells to escape more readily from a primary tumor and establish secondary tumors, although angiogenesis may often not be necessary for metastasis.

[Table 10.2](#) provides a summary, listing 10 biological characteristics cancer cells acquire that have been proposed as hallmarks of cancer. We expand on some points in [Section 10.2](#), but also provide details on some other of the points in later sections.

TABLE 10.2 TEN ACQUIRED BIOLOGICAL CAPABILITIES PROPOSED AS HALLMARKS OF CANCER BY DOUGLAS HANAHAN AND ROBERT WEINBERG

Acquired biological capability	Examples of how the biological capability is acquired
Self-sufficiency in growth signaling	Activate cellular oncogene
Insensitivity to signals suppressing growth	Inactivate <i>tp53</i> to avoid p53-mediated cell cycle arrest
Ability to avoid apoptosis	Produce IGF survival factor
Replicative immortality	Switch on telomerase (Box 10.1)

IGF, insulin growth factor; TGF β , transforming growth factor β ; VEGF, vascular endothelial growth factor.

(Adapted from [Hanahan D & Weinberg R \[2011\] Cell 144:646–674; PMID 21376230](#). With permission from Elsevier.). Note: in addition to these hallmarks, others have been proposed such as epigenetic dysregulation, including dedifferentiation (see PMID 33465324 for a recent review).

Acquired biological capability	Examples of how the biological capability is acquired
Genome instability	Inactivate certain genes involved in DNA repair
Induction of angiogenesis	Produce factor that induces VEGF
Tissue invasion and metastasis	Inactivate e-cadherin
Ability to avoid immune destruction	Paralyse infiltrating cytotoxic T lymphocytes and natural killer cells by secreting TGF or other immunosuppressive factor
Induction of tumor-promoting inflammation	Redirect inflammation-causing immune system cells that infiltrate the tumor so that they help in various tumor functions (see Table 10.3)
Reprogramming energy metabolism	Induce aerobic glycolysis

IGF, insulin growth factor; TGF β , transforming growth factor β ; VEGF, vascular endothelial growth factor. (Adapted from [Hanahan D & Weinberg R \[2011\]](#) *Cell* 144:646–674; PMID 21376230. With permission from Elsevier.). Note: in addition to these hallmarks, others have been proposed such as epigenetic dysregulation, including dedifferentiation (see PMID 33465324 for a recent review).

The initiation and multi-stage nature of cancer evolution and why most human cancers develop over many decades

Epidemiology studies have shown that age is a very large factor in cancer incidence (the rate at which it is diagnosed). For example, the age-incidence plots for epithelial cancers suggested that the risk of death from this cause increases roughly as the fifth or sixth power of elapsed lifetime. That observation suggested that perhaps six to seven independent events might be required for an epithelial cancer to develop (if the probability of an outcome is a function of some variable raised to the power n , a total of $n + 1$ independent events, each occurring randomly, are required for the outcome to be achieved).

The epidemiology studies provided an early indication of the multistage nature of cancer, and a suggestion of the number of critical steps involved. Now we know that as normal cells evolve to become cancer cells, they pick up many somatic changes—both genetic and epigenetic. A small subset of the genetic changes, known as **driver mutations**, result in altered expression of certain key

genes (those regulating cell proliferation and apoptosis) so that a growth advantage is conferred on their descendants. Driver mutations are positively selected and causally implicated in cancer development. The remaining mutations are *passenger mutations*.

A cell with a driver mutation that its neighbors lack usually possesses a small growth advantage, of the order of just a 0.4 % increase in the difference between new cell formation and cell death. The growth advantage is small because we have multiple layers of defense against cancer. Many tumor cells succumb to our natural defenses, or are disadvantaged by certain karyotype changes. Despite the high attrition rate of tumor cells, the small growth advantage can ultimately lead to a large mass, containing billions of cells, but that usually takes many years.

Clonal expansion and successive driver mutations

Tumors are monoclonal: all the cells descend from a single starting cell. Strong evidence for that supposition came from studies of B-cell lymphomas. Recall from [Section 4.5](#) that individual B cells in a person make different immunoglobulins, but the cells in individual B-cell lymphomas all make the same type of immunoglobulin.

Preferential clonal expansion of the mutant cells produces an expanded target (more cells) for a second driver mutation to occur in one of the mutant cells. As the process continues, cells progressively acquire more mutations ([Figure 10.4A](#)) and epigenetic dysregulation, causing them to become ever more like a cancer cell.

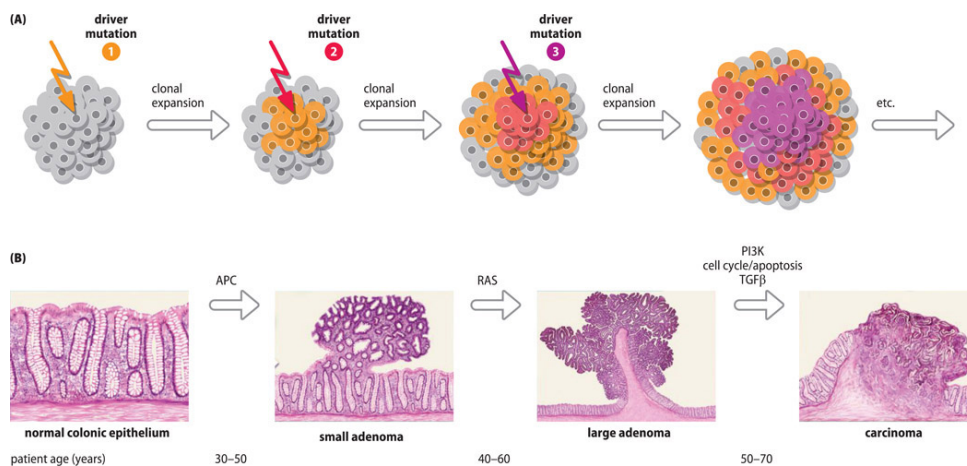


Figure 10.4 Driver mutations in the multistage evolution of cancer. (A) *General process.*

Each successive driver mutation gives the cell in which it occurs a growth advantage, so that it forms an expanded clone and thus presents a larger target for the next mutation. Orange cells carry driver mutation 1; red cells have sequential driver mutations 1 and 2; and purple cells have driver mutations 1, 2, and 3. (B) *Genetic alterations and the progression of colorectal cancer.*

The major signaling pathways that drive tumorigenesis are shown at the transitions between each tumor stage. One of several driver genes that encode components of these pathways can be altered in any individual tumor. Small and large adenomas appear as intestinal polyps that are benign but can progress to become carcinomas, cancers that invade the underlying tissue. Patient age indicates the time intervals during which the driver genes are usually mutated. Note that this model may not apply to all tumor types. PI3K, phosphoinositide 3-kinase pathway; TGFb, transforming growth factor b pathway. (B, from [Vogelstein B et al. \[2013\] Science 339:1546–1558; PMID 23539594](#). With permission from the AAAS.)

In some cases, just a few driver mutations are required. [Figure 10.4B](#) illustrates a classic example: the gradual transformation from normal epithelium to carcinoma in the development of colon cancer. The initial driver mutation is almost always one that affects the Wnt signaling pathway usually through loss of function of the *APC* gene at 5q21, but there can be more flexibility in the order of the subsequent genetic changes.

Cancer develops by accelerating mutation in two major ways

The average rate of mutation in human cells is low (about 10^{-6} per gene per cell) and the majority of cancer-causing mutations are recessive at the cellular level—both alleles need to be mutated. Cancer might therefore be expected to be highly improbable: the chance that any cell would receive successive mutations, often in both alleles, at several cancer-susceptibility loci would normally be vanishingly small.

Cancer nevertheless is common, and altered expression at a few cancer-susceptibility loci can be sufficient. Cancer is common largely because driver mutations greatly increase the probability of later mutations and epigenetic changes. They do this in two major ways, as listed below.

- *Conferring a growth advantage on cells.* If cells with a driver mutation have an increased growth rate, they will produce more progeny than other cells. Simply by producing an expanded target of mutant cells the probability of a subsequent mutation is increased (see [Figure 10.4A](#)).
- *Destabilizing the genome.* This increases the likelihood of later mutations in cancer. Chromosome instability is a feature of most tumor cells, producing grossly abnormal karyotypes with abnormal numbers of chromosomes and frequent structural arrangements that can activate oncogenes or cause a loss of tumor suppressor genes. In some cancers, a form of global DNA instability occurs: mutations in key DNA repair genes result in greatly elevated mutation rates. Increased overall mutations may mean, too, that genes regulating epigenetic modifications are also affected, resulting in altered expression at these types of cancer-susceptibility loci.

Additionally, some types of epigenetic change cause genome instability. In [Section 10.3](#) we explore the detail of genome instability and epigenetic dysregulation, and how they interact in cancer.

The generally late age of cancer onset

Tumors gradually acquire mutations to evolve from benign to malignant lesions. Because that takes some time, cancer is primarily a disease of aging. In self-renewing tissues—such as epithelial cells lining the gastrointestinal tract and genitourinary epithelium—the cells contain DNA that has progressively accumulated mutations through multiple DNA replication cycles in progenitor cells (recall that errors in DNA replication and post-replicative DNA repair are frequent causes of mutations). Thus a colorectal tumor in a person in their 80s or 90s will have nearly twice as many somatic mutations, mostly inconsequential, as in a morphologically identical colorectal tumor in a person half their age. The difference in ages when the same type of tumor presents will reflect when crucially important driver mutations occurred.

Cells in tissues associated with some other cancers do not replicate, and the tumors associated with these cells have fewer mutations, such as in glioblastomas (advanced brain tumors formed from nonreplicating glial cells) and pancreatic cancers (epithelial cells of the pancreatic duct also do not replicate). Initiating

driver mutations in these cases must occur in cells that have had comparatively few mutations.

Childhood cancers and cancers arising in embryonic cells

Some types of cancer commonly arise in childhood. Pediatric tumors often occur in tissues that do not self-renew, and such tumors typically have fewer mutations than adult tumors. However, leukemias and lymphomas, which are diseases of self-renewing blood cells, can also often develop early in life. Here the precursor cells are already mobile and invasive and are thought to require fewer DNA changes than in solid tumors, in which the tumor cells require additional mutations to confer these biological capabilities.

Some childhood cancers—including retinoblastoma, medulloblastoma, neuroblastoma, and Wilms tumor—can arise from an initiating mutation that arises in embryonic cells. Progenitor cells in the embryo resemble cancer cells—they are poorly differentiated and rapidly dividing. If they receive a cancer-predisposing mutation, they are much more likely to develop into tumors at an early stage than more differentiated cells with the same mutation.

Intratumor heterogeneity arises through cell infiltration, clonal evolution, and differentiation of cancer stem cells

Although tumors are considered to be monoclonal (composed of cells derived from a single ancestral cell), that does not mean that the cells in a tumor are the same. Instead, tumors often resemble organs, having quite complicated tissue architectures and being composed of functionally different cells.

A first level of intratumor heterogeneity exists because tumors are usually made up of both tumor cells proper (that originate from a single cell, and so are monoclonal), and also various unrelated cells that infiltrate the tumor from the surrounding environment. Different types of stromal cells, including immune infiltrating cells, become part of the tumor microenvironment and can be redirected to support tumor activities ([Figure 10.5A](#) and [Table 10.3](#)).

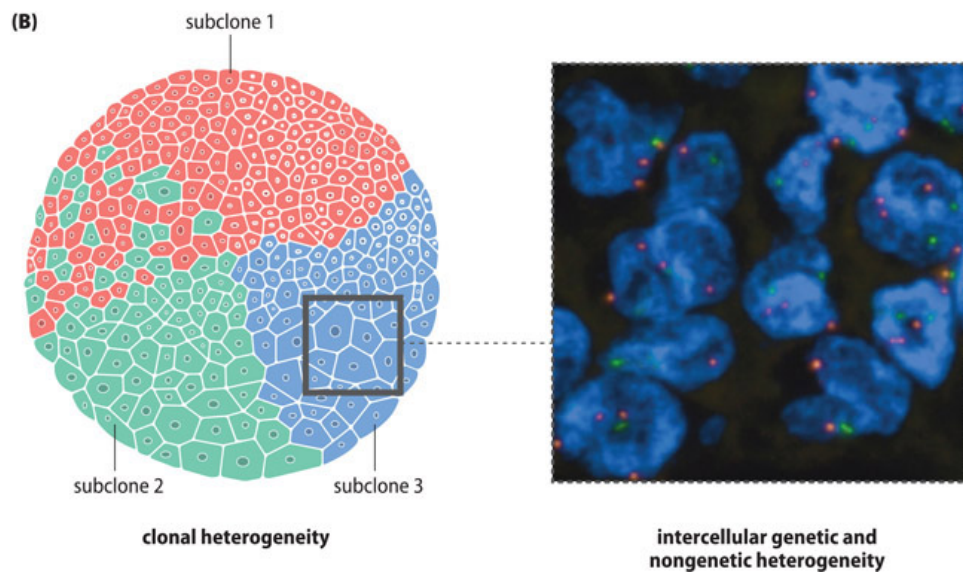
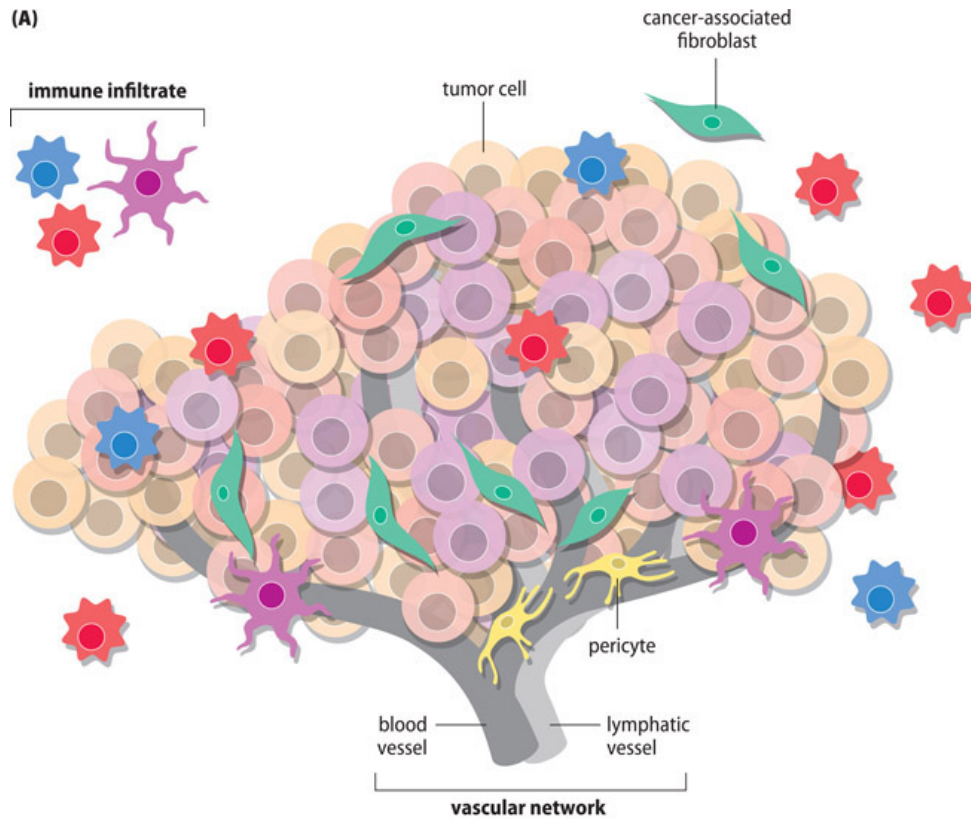


Figure 10.5 Cell heterogeneity within tumors. (A) Tumors as organs. Tumor formation involves co-evolution of neoplastic and non-neoplastic cells in a supportive and dynamic microenvironment that includes different stromal cells—cancer-associated fibroblasts, vascular endothelial cells (including pericytes), and diverse infiltrating immune cells—and the

extracellular matrix. The tumor microenvironment offers structural support, access to growth factors, vascular supply, and immune cell interactions. The immune cells include cell types normally associated with tumor-killing abilities as well as immune cells that can have tumor-promoting properties (see [Table 10.3](#)). For an account of the properties of support cells in tumors, see PMID 22439926. (B) Tumor subclones. Different tumor subclones may show differential gene expression due to both genetic and epigenetic heterogeneity. Cells from some subclones may intermingle (subclones 1 and 2) or be spatially separated (subclone 3), sometimes by physical barriers such as blood vessels. Within a subclonal population of tumor cells there may be intercellular genetic and nongenetic variation. For example, in the expanded square (which represents a section taken from a spatially separated subclone), differences in chromosome copy number between cells are revealed by the hybridization signals obtained with fluorescent probes for the centromeres of chromosome 2 (red) and chromosome 18 (green), against a background stain of blue for DNA. (A, From Junttila MR & de Sauvage FJ [2013] *Nature* 501:346–354; PMID 24048067. With permission from Macmillan Publishers Ltd. B, From [Burrell RA et al. \[2013\]](#) *Nature* 501:338–345; PMID 24048066. With permission from Macmillan Publishers Ltd.)

A second level of intratumor heterogeneity exists because the tumor cells proper within a tumor can become functionally distinct from each other. That can happen as a result of differential genetic changes. In addition, differential epigenetic changes can occur and some tumors can clearly be seen to have cells at different stages of differentiation.

Cells that have descended from the originating cancer cell can acquire new mutations conferring some additional growth advantage or other advantageous tumor-associated biological capability. A cell with an advantageous mutation such as this can form a subclone that competes with and outgrows the other cells. The process continues with new subclones competing against previous subclones. Subclones may intermingle or they can be spatially distinct ([Figure 10.5B](#)).

After the appearance of successive subclones, a tumor might be dominated by cells from a recent particularly successful subclone (a *clone sweep*) but still contain some cells from previously dominant subclones. Clonal evolution by acquisition of new mutations—both driver and passenger mutations—might therefore explain how functionally different types of tumor cell could arise within the same tumor. Despite being functionally divergent, the cells in the different subclones may or may not be recognizably different in appearance.

Subclones of a primary tumor can also undergo mutations that will drive genetic divergence leading to metastases. A paper published by Wu et al. in 2012 (PMID 22343890) gives the example of clonal selection driving genetic divergence of metastases in medulloblastoma.

In addition to clonal evolution, the concept of cancer stem cells has been invoked to explain intratumor heterogeneity. That is, self-renewing tumor cells have been proposed to give rise to all the different types of tumor cell within a tumor by progressive differentiation ([Box 10.2](#)). Because cancer stem cells are very long-lived and can potentially regenerate tumors or seed metastases starting from a single cell, there are important implications for cancer therapy.

Although the concept of cancer stem cells and clonal evolution might seem to provide alternative explanations for intratumor cell heterogeneity, they are not mutually exclusive. There is some evidence that even the stem cells within some tumors are heterogeneous as a result of mutation-induced divergence.

TABLE 10.3 DIFFERENT CATEGORIES OF STROMAL CELL TYPES CAN SUPPORT THE TUMOR MICROENVIRONMENT

Stromal cell category	Examples of cell types	Functions in support of:			
		Mitogenesis	Angiogenesis	Tissue invasion	Metastasis
Infiltrating immune cells	macrophages	+	+	+	+
	mast cells	+	+	+	
	neutrophils	+	+	+	+
	T cells (notably of the Th2-CD4 class and regulatory T cells); B cells	+			
Cancer-associated	activated tissue fibroblasts	+	+	+	+

The table gives a quite selective list, both of the different stromal cell types that support tumors and of their functions. For a fuller account, see Table 2 of Hanahan D & Coussens LM (2012) *Cancer Cell* 21:309–322; PMID 22439926.

Stromal cell category	Examples of cell types	Functions in support of:			
		Mitogenesis	Angiogenesis	Tissue invasion	Metastasis
fibroblastic cells					
Endothelial cells	endothelial tip, stalk, tube		+		
Pericytes	mature/immature pericytes		+		

The table gives a quite selective list, both of the different stromal cell types that support tumors and of their functions. For a fuller account, see Table 2 of Hanahan D & Coussens LM (2012) *Cancer Cell* 21:309–322; PMID 22439926.

BOX 10.2 CANCER AS A DISEASE OF STEM CELLS

Somatic cells may have rather short lives—only about a week or so, on average, in the case of intestinal epithelial cells. Short-lived cells need to be replaced periodically by new cells produced ultimately from the relevant tissue stem cells. The stem cells are capable of two types of cell division: symmetrical cell division and asymmetrical cell division. If the numbers of stem cells get too low for any reason, they can quickly regenerate by multiple successive symmetrical cell divisions, each producing daughter cells identical to the parent stem cell.

Asymmetrical stem cell divisions are reserved for making differentiated cells. In this case, when the stem cell divides it gives rise to one daughter cell that is identical to the parent cell (step a in [Figure 1](#)), plus a more differentiated, **transit-amplifying cell** (step b in [Figure 1](#)).

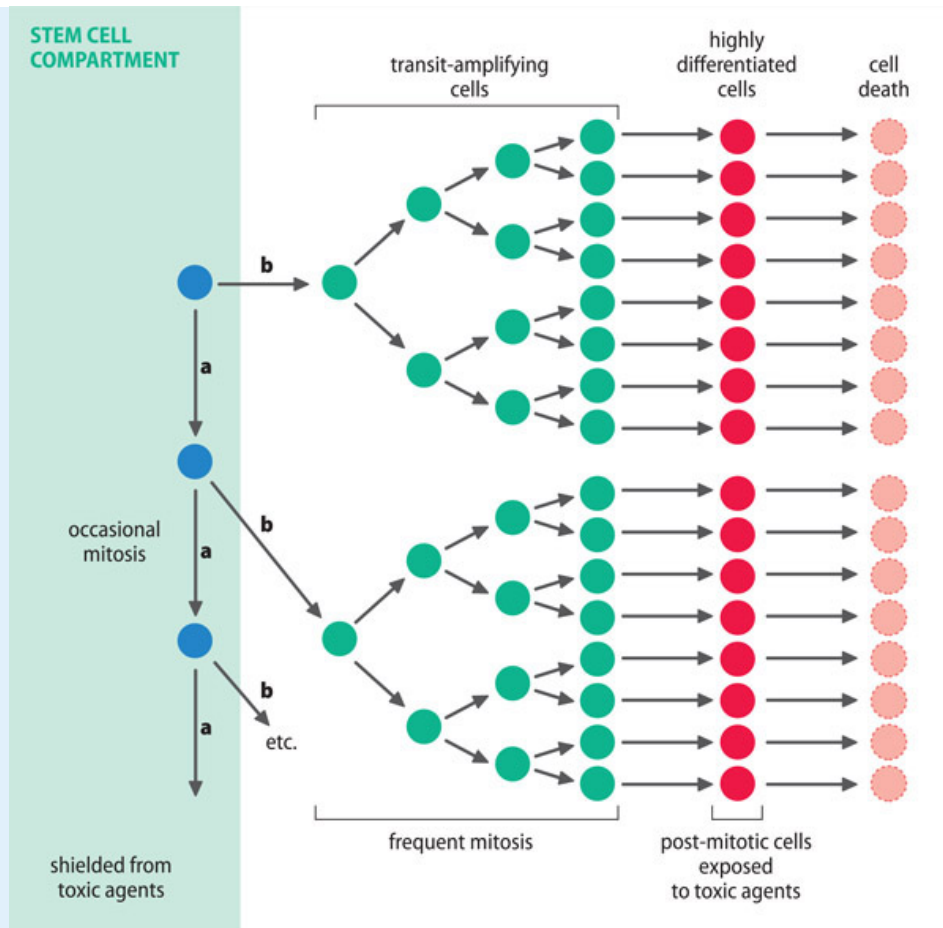


Figure 1 Epithelial tissue as an example of cell differentiation from stem cells and protection of the stem cell genome. Each stem cell (*blue*) divides only occasionally in an asymmetric fashion (steps a and b) to generate a new stem cell daughter and a more differentiated transit-amplifying daughter cell (*green*). Transit-amplifying cells undergo repeated rounds of growth and division, leading to exponential increase in cell numbers. Eventually, the products of these cell divisions further differentiate into post-mitotic highly differentiated cells (*red*). The highly differentiated cells are often in direct contact with various toxic agents, and are frequently shed (so that any harmful mutations that arise in these cells are quickly lost from the tissue). The stem cells are protected from the potentially mutagenic effects of toxic agents because they are shielded by an anatomical barrier. (Adapted from [Weinberg RA \[2014\]](#) *Biology of Cancer*, 2nd edn., Garland Science.)

Newly formed transit-amplifying cells go through multiple symmetrical cell divisions to produce very large numbers of cells that subsequently undergo differentiation to give rise to the highly differentiated, comparatively short-lived

cells. Because the latter cells have short lives, mutation in them could never lead to cancer. Instead, cancer must arise in a longer-lived progenitor cell.

Because the lineage of stem cells represents the only stable repository of genetic information within the tissue, the genomes of stem cells need to be protected as far as possible from mutation. First, the stem cell compartment is physically separated to reduce contact with potential mutagens. For intestinal epithelial cells, for example, the stem cell compartment lies at the base of the intestinal crypts (see [Figure 8.11B](#) on page 262 for the latter); here they are far removed from the epithelial cell lining that comes into contact with potentially hazardous mutagens in our diet. Secondly, because transit amplifying cells can expand exponentially from a single stem cell, stem cells may often be comparatively quiescent: by rarely needing to divide, mutations arising from DNA replication errors are minimized (but intestinal epithelial stem cells divide quite regularly).

Despite efforts to maintain their genomes, the very long lifetimes of stem cells make them targets for mutagenesis to form cancer stem cells. Additionally, mutations can be conveyed indirectly into the stem cell pool after mutated transit amplifying cells are dedifferentiated by epigenetic changes to become cancer stem cells. By being relatively resistant to cytotoxic chemicals and radiation, and able to keep on regenerating the more differentiated tumor cells, cancer stem cells pose problems for cancer treatment.

There is now substantial evidence for stem cells in many cancers, including in solid tumors. One of the early pieces of supportive evidence came from analyses of many types of leukemia. For example, the Philadelphia chromosome (a specific chromosomal translocation predisposing to chronic myeloid leukemia [CML; see below]) is often seen in different types of blood cells (B and T lymphocytes, neutrophils, granulocytes, megakaryocytes, and so on) in CML patients. The Philadelphia chromosome is presumed to arise in a precursor of all those different blood cell types, a hematopoietic stem cell. The idea of cancer stem cells is also supported by hierarchical cell organizations for certain types of cancer. In some types of neuroblastoma and myeloid leukemia, for example, the cancer evolves so that some tumor cells differentiate, and have limited capacity for proliferation despite retaining the oncogenic mutations of their malignant precursors.

Further evidence came from flow cytometry, which allows separation of phenotypically distinct subpopulations of live cancer cells that can be studied after transplanting them into immunocompromised mice. Using this approach, it became clear that only a small proportion of cancer cells in leukemia and breast cancer proliferate extensively, and they express specific combinations of cell surface markers. Breast-cancer-initiating cells, for example, are found to be a minority population of CD44⁺CD24⁻ cells and leukemia initiating cells are a minority population of CD34⁺CD38⁻ cells. Lineage studies have also supported the existence of cancer stem cells—we describe in [Section 10.5](#) how single-cell genomics is transforming cancer research.

10.2 ONCOGENES AND TUMOR SUPPRESSOR GENES

By 2020, 568 human genes had been identified as mutational cancer driver genes, having a causal role in cancer as a result of mutation. The relevant data were obtained by different approaches: analyzing tumor-associated chromosomal rearrangements (notably translocations), identifying tumor-specific changes in gene copy number, and identifying tumor-specific mutations (after comparing tumor DNA sequences with the corresponding DNA sequence in normal cells from the same individual).

Two fundamental classes of cancer gene

The key cancer-susceptibility genes—those in which driver mutations occur—can be grouped into two fundamental classes, according to how they work in cells. Some are dominant at the cellular level: mutation of a single allele is sufficient to make a major, or significant, contribution to the development of cancer. Others are recessive: both alleles need to be inactivated to make a significant contribution to cancer.

Oncogenes are the exemplars of dominantly acting cancer-susceptibility genes. In our cells the normal copies of these genes (sometimes called protooncogenes) often function in growth signaling pathways to promote cell proliferation or inhibit apoptosis, but as we describe below they can also work in other cellular functions. An activating mutation in a proto-oncogene can result in inappropriate

constitutive high-level expression (instead of being switched on just when needed). An activating mutation like this in just a single allele of a proto-oncogene can make a significant contribution to the tumorigenesis process.

Tumor suppressor genes are the exemplars of recessive cancer-susceptibility genes. Normal copies of classical tumor suppressor genes work in the opposite direction to oncogenes—to suppress cell proliferation (by inducing cell cycle arrest) or to promote apoptosis of deviant cells. When both alleles of a classical tumor suppressor gene are inactivated, that locus can make a significant contribution to cancer development.

A common analogy imagines an oncogene as the accelerator of a car and a tumor suppressor as the brake. The car will run out of control if the accelerator is jammed on (inappropriately activated) or if the brake fails. The cell is more complicated than this analogy allows: it has many different types of accelerator and brake to regulate cell growth and turnover, and usually several of the cell's accelerators and brakes need to be faulty to cause real damage.

As well as standard oncogenes and tumor suppressor genes, various other types of cancer-susceptibility gene have been identified that, when mutated, can assist tumor development. As described below, some work in DNA repair and genome maintenance. Some others support certain biological capabilities of cancer cells; they include, among others, genes encoding telomerase and proteins involved in energy metabolism and angiogenesis.

Viral oncogenes and the natural roles of cellular oncogenes

Oncogenes were discovered after it became clear that certain cancers in chickens and rodents were induced by viruses. (Note that most human cancers are not caused by viruses. Nevertheless, some viruses are implicated in specific human cancers: Epstein–Barr virus in nasopharyngeal carcinoma and lymphomas; some papillomaviruses in cervical and oropharyngeal squamous cell carcinoma; chronic hepatitis B virus infection in hepatocellular carcinoma; acute transforming human T-cell lymphotropic virus in acute T-cell leukemia; and human herpesvirus-8 in Kaposi's sarcoma.)

Among the viruses found to cause animal cancers were types of acute transforming retrovirus (also called oncoretroviruses) that could make cells in culture change their normal growth pattern to resemble that of tumors (a process known as **transformation**). Whereas normal versions of these retroviruses had the

three standard transcription units (*gag*, *pol*, and *env*), the oncoretroviruses had an altered genome in which part of the viral genome had been replaced by an altered copy of a cellular gene (a *proto-oncogene*). The copy of the cellular proto-oncogene in an oncoretrovirus is located close to powerful viral promoter/enhancer sequences that ensure inappropriate high-level expression that drives abnormal cellular proliferation, leading to cancer.

The normal cellular gene, the *proto-oncogene*, plays a role in growth-signaling pathways (as a growth factor, growth factor receptor, cell cycle regulator, transcription factor or some other protein working in signal transduction), or by inhibiting apoptosis. It normally promotes cell proliferation only when there is a natural need for cell proliferation. But it too becomes an oncogene when inappropriately expressed, as detailed in the following section

How normal cellular proto-oncogenes are activated to become cancer genes

Proto-oncogenes are activated by a DNA change that is dominant at the cellular level (and normally affects just a single allele). In the subsections below, we describe the three ways in which this can occur. Two of the three types of DNA change result in enhanced gene expression, so that the gene involved does not respond to normal inhibitory signals. The third class is made up of activating point mutations that alter how the protein behaves (see [Figure 10.6](#)).

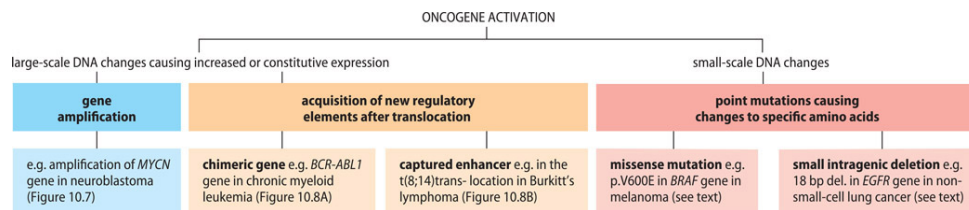


Figure 10.6. Three major ways in which cellular proto-oncogenes are activated to become cancer genes.

Note that cells have multiple anti-cancer defense systems, and the activation of a single cellular proto-oncogene is usually not oncogenic by itself. If we experimentally activate a cellular proto-oncogene in cultured cells, the usual effect is to induce cell cycle arrest (the abnormal proliferative signals usually induce cellular defense mechanisms that shut down cell proliferation); multiple genetic (and epigenetic) changes are needed to induce cancer.

Activation by gene amplification

Tumor cells often contain abnormally large numbers—often hundreds of copies—of a structurally normal oncogene. The *MYCN* oncogene, for example, is frequently amplified in late-stage neuroblasts ([Figure 10.7A](#)) and in rhabdomyosarcomas; *ERBB2* (also called *HER-2*) is often amplified in breast cancers. The gene amplification mechanism is not simple tandem amplification; instead, there seem to be complex rearrangements that bring together sequences from several different chromosomes. The amplification may manifest itself in two forms:

- *double minutes*, an extrachromosomal form made up of tiny, paired acentric chromatin bodies that are separated from chromosomes and contain multiple copies of just a small set of genes ([Figure 10.7B](#))
- *homogeneously staining regions*, a corresponding intrachromosomal form in which multiple repeated copies integrate into chromosomes.

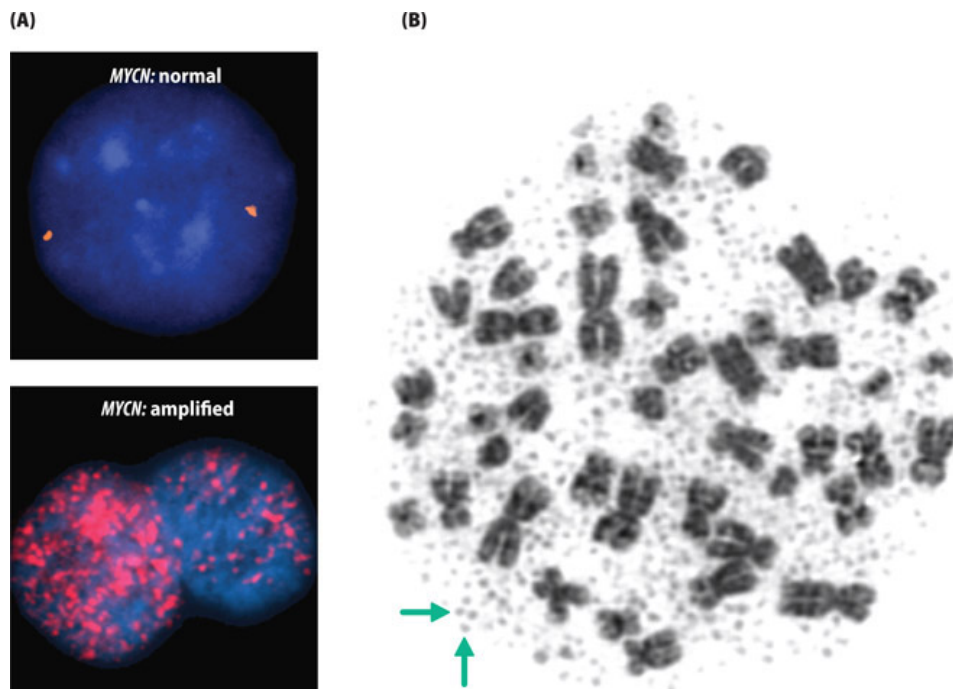


Figure 10.7 Amplification of the *MYCN* gene and formation of double minutes in neuroblastoma cells. (A) Fluorescence *in situ* hybridization (FISH) images using a labeled *MYCN* gene probe, showing two copies of the gene (red signals) in normal cells against a background of DNA staining (shown in blue). In neuroblastoma cells, the *MYCN* gene can undergo extensive amplification to produce many dozens or even hundreds of *MYCN* gene

copies (as shown at the bottom). (B) A metaphase chromosome preparation from a neuroblastoma tumor sample, showing *double minute* chromosomes (which appear as a cloud of very small dots; arrows indicate two of them). (A, Courtesy of Nick Bown, NHS Northern Genetics Service, Newcastle upon Tyne, UK; B, Courtesy of Paul Roberts, NHS Cytogenetics Service, Leeds, UK.)

Translocation-induced oncogene activation

Chromosomal translocations occur when DNA molecules receive double-strand breaks and are then rejoined incorrectly so that pieces of different DNA molecules are joined together. When that happens, an oncogene is often inappropriately transcriptionally activated and so there can be a selective growth advantage.

Translocations that activate oncogenes are common in cancer (more than 300 cancer-associated translocations are listed within the Cancer Gene Census section of the COSMIC database described in [Section 10.4](#)). In many cases, the translocations result in the formation of clearly chimeric genes that result in the constitutive expression of oncogene sequences. In other cases, the oncogene sequence is not interrupted by a breakpoint; instead it is simply brought into close proximity to regulatory sequences in another gene that is actively expressed (see [Table 10.4](#) for some examples).

TABLE 10.4 EXAMPLES OF ONCOGENE ACTIVATION BY TRANSLOCATION

Tumor type	Oncogene (location)	Interacting gene (location)
Acute lymphoblastoid leukemia (ALL)	<i>MLL</i> (11q23)	<i>AF4</i> (4q21), <i>AF9</i> (9p22) <i>AFX1</i> (Xq13), <i>ENL</i> (19p13)
Acute myeloid leukemia (AML)	<i>FUS</i> (16p11)	<i>ERG</i> (21q22)
Acute promyelocytic leukemia	<i>PML</i> (15q24)	<i>RARA</i> (17q21)

Note that certain oncogenes such as *MLL* participate in translocations with many other genes and that immunoglobulin genes (such as *IGH*), and T-cell receptor genes (such as *TRD*) are frequently involved in oncogene-activating translocations to cause B- or T-cell cancers. For a complete list, go to the Cosmic database at <https://cancer.sanger.ac.uk/census/> then move to the Breakdown list and select Translocations.

* See [Figure 10.8A](#).

Tumor type	Oncogene (location)	Interacting gene (location)
Burkitt's lymphoma	<i>MYC</i> (8q24)	<i>IGH</i> (14q32)
Chronic myeloid leukemia (CML)	<i>ABL</i> (9q34)*	<i>BCR</i> (22q11)*
Ewing sarcoma	<i>EWS</i> (22q12)	<i>FLI1</i> (11q24)
Follicular B-cell lymphoma	<i>BCL2</i> (18q21)	<i>IGH</i> (14q32)
Tcell leukemia	<i>LMO1</i> (11p15), <i>LMO2</i> (11p13), <i>L1</i> (1p32)	<i>TRL1</i> (14q11)

Note that certain oncogenes such as *MLL* participate in translocations with many other genes and that immunoglobulin genes (such as *IGH*), and T-cell receptor genes (such as *TRD*) are frequently involved in oncogene-activating translocations to cause B- or T-cell cancers. For a complete list, go to the Cosmic database at <https://cancer.sanger.ac.uk/census/> then move to the Breakdown list and select Translocations.

* See [Figure 10.8A](#).

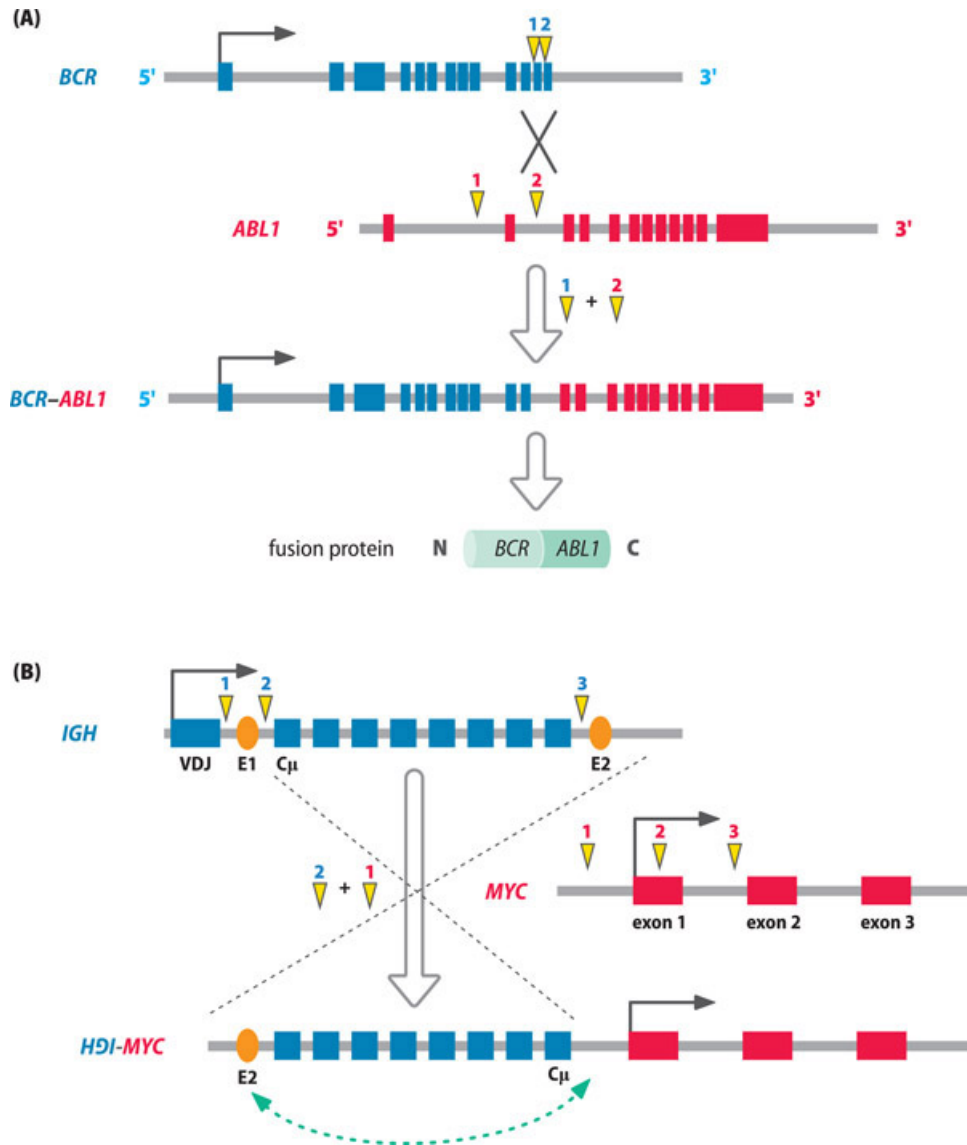


Figure 10.8 Two ways in which translocation can result in oncogene activation by ensuring inappropriate expression. (A) Chimeric gene formation. This shows formation of the chimeric *BCR-ABL1* gene in acute myeloid leukemia, permitting inappropriate expression of the *ABL1* oncogene as a fusion *BCR-ABL1* protein expressed through the promoter and regulatory sequences provided by the *BCR* gene. Blue and red vertical bars represent exons. Vertical yellow darts indicate observed breakpoints in patients; here we show recombination at breakpoint 1 in the *BCR* gene, and at breakpoint 2 in the *ABL1* gene. The resulting *BCR-ABL1* chimeric gene produces a large protein with constitutively active tyrosine kinase activity, which does not respond to normal controls. (B) Enhancer capture. The common t(8;14) translocation in Burkitt's lymphoma brings a B-cell specific enhancer (yellow oval) from the *IGH* (immunoglobulin heavy chain) gene into close proximity to the *MYC* gene so that *MYC* is inappropriately activated in B-cells. Here we show the chromosome that results after the use of breakpoint 2 in

the *IGH* gene and breakpoint 1 in the *MYC* gene. Note that on the translocation chromosome the two sense strands of the genes are on opposing DNA strands so that the 3' end of the *IGH* gene is distant from the 5' end of the *MYC* gene. The E2 B-cell enhancer is nevertheless close enough to the promoter of the *MYC* gene so that they interact (dashed green arrow), driving strong inappropriate expression of the *MYC* gene in B cells.

The Philadelphia (Ph¹) chromosome, occurring in 90 % of individuals with chronic myeloid leukemia, illustrates how a translocation gives rise to cancer via a chimeric gene. It results from a balanced reciprocal translocation with breakpoints near the start of the *ABL1* oncogene at 9q34 and close to the end of *BCR* gene at 22q11 ([Figure 10.8A](#)). The resulting *BCR-ABL1* fusion gene on the Philadelphia chromosome (with the *ABL1* coding sequence positioned downstream of the *BCR* gene sequence and *BCR* promoter) produces a large protein that carries the *ABL1* polypeptide sequence at its C-terminal end. This fusion protein acts as a growth-stimulating tyrosine kinase that is *constitutively* active and so drives cell proliferation.

Tumors of B and T cells, including various lymphomas and leukemias, often result from translocations with breakpoints in an immunoglobulin heavy-chain or light-chain gene, notably *IGH*, or in a T-cell receptor gene such as *TRA*, *TRB*, or *TRD*. Recall from [Section 4.5](#) that developing B and T cells are quite exceptional cells because of the requirement for programmed DNA rearrangements that rearrange, respectively, immunoglobulin (Ig) and T-cell receptor (TCR) genes in order to make cell-specific Ig or TCR chains. Because of the natural need to produce double-strand breaks in the Ig or TCR genes, there is a higher chance that these genes will participate in translocations.

Because the large Ig and TCR genes contain many different enhancer sequences, translocations can often result in the transcriptional activation of an oncogene that lies close to the reciprocal breakpoint. As a result, some translocations activate an oncogene simply by bringing a T- or B-cell specific enhancer in close proximity to the oncogene promoter (see [Figure 10.8B](#))

Gain-of-function mutations

Oncogenes can also be activated by certain point mutations that make a specific change at one of a few key codons (often a missense mutation, but sometimes

small deletions of a few codons can change the behavior of the relevant protein). Activating mutations in some cellular oncogenes are particularly common, especially when the genes make a product that links different biological pathways connected to cell proliferation and growth.

Take, for example, the human Ras oncogenes—*HRAS*, *KRAS*, and *NRAS*—that make highly related 21 kDa Ras proteins with 188 or 189 amino acids. The Ras proteins work as GTPases and mediate growth signaling by receptor tyrosine kinases in mitogen-activated protein (MAP) kinase pathways. Heavily implicated in cancer, they act as signaling hubs (a single Ras protein interacts with multiple intercellular signaling proteins and can transmit a signal from a receptor tyrosine kinase to various downstream signaling pathways, thereby affecting multiple processes). About one in six human cancers has activating mutations in a RAS gene, most commonly in *KRAS* (which is naturally expressed in almost every tissue). More than 99 % of the activating Ras mutations are in one of only three key codons, codon 12 (Gly), codon 13 (Gly), and codon 61 (Gln).

The bias toward missense mutations, and the very narrow distribution of where the mutations occur, distinguishes oncogenes from tumor suppressor genes (see [Figure 10.9](#) for examples). In some cases, small intragenic deletions that remove a few codons are observed that can also result in a change of function. For example, the c.2240_2257del18 mutation in the *EGRF* gene is commonly found in non-small-cell lung cancer, and replaces a heptapeptide sequence of the EGFR protein by a serine. The mutation affects an ATP-binding pocket and the effect is to enhance signaling, a gain of function.



Figure 10.9 Oncogenes differ from tumor suppressor genes in the distribution and range of cancer-associated mutations. The distributions of cancer-associated missense mutations (red arrowheads) and mutations introducing a premature termination codon (PTC; blue arrowheads) are mapped to the corresponding regions of protein products for two representative oncogenes (*PIK3CA* and *IDH1*) and two tumor suppressor genes (*RB1* and *VHL*). Colored bars on the pale green background represent functional domains and motifs. The data were collected from genome-wide studies annotated in the COSMIC database (release version 61). For *PIK3CA* and *IDH1*, mutations obtained from the COSMIC database were randomized, and the first 50 are shown. For *RB1* and *VHL*, all mutations recorded in COSMIC were plotted. Note the predominance of missense mutations in the oncogenes and how they are restricted to just a very few codons. Abbreviation: aa, amino acid residues. (From Vogelstein B et al. [2013] *Science* 339:1546–1558; PMID 23539594. With permission from the AAAS.)

It should be noted that even advanced cancers retain some characteristics of their tissue of origin, and so a gene that might behave as an oncogene in one type of tumor may behave differently in a tumor originating from a different tissue. Thus, for example, the frequent observation of specific missense mutations in the *NOTCH1* gene in lymphomas and leukemias indicate that here *NOTCH1* behaves as an oncogene. But in squamous cell carcinomas, *NOTCH1* mutations are nonrecurrent and usually inactivating, suggesting that in these tumors *NOTCH1* might behave as a tumor suppressor.

Tumor suppressor genes: normal functions, the two-hit paradigm, and loss of heterozygosity in linked markers

Tumor suppressor genes make products that keep cells under control by restraining cell proliferation either directly or indirectly, and have been classified into different groups as listed below.

- *Gatekeeper genes* directly restrain cell proliferation. Their products regulate cell division—by regulating the cell cycle and inducing cell cycle arrest, as required, or by working in upstream growth signaling pathways—or they promote apoptosis.
- *Caretaker genes*, indirectly restrain cell proliferation by helping to maintain the integrity of the genome
- *Landscaper genes* indirectly restrain cell proliferation by controlling the stromal environment in which the cells grow.

Unlike oncogenes, a tumor suppressor gene contributes to cancer when the gene is lost or inactivated in some way. Whereas mutated oncogenes act in a dominant manner at the cellular level, mutated tumor suppressor genes often act in a recessive manner. For classical tumor suppressor genes, inactivation of one copy of a tumor suppressor gene has little effect; the additional loss or inactivation of the second gene is required in the tumorigenesis process. For these genes, the tumor suppressor locus needs to sustain two “hits” to make a significant contribution to tumorigenesis.

Familial cancers and the two-hit paradigm

From what we have described so far, the idea of familial cancers might seem strange; nevertheless, they do account for a minority of cancers. Familial cancers nearly always involve inheritance of a loss-of-function allele in a tumor suppressor gene (but see below for the example of inherited mutations in the *RB1* oncogene).

The two-hit hypothesis proposed by Alfred Knudson explained why certain tumors can occur in hereditary or sporadic forms. In the hereditary form, one inactivating mutation (the first hit) in a tumor suppressor gene is inherited and the second hit occurs in the somatic cancer progenitor cell; in the sporadic form, two successive inactivating hits, one in each allele, occur in a somatic cell to initiate tumorigenesis (see [Figure 10.10](#)).

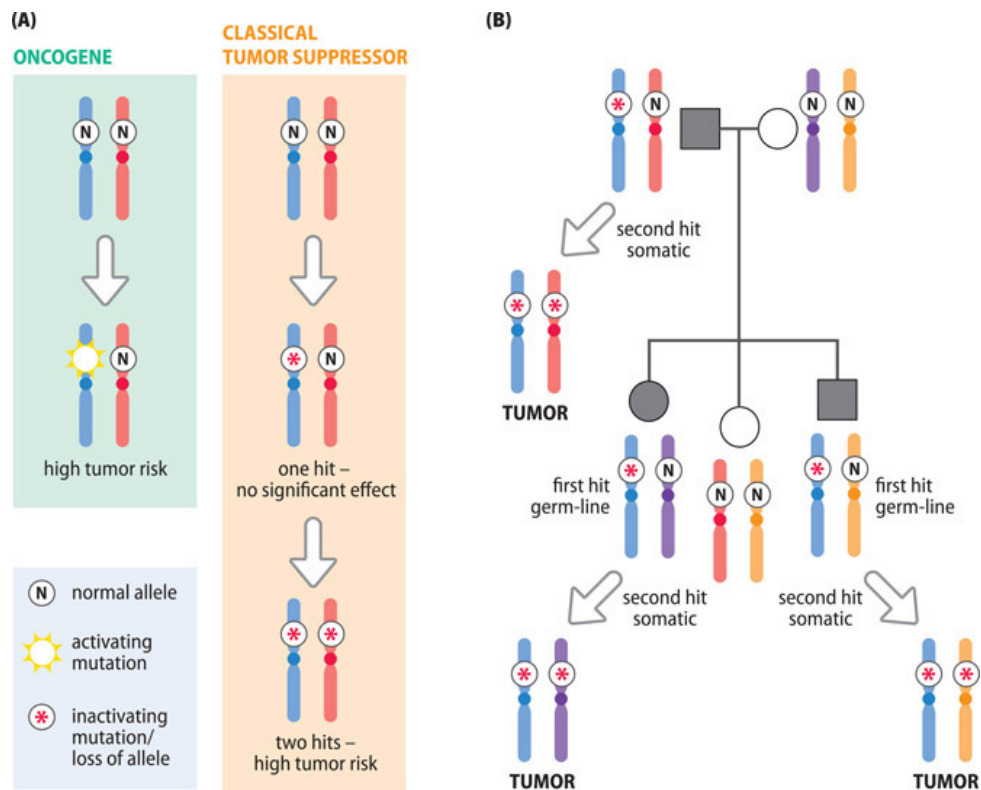


Figure 10.10 Classical tumor suppressor genes and the two-hit hypothesis. (A) Activating mutations in a single allele of an oncogene are sufficient to confer a high risk of tumorigenesis. For a classical tumor suppressor locus to make a significant contribution to tumorigenesis, both alleles need to lose their function (the loss of function may occur through mutational inactivation or loss of the allele, or sometimes epigenetic silencing). Some tumor suppressors do not follow this simple model. (B) Cancers due to mutations at a classical tumor suppressor locus are recessive at the cellular level (both alleles need to be inactivated) but cancer susceptibility

can still be dominantly inherited. Inheritance of a single germline mutation (first hit, on the pale blue chromosome here) means that each cell of the body already has one defective allele and there is a very high chance of some cells receiving a second (somatic) hit. In sporadic forms of the disease, tumors are thought to arise by two sequential somatic mutations in the same cell.

Retinoblastoma, a cancer of the eye that represents 3 % of childhood cancers, provided the first support for the two-hit hypothesis. In retinoblastoma, tumors can occur in both eyes or in one eye. People with bilateral tumors often transmit the disorder to their children, but the children of a person with a unilateral retinoblastoma usually do not have retinoblastoma.

Statistical modeling indicated that hereditary cases of retinoblastoma probably developed after only one somatic mutational event. People with bilateral retinoblastomas were postulated to have inherited an inactivating mutation in one copy of a retinoblastoma-susceptibility locus, now called *RB1*; in that case each nucleated cell in the body would have one inactive *RB1* allele. Retinoblastomas develop from many poorly differentiated retinoblast progenitor cells that proliferate rapidly. There is therefore a high chance that within a population of a million or so retinoblasts carrying an inactivated *RB1* allele, more than one cell sustains an additional inactivating mutation in the second *RB1* allele. If multiple tumors can form, bilateral tumors are likely to occur.

If, however, two normal *RB1* alleles have been inherited, each tumor must occur by two successive hits at the *RB1* locus in one somatic cell. Unless the first somatic mutation just happened to occur very early in embryogenesis, the chances that two sequential somatic mutations would cause a loss of function of both *RB1* copies in more than one cell would be expected to be very rare. That makes unilateral retinoblastoma the expected outcome; the age of onset is generally later than in cases with inherited *RB1* mutations.

Note that while people with bilateral tumors can be confidently expected to have inherited a germline *RB1* mutation, a minority of people with unilateral tumors also have a germline mutation (by chance, tumor formation has only occurred in one eye). Inheritance of retinoblastoma susceptibility is dominant, but incompletely penetrant.

The two-hit paradigm explains why the cancer can be transmitted in a dominant fashion, even although the phenotype is recessive at the cellular level (both alleles of a tumor suppressor gene need to be inactivated or silenced); see [Figure 10.10B](#).

It also applies well to some other cancers that exist in both familial and sporadic forms but not to some other cancers, as described below.

Loss of heterozygosity

For tumor suppressor genes, the initial hit is typically confined to the tumor suppressor locus—usually an inactivating point mutation. Inactivation of the second allele can also occur by a locus-specific DNA change—a point mutation, gene deletion, or gene conversion—or sometimes by epigenetic silencing.

Often, however, the second allele is inactivated by large-scale DNA changes (loss of the whole chromosome, or a substantial part of it) or by mitotic recombination (loss of the whole chromosome, or a substantial part of it) or by mitotic recombination (Figure 10.11). In that case *loss of heterozygosity* will be evident: linked DNA markers that are constitutionally heterozygous in normal blood cells from an individual contain just a single allele in the tumor sample.

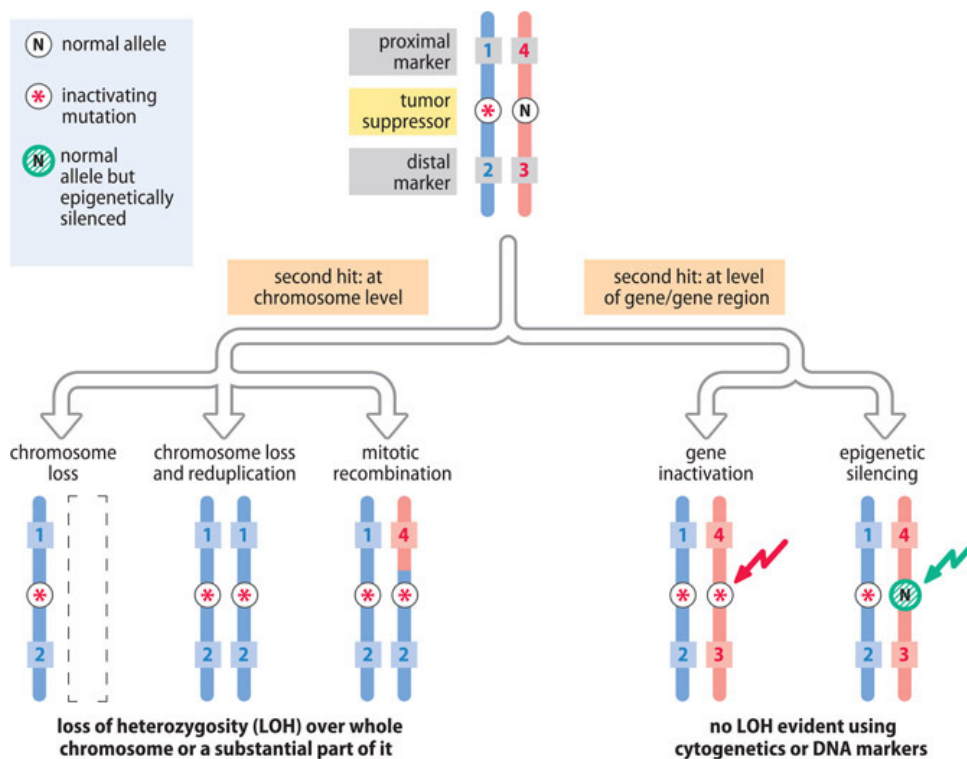


Figure 10.11 Different types of second hit at a tumor suppressor locus, some readily detectable by screening for loss of heterozygosity, and others not. Here the first hit at the tumor suppressor locus is shown as a small-scale inactivating mutation on the blue chromosome. A second hit that involves a large-scale (chromosomal) change (such as loss of the orange

chromosome or loss of a part of that chromosome by mitotic recombination) can result in obvious *loss of heterozygosity* (readily detectable at the level of cytogenetic or DNA marker analysis). Sometimes, however, the second hit can be an inactivating mutation at the second allele or an epigenetic silencing event encompassing the tumor suppressor locus. In these cases, both alleles are unable to be expressed but loss of heterozygosity would not be evident by either cytogenetic analyses or DNA analyses using flanking markers.

Loss of heterozygosity has been used as a way of mapping tumor suppressor genes. Paired samples of blood and the relevant tumor from individuals are screened with DNA markers from across the genome to identify chromosomes and, more profitably, chromosomal regions that show convincing loss of constitutional marker heterozygosity in the tumor samples. Analysis of multiple different tumors might lead to the identification of a quite small subchromosomal region defined by different mitotic crossovers or other breakpoints observed.

The key roles of gatekeeper tumor suppressor genes in suppressing G₁-S transition in the cell cycle

Understanding how cell division is regulated is of paramount importance in understanding cancer. Protein complexes made up of cyclins and cyclin-dependent kinases (CDKs) have key roles in regulating the cell cycle at certain cell cycle *checkpoints*.

The regulation of G₁, the phase of the cell cycle when cells make the decision whether or not to divide, is pivotal in tumorigenesis. A principal checkpoint occurs late in G₁, close to the G₁/S boundary, and is subject to intense regulation. A complex of cyclin E and the CDK2 protein works at this checkpoint to promote the transition from G₁ to S phase (which will usually commit the cell to cell division).

The CDK2–cyclin E complex is in turn regulated by interconnecting pathways. The control system has two arms in which the RB1 and p53 tumor suppressor proteins have commanding roles; another three tumor suppressor proteins, p14, p16, and p21 (the numbers refer to initially estimated molecular weights in kDa) support p53 and RB1 in putting a brake on cell division ([Figure 10.12](#)).

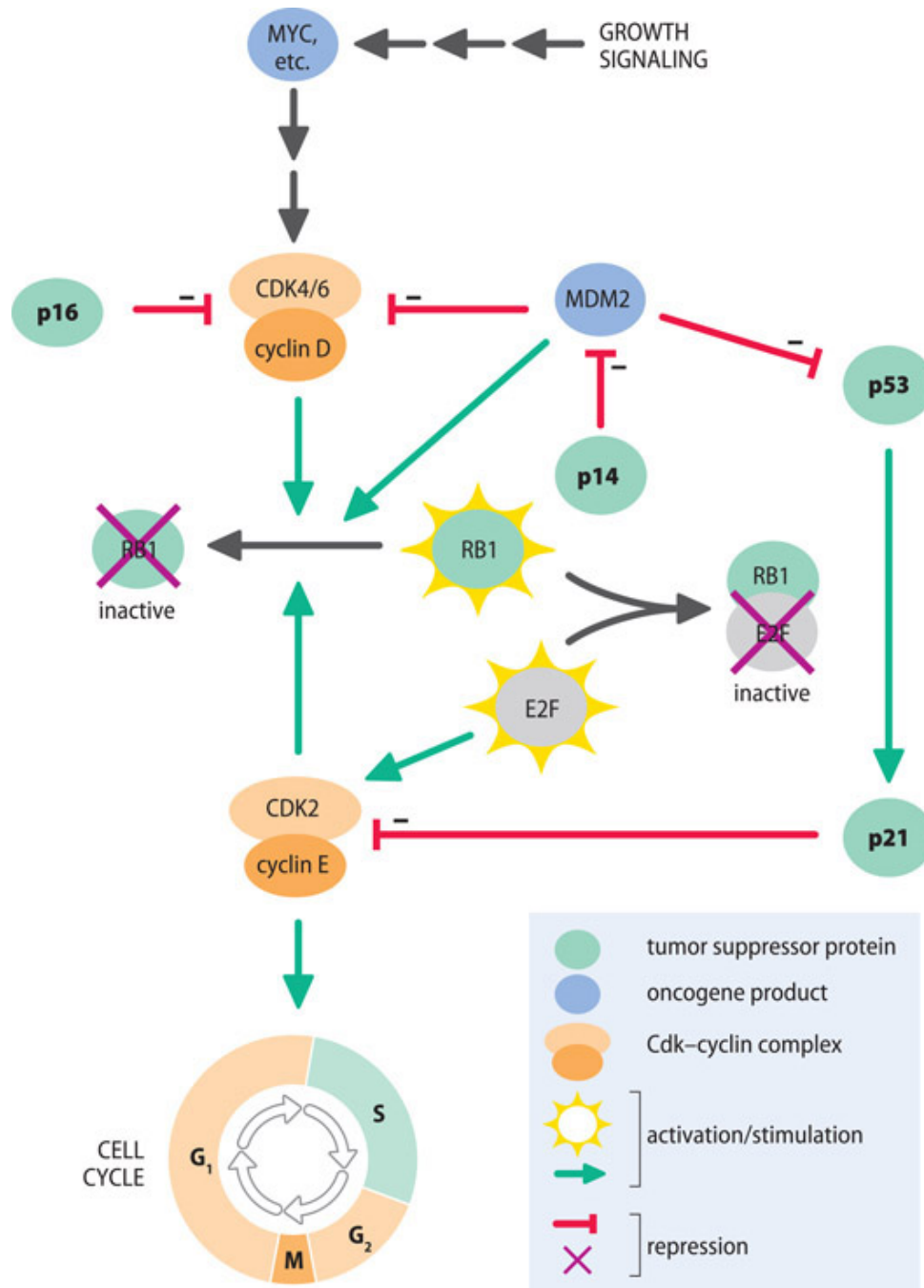


Figure 10.12 Major roles for p53, RB1, and accessory tumor suppressor proteins as brakes on cell growth. To permit cell growth, the CDK2–cyclin E complex promotes the G₁-S transition and is stimulated to do so by the E2F transcription factor. (E2F activates the transcription of multiple genes whose products are required for progression to S phase, notably cyclin E.) Five tumor suppressor proteins work in the opposite direction, as brakes on cell growth. RB1 inhibits the E2F transcription factor by binding to it to keep it in an inactive form. It, in turn, is repressed by CDK4/6–cyclin D and MDM2 but is assisted by proteins that repress

its inhibitors: p16 (also called INK4a because it inhibits cyclin-dependent kinase 4) and p14 (also called ARF; it inhibits MDM2). p53 and p21 work in a pathway that bypasses E2F to inhibit the CDK2-cyclin E complex directly. Normally, p53 concentrations are kept low in cells but are increased in response to severe DNA damage. Elevated p53 both suppresses cell division (as shown here) and also stimulates apoptosis pathways (as shown in [Figure 10.13](#)).

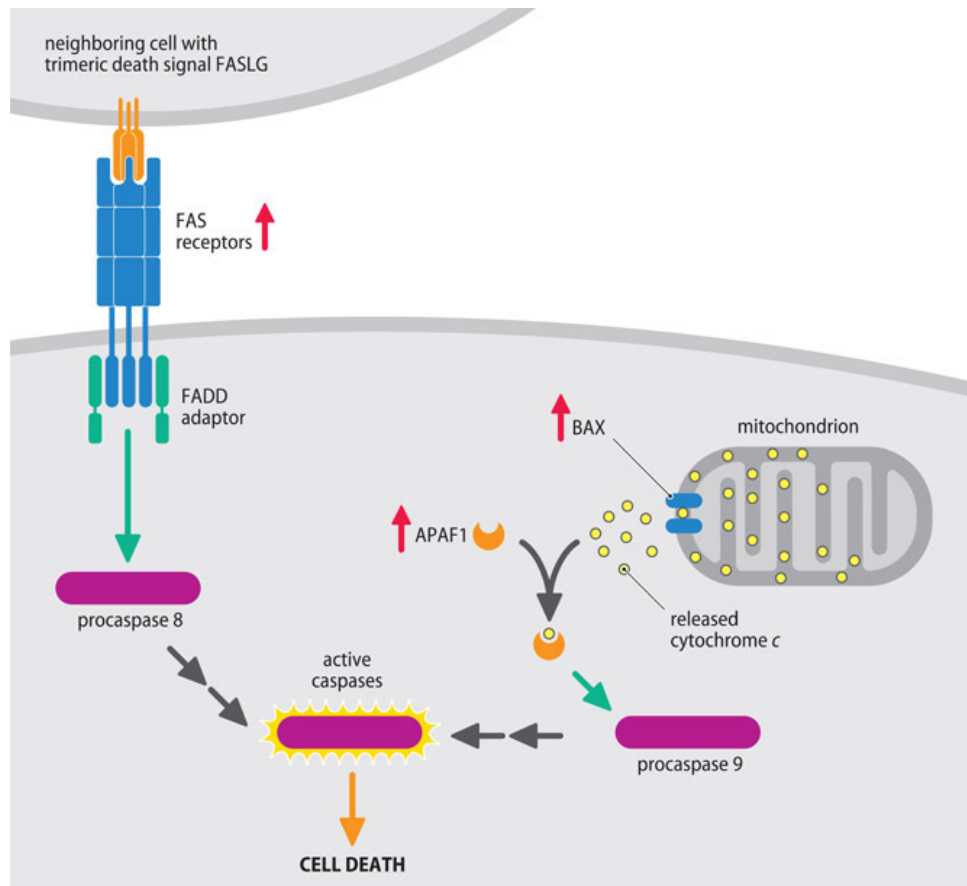


Figure 10.13 Regulation of different apoptosis pathways by p53. When actively expressed at high concentrations, p53 stimulates the transcription of various genes to produce increased quantities of apoptosis-promoting proteins (indicated by vertical red arrows). They include cell surface receptors that are able to recognize death signals from neighboring cells, such as FAS receptors, and regulators of the mitochondrial apoptosis pathway, notably BAX and APAF1. FAS receptors are monomers, but when contact is made with the trimeric FAS ligand (FASLG) they form trimers. The FAS trimers recruit an adaptor (FADD), forming a platform for binding and activating procaspase 8. The BAX1 protein forms oligomers within the mitochondrial outer membrane that act as pores, allowing the release of cytochrome *c* into the cytosol. The released cytochrome *c* binds and activates APAF1, which in turn binds and activates procaspase 9. Activated procaspases 8 and 9 ultimately lead to mature effector caspases that destroy the cell.

Growth signaling pathways can induce a loss of RB1 function by stimulating CDK4–cyclin D or CDK6–cyclin D complexes that inactivate RB1 (by phosphorylating it) and the negative regulator MDM2 (which adds ubiquitin residues to target RB1 for destruction to keep RB1 levels low when growth is needed). Otherwise, p16 and p14 work to suppress the inhibition of RB1 and so put a brake on cell growth ([Figure 10.12](#)).

Elevated levels of p53 protein can stimulate p21 to inhibit the CDK2-cyclin E complex so as to induce cell cycle arrest. However, p53 is normally kept at relatively low concentrations in cells, mostly because MDM2 binds to p53 and adds ubiquitin groups to it, targeting it for destruction.

Signals from different sensors that detect cell stress—such as sensors of DNA damage—result in phosphorylation of p53. Phosphorylated p53 is not bound by MDM2 and so p53 levels increase. This can lead to cell cycle arrest (see [Figure 10.12](#)), which provides the opportunity to repair DNA, or to apoptosis when the DNA damage is too severe to repair. As detailed below, p53 has a pivotal role in cancer and is a rather unconventional tumor suppressor.

Note that although very different in sequence, the p14 and p16 tumor suppressors are both made from alternative splicing of a single gene, *CDKN2A* (see [Figure 6.8B](#) on page 146), and loss-of-function mutations in this one gene can inactivate both the RB1 and p53 arms of the cell cycle control system. Not surprisingly, *CDKN2A* mutations are important in tumorigenesis, and homozygous deletion or inactivation of this gene is quite common in cancers.

The additional role of p53 in activating different apoptosis pathways to ensure that rogue cells are destroyed

Cells that are unwanted, heavily damaged, or actively dangerous are normally induced to commit suicide through **apoptosis** (programmed cell death) pathways. Some apoptosis pathways work through a cell surface receptor that receives a “death signal” from neighboring cells (examples include FAS receptors and other members of the tumor necrosis factor receptor superfamily). Other pathways, such as the mitochondrial apoptosis pathway, respond to certain types of internal damage, such as that caused by harmful reactive oxygen species or exposure to dangerous levels of ionizing radiation.

In most cases the apoptosis pathway ends by triggering the cell to produce certain caspases, proteolytic enzymes that wreak havoc by inactivating all kinds of

important proteins in the cell; an endonuclease is also activated that cleaves DNA into small fragments. Because each of our normal cells has the potential to commit suicide, apoptotic pathways need to be very tightly regulated.

Various cancer-associated genes make products that regulate apoptosis. They include some tumor suppressors, notably *TP53*. When an unexpected double-strand break occurs in DNA, the DNA damage response activates high-level expression of p53. In response, p53 may activate transcription of various apoptosis-promoting genes in different apoptosis pathways ([Figure 10.13](#)).

From the above, we can see that p53 has dual central roles: it inhibits excessive cell proliferation, and it also acts as a “guardian of the genome” by inducing apoptosis in response to double-strand DNA breaks (which are common in cancer cells). To promote cell proliferation and inhibit apoptosis, cancers frequently seek to inactivate both *TP53* alleles, and *TP53* is the most commonly mutated gene in cancer.

Note that oncogenes also have a role in inhibiting apoptosis. For example, the oncogene *BCL2* works in the mitochondrial apoptosis pathway, where its protein product inhibits cytochrome *c* release from mitochondria and is inhibited in turn by the BAX protein. In cancer cells, over-expression of certain oncogenes, such as *BCL2*, inhibits apoptosis.

Tumor suppressor involvement in rare familial cancers and non-classical tumor suppressors

Familial cancer is comparatively infrequent. Some rare examples are known of heritable oncogene mutations that cause cancer. For example, germline mis-sense mutations in the *RET* proto-oncogene are found in familial thyroid cancer. However, the great majority of familial cancers have germline mutations in tumor suppressor genes, including both gatekeeper genes such as *RB1* (with normal roles in restraining cell proliferation and/or promoting apoptosis) and caretaker genes (with genome maintenance roles, notably in DNA repair); [Table 10.5](#) gives some examples.

TABLE 10.5 EXAMPLES OF FAMILIAL CANCERS RESULTING FROM GERMLINE MUTATIONS IN TUMOR SUPPRESSOR GENES

Familial cancer type	Gene*	Normal function of gene product(s)
----------------------	-------	------------------------------------

Familial cancer type	Gene*	Normal function of gene product(s)
DEFECT IN GATEKEEPER GENE		
Familial adenomatous polyposis coli	<i>APC</i>	multiple functions, notably in signal transduction (Wnt pathway)
Familial melanoma	<i>CDKN2A</i>	two unrelated protein products, p14 and p16, facilitate p53-mediated cell cycle arrest (see Figure 10.12)
Gastric carcinoma	<i>CDH1</i>	regulator of cell-cell adhesion
Gorlin syndrome (basal cell carcinoma, medulloblastoma)	<i>PTCH</i>	sonic hedgehog receptor
Juvenile polyposis coli	<i>DPC4</i>	signal transduction (TGF β pathway)
	<i>SMAD4</i>	
Li-Fraumeni syndrome (multiple different tumors)	<i>TP53</i>	the p53 transcription factor induces cells to undergo cell cycle arrest (see Figure 10.12) or apoptosis (see Figure 10.13)
Neurofibromatosis type 1 (NF1)	<i>NF1</i>	negative regulation of Ras oncogene
Neurofibromatosis type 2 (NF2)	<i>NF2</i>	cytoskeletal protein regulation
Retinoblastoma	<i>RB1</i>	acts as a brake on the cell cycle (see Figure 10.12)
Wilms tumor (childhood kidney tumor)	<i>WT1</i>	a transcriptional repressor protein with multiple functions including regulating the fetal mitogen insulin-like growth factor
DEFECT IN CARETAKER GENE		
Familial breast/ovarian cancer	<i>BRCA1</i>	makes product that interacts with double-strand DNA repair complex/components
	<i>BRCA2</i>	
Hereditary non-polyposis colorectal cancer (Lynch syndrome)	<i>MLH1</i>	DNA mismatch repair
	<i>MSH2</i>	

Familial cancer type	Gene*	Normal function of gene product(s)
----------------------	-------	------------------------------------

Gatekeeper genes include classical tumor suppressors that work in regulating cell division or upstream growth signaling pathways. Caretaker genes include other tumor suppressors that work in DNA repair or DNA damage responses.

* Predisposing locus that shows germline mutations.

In retinoblastoma, few driver mutations are thought to be required for tumorigenesis (embryonic retinal progenitor cells are both poorly differentiated and rapidly proliferating, and so these cells already have two important tumor cell characteristics). The two-hit paradigm applies to additional types of cancer in which investigation of rare familial forms led to the identification of a tumor suppressor gene that was then found to be mutated in more common sporadic forms.

Some cancers that exist in both heritable and sporadic forms do not, however, readily fit the classical two-hit tumor suppressor paradigm. Major tumor suppressor genes implicated in the common sporadic tumors are often different from those involved in familial forms. This can be explained at least in part by disease heterogeneity. For example, *BRCA1*, the principal tumor suppressor gene implicated in familial breast cancer, is inactivated in only 10–15 % of sporadic breast cancers. The latter form a recognizably distinct subset of sporadic breast cancers (and in these cases any second hit occurs by epigenetic silencing). Other data from many cancers have prompted the need for a radical overhaul of the classical two-hit suppressor hypothesis, as described in the next subsections.

Non-classical tumor suppressors

Some cancer-susceptibility genes seem to lose the function of one allele but the second allele seems perfectly normal at the DNA level. Sometimes the second allele is *epigenetically silenced*. In other cases, however, inactivating a single allele seems to be sufficient to induce a tumorigenic change; that is, a significant contribution to tumorigenesis can be made by heterozygous loss of function, or *haploinsufficiency*.

Examples include some tumor suppressor genes involved in genome stability for which homozygous inactivation would be expected to lead to cell death (but

can be averted by a third hit such as mutation in *TP53*). And mutation of just a single allele of certain tumor suppressor genes, such as *BRCA1*, has been shown to lead to genome instability in cultured cells and animal models.

Gain-of-function mutations can also occur in some tumor suppressor genes; in that case, a single mutated tumor suppressor allele can behave like an onco-gene. For example, missense mutations are very common in *TP53*, which makes the p53 tumor suppressor, and the resulting mutant p53 proteins can behave in a dominant-negative fashion ([Box 10.3](#)).

Partial inactivation of tumor suppressors can also make vital contributions to tumorigenesis; even quite subtle changes to the dosage of some tumor suppressors can sometimes make a substantial difference. The dosage effects can be highly tissue-specific and dependent on the context, such as the genetic background—for the example of the *PTEN* tumor suppressor, see the review by [Berger et al. \(2011\)](#) under Further Reading.

BOX 10.3 A CENTRAL ROLE IN CANCER FOR THE *TP53* SUPPRESSOR GENE THAT MAKES A NON-CLASSICAL TUMOR SUPPRESSOR, p53

The *TP53* gene at 17p13 has a central role in cancer, being mutated in nearly half of all tumors. The gene product, p53, has many roles and is involved in numerous different features of cancer. However, much of its importance comes from its role as a “guardian of the genome”—it connects DNA damage, a common feature in cancer cells (which frequently undergo genome instability as described in [Section 10.3](#)), to decisions to induce cell cycle arrest (see [Figure 10.12](#)) or apoptosis (see [Figure 10.13](#)).

The p53 control mechanism that seeks to nip tumor-igenesis in the bud can never be a failsafe mechanism; as a back-up, two p53-related proteins, p63 and p73, are produced with functions that partly overlap those of p53. Nevertheless, p53 has the dominant role.

As befits its crucial role, p53 is expressed in all cells. Germline mutations in *TP53* underlie Fraumeni syndrome (OMIM 151623), a dominantly inherited disorder in which those affected within a family can present with different early-onset tumors ([Figure 1](#)).

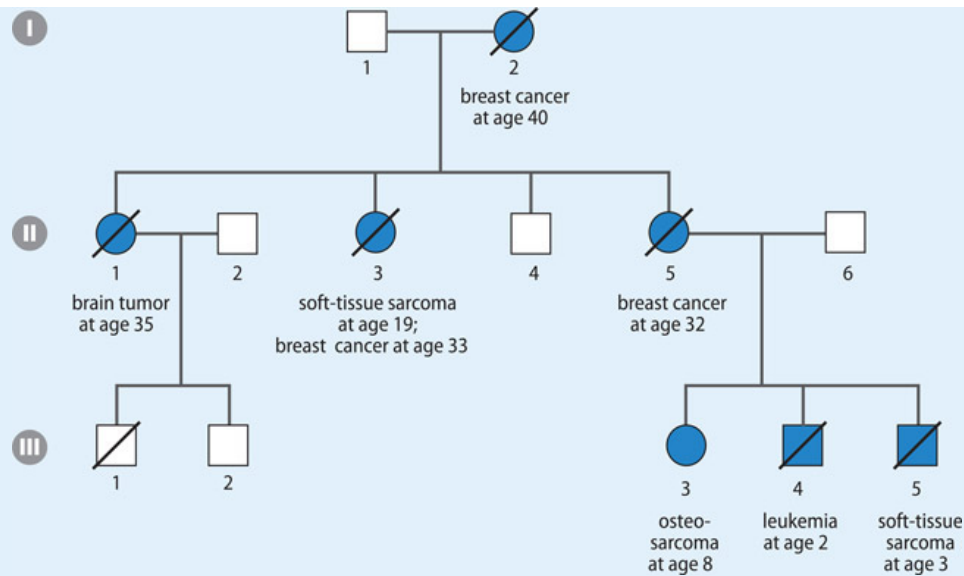


Figure 1 A typical pedigree of Li-Fraumeni syndrome. (Adapted from Malkin D [1994] *Annu Rev Genet* 28:443–465; PMID 7893135. With permission from Annual Reviews.)

p53 AS A NON-CLASSICAL TUMOR SUPPRESSOR

In several ways, p53 does not behave as tumor suppressor. In most tumor suppressors (such as *RB1*, *APC*, *NF1*, *NF2*, and *VHL*) the primary mutations are mostly deletion or nonsense that result in little or no expression of the respective proteins. *TP53* is different: the great majority of small-scale cancer-associated mutations are single nucleotide missense mutations that are very largely clustered within the central DNA-binding domain.

Six codons are predominantly mutated within the DNA-binding domain, and the missense mutations fall into two classes ([Figure 2](#)). In the DNA contact class, the missense mutation alters an amino acid that is normally used to make direct contact with the DNA of genes regulated by p53. The conformation class of mutations disrupt the structure of the p53 protein.

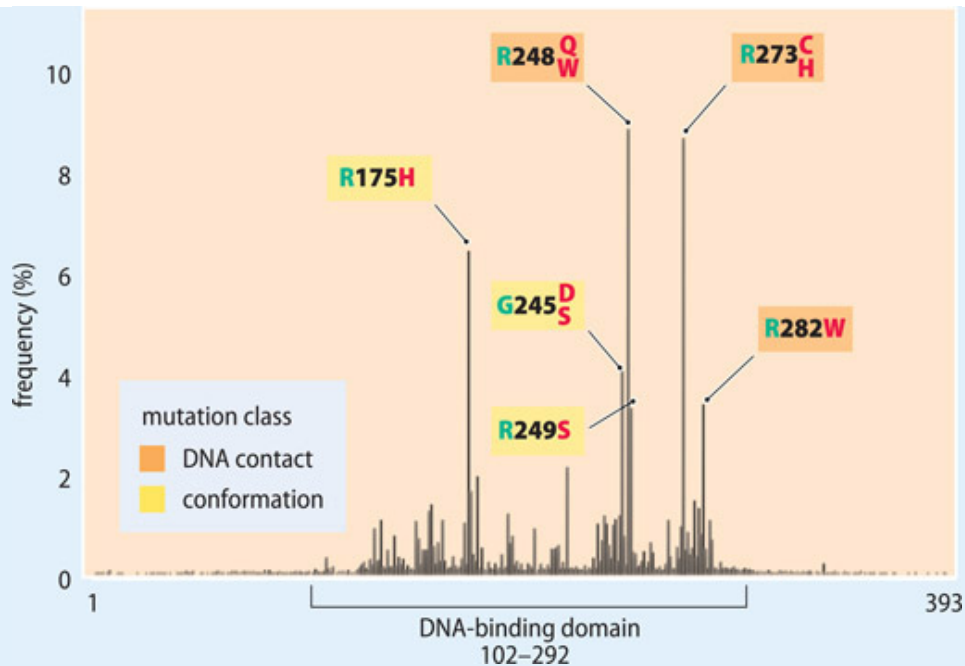


Figure 2 *TP53* missense mutations are very largely confined to the DNA-binding domain, with six hotspots. Vertical black lines indicate frequencies of missense mutations at each of the 393 codon positions. Two types of amino acid replacement are seen at codons 245, 248, and 273 (for example, at codon 273 arginine is replaced by cysteine or histidine). (Adapted from [Freed-Pastor WA & Prives C. \[2012\]](#) *Genes Dev* 26:1268–1286; PMID 22713868. With permission from Cold Spring Harbor Laboratory Press.)

The mutated p53 proteins have multiple properties that distinguish them from wild-type p53. First, unlike wild-type p53, mutant p53s do not participate in a self-limiting regulation. In normal cells, the amount of p53 is kept low because p53 is negatively regulated by MDM2 (and MDM4), and p53 positively regulates the production of its major antagonist MDM2; in cells with missense mutations in *TP53*, large amounts of mutant p53 are produced because mutant p53 fails to stimulate the production of MDM2. Mutant p53 can work in a dominant-negative fashion. It suppresses wild-type p53 and also the related p63 and p73 transcription factors (which show high sequence homology to p53 in some domains), and it antagonizes the interaction of wild-type p53 and the recognition sequences it must bind in its target genes (**Figure 3**). Instead, mutant p53 works as a rather different type of transcription factor by stimulating transcription of quite different target genes, including many genes that stimulate cellular proliferation or that inhibit apoptosis; see the review by [Freed-Pastor & Prives \(2012\)](#) under Further Reading.

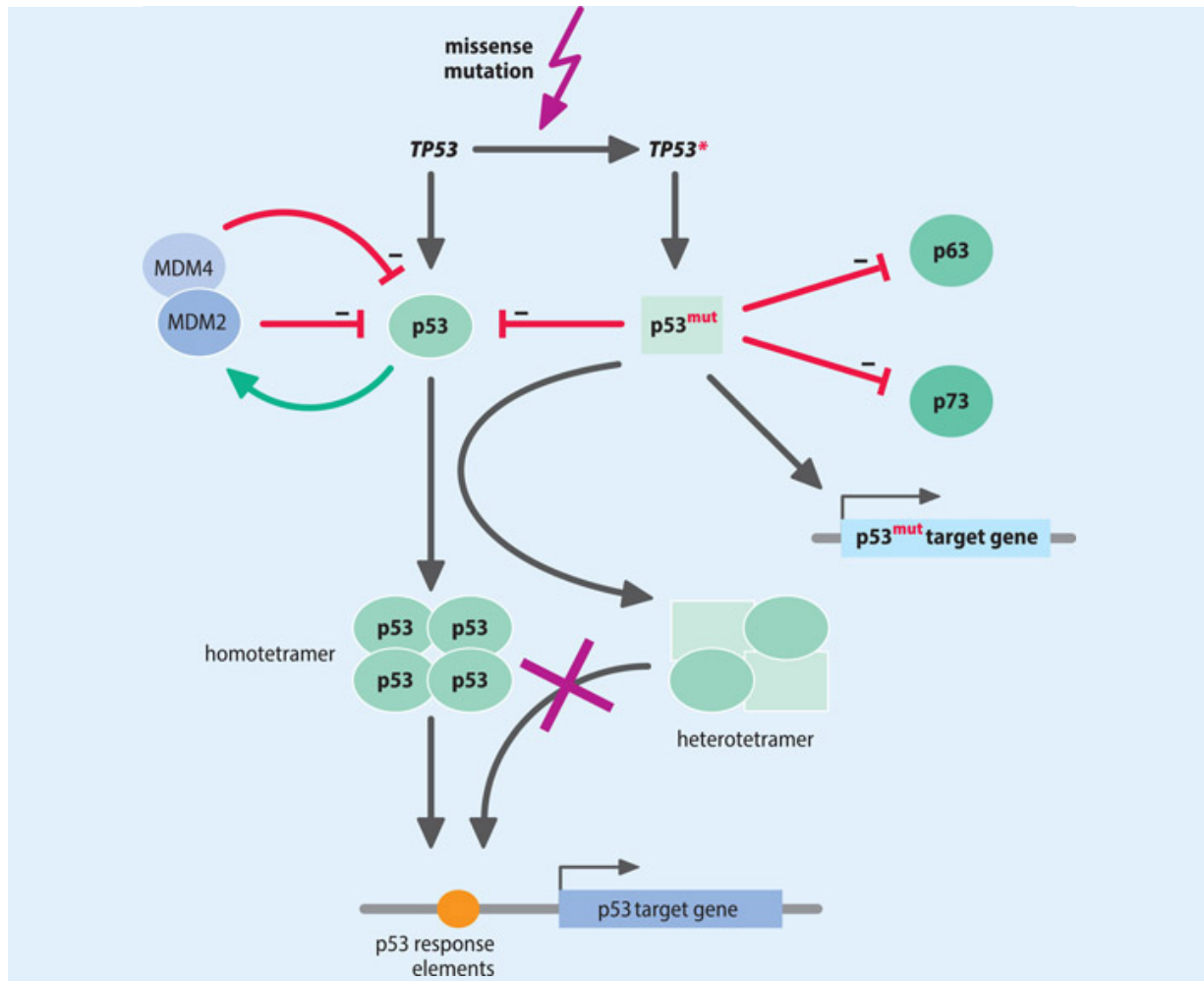


Figure 3 Missense p53 mutants have multiple novel properties and can show dominant-negative interactions with wild-type p53. Wild-type p53 works as a homotetramer to recognize and bind DNA sequences with specific motifs (p53 response elements) in the control regions of the p53 target genes. The p53 missense mutants suppress both wild-type p53 and the related p63 and p73 transcription factors. Mutant p53 is produced in very large amounts (unlike wild-type p53, it is not subject to self-regulation through stimulation of the MDM2 repressor) and interferes with normal p53-mediated transcription by interacting with wild-type p53 to form unproductive heterotetramers. Instead, mutant p53 stimulates the transcription of different genes.

The significance of miRNAs and long noncoding RNAs in cancer

Irrespective of the class of cancer-susceptibility gene, the normal products are almost always proteins. However, hundreds of different noncoding RNAs are also

known to be aberrantly expressed in cancer, and not surprisingly, given the widespread involvement of miRNAs in controlling gene expression, aberrant miRNA expression is very common in cancer.

Certain miRNAs are known to have important regulatory roles in processes relating to cancer, such as cell cycle control, cellular senescence, apoptosis, and DNA damage responses. Dysregulation of miRNA expression is frequent in cancer; certain miRNA genes can be lost in cancer cells and other miRNAs are known to be overexpressed in certain tumors.

On the basis of the above observations alone, various miRNAs have been viewed as behaving as tumor suppressors or oncogenes. For example, the *MIR15A* and *MIR16-1* genes at 13q14 have been regarded as tumor suppressors on the basis that they normally induce apoptosis by targeting BCL-2, but are frequently deleted or down-regulated in chronic lymphocytic leukemia. The *MIR21* gene, which regulates the *PTEN* and *PDCD4* genes and is over-expressed in many solid tumors, might be an example of a miRNA gene that can behave as an oncogene.

The dysregulated expression of miRNA and the loss of certain miRNA genes in cancers may be important events in cancer progression. However, it is difficult to evaluate the contributions of individual miRNAs: a single miRNA can regulate many different mRNAs, and a single mRNA may be regulated by many different miRNAs. At the time of writing, there seems little direct evidence that miRNA genes have unambiguous roles as oncogenes or tumor suppressor genes. Nevertheless, miRNA expression patterns may help us to dissect different disease subgroups, and there is interest in using miRNAs as therapeutic targets and as cancer biomarkers.

More recently, the possible roles of long noncoding RNAs in cancer have begun to be investigated. They have been less well studied (and, for example, have not been covered in whole exome sequencing studies), but there is strong evidence that some are important in cancer. In mice, for example, the *Xist* gene is not just involved in X-chromosome inactivation but also suppresses cancer *in vivo*; if *Xist* is deleted in blood cells, mutant females develop a highly aggressive myeloproliferative neoplasm and myelodysplastic syndrome with 100 % penetrance. Readers who may be interested in the role of noncoding RNAs in cancer can find a recent review at PMID 31730848.

10.3 GENOMIC INSTABILITY AND EPIGENETIC DYSREGULATION IN CANCER

The evolution of cancer cells is driven not just by a series of changes to the genome, but also by epigenetic changes. Natural selection at the cell level drives cells to relax normal controls on cell proliferation and apoptosis. Achieving this involves efforts to subvert both genome and epigenome stability. As we describe below, epigenetic changes may sometimes initiate the process of tumorigenesis.

An overview of genome and epigenome instability in cancer

Genome instability is an almost universal characteristic of cancer cells, and frequently results from defects in chromosome segregation or DNA repair. By weakening the capacity to maintain the integrity of the genome, more DNA changes will be generated for natural selection to work on to drive tumor formation. Eventually, a cell can build up a sufficient number of DNA changes to become an invasive cancer cell. Genomic instability can manifest itself at two levels:

- *at the chromosomal level.* Chromosomal instability (sometimes abbreviated as CIN) is a particularly common form of genome instability. Tumor cells typically have grossly abnormal karyotypes (extra or missing chromosomes and many structural rearrangements), and they often show chromosomal instability in culture
- *at the DNA level.* The instability may be genome-wide or localized. As detailed below, a genome-wide form of DNA instability is especially evident in some types of colon cancer. Sporadic colorectal tumors either show chromosome instability (in most cases) or global DNA instability (in about 15 % of cases), but not both: the instability seems to be the result of natural selection. More localized DNA instability is exemplified by the phenomenon known as *kataegis*, a form of clustered hypermutation first reported in 2012. We provide details within the context of cancer genomics in [Section 10.4](#).

Epigenetic dysregulation is a feature of all cancer cells, ranging from apparently normal precursor tissue to advanced metastatic disease. As well as being important in cancer progression (and helping cancer cells achieve each of the 10 characteristic biological capabilities listed in [Table 10.2](#) above), epigenetic dysregulation can be a key step in the initiation of cancer. And as we describe below, certain types of epigenetic dysregulation also cause chromosome instability and accelerated genetic changes in tumor cells.

The reader may justifiably wonder how cancer cells with their often bizarre chromosome constitutions, plus both DNA and epigenetic stability and inefficient anaerobic glycolysis, would be able to grow and proliferate as much as they do. But somehow these features are the outcome of the relentless drive by natural selection to enable cells to escape many layers of controls on cell division and programmed cell death.

Different types of chromosomal instability in cancer

Chromosome abnormalities are important in accelerating tumorigenesis: oncogenes can be activated by rearrangements such as translocations, and tumor suppressor alleles can be lost through deletions, whole chromosome loss, or recombinations. Standard cytogenetic methods are often difficult to carry out on tumor cells, but various DNA-based methods can be used to study chromosome instability in cancer, and they can have quite a high resolution. They include two microarray-based DNA hybridization methods—one based on comparative genome hybridization, and one on SNP (single nucleotide polymorphism) analyses—which we introduce within the general context of DNA-based diagnosis in [Section 11.1](#). Another method is *spectral karyotyping*, a type of multicolor chromosome FISH (fluorescence *in situ* hybridization). Unlike the microarray hybridization-based methods, it can reveal balanced chromosome abnormalities (in which there is no net loss or gain of DNA), as well as unbalanced chromosome abnormalities; see [Figure 10.14](#) for an application.

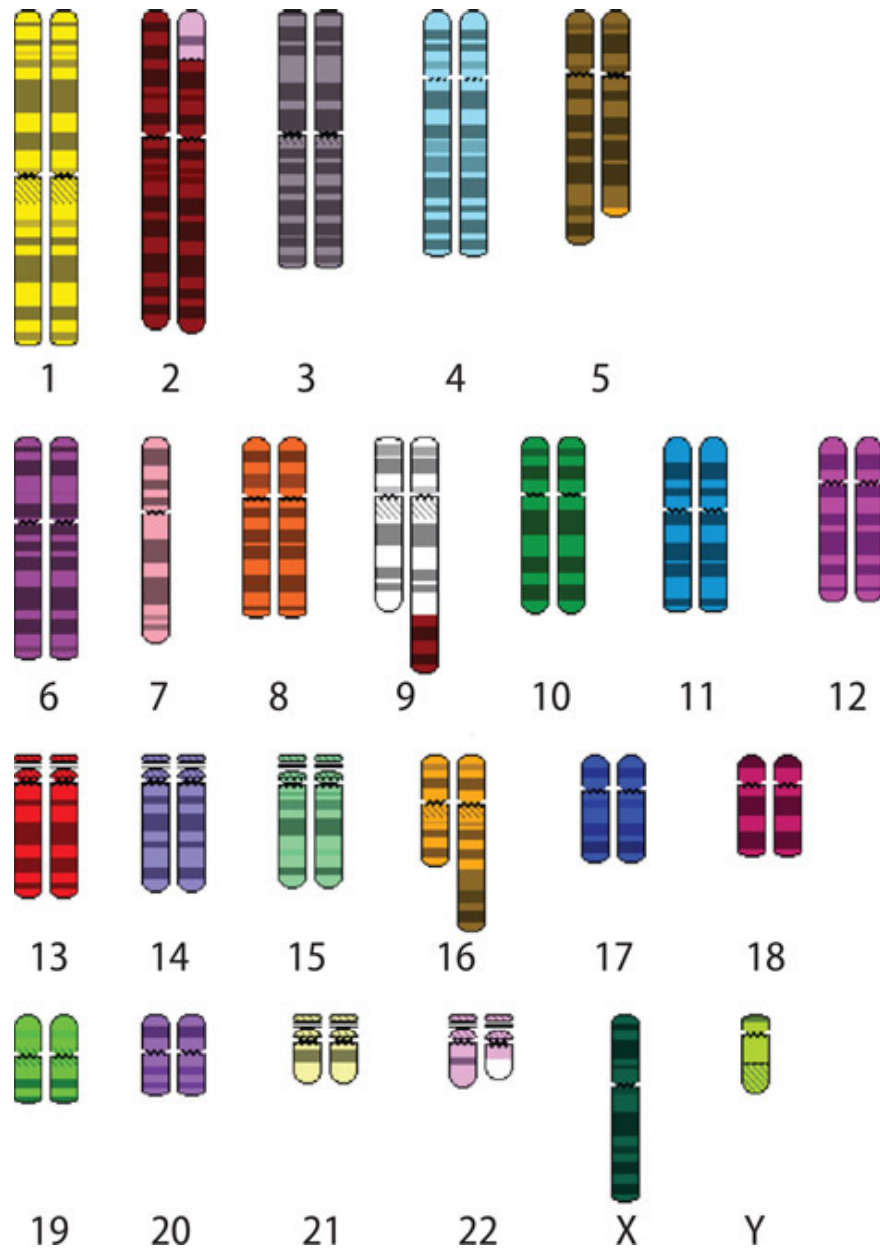


Figure 10.14 An example of using spectral karyotyping to analyze chromosomes in tumor cells. Spectral karyotyping (SKY) is a variant of chromosome fluorescence *in situ* hybridization (FISH) in which cocktails of many fluorescently labeled DNA probes from different regions of chromosomes are used to “paint” chromosomes so that entire chromosomes are labeled with a specific fluorochrome and become fluorescent. Different chromosomes are painted with different combinations of multiple fluorescent labels. An image analyzer scans the fluorescent signals and can discriminate between the different fluorescence signals used for each of the 24 different chromosomes. To help us visualize the result it assigns artificial (“false”) colors for each chromosome signal. In this example, there is a three-way variant of the standard 9;22 translocation (involving chromosome 2), plus an additional 5;16 translocation and the loss of

one copy of chromosome 7. The karyotype is interpreted as 45,XY,t(2;9;22)(p21;q34;q11),t(5;16)(q31;q24),-7.

A major source of aneuploidies in cancer cells is defects in the spindle checkpoint, the cell cycle control mechanism that checks for correct chromosome segregation (it normally ensures that the anaphase stage of mitosis cannot proceed until all chromosomes are properly attached to the spindle). Extra centrosomes are often seen in cancer cells and may trigger the formation of abnormal spindles and unequal segregation of chromosomes into the daughter cells.

Structural chromosome abnormalities in cancer cells can arise in different ways. The most common source is an abnormal response to unrepaired DNA damage. As detailed in [Section 4.2](#) we have complex DNA repair systems that can never be 100 % efficient. Normally DNA damage responses act as a backup: they trigger apoptosis if the DNA damage is severe, or they arrest the cell cycle so that an unrepaired defect can be repaired. Defects in DNA repair of DNA damage responses allow unrepaired or damaged DNA to be passed on to daughter cells.

Failure to repair double-strand DNA breaks is an important source of structural chromosome abnormalities and can be precipitated by inactivation of key caretaker genes that function in this repair pathway, including the breast-cancer-associated *BRCA1* and *BRCA2* genes, which work in homologous recombination-mediated DNA repair. Proteins such as the ATM (ataxia–telangiectasia mutated) protein kinase work as sensors to detect unprogrammed double-strand DNA breaks. They then activate signaling mediators, which in turn recruit effectors to repair the damage. As well as DNA repair, the DNA damage response involves arresting the cell cycle, notably by activating p53 ([Figure 10.15](#)).

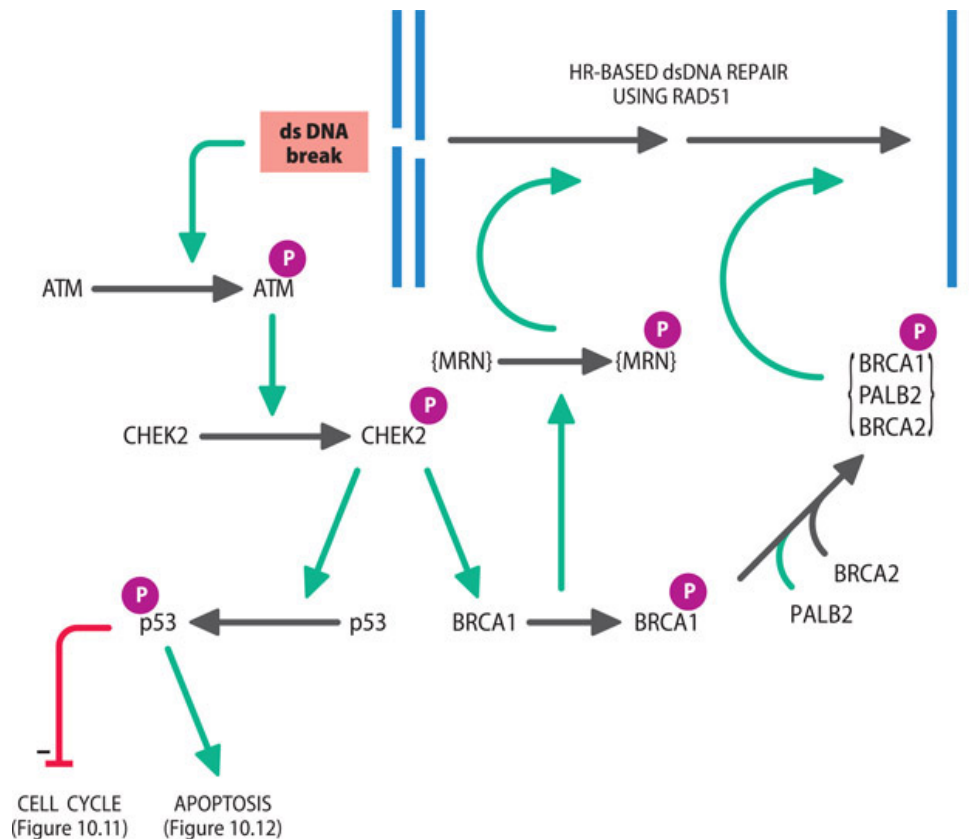


Figure 10.15 Some cellular signaling responses to double-strand DNA breaks and different roles of BRCA1 and BRCA2 in homologous recombination-based DNA repair. Homologous recombination (HR) appears to be the major mechanism for repairing double-strand breaks in proliferating cells. Green arrows indicate stimulatory reactions; the red T-bar indicates inhibition. The ATM protein kinase is a prominent sensor of DNA damage. It is activated by phosphorylation (P), and in turn causes phosphorylation-mediated activation of CHEK2, which similarly activates p53 and BRCA1. Phosphorylated BRCA1 has multiple roles including activating protein complexes that are directly involved in HR-mediated double-strand (ds) DNA repair. These complexes—shown here by curly brackets—include the MRE11–RAD50–NIBRIN (MRN) complex, and also a complex in which activated BRCA1 recruits BRCA2 through an intermediary binding protein PALB2. As well as activating DNA repair, the DNA damage response may initiate cell cycle arrest, notably by activating p53 (which works at the G₁–S checkpoint); if DNA damage cannot be readily repaired, it can also activate apoptosis through enhanced p53 production.

Chromothripsis and chromoplexy

Sometimes chromosome breakage can involve an extensive localized rearrangement of chromosomes. In the process of *chromothripsis* large numbers of chromosomal rearrangements are generated in what appear to be single catastrophic events. The chromosome rearrangements may occur by chromosome shattering and aberrant rejoining of fragments by error-prone end-joining DNA repair pathways, or by aberrant DNA replication-based mechanisms. Chromothripsis may not be common in many cancers, but it is significantly more frequent in cells with mutated p53. Interested readers can find a recent review at PMID 28899600.

Another, somewhat bizarre type of chromosome rearrangement, known as *chromoplexy*, can occur in tumors where chains of linked chromosomes form by serial translocations. An initial translocation might arise between chromosomes A and B, but unlike in conventional reciprocal translocation, the broken ends are not joined together to form hybrid A–B and B–A chromosomes. Instead, they may engage in new translocation with chromosomes C and D, which then generates another pair of broken ends that engage in further translocation with chromosomes E and F and so on. Interested readers can find a recent review at PMID 23680143.

Telomeres and chromosome stability

In human cells, the telomeres shorten at cell division (usually by about 30–120 nucleotides at each cell division). By inactivating normal controls on cell growth, cancer cells can reach a stage where some telomeres become so short that the cell can misinterpret the ends of seriously shortened telomeres as breaks in double-strand DNA. That alerts a DNA repair pathway that attempts a repair by fusing chromosomes at their ends. The resulting chromosomes with two centromeres may be pulled in opposite directions at mitosis, causing further broken ends and new cycles of chromosome fusion and breakage. Cancer cells seek to avoid this type of chromosome instability, and most frequently the solution involves ensuring that telomerase somehow becomes expressed (as previously detailed in [Box 10.1](#)).

Deficiency in mismatch repair results in unrepaired replication errors and global DNA instability

Mutation in genes involved in different types of DNA repair leads to cancer. Defective homologous recombination-based DNA repair is associated with various types of cancer, notably breast, ovarian, and pancreatic cancer. Genetic deficiency in components of nucleotide excision repair produces syndromes with increased cancer susceptibility, notably xeroderma pigmentosum (OMIM 278700). Genetic defects in base excision repair are associated with certain neurological disorders and occasionally cancer.

Germline mutations in both copies of the *MUTYH* gene (which is involved in the repair of adenines that are inappropriately base paired with guanine, oxoguanine, or cytosine) result in an autosomal recessive form of familial adenomatous polyposis (FAP), a type of hereditary colon cancer in which multiple polyps (adenomas) develop. Deficiency in **mismatch repair**, which corrects errors in replication that for some reason have not been detected by the proofreading activity of a DNA polymerase, results in a global form of DNA instability and is most commonly associated with colon cancer.

The mismatch repair mechanism

The mismatch repair (MMR) components work closely with the DNA replication machinery. In human cells, three types of protein dimer carry out most of the repairs ([Figure 10.16A](#)). Two of them—hMutSa and hMutSb—are needed to identify base mismatches. hMutSa identifies base-base mismatches but can also handle mismatching due to single-nucleotide insertions or deletions. hMutSb can spot base mismatching for different sizes of very short insertions or deletions (which frequently occur at short tandem repeats as a result of replication slippage, the tendency for DNA polymerase to stutter or skip forward at tandem repeats as shown in [Figure 4.6](#)).

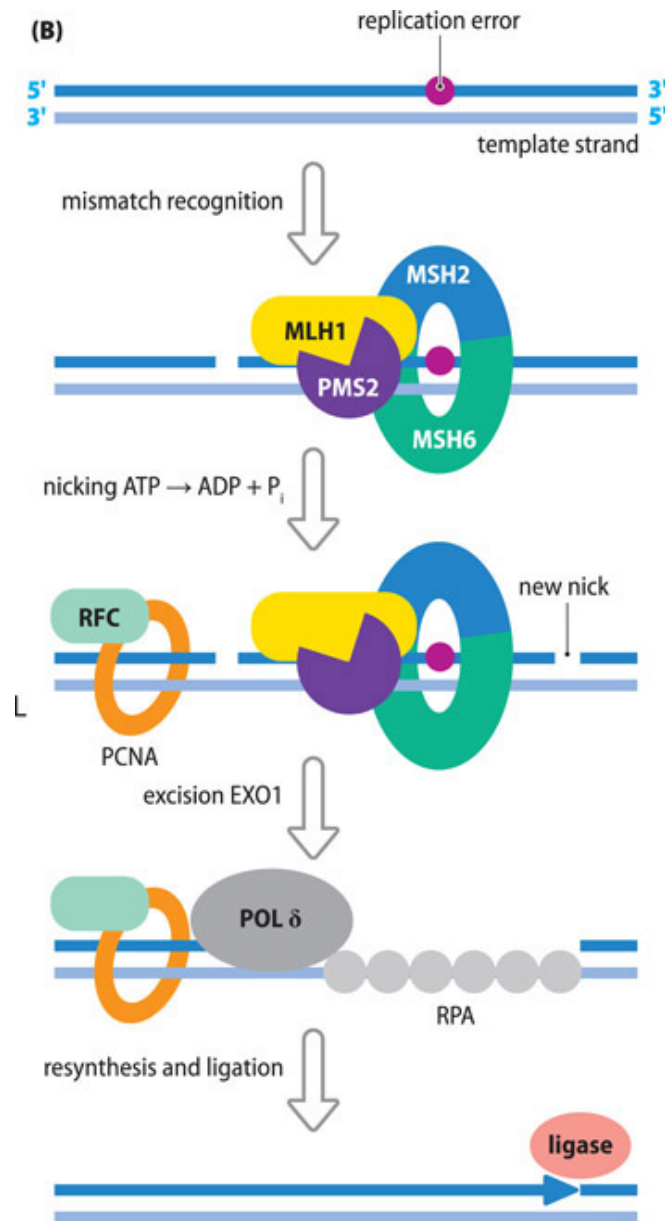


Figure 10.16 Mismatch repair for correcting replication errors. (A) Major classes of MutS or MutL dimers in human mismatch repair. (B) Mechanism of 5'-directed mismatch repair in eukaryotic cells. Replication errors on a newly synthesized strand result in base mismatches that can be recognized by a MutS–MutL complex. The MutS component works as a clamp that can slide along the DNA, allowing it to scan for a base–base mismatch (MutSa) or unpaired insertion/deletion loop (often MutSb). MutLa, which has an endonuclease function, can form a ternary complex with MutS and DNA. After the newly replicated DNA has been identified (by having a preexisting nick in the DNA), PCNA (proliferating cell nuclear antigen) and RFC (replication factor C) are loaded onto the newly replicated DNA, where they help trigger the endonuclease function of PMS2 to make a new nick close to the replication error. EXO1

exonuclease is recruited to excise the sequence containing the replication error, making a gapped DNA. The resulting stretch of single-stranded DNA (stabilized by binding the RPA protein) is used as a template for the resynthesis of the correct sequence using high-fidelity DNA polymerase δ , followed by sealing with DNA ligase I. (Adapted from Geng H & Hsieh P [2013] In *DNA Alterations in Lynch Syndrome: Advances in Molecular Diagnosis and Genetic Counseling* [M. Vogelsand, ed.]. With permission from Springer Science and Business Media.)

The MMR machinery cannot simply repair one of the two strands at random: there has to be a way of distinguishing the original (correct) strand from the newly replicated strand with the incorrect sequence that needs to be repaired. Before being repaired by DNA ligase, nicks (single-strand breaks) are common on a freshly replicated DNA strand, and in human (and eukaryotic) cells the strand distinction is achieved by identifying a nearby nick on the newly replicated DNA strand. Then hMutLa cleaves the newly replicated strand close to the mismatch and recruits an exonuclease to excise a short stretch of DNA containing the replication error so that the DNA can be resynthesized and repaired ([Figure 10.16B](#)).

Consequences of defective mismatch repair (MMR)

Loss of function for both alleles of a mismatch repair gene can result in a form of global DNA instability (in which replication errors in newly synthesized DNA go uncorrected). That may be apparent in some tumors and can readily be detected by testing for global *microsatellite instability*. To do this, a selection of standard microsatellite DNA markers from across the genome are tested in tumor DNA to see if they show higher frequencies than normal for minor additional bands. Tumors demonstrating **microsatellite instability** are described as being MSI-positive (or sometimes MIN-positive) tumors.

Lynch syndrome (also called hereditary nonpolyposis colon cancer; OMIM 120435) is a type of familial cancer in which an inactivating allele in a mismatch repair gene is transmitted by heterozygotes, predisposing to colorectal cancers and certain other cancers, including cancers of the endometrium. When investigating possible Lynch syndrome cases, if DNA analyses have not revealed an immediately obvious mutation, back-up immunohistochemistry analyses can be used as in the case study reported in [Clinical Box 13](#).

In cells in which the MMR machinery is defective, the mutation rate increases about 1000-fold and so generates large numbers of mutations to drive carcinogenesis. In coding sequences, inefficient repair of base mismatches and replication slippage errors can result in gene inactivation or mutant proteins: long runs of a single nucleotide are particularly vulnerable to frameshifting insertions or deletions, and nucleotide substitutions can result in nonsense or missense mutations.

Defective mismatch repair can occur occasionally in other types of tumor, but it is particularly associated with colon cancer. Why should that be? One explanation is that MMR deficiency sabotages a key defense system that protects against colorectal cell proliferation. In the colorectum, transforming growth factor β (TGF β) is a particularly strong inhibitor of cell proliferation, and it specifically binds to a receptor on the surface of the cells of which the TGFBR2 protein is a key component. However, the *TGFBR2* gene is readily inactivated as a result of mismatch repair deficiency because it has a long sequence of adenines that make it vulnerable to frameshifting insertions and deletions ([Figure 10.17](#)). Somatic mutations in *TGFBR2* are found in about 30 % of sporadic colorectal cancer but are very frequent in MSI-positive colorectal cancer.

121										130
TGC	ATT	ATG	AAG	GAA	AAA	AAA	AAG	CCT	GGT	
C	I	M	K	E	K	K	K	P	G	

Figure 10.17 A long homopolymeric region in TGFBR2-coding DNA is a weak spot in the defense. The nucleotide sequence of codons 121–130 within exon 3 of the TGFBR2 (transforming growth factor receptor β -2) gene is shown with predicted amino acids below. The sequence contains a perfect run of 10 adenines that is vulnerable to insertion or deletion by replication slippage ([Figure 4.6](#)), especially when cells become defective at mismatch repair. Resultant frameshift mutations lead to a failure to make the 567-residue TGFBR2 protein, with adverse consequences for TGF β signaling.

CLINICAL BOX 13 CASE STUDY: LYNCH SYNDROME

Margaret presented with endometrial adenocarcinoma at 46 years of age. She had been adopted and no wider family history was available. She developed right-sided colorectal adenocarcinoma at age 47 and died at an early age. No

blood sample was kept at that time, but tumor tissue sections were stored in the laboratory.

Margaret had had two children, Julia and Simon, and almost two decades after her mum died, Julia sought clinical advice in her mid-40s, concerned about the history of cancers in her mum. Analysis of archived tumor DNA samples from Margaret showed evidence of microsatellite instability suggesting a problem with mismatch repair.

A battery of tests was carried out on Julie's germline DNA sample. DNA sequencing was performed on all exons (plus flanking sequences) in the four mismatch repair genes, *MLH1*, *PMS2*, *MSH2*, and *MSH6*, and in the 3'-UTR of the *EPCAM* gene (which neighbors the 5' end of the *MSH2* gene). MLPA assays (multiplex ligation-dependent probe amplification; described in [Section 11.3](#)) checked for copy number changes in relevant exons. The sequencing results did not identify a likely pathogenic point mutation, and no exon deletions or duplications were recorded by MLPA.

Subsequently, after Julia had a total hysterectomy at 46 years of age, pathology results showed an early (incidental) endometrial adenocarcinoma. Simon had no clinical symptoms but the clear family history and their mum's colorectal adenocarcinoma and tumor micro-satellite instability raised the possibility of Lynch syndrome, despite the negative DNA sequencing and MLPA results. Immunohistochemistry studies were therefore carried out on tumor samples from Julia and confirmed Lynch syndrome (see [Figure 1](#)).

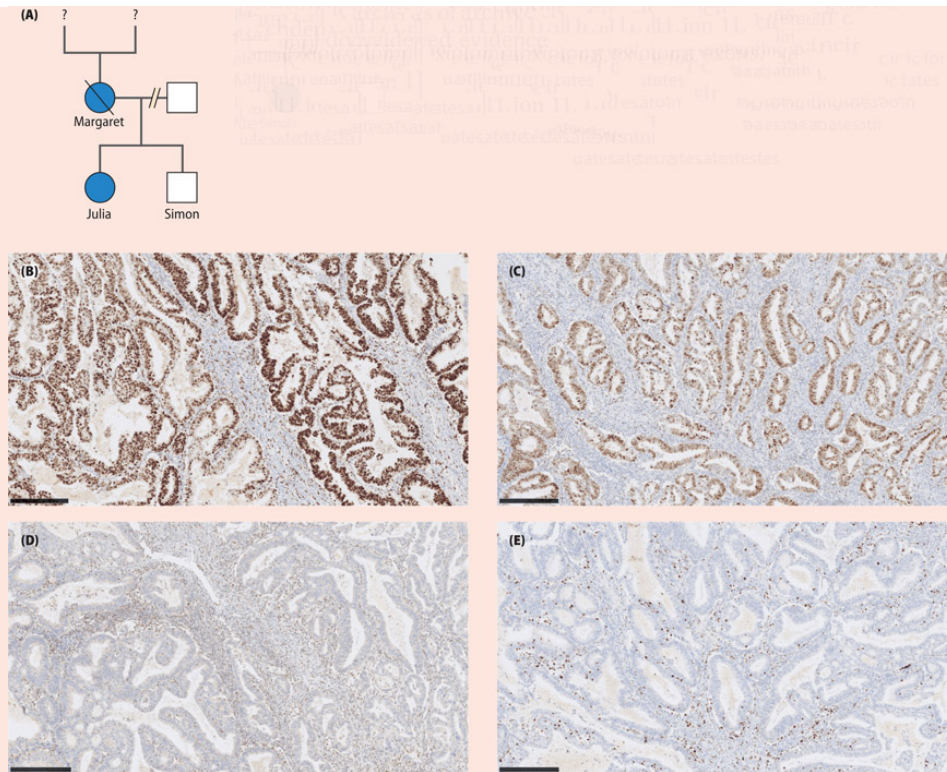


Figure 1 Family pedigree (A) and immunohistochemistry analyses on samples from **Julia's tumor (B-E)**. Question marks indicate lack of information about Margaret's antecedents. Immunohistochemistry was performed with monoclonal antibodies specific for MLH1(B), PSM2 (C), MSH2 (D) and MSH6 (E). The strong brown staining in (B) and (C) indicates the presence of MLH1 and PMS2; the lack of brown staining in (D) and (E) shows that both MSH2 and MSH6 are not expressed.

As shown in [Figure 10.16](#), the human MSH2 and MSH6 proteins normally work as a heterodimer (hMutSa) in mismatch repair. The two proteins are not, however, equal partners: an inactivating mutation in the *MSH2* gene that prevents production of the MSH2 protein also suppresses production of the MSH6 protein, but by contrast the MSH2 protein continues to be made after the *MSH6* gene is inactivated. Julia appears to have inherited from her mum an unidentified inactivating mutation in the *MSH2* gene, most likely a point mutation in a regulatory sequence within an intron or nearby noncoding DNA sequence. As a result, Julia was advised a clinical follow-up program with ongoing two-yearly colonoscopies.

Different classes of cancer susceptibility gene according to epigenetic function, epigenetic dysregulation, and epigenome–genome interaction

Recall that in somatic cells much of the genome is transcriptionally silenced (the heterochromatic regions and a significant, but variable, fraction of the euchromatin). This is achieved by epigenetic modifications—DNA methylation, histone modifications, and nucleosome repositioning—that attract specific proteins to compact the DNA and deny access to the transcription machinery.

The epigenetic modifications ensure that cells have distinctive chromatin patterns. They allow specific gene expression patterns to be established that determine the identity of a cell (so that it behaves as a T cell or a cardiomyocyte, for example). And they help to maintain genome stability (by maintaining the stability and function of centromeres and telomeres, and by suppressing excess activity by transposons).

One rationale for epigenetic dysregulation in tumors is that it allows cancer cells to revert to less differentiated states, permitting more flexibility to adapt to changing environments, and to assist the transformation required for the progression to cancer. It was initially thought to result simply from genetic changes in genes controlling epigenetic regulation but more recently, as detailed below, it has become clear that epigenetic changes can also initiate cancer formation.

Classifying cancer genes by epigenetic function

Previously we have classified cancer genes at two levels: by the dominant or recessive effect of genetic mutations on the phenotypes of cells (oncogenes versus tumor suppressor genes) and by selection (*driver genes*—where mutation or aberrant expression is subject to selection during tumorigenesis—and *passenger genes* that are not subject to selection towards advancing tumorigenesis). A third way of classifying cancer genes, proposed by Andrew Feinberg in [2016](#) (see Further Reading), is on the basis of epigenetic function, where three categories have been envisaged, as listed below.

- *Epigenetic modulator.* A gene, mutated or not, that activates or represses the epigenetic machinery in cancer. Examples include the *IDH1* and *IDH2* genes that work in the TCA (Krebs) cycle metabolism as isocitrate dehydrogenases, as detailed later. Others include the *CTCF* gene that makes a protein involved in regulating chromatin architecture and transcription, and genes making various types of cell signaling components, including the *KRAS*, *APC*, *TP53*, and *YAP1* genes.
- *Epigenetic modifier.* A gene, mutated or not, that modifies DNA methylation or chromatin structure or its interpretation in cancer. Examples include *DNMT3A* (DNA methylation), *SMARCA4* (chromatin remodeling), and *EZH2* (makes the enzymatic component of the Polycomb Repressive Complex 2 responsible for epigenetic maintenance of genes regulating development and differentiation).
- *Epigenetic mediator.* A usually unmutated gene that is regulated by an epigenetic modifier in cancer, and that increases pluripotency or survival. Examples include classic genes associated with pluripotency, such as *OCT4*, *NANOG*, *SOX2*, *LIN28*, and *KLF4*.

DNA methylation profiles of cancer cells and their effects on gene expression

Epigenetic profiling of cancer cells has been limited because histone modification profiles are difficult to obtain from solid tumors; for that reason, much of our information on the epigenetic profiles of tumors has come from studies on DNA methylation. In human cells, DNA methylation is almost exclusively restricted to certain cytosines that have a neighboring guanine within the dinucleotide CG (or CpG, as it is often called). Methylated cytosines can be distinguished from unmethylated cytosines by treating DNA with sodium bisulfite. (Sodium bisulfite changes all unmethylated cytosines to uracils, which become thymines in replicated DNA; methylated cytosines do not react and are unchanged. We describe the method in detail in [Figure 11.15](#) on page 449.)

In somatic mammalian cells, about 70–80 % of the cytosines present in CG dinucleotides are present as 5-methylcytosine. The 5-meCG sequences are recognized and bound by specific proteins that are important in helping to organize the chromatin into compact formations that lead to transcriptional

silencing. In cancer, however, the DNA methylation patterns are changed in two ways: extensive hypomethylation and selective hypermethylation.

Across the genome as a whole, cancer cells typically show significantly reduced DNA methylation (*hypomethylation*). That includes very many genes; long blocks of sequences, enriched in repetitive DNA but containing about one-third of transcriptional start sites, are hypomethylated.

Loss of methylation in constitutively heterochromatic regions may produce aberrant transcriptional expression of highly repetitive DNA sequences, resulting in widespread chromosomal instability. That seems to be a very common event in early adenomas, for example, occurring shortly after the disturbance to the Wnt signalling pathway (mutations in *APC* or equivalent) shown in [Figure 10.4B](#) above. There is uncertainty about how this happens. One hypothesis is that the demethylation of highly repetitive DNA sequences allows normally silenced retrovirus-like elements and other related transposable elements in the genome to become active and jump to new locations in the genome, creating havoc. Constitutional DNA hypomethylation and chromosomal instability are also features of some human disorders, such as ICF1 (immunodeficiency with centromeric instability and facial anomalies; OMIM 242860), an autosomal recessive disorder that often results from mutations in the *DNMT3B* DNA methyltransferase gene.

DNA *hypermethylation* commonly occurs at the promoters of a few hundred genes in cancer cells, including tumor suppressor genes, DNA repair genes, and genes encoding certain transcription factors that are important in differentiation. Some tumor suppressor genes, such as *CDKN2A* and *MGMT*, are frequently silenced in a wide range of tumors; for others, silencing is limited to certain types of cancer: *VHL* in renal cancer, for example, and *BRCA1* in breast and ovarian cancer. A more extensive form of DNA hypermethylation occurs in some cases for certain cancers; we discuss this later in the chapter, in the context of links between metabolic and epigenetic dysregulation in cancer.

Genome–epigenome interactions and epigenetic initiation of tumorigenesis

Genetic and epigenetic alterations in cancer used to be regarded as separate mechanisms. Now we know that they work closely together, complementing each

other towards promoting cancer development. As a result, cancer cells can attain greater *plasticity*; by accelerating the normal rate of genetic changes, and promoting dedifferentiation, they can escape normal cellular controls.

Table 10.6 lists some major genome–epigenome interactions in cancer. In addition to genomic changes causing changes to the epigenome, epigenetic changes can provoke major changes to the genome, including increased chromosome instability and silencing of tumor suppressor genes. Because epigenetic changes can be environmentally induced, it is conceivable that epigenetic changes may sometimes initiate tumorigenesis. Take the example of chronic inflammation that can arise out of certain diseases or certain bacterial infections. There is a tight association between chronic inflammatory bowel diseases such as ulcerative colitis and Crohn’s disease and the subsequent development of colorectal cancer. And chronic inflammation arising from infection with the bacterium *H. pylori* is also the biggest risk factor for developing gastric cancer. Signaling molecules in inflammatory-associated signaling pathways such as the NF-κB pathway can induce epigenetic changes involved in tumorigenesis.

TABLE 10.6 EXAMPLES OF GENOME-EPIGENOME INTERACTIONS IN CANCER				
Genome			Epigenome	
Change	Mutation in genes encoding an epigenetic modulator	⇒	Epigenetic dysregulation	Effect
	Mutation in genes encoding an epigenetic modifier [*]	⇒	Epigenetic dysregulation	
Effect	C → T substitutions	⇐	Deamination of 5-meC ^{**}	Change
	Chromosome instability	⇐	Hypomethylation of highly repetitive DNA	
	Silencing of tumor-suppressor genes	⇐	Induced epigenetic changes ^{***}	

^{*} See [Figure 10.21](#) for the example of mutation of isocitrate genes that results in dedifferentiation.

^{**} By one estimate, over 60% of the point mutations in the genomes of tumors of internal organs (in tissues that are shielded from UV radiation) arise in CpG sequences.

^{***} As a result of genetic changes (top two rows) or non-genetic changes, such as altered cell signaling in inflammation, and so on.

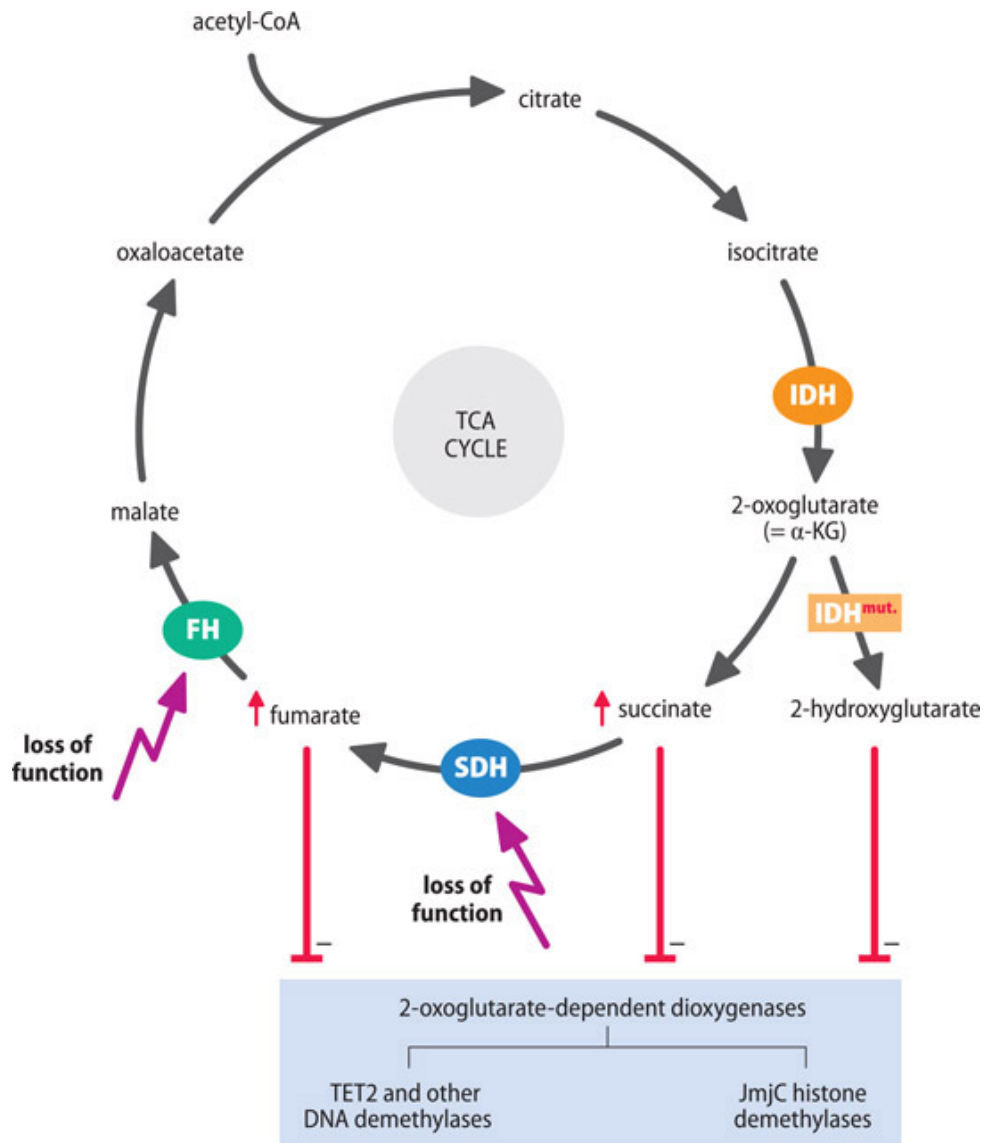


Figure 10.21 Mutation of certain genes encoding enzymes of the tricarboxylic acid (TCA) cycle can cause epigenetic modifications that contribute to cancer. Normal alleles of the *IDH1* and *IDH2* genes produce an isocitrate dehydrogenase enzyme that converts isocitrate to 2-oxoglutarate (also called α-ketoglutarate; α-KG). In certain cancers, certain missense mutations in *IDH1* (R132H, R132C) or in *IDH2* (R140Q, R172K) result in a mutant isocitrate dehydrogenase (IDH^{mut.}) that can convert the 2-oxoglutarate made by a normal IDH allele to 2-hydroxyglutarate. This abnormal oncometabolite changes the epigenetic profile of the cell, reversing differentiation to make the cell more like a stem cell. It does that by inhibiting multiple enzymes that depend on using 2-oxoglutarate as a cofactor, including some DNA demethylases such as TET2 and certain histone demethylases of the JumanjiC (JmjC) class. High levels of succinate and fumarate can also inhibit the 2-oxoglutarate-dependent enzymes, as when two

loss-of-function alleles result in a genetic deficiency of fumarate dehydrogenase (FH) or succinate dehydrogenase (SDH), causing a buildup of substrate (red arrows).

10.4 NEW INSIGHTS FROM GENOME-WIDE STUDIES OF CANCERS

Until quite recently, molecular genetic studies of cancer cells had focused on individual genes of interest. Databases were established to store information on DNA changes associated with cancer-associated changes in important cancer-susceptibility genes, such as the International Agency for Research on Cancer's *TP53* database (<http://p53.iarc.fr/>).

Once the sequence of the human (euchromatic) genome had been obtained, the age of *cancer genomics* could begin. Different genome-wide screens were devised to get comprehensive data from cancer cells, beginning with microarray studies that reported the relative abundance of transcripts from thousands of different human genes. To seek out novel cancer-susceptibility genes, whole genome association studies of the kind described in [Section 8.2](#) have been used, and have been useful (interested readers can find an example at PMID 32424353), but whole exome and whole genome sequencing have been especially fruitful. More recently, high-resolution genome-wide DNA methylation screens have been carried out for some types of cancer, and multiple single-cell genomics and transcriptomic studies have been carried out.

After the launch of the Cancer Genome Project in the UK in 2000, and The Cancer Gene Atlas (TCGA) in the USA in 2006, the International Cancer Genome Consortium (ICGC) was created in 2007 to coordinate efforts on a global scale. The burgeoning data coming out of the cancer genome projects are stored in dedicated databases and can be navigated with dedicated Web browsers ([Table 10.7](#)). In addition to transforming our understanding of cancer—we describe below some examples of new insights that have emerged—the new data will also have important consequences for cancer diagnosis and treatment.

TABLE 10.7 EXAMPLES OF DATABASES, WEB BROWSERS, AND NETWORKS IN CANCER GENOMICS

Electronic resource	Description	Website URL
---------------------	-------------	-------------

Electronic resource	Description	Website URL
COSMIC* database	stores and displays somatic mutation information and related details and contains information relating to human cancers; includes a census of human cancer genes at http://cancer.sanger.ac.uk/census/	http://cancer.sanger.ac.uk/cosmic/
International Cancer Genome Consortium (ICGC)	applications can be made to access controlled data from the ICGH whose goal is to get a comprehensive description of genomic, transcriptomic, and epigenomic changes in multiple different tumor types and/or subtypes that are of clinical and societal importance across the globe	https://daco.icgc.org
The Cancer Genome Atlas (TCGA)	cancer genomics network of research centers in the US	http://cancergenome.nih.gov

* [Catalog of Somatic Mutations in Cancer](#)

Genome sequencing has revealed extraordinary mutational diversity in tumors and insights into cancer evolution

Massively parallel DNA sequencing (also called *next-generation sequencing*) is transforming genetics because it offers a huge step up in DNA sequencing output. Because of the extraordinary complexity of cancer evolution it has been extensively applied to sequencing cancer genomes.

Initially, exome sequencing was used to analyze the DNA of cancer cells (the great majority of cancer genes make proteins, and many of the mutations occur within exons). Exome sequencing cannot readily detect copy number variation,

and so cancer exome sequencing projects have been supplemented by genome-wide screens for copy number variation. More recently, sequencing of whole (euchromatic) cancer genomes has also been carried out to reveal all classes of change in somatic DNA in a tumor (when referenced against a corresponding normal tissue genome from the individual).

The first whole cancer genome to be sequenced—an acute myeloid leukemia genome—was reported in 2008. Since then, large numbers of whole cancer genome sequences have been determined. Using cloud computing, the Pan-Cancer Analysis of Whole Genomes (PCAWG) Consortium has facilitated international data sharing. In 2020 it reported the analysis of 2658 whole-cancer genomes and matching normal tissues across 38 tumor types, representing most common cancers.

The extraordinary volume of data pouring out of cancer genome sequencing is collated in different databases, notably the COSMIC database (see [Table 10.7](#)). Working out what the huge amount of sequence data means is inevitably a challenge, but already some valuable insights have been revealed.

Mutation number

How many mutations are there in a cancer? There are certainly more than we used to think. From multiple sequenced cancer genomes we now know that adult cancers often have between 1000 and 10 000 somatic substitutions across the genome. However, some types of cancer—medulloblastomas, testicular germ cell tumors, and acute leukemias, for example—have relatively few mutations; others, such as lung cancers and melanomas, have many more mutations (sometimes more than 100 000). If we focus on just the coding sequence (1.2 % of the genome) and consider only the nonsynonymous mutations (which, by changing an amino acid, are more likely to have an effect on cell function than, say, synonymous mutations), the number of nonsynonymous mutations per tumor continues to show a clear dependence on the type of tumor ([Figure 10.18](#)).

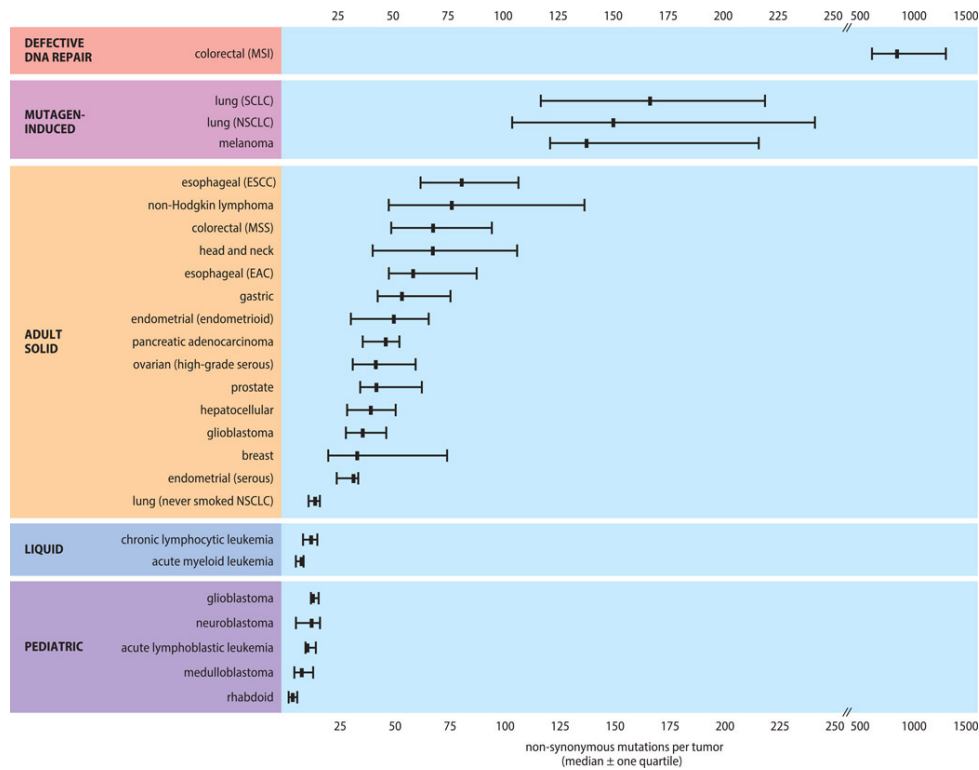


Figure 10.18 Variation in the number of somatic nonsynonymous mutations per tumor in representative human cancers. The median number of nonsynonymous mutations per tumor is estimated from genome-wide sequencing in tumors. Horizontal bars indicate the 25% and 75% quartiles. MSI, microsatellite instability; SCLC, small-cell lung cancers; NSCLC, non-small-cell lung cancers; ESCC, esophageal squamous cell carcinomas; MSS, microsatellite stable; EAC, esophageal adenocarcinomas. (Data from [Vogelstein B et al. \[2013\] Science 339:1546–1558; PMID 23539594.](#))

How can we explain the differences in mutation number? In part this is because different cancers can vary in the number of cell divisions separating the fertilized egg and the cancer cell. And differences in mutation rate at the cell divisions from the fertilized egg cell to the cancer cell must be a factor. Tumors in children or young adults might have lower mutation prevalence simply because the cancer cell has been through comparatively few mitoses. The high mutation prevalence in lung cancers and melanomas most probably reflects an exceptionally high exposure or vulnerability to specific mutagens (tobacco carcinogens and UV radiation, respectively). Inevitably, perhaps, the highest mutation frequencies are found in cancers having a mutation that causes a defect in mismatch repair; see [Figure 10.18](#)).

Mutational processes and cancer evolution

Cancer genome and exome sequencing has permitted comprehensive studies of the mutational processes involved in the evolution of a cancer. In 2012 a series of papers provided the first comprehensive dissection of breast cancer. To the cancer surgeon, one of the most striking things about breast cancer is how it progresses differently in each patient, and how each patient responds differently to therapy. This is where molecular genetics might make a difference by helping identify subclasses of tumors with distinct properties that allow different treatment options to be applied, depending on the tumor subtype. The breast cancer studies revealed extraordinary mutational diversity, with multiple independent *mutational signatures*; they also indicated that in most such cancers more than one mutational process has been operative.

Specific mutational signatures found in some cancers simply reflect excessive exposure to specific environmental mutagens that preferentially cause particular types of mutation (for example, UV radiation causes preferential C:G → T:A transitions in melanoma, and tobacco carcinogens cause preferential C:G → A:T transversions in lung cancer). Splicing mutations are particularly common in some types of cancer, notably myelodysplastic syndrome (MDS) and chronic lymphocytic leukemia (CLL). This happens because in these cancers some genes that make components of the RNA splicing machinery (such as *U2AF1* in MDS, and *SF3B1* in both MDS and CLL) are frequently mutated.

The comprehensive studies of breast cancer were the first to illustrate just how complex the mutational processes are. One novel mutation process initially discovered from sequencing breast cancer genomes is a type of hypermutation called *kataegis* (from the Greek word for thunderstorm). If a breast cancer genome shows, say, 10 000 tumor-specific mutations, one might expect that the mutations would be mostly randomly distributed across the genome. In that case the average density of the mutations would be 10 000 per 3 Gb or roughly 1 mutation in every 300 kb of DNA. But sometimes highly clustered mutations of the same type are seen, such as C → T mutations ([Figure 10.19](#)). This type of hypermutation appears to result from excess activity of cellular APOBEC proteins (which naturally act as cytidine deaminases in processes such as antibody diversification and RNA editing). By promoting excess activity by these enzymes, tumors find

yet another way of generating multiple mutations that natural selection can work on to promote cancer development.

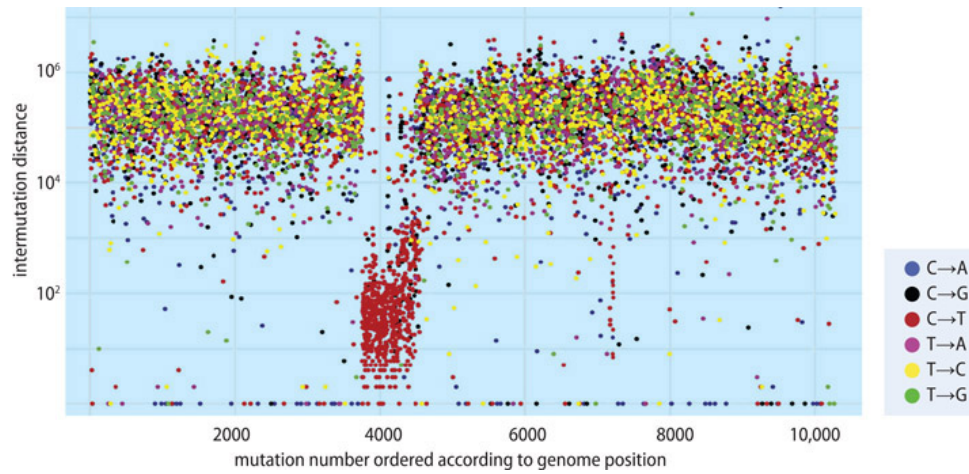


Figure 10.19 A rainfall plot showing an example of kataegis, a form of clustered hypermutation, in breast cancer. Cancer-specific mutations (a total of more than 10 000 from across the genome in this case) are ordered on the horizontal axis according to their position in the genome, starting from the first variant on the short arm of chromosome 1 (at the extreme left) to the last variant on the long arm of chromosome X (position number about 10 500) and are colored according to mutation type (see the color key at the right). The vertical axis shows the distance between each mutation and the one before it (the intermutation distance), plotted on a logarithmic scale. Most mutations in this genome have an intermutation distance of about 10^5 bp to about 10^6 bp, but the plot clearly shows a major region of hypermutation roughly centered on mutation position 4000 (corresponding to a 14 Mb region on the long arm of chromosome 6), where there is an extraordinary clustering of C \rightarrow T mutations (red dots) that are spaced from their nearest neighbors by very short distances (often 100 bp or less). Within this region there are defined very short regions of intense C \rightarrow T mutation clustering as shown in Figure 4 of the original article. (From Nik-Zainal S et al. [2012] *Cell* 149:979–993; PMID 22608084. With permission from Elsevier.)

Intertumor and intratumor heterogeneity

Genome sequencing has provided the first full understanding of mutational differences between different tumors of the same type, and of mutational differences within tumors. Most differences are due to passenger mutations. When coding sequences were scanned, driver mutations were found to be common in

key cancer-susceptibility genes. For example, mutations in *APC*, *TP53*, and *KRAS* were found to be frequent in colorectal tumors, and the *BRAF* gene was implicated in more than 60 % of melanoma tumors.

The first comprehensive insights into intratumor heterogeneity came from a study of renal cancer by [Gerlinger et al. in 2012](#) (PMID 22397650) in which exome sequencing analyses were carried out on multiple biopsies taken from a primary renal carcinoma and associated metastases from a single individual. Only about one-third of the 128 somatic mutations detected by exome sequencing in the different biopsies from this individual were present in all regions and only one driver gene—*VHL*, the von Hippel–Lindau tumor suppressor gene—was mutated in all analyzed regions. Another driver gene, *SETD2* (which encodes the histone H3K36 methyl-transferase), showed three distinct mutations associated with different regions, and selection pressure was inferred to have found three different ways of inactivating the *SETD2* gene to produce similar tumor phenotypes.

In the above study, in addition to the differences in mutation profiles between biopsies from tumors from different regions, even a single biopsy appeared to consist of two different clonal populations. The cells that seeded metastases seem to have diverged at an early stage from those that formed the primary tumor; the two groups had a differentiating series of mutations, and the cells that would form metastases lacked a mutation in a driver gene, *MTOR* (which encodes the mammalian target of rapamycin kinase).

Defining the landscape of driver mutations in cancer and establishing a complete inventory of cancer-susceptibility genes

As described below, proteins made by known cancer genes have become targets for successful anti-cancer drug development. Identifying new cancer-susceptibility genes has therefore been a key goal of cancer genome studies. Until quite recently, most known cancer genes had been identified by three approaches: analyses of associated chromosome abnormalities (notably translocations) in which breakpoints could be identified by FISH; candidate gene studies (using information from experimental model organisms); and studies of copy number variation (by detecting oncogene amplification or by scanning for loss of heterozygosity). Linkage analyses had also been important in defining some genes that underlie inherited cancers, such as the *BRCA1* and *BRCA2* genes that are important susceptibility genes in breast and ovarian cancer.

To identify novel cancer-susceptibility genes, genomewide association (GWA) studies were initially employed, but they have had limited success. Instead, genome or exome sequencing of multiple tumors has become the preferred approach: the sequences are referenced against normal somatic cells from the relevant individuals to identify tumor-specific mutations.

How can genome-wide sets of tumor-specific mutations allow us to identify cancer-susceptibility genes? The driver mutations for any type of cancer might be expected to be confined to a comparatively small set of key cancer genes that are frequently mutated in that type of cancer (and which are presumably crucial to its evolution). Passenger mutations, by contrast, might be expected to be rather randomly distributed across the genome and to be somewhat different in unrelated tumors of the same cancer type. By looking at multiple tumors of the same type, one might expect to quickly discriminate between driver and passenger mutations. (However, it is not always so straightforward: a cluster of somatic mutations may also be attributable to an increased local mutation rate; in that case, passenger mutations may initially be confused with driver mutations.)

Cancer gene and driver mutation distribution

Genome-wide sequencing approaches have been enormously successful, not just in identifying novel cancer-susceptibility genes but also in defining the distribution of the cancer-susceptibility genes in different cancers and the profile of associated driver mutations. In a study of 100 breast cancer tumors, for example, a total of 250 driver mutations were found; the number of driver mutations per tumor ranged up to a maximum of six, with an average of 2.5 ([Figure 10.20](#)).

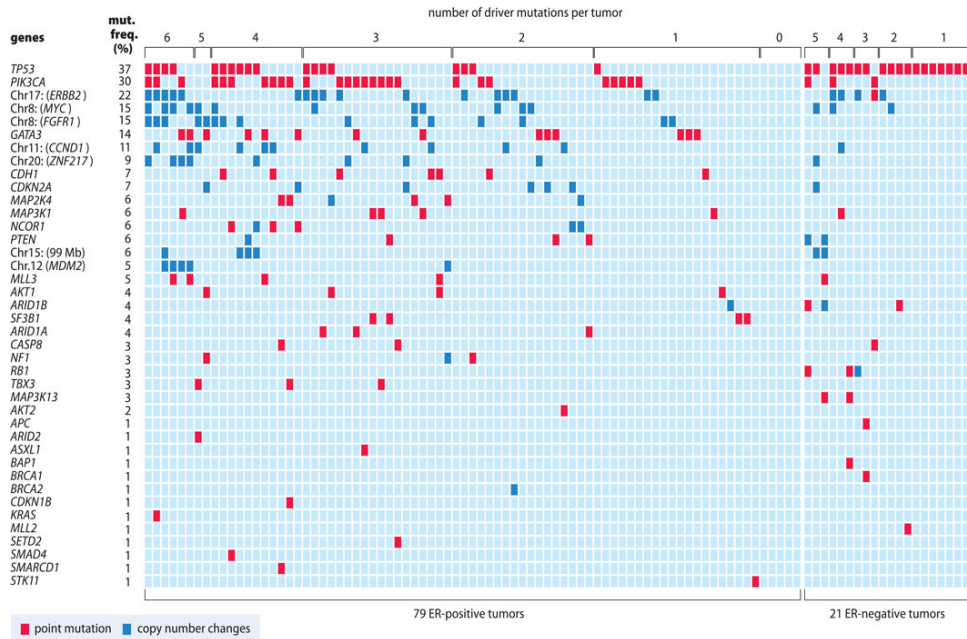


Figure 10.20 The landscape of driver mutations in a study of 100 primary breast cancers.

Of the 100 cancers, 79 expressed estrogen receptor (ER-positive), and 21 were ER-negative. By referencing against control DNA samples from normal cells, somatic mutations were identified in 40 cancer-susceptibility genes (shown in the left-hand column). Point mutations were identified by whole exome sequencing. Changes in copy number (shown in blue) include amplification of oncogenes and loss of alleles in tumor suppressors; they were identified by hybridization to whole genome SNP (single nucleotide polymorphism) arrays (described at http://www.sanger.ac.uk/genetics/CGP/CopyNumberMapping/Affy_SNP6.shtml). At least one of these genes or loci was mutated in all of the tumors, except in five ER-positive tumors. A maximum of six of the genes were mutated in any one tumor. The most significant (frequent) cancer genes were found to be *TP53* (mutated in 37 % of all tumors, and in close to 90 % of ER-negative tumors) and *PIK3CA* (mutated in 30 % of the tumors). Mut. Freq., mutation frequency. (Adapted from Stephens PJ et al. [2012] *Nature* 486:400–404; PMID 22722201. With permission from Macmillan Publishers Ltd.)

In the above study, seven genes—*TP53*, *PIK3CA*, *ERBB2*, *MYC*, *FGFR1/ZNF703*, *GATA3*, and *CCND1*—were found to be mutated in 10 % or more of tumors, and collectively these genes were the source of almost 60 % of the driver mutations detected in coding sequences (see [Figure 10.20](#)). In a parallel exome sequencing study of 510 breast cancers, also published in 2012, *TP53* and *PIK3CA* were also found to be the most frequently mutated genes. More recently, whole genome analyses have extended the hunt for driver mutations into

noncoding DNA. Although some noncoding driver mutations are common (such as in the *TERT* promoter to drive expression of telomerase), data from the Pan-Cancer Analysis of Whole Genomes suggest that noncoding driver mutations are rare compared to driver mutations in coding DNA, possibly as a result of a lack of discovery power.

Novel cancer-susceptibility genes

The genome-wide sequencing projects have recently delivered many novel cancer-susceptibility genes. However, the genes being discovered are ones that are infrequently mutated; as with other complex diseases, the major cancer-susceptibility genes have previously been identified. It may be that there will be a considerable tail of low-frequency cancer-susceptibility genes.

By March 2022 the Cancer Gene Census within the Cosmic database (at <http://cancer.sanger.ac.uk/census/>) had listed a total of 578 identified human cancer susceptibility genes. Until recently, the focus has very much been on looking at coding sequences. However, a study of regulatory regions in melanomas has emphasized the need to look at noncoding regions: whole genome sequencing found that highly recurrent somatic mutations occur at two specific nucleotides in the promoter of the telomerase reverse transcriptase gene, *TERT*. Further studies show that the effect of the mutations is to generate a binding site for the ETS transcription factor that upregulates *TERT* expression. The *TERT* promoter mutations were found to occur in more than 70 % of melanomas and about one in six of the other types of tumor examined in the study.

Novel cancer-susceptibility genes are going to be drawn from the genes that support the different biological capabilities of cancers. As we explain in the next subsection, many might not be the conventional oncogenes and tumor suppressor genes that we have become familiar with.

Non-classical cancer genes linking metabolism to the epigenome

One of the surprises emerging from cancer genome sequencing has been the extent to which genes that work in metabolism are important in cancer. These

genes can be non-classical oncogenes or tumor suppressor genes, and many of them have been linked to epigenetic regulation.

Take, for example, the *IDH1* and *IDH2* genes that make respectively cytosolic and mitochondrial isocitrate dehydrogenase, enzymes that work in the tricarboxylic acid (Krebs) cycle to convert isocitrate to 2-oxoglutarate (also known as α -ketoglutarate). One of these two genes is (heterozygously) mutated in 80–90 % of adult grade II/III gliomas and secondary glioblastoma, in more than 50 % of chondrosarcomas, in a significant proportion of acute myeloid leukemias, and in some other cancers. In terms of mutation types and distribution, the genes clearly fall in the oncogene camp (as previously noted for *IDH1* in Figure 10.9).

The predominant *IDH1/IDH2* cancer-associated mutations are specific missense mutations producing mutant enzymes that convert 2-oxo-glutarate (produced by the normal allele) to 2-hydroxyglutarate. At high concentrations, 2-hydroxyglutarate inhibits multiple enzymes that depend on 2-oxoglutarate as a cofactor and work in epigenetic modification, including certain DNA demethylases, such as TET2, and various histone demethylases. That can cause reprogramming of the cell to make it less differentiated (**Figure 10.21**).

As well as oncogenes, tumor suppressor genes regulate the epigenetic-metabolic link in cancer cells. The paper published by [Sebastian et al. in 2012](#) (PMID 23217706) gives the example of the SIRT6 tumor suppressor, a histone deacetylase that normally suppresses aerobic glycolysis.

Tracing the mutational history of cancers: just one of the diverse applications of single-cell genomics and transcriptomics in cancer

The genomics revolution has recently produced an extraordinary technological achievement: the capacity to carry out simultaneous single-cell analyses of genomes and transcriptomes in populations of cells. Not unexpectedly, given the capacity for single body cells to mutate and evolve into rogue cancer cells, applications in oncology have been right at the forefront. **Figure 10.22** illustrates the diverse applications in cancer research.

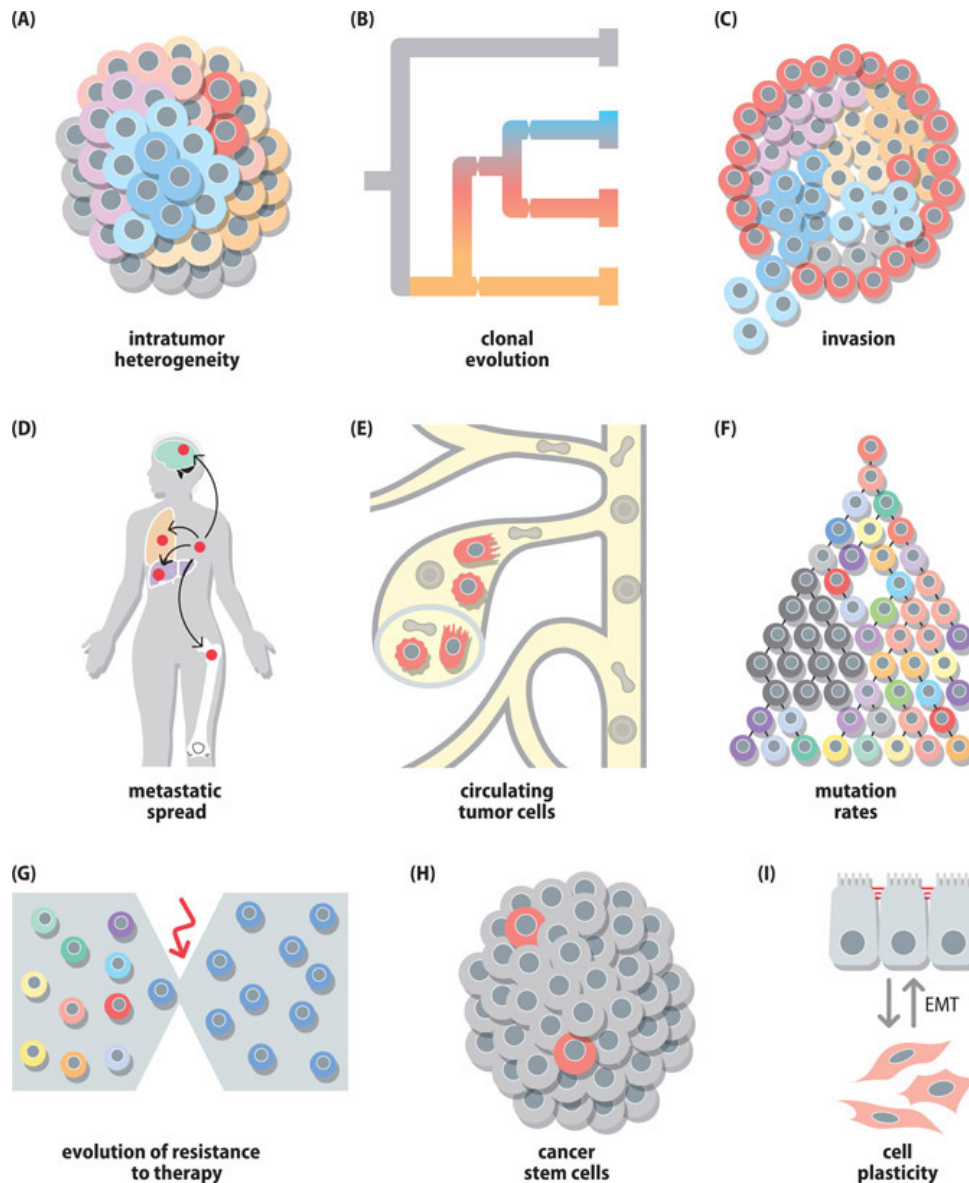


Figure 10.22 Applications of single-cell genomics and transcriptomics in cancer research.

EMT, epithelial-mesenchymal transition. [Adapted from Lindau et al. (2015) *Nature* 526:525–530; PMID 26466571]. With permission from Springer Nature Copyright© 2015.

As an illustration of the power of single-cell genomics, consider how it is applied to study tumor evolution. One can always get an idea of the general evolution of a tumor class by comparing early-stage tumors with pre-cancerous lesions and late-stage tumors. But there is the problem of heterogeneity within tumors: a tumor is not simply a clonal colony of cells. Rather, because increasing genomic instability drives accelerated mutation, a tumor is a heterogeneous

population of cells related by mutational branchpoints of the type shown in Figure 10.22B.

By carrying out single-cell genomic DNA sequencing in tumors, we can position cells of the tumor in phylogenetic lineages to reveal the nature and order of successive driver mutations. Since the mutational landscape in leukemia is relatively simple (in comparison with that of most solid tumors), a single blood sample is enough to allow the evolution of a tumor to be reconstructed. If driver A is found in all leukemic cells and driver B is present in some cells only, for example, one can infer that mutation A appeared first and subsequently mutation B arose and was transmitted to a subpopulation. For an example of tracing the mutational history of cancers, see [Figure 10.23](#). Single-cell transcriptomics in cancer sequencing is also used in various ways to illuminate cancer processes, and we consider them against the broader background of cancer transcriptomics in the next section.

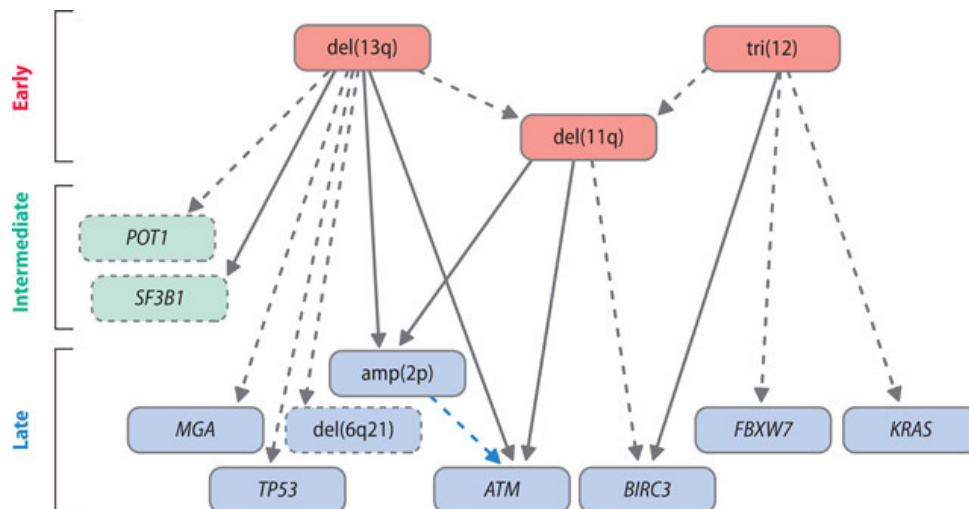


Figure 10.23 Mutational history of chronic lymphocytic leukemia inferred from the clonality of mutations after single-cell sequencing analyses on a single blood sample. Early events are a few types of large copy number changes due to chromosomal instability, but there is substantial diversity in the late driver mutations. (Reprinted from Landau et al. [2015] *Nature* 526:525–530; PMID 26466571.)

Genome-wide RNA sequencing enables insights into the link between cancer genomes and cancer biology and aids tumor classification

Following on from the huge effort in cancer genome analysis, genome-wide analyses have been carried out to study how genetic changes in tumor cells are

expressed to alter the biology of the cells. Most of the effort has been expended on studying RNA transcripts from cancer cells. The first genome-wide studies of cancer transcriptomes used microarray-based expression analyses, but modern analyses use **RNA-Seq** (RNA transcripts are first converted to DNA using reverse transcriptase, then sequenced).

The interest in broadening genome-wide studies of cancer to include additional “omics”, notably transcriptomics and proteomics, has been driven by two major aims. The first is to be able to link genome changes to cancer biology. In this regard, the [Paull et al. \(2021\)](#) reference under Further Reading has been a major advance, linking alterations to the genome to the transcriptional profiles of diverse types of cancer cells. It identified 407 proteins as master regulators that, by working together in groups with variable modular combinations, convert the genomic changes into 112 transcriptionally distinct tumor subtypes.

The second major aim has been to develop some molecular classification of tumors that might improve on conventional methods (which largely depend on how tumors appear when pathologists examine them under the microscope). Improved classification of tumors into subgroups has the potential for more efficiently targeting clinical actions such as prognosis and treatment. New molecular classification methods have recently become available for some types of cancer, such as a study reported in 2021 that identified four molecular subtypes of small cell lung cancer (PMID 33482121). For other cancers such as breast cancer, previous knowledge has been built upon using new molecular subtyping arising from various genome-wide “omics” technologies as reported in the [Parsons and Francavilla \(2020\)](#) review under Further Reading. This area is an evolving one, and clinical benefit from molecular subtyping will be progressively accrued.

The need to focus on biological pathways important in cancer cell evolution

If we view the genome changes in cancers to be complex, linking them in logical ways to diverse changes in the transcriptome and proteome takes the level of complexity to new levels, and will keep researchers busy for quite some time.

To devise clinical benefit from all these studies, it may be productive to concentrate on the effects of the genetic changes within cancer cells. The huge

complexity of all the individual mutations that contribute to cancer can be reduced if we focus not just on the (still very heterogeneous) collection of individual cancer susceptibility genes, but on how key regulators work in biological pathways that are important in cancer cells. Whatever altered genomic states and altered transcriptional states arise as cancers evolve, the functional endpoint is significant changes in key proteins working in various cell signaling pathways (there is little evidence of driver mutations in genes that make noncoding RNAs). In an overview of molecular events in cancer reported in 2013, Bert Vogelstein and colleagues identified 64 high-penetrance oncogenes and 74 high-penetrance tumor suppressor genes, but reported that they all act through one or more of just 12 cell signaling pathways (see [Figure 20.24A](#)).

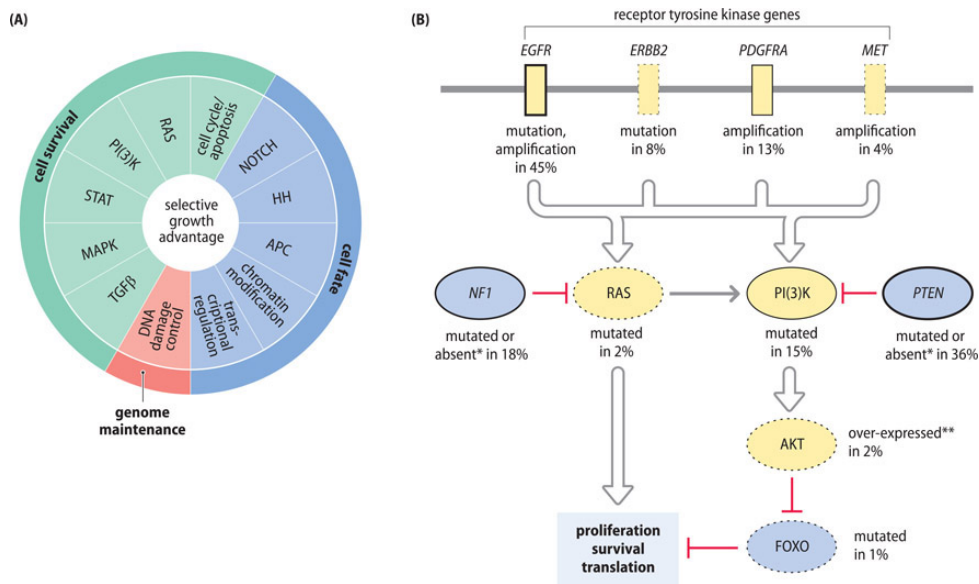


Figure 10.24 Reducing complexity by focusing on signaling pathways important in cancer.

(A) Twelve cancer cell signaling pathways and the processes they regulate. All of the pathways confer a selective growth advantage. The pathways can be organized into three core cellular processes (outer ring). Note that *TP53* encoding the p53 master regulator functions in both the genome maintenance and cell cycle/apoptosis pathways. From Vogelstein B et al. [2013] *Science* 339:1546–1558; PMID 23539594. Reprinted with permission from the AAAS. (B) The importance of the RAS/PI(3)K pathway in glioblastoma multiforme. Oncogene products are shown in yellow shading; tumor suppressor proteins are in blue shading. Red T-bars indicate inhibition. The four upstream receptor tyrosine kinase genes and the genes making each of the six downstream proteins are all mutated, to different extents, in glioblastoma multiforme tumors.

*Via homozygous gene deletion. **Via gene amplification. Data from the Cancer Genome Atlas Research Network [2008] *Nature* 455:1061–1068; PMID 18772890.

For some cancers, what seems a hugely complex series of associated mutational changes and even of genes conferring susceptibility to that cancer, the picture may be altogether simpler at the level of cell signaling pathways. Take the example of glioblastoma multiforme where tumors show high levels of genetic heterogeneity. Despite the large number of genes involved, they all work in two major signaling pathways, a RAS-PI(3)IK pathway (see Figure 10.24B) plus the cell cycle-apoptosis pathway.

Single-cell transcriptomics

Classification of tumor cell diversity and the ability to identify rare tumor cell populations is necessarily limited when analyzing bulk tumor cell populations. The development of droplet-based microfluidic single-cell RNA-Seq enables high-throughput capture and molecular barcoding of individual cells that can be analyzed rapidly. This field is still young, but the review published by [Kim et al. in 2020](#) under Further Reading provides an example of the kinds of applications that are being found.

10.5 GENETIC INROADS INTO CANCER THERAPY

As described in [Section 8.3](#), complex diseases are caused by a combination of genetic and environmental factors, and cancers are no different in this respect. Before we go on to look at therapeutic approaches directed at genetic control points in cancer, it is important to acknowledge the huge effect of environmental factors in cancer. In addition to well-established connections between UV radiation and melanoma, and between tobacco carcinogens and lung cancer, many other cancers are strongly determined by environmental factors. Rates of colon cancer, for example, vary as much as 20-fold between countries; the dramatic differences are due to environmental factors, specifically dietary components, rather than genetic susceptibility. That is evident when a population migrating

from one country to another exhibits a colon cancer rate typical of the new country within one or two generations of settling there. Microbial infections play their part too, and not just viral infections associated with cancers but also certain chronic bacterial infections. The outstanding example is *Helicobacter pylori*, a gastric pathogen that colonizes ~50 % of people across the world and persists for the lifetime of the host. Infection with *H. pylori* causes chronic inflammation and is the strongest known risk factor for gastric cancer, the second most frequent cause of cancer-related deaths worldwide.

How can the burgeoning knowledge of the underlying genetics of cancer have a clinical impact? As previous sections of this chapter testify, the revolution in cancer genomics has made clear the complexity of cancer evolution, and also the extraordinary degree of both intratumor and intertumor heterogeneity. This can pose difficulties in validating biomarkers for the oncogenic process: biopsies from the same tumor may show different genetic profiles. And, because of intra-tumor heterogeneity, natural selection can be expected to propel the growth of drug-resistant clones.

Treatment or prevention?

Faced with these problems, should we simply accept that treating cancers is never going to be anything more than damage limitation, disease management rather than cure? Maybe, but as described below, there may yet be grounds for optimism. But, undeniably, genetics—and especially genomics—has shone a bright torch into the gloom that used to shroud the inner workings of many cancers. The result is a much more informed understanding of the fine, granular detail of the underlying mutation mechanisms, a greater appreciation of the molecular characteristics of cancers, and detailed insights into how cancers evolve. Once we have fully understood the molecular pathways of cancer and cancer evolution in fine detail, we may be in a much better position to devise novel treatments.

Targeted anticancer therapies are directed against key cancer cell proteins involved in oncogenesis or in escaping immunosurveillance

Traditional cancer treatments have been blunt tools: surgery to excise tumors and chemotherapy or radiotherapy to kill them. The latter two methods are simply designed to kill actively dividing cells; the problem, of course, is that they also kill actively dividing normal cells, adversely affecting the health of the patient.

Despite their limitations, the long-established triad of blunt-tool methods are still frequently used in cancer treatment today.

Genetic and especially genomic analyses have recently identified many cancer-susceptibility genes, enabling multiple opportunities to develop *targeted* anti-cancer therapies, directed at specific proteins directly or indirectly involved in oncogenesis. Good targets should be present in cancer cells but not normal ones, or should be strongly upregulated in cancer cells when compared to normal cells. As described in the subsections below, targeted anticancer therapies involve designing molecules that specifically bind to and inhibit key proteins involved in oncogenesis, or that can selectively kill cancer cells (either directly, or indirectly through the intervention of T cells). We cover targeted therapies using small molecular drugs or monoclonal antibodies here. A third method, using genetically engineered T cells, is described in the section following the next one.

Targeted therapies using small molecule drugs

Conventional small molecule drugs produced by the pharmaceutical industry can be screened for evidence of binding to a specific protein of interest. Often a molecule like this can fit snugly in a cleft in the protein, sometimes disrupting the function of the protein.

The first successful targeted anticancer therapy was achieved using this approach more than three decades ago. It was prompted by the *BCR-ABL* chimeric oncogene on the Philadelphia chromosome frequently found in chronic myeloid leukemia (CML), as shown in Figure 10.8A. The rationale was this: if the *BCR-ABL* gene is present in tumor cells but not in normal cells, can we find a drug to specifically bind to the BCR-ABL fusion protein and stop it working? And if so, shouldn't the drug selectively stop the tumor cells proliferating? The answer to both questions was yes. Imatinib (marketed as Gleevec^R) obtained FDA approval in 2001 to treat CML patients with the Philadelphia chromosome and was seen to be a resounding success. Subsequently there has been a variety of other successes with small-molecule drugs. Some of these have been targeted to bind key proteins implicated in different cancers—see [Table 10.8](#).

TABLE 10.8 EXAMPLES OF TARGETED CANCER THERAPIES

Cancer	Drug/MAb	Protein target	Mode of action
--------	----------	----------------	----------------

Cancer	Drug/MAb	Protein target	Mode of action
SMALL MOLECULE DRUGS			
Breast	Tamoxifen	Estrogen receptor in ER-positive breast cancers	Blocks ER, preventing growth signals
Leukemia, (AML, Ph+)	Imatinib	BCR-ABL1 fusion protein	Inhibits abnormal signaling by fusion protein tyrosine kinase
Leukemia (AML, CLL)	Venetoclax	BCL2, a key inhibitor of apoptosis	Binds to BCL2 to disrupt its function, thereby stimulating apoptosis
Lymphoma (SLL)			
Melanoma	Verumafenib	BRAF V600E mutant protein	Specifically inhibits V600E mutant BRAF, triggering apoptosis
Non-small cell lung cancer	Crizotinib	EML4-ALK fusion protein	Inhibits abnormal signaling by fusion protein tyrosine kinase
Ovarian (advanced, BRCA1/2 minus)	Olaparib [*]	PARP1 enzyme	Blocks repair of DNA breaks in BRCA1 - mutant cancers
Various (advanced)	Becacizumab (Avastin)	VEGF (vascular endothelial growth factor)	Inhibits angiogenesis
MONOCLONAL ANTIBODIES			
Breast	Trastuzumab	EGF receptor (EGFR) on HER2-positive cells	Attaches to receptor; identifies the cell as a target for the immune system
Leukemia	Rituximab	CD20 B-cell surface protein	Binds to CD20; identifies cells as targets for NK cells.

Cancer	Drug/MAb	Protein target	Mode of action
Lymphoma	Ibritumomab**	CD20 B-cell surface protein	Binds to CD20; carries a radioactive payload to kill cells it binds to
Melanoma	Ipilimumab	CTLA4 T-cell inhibitor	Blockade of CTLA4 or PD1 allows T cells to renew attack on cancer cells
	Nivolumab	PD1 T-cell inhibitor	
Prostate (advanced)	Sipuleucel-T***	Prostatic acid phosphatase (PAP)	Stimulates T-cell response against PAP

AML, acute myeloid leukemia. CLL, chronic lymphocytic leukemia. EGF, epidermal growth factor. SLL, small lymphocytic lymphoma. Ph+, Philadelphia chromosome present. PARP, Poly(ADP-Ribose).

* Demonstrates the potential of *synthetic lethality*, where a combination of two nonlethal deficiencies may result in a lethal effect. PARP activates repair of single-strand DNA breaks. When PARP is absent, these breaks have to be repaired by BRCA1/2-mediated homologous recombination, and so when tumor cells lack BRCA1/2, the error-prone nonhomologous end joining method of DNA repair is used, often leading to cell death.

** Ibritumomab is radiolabeled before use by attaching a yttrium 90 radioisotope.

*** Uses a proprietary protein (PAP fused to GM-CSF to stimulate the patient's own leukocytes *ex vivo*).

Note that the target proteins in some targeted therapies may play a supportive role common to multiple cancer types. For example, to fuel their growth, advanced tumors often develop a vascular supply ([Figure 10.5A](#)), providing oxygen and nutrients. Treatment with avastin^R, which inhibits *angiogenesis*, the process by which new blood vessels form, inhibits the outgrowth of metastases.

While there have been promising successes for targeted therapies at early stages in treatment, relapses are common as tumors mutate to become resistant to treatment, and we consider this problem below.

Targeted therapies using monoclonal antibodies

Monoclonal antibodies raised against a specific protein target provide an alternative type of drug that can be used as a way of killing cancer cells. Occasionally, a specific monoclonal antibody, radiolabeled with a cytotoxic

radioisotope, permits direct killing of cancer cells—see the example of ibritumomab in [Table 10.8](#). Usually, however, the object is a type of **immunotherapy** to encourage some immune response against cancer cells (see [Table 10.8](#) for examples).

An important example of immunotherapy is *immune checkpoint therapy*. This field was developed independently by Tasuku Honjo and James Allison who shared the Nobel Prize in Physiology or Medicine in 2018 for their work on engineering antibodies to, respectively, the PD1 and CTLA4 cell surface receptors. As part of the self-nonsel self recognition system, T cells have certain brakes to ensure that they are not inappropriately activated, notably by ligand activation of PD1 and CTLA4. When a T cell encounters a cell with a ligand protein on its surface for the PD1 and/or CTLA4 receptor, the resulting receptor-ligand interaction conveys a signal to inhibit T cell responses. In order to escape immune surveillance, cancer cells take advantage by expressing high amounts of the PD1 and/or CTLA4 ligands on their cell surfaces. The ipilimumab and nivolumab monoclonal antibodies bind specifically to the PD1 and CTLA4 receptors, respectively, blockading them from interaction with ligand proteins, thereby reactivating the capacity of T cells to kill cancer cells.

CAR-T Cell therapy and the use of genetically engineered T cells to treat cancer

In addition to standard small molecule drugs and monoclonal antibody drugs, novel targeted anticancer therapies have been developed using genetically modified T cells, building upon the natural roles of T cells in tumor immunosurveillance. (T cells are able to detect and kill tumor cells arising from both virus infection and from non-viral genetic and epigenetic changes to cells of the body. In the latter case, the tumor antigens with highest specificity are derived from new peptides created through chromosomal translocations and frameshifting mutations.)

Cytotoxic T cells naturally recognize peptide antigens on the surface of cells only after they have been bound by an HLA protein (MHC restriction—see [Box 8.3](#) on page 265), and successful recognition is also dependent on binding of ligands to additional co-receptors on T cells. In the laboratory, T cells can be genetically engineered so that they can bind a specific protein of interest without the need for an associated HLA protein. To do that cultured T cells are transfected

with a gene construct that is designed to make an artificial trans-membrane *chimeric antigen receptor* (CAR).

The extracellular domain of a T-cell chimeric antigen receptor is designed for antigen recognition. It is typically largely composed of an scFv (single-chain variable fragment) antibody sequence, that is, the variable regions on heavy (V_H) and light (V_L) chains of a standard monoclonal antibody connected by a short linker peptide. An additional hinge region is purely for structural reasons, enhancing the flexibility of the antigen-binding head of the scFv domain (see [Figure 10.25](#)).

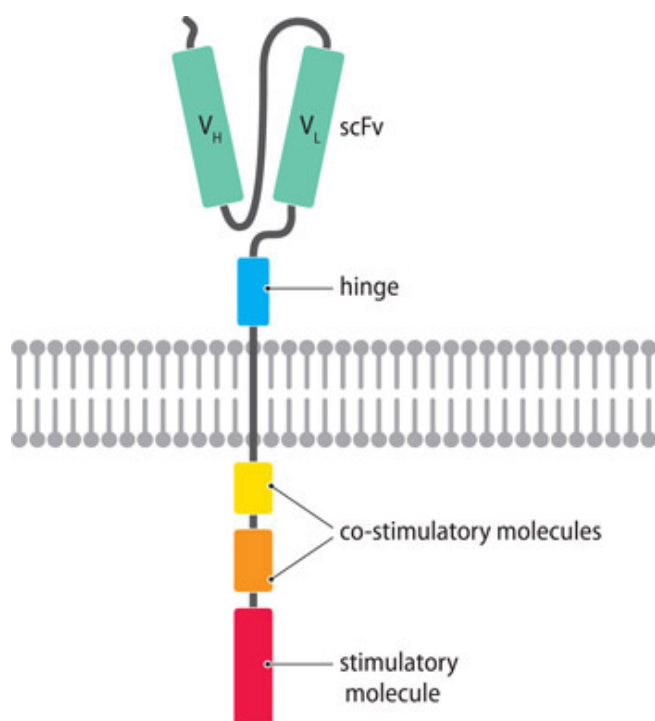


Figure 10.25 Structure of a third-generation chimeric antigen receptor, as deployed in CAR-T cell therapy. The extracellular region is composed of a scFv domain component used in antigen recognition (with the distal ends of the V_H and V_L variable Ig chains forming the antigen binding site) and a short hinge to allow flexibility. The intracellular region is composed of a stimulatory molecule (often a CD3 zeta chain from a T cell receptor) and two or more co-stimulatory molecules (such as CD27, CD28, OX40, 4-1BB, and ICOS).

The intracellular region is used in signal transduction. It is formed by bringing together two types of molecule: stimulatory molecules permit downstream activation of the T cells to release cytokines (ultimately leading to death of the

target cell) and interleukins (needed for proliferation), and costimulatory domains serve to enhance the immune response.

Because of *MHC restriction*, recognition of other cells by T cells requires that the cells present antigens bound to HLA proteins, and so cancer cells can switch off HLA expression to escape immunosurveillance by T cells. However, CAR-T cells and other types of genetically engineered T cells remove the requirement for HLA recognition. Treatment with engineered T cells such as these are the equivalent of giving patients a “living drug” that targets a specific protein of choice.

In an early application CAR-T cells were designed to recognize the B-cell antigen CD19 and used to treat B-cell lymphomas and some leukemias with very considerable success. Initially, *autologous* T cells were used: T cells would be removed from a patient then genetically engineered in the lab to make desired CAR-T cells that would be infused back into the patient. For greater convenience and reduced costs, there has been a move towards producing universal, off-the-shelf CAR-T cells; because they would be genetically different to the cells of a patient, there is a requirement for immunosuppressant drugs during treatment.

Despite the very significant clinical benefits, there have been drawbacks in CAR-T cell therapy, notably the propensity to induce “cytokine storms”, wherein a massive release of cytokines has resulted in some fatalities in clinical trials. Fine-tuning the protein engineering has become a priority.

The molecular basis of tumor recurrence and the evolution of drug resistance in cancers

Although the initial results of anticancer targeted therapies using small molecule drugs, monoclonal antibodies, or engineered T cells can be very positive, and though they can have significant side effects, the treatments are generally well tolerated when compared to chemotherapy. Applicability has, however, been variable. The therapies have worked well in leukemias, and chronic myelogenous leukemia has been especially amenable given that more than 90 % of patients have the Philadelphia chromosome and produce the BCR-ABL fusion protein that can be treated effectively with Gleevec® (imatinib). But in general, leukemias are amenable to treatment, showing comparatively limited genomic instability and often being identified at an early stage in tumor development. By contrast,

epithelial cancers are not so easily treated, showing much greater genomic instability and not usually being caught until later stages in tumor development.

The basis of tumor recurrence

A general problem in treating cancer is that some cancer cells can survive treatment and tumors frequently recur quite quickly. That poses the question: why? Two possible answers are that cancer stem cells are especially resistant to drug treatment or that subpopulations within a tumor can survive treatment. Support for especially resistant stem cells includes a study on a genetically engineered mouse model of glioblastoma in which a relatively quiescent subset of endogenous glioma cells, with properties resembling cancer stem cells, was found to be responsible for sustaining long-term tumor growth (by producing transient populations of highly proliferative cells). If cancer stem cells are relatively resistant to therapy, they might survive to repopulate a vastly shrunken tumor. If so, the problem becomes how to effectively target and kill populations of cancer stem cells about which we know little.

The possibility of tumor heterogeneity has been amply demonstrated in various studies. If a malignant tumor consists of genetically different populations, some cells might survive drug treatment and natural selection could foster the development of tumor subclones with mutations that render the therapeutic drug ineffective in some way. (There are parallels, therefore, with infectious diseases and the evolution of drug resistance in microbes.)

The evolution of drug resistance

The evolution of drug resistance in targeted cancer therapy can occur in different ways. Sometimes, mutations develop in the gene encoding the drug target itself. For example, in the treatment of chronic myeloid leukemia with imatinib, tumor subclones develop imatinib resistance by developing point mutations that alter the kinase domain of the BCR-ABL1 fusion protein. The mutant kinase retains the catalytic activity required for tumor formation, but its altered structure means that imatinib can no longer bind to it effectively to inhibit it. Drug resistance for many other kinase inhibitors works by a similar mechanism: often the mutations confer

resistance by blocking interactions between drug and target through steric hindrance.

An alternative way of developing drug resistance occurs when the tumor mutates to amplify the drug target gene. Occasionally, for example, resistance to kinase inhibitors in chronic myeloid leukemia is achieved when tumors succeed in amplifying the *BCR-ABL1* gene. Prostate cancers often acquire resistance to drug-mediated androgen deprivation by amplifying the androgen receptor gene.

Yet another option for a tumor to develop drug resistance is to find a way of bypassing the primary drug target (which remains unaltered, and continues to be inhibited by the drug). This can take the form of mutating a downstream effector in the same pathway to render cells insensitive to drug inhibition of a cell surface receptor, for example; or an alternative pathway is activated. For example, the monoclonal antibody trastuzumab is designed to treat breast cancer by binding to and interfering with the human epidermal growth factor receptor 2 (HER2), but tumors can bypass the effects of the drug by activating expression of an alternative receptor, such as HER3.

Of course, it would be highly desirable to monitor tumors so as to detect emerging resistance clones as soon as possible to hopefully permit changes in treatment to stop them in their tracks. The recent development of “liquid biopsies” may be an important advance (see [Clinical Box 14](#)).

CLINICAL BOX 14 LIQUID BIOPSIES IN CANCER: TOWARDS CLINICAL PRACTICE

Diagnosing and monitoring cancer are aided by taking cell samples for examination. While leukemias are conveniently diagnosed and monitored using blood tests, solid tumors have routinely been accessed through invasive biopsies. Most use hollow needles following local anesthetic, but less easily accessed tumors have often required the use of cutting tools attached to an endoscope, or even open or laparoscopic surgery. Monitoring cancer using serial biopsies is generally not a very attractive option, therefore, especially if patients are elderly and frail.

The attractive possibility of **liquid biopsies** in cancer has been made possible by the observation that both circulating tumor cells (CTC) and cell-free

circulating tumor DNA (ctDNA) are present in the peripheral blood of cancer patients (see [Figure 1](#)).

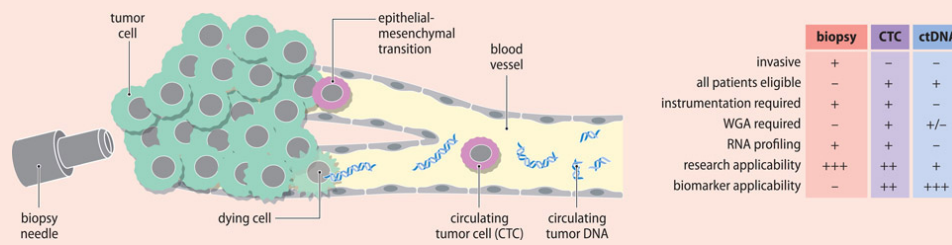


Figure 1 Comparison of liquid biopsies versus needle biopsies for investigating cancer.

CTC, circulating tumor cells; ctDNA, circulating (cell-free) tumor DNA; WGA, whole genome amplification. (Reproduced from Wyatt AW & Gleave ME [2015] *EMBO Mol Med* 7:878–894; PMID 25896606. © 2015 The Authors. Published under the terms of the CC BY 4.0 license.)

Retrieving and analyzing tumor DNA from peripheral blood is not straightforward because of the limited quantity of tumor material in the circulation (1 ml of blood has 10 million leukocytes but maybe only 1 CTC; and the fraction of circulating cell-free DNA that originates from tumors may often be very low). Currently, specialized techniques such as droplet digital PCR are required to amplify the DNA prior to sequence analysis.

Of the numerous applications of liquid biopsies in cancer, a very important one will be in monitoring the evolution of treatment resistance to detect, for example, secondary mutations that arise in tumors to prevent a treatment drug working. By identifying the drug-resistance mutation before clinical signs of disease progression develop, it may be possible to quickly make a compensatory adjustment to the treatment.

The promise of combinatorial drug therapies

Because of tumor heterogeneity, is targeted drug therapy always doomed to eventual failure? One approach that has great potential is combinatorial therapy, that is, using combinations of treatments that act differently. It conceivably might even have the potential to lead to actual cures for some cancer patients, rather than simply temporary remissions. A successful template has already been provided by the recent success in treating human immunodeficiency virus (HIV). Just like tumor cells, the HIV virus is highly mutable and can quickly mutate to resist any

individual antiviral drug. But the highly active antiretroviral therapy (HAART) strategy used a combination of different antiretroviral drugs. Because the chances that individual viruses could mutate to become simultaneously resistant to two or three drugs are low, HAART has been much more successful than previous HIV treatments.

Future precision oncology might involve the simultaneous use of multiple agents and drugs that target diverse vulnerabilities of cancer cells before resistance has a chance to develop. There are potentially huge numbers of possible permutations. One might target upstream and downstream components in a pathway known to be important in the development of specific cancers, or components in parallel pathways. There are different types of treatment, including standard drug therapies and immunotherapies. By 2020 more than 5000 clinical trials were ongoing globally to assess the clinical benefits from new combination therapies. Because the possibilities to combine treatments dramatically outnumber the patients available to enroll in clinical trials there is an urgent clinical need for rational cancer treatment combinations.

SUMMARY

- Cancers are diseases in which there is an unregulated increase in cell growth that leads to cells invading neighboring tissues and spreading to distant sites in the body.
- The genetic contribution to cancers predominantly occurs through somatic mutations. Germline mutations may result in inherited cancers, but even in these cases additional somatic mutations are required for cancers to form.
- Cancer development occurs only after a series of successive regulatory controls have gone wrong in cells, leading to increased cell proliferation or reduced apoptosis.
- Cancers are primarily diseases of later life, because it takes time for multiple cell controls to be disrupted.
- Tumors originate ultimately from a single cell but they are genetically heterogeneous. Descendants of the founder cell can acquire genetic mutations that afford a growth advantage; they form

a dominant subclone that is then surpassed in growth by successive subclones (which acquire additional mutations conferring further growth advantages).

- Cancers develop by accelerating mutation in two different ways: by conferring a growth advantage in cells, and by destabilizing the genome to increase the probability of later mutations.
- In some types of cancer, undifferentiated cells are found with stem cell properties; they can self-replicate and also give rise to more differentiated cells within the tumor. Genetically different cancer stem cells may also arise by clonal evolution.
- Tumors most likely originate from cells that already have a high proliferative capacity, such as stem cells or rapidly multiplying and poorly differentiated embryonic tissues. But genetic and epigenetic changes in differentiated cells may also cause these cells to become progressively more plastic (less differentiated) and progressively acquire other characteristics of cancer cells.
- Intratumor heterogeneity includes not just genetically different descendants from a single founding cell, but also non-tumor cells that are recruited to the tumor microenvironment, including some types of infiltrating immune cell.
- Cancer is a contest between Darwinian natural selection operating at the level of the individual (over generations) and at the level of the cell (within a single individual). Although cancer cells can successfully proliferate and form tumors within a person, they cannot leave progeny beyond the life of their host; tumorigenesis processes must start afresh in a new individual.
- Cancer cells usually contain thousands of somatic mutations. A small number, often from one to five or six, are driver mutations that are crucially important in cancer development and are positively selected. The rest are chance (passenger) mutations resulting from genomic instability.
- Cancer genes can be grouped into two classes according to how they work in cells. In some cases, mutation of a single allele is sufficient to make a major contribution to the development of cancer. For other

cancer genes, both alleles need to be inactivated to make a significant contribution to cancer.

- Oncogenes are dominantly acting cancer genes that arise through some activating mutation in one allele of a normal cellular “proto-oncogene”. Classical protooncogenes typically work in growth signaling pathways to promote cell proliferation or inhibit apoptosis.
- Proto-oncogenes can be activated to become onco-genes by acquiring gain-of-function mutations; by being over-expressed as a result of gene amplification; or through activated expression resulting from a trans-location (which repositions a transcriptionally silenced gene so that it comes under the control of transcription-activating regulatory elements).
- Classical tumor suppressor genes are recessively acting cancer genes in which the inactivation of both alleles promotes cell proliferation or inhibits apoptosis. Additional tumor suppressors work in other areas such as in genome maintenance.
- The two-hit hypothesis describes how cancer develops from two successive inactivating mutations in a tumor suppressor gene. It explains why dominantly inherited cancers are recessive at the cellular level (the first mutation occurs in the germ line and so there is a very high chance that the second allele is inactivated in at least one cell in the body to form a tumor). In sporadic cancers of the same type, both the first and second inactivating mutations occur in a somatic cell.
- Genome instability ensures additional mutations for natural selection to work on to drive tumor formation. It often manifests as chromosomal instability (resulting in aneuploidies, translocations, and so on) but can also be apparent at the DNA level as microsatellite instability (resulting from mutations in genes that work in mismatch DNA repair).
- Epigenetic dysregulation is important in both cancer initiation and cancer progression. It can be induced by genetic changes (notably mutation in genes that make epigenetic regulators) or by tissue inflammation causing altered cell signaling that results in altered chromatin states.

- Aberrant chromatin states produced by epigenetic dysregulation can allow cancer cells to become unspecialized (poorly differentiated) and can silence alleles of cancer-susceptibility genes. Additionally, DNA hypomethylation can result in widespread chromosome instability.
- Genome-wide gene expression profiling of tumors can subdivide cancers of the same type, such as breast carcinomas, into different groups with different biological characteristics and different drug responses.
- Two tumors of the same type show very different mutational spectra—the great majority of passenger mutations are often distributed randomly across the genome; although some key cancer genes might be mutated in both tumors, other driver mutations may be located in different cancer-susceptibility genes.
- Tumors evolve, so cells in different regions of the same tumor can show regional mutational differences; metastatic cells typically share mutations that distinguish them from the primary tumor.
- Human cancer-susceptibility genes have been identified by analyzing associated chromosome breakpoints or associated changes in copy number (oncogene amplification, or loss of heterozygosity in the case of tumor suppressor genes); by studying candidate genes suggested by analyses of experimental organisms; and by exome or genome sequencing.
- In targeted cancer therapies, a drug or other treatment agent is directed at counteracting the effects of a specific genetic mutation that is known to be crucial for development of the cancer.
- Recurrence of tumors may be driven by cancer stem cells that are comparatively resistant to therapy.
- After initial success in shrinking tumors, cancer therapies often fail, causing a clinical relapse. Tumor cells evolve to become resistant to the drug as a result of natural selection (which promotes the growth of tumor cells that develop mutations to combat the effects of the drug).

- Tumors often develop drug resistance by changing the conformation of the drug target so that the drug is sterically hindered from binding to it, by amplifying the gene encoding the drug target or by activating an alternative pathway that bypasses the effect on the drug target.

QUESTIONS

Questions can be downloaded by visiting the following link, under Support Materials: www.routledge.com/9780367490812.

FURTHER READING

Cancer biology

[Weinberg, RA](#) (2014) *The Biology of Cancer*, 2nd ed., Garland Science.

General molecular characteristics of cancer

[Hanahan D & Weinberg R](#) (2011) Hallmarks of cancer: the next generation. *Cell* 144:646–674; PMID 21376230.

Shay JW & Wright WE (2011) Role of telomeres and telomerase in cancer. *Semin Cancer Biol* 21:349–353; PMID 22015685.

Trybek T (2020) Telomeres and telomerase in oncogenesis. *Oncol Lett* 20:1015–1027; PMID 32724340.

Vaupel P, Multhoff G (2021) Revisiting the Warburg effect: historical dogma versus current understanding. *J Physiol* 599:1745–1757; PMID 33347611.

Cancer evolution, cancer stem cells, and intratumor heterogeneity

[Burrell RA](#) (2013) The causes and consequences of genetic heterogeneity in cancer evolution. *Nature* 501:338–345; PMID 24048066.

Clevers H (2011) The cancer stem cell: premises, promises, and challenges. *Nature Med* 17:313–319; PMID 21386835. [For a follow-up in 2017 see also PMID 28985214.]

[Gerlinger M](#) (2012) Intratumor heterogeneity and branched evolution revealed by multiregion sequencing. *N Engl J Med* 366(10):883–892; PMID 22397650.

Greaves M & Maley CC (2012) Clonal evolution in cancer. *Nature* 481:306–313; PMID 22258609.

Hausser J & Alon U (2020) Tumor heterogeneity and the evolutionary trade-offs of in cancer. *Nat Rev Cancer* 20:247–257; PMID 32094544.

Magee JA (2012) Cancer stem cells: impact, heterogeneity and uncertainty. *Cancer Cell* 21: 283–296; PMID 22439924.

Pon JR & Marra MA (2015) Driver and passenger mutations in cancer. *Annu Rev Pathol Mech Dis* 10: 25–50; PMID 25340638.

Reiter JG (2019) An analysis of genetic heterogeneity in untreated cancers. *Nat Rev Cancer* 19:639–650; PMID 31455892.

Trumpp A, Haas S (2014) Cancer stem cells: the adventurous journey from hematopoietic to leukemic stem cells. *Cell* 185:1266–1270; PMID 35385684.

Oncogene activation and chromosome aberrations in cancer

Hnisz D (2016) Activation of proto-oncogenes by disruption of chromosome neighborhoods. *Science* 351:1454–1458; PMID 26940867.

Mertens F (2015) The emerging complexity of gene fusions in cancer. *Nat Rev Cancer* 15:371–381; PMID 25998716.

Mitelman Database of Chromosome Aberrations and Gene Fusions in Cancer.
<https://mitelmandatabase.isb-cgc.org/>

Roukos V (2013) The cellular etiology of chromosome translocations. *Curr Opin Cell Biol* 25:357–364; PMID 23498663.

Storlazzi CT (2010) Gene amplification as double minutes or homogeneously staining regions in solid tumors: origin and structure. *Genome Res* 20:1198–1208; PMID 20631050.

Tumor suppressor genes

[Berger AH](#) (2011) A continuum model for tumor suppression. *Nature* 476:163–169; PMID 21833082.

Berger AH & Pandolfi PP (2011) Haplo-insufficiency: a driving force in cancer. *J Pathol* 223:137–146; PMID 21125671.

[Freed-Pastor WA & Prives C](#) (2012) Mutant p53: one name, many proteins. *Genes Dev* 26:1268–1286; PMID 22713868.

Knudson AG (2001) Two genetic hits (more or less) to cancer. *Nat Rev Cancer* 1:157–162; PMID 11905807. [A historical perspective of the development of the two-hit tumor suppressor hypothesis.]

[Sebastian C](#) (2012) The histone deacetylase SIRT6 is a tumor suppressor that controls cancer metabolism. *Cell* 151(6):1185–1199.

Solimini NL (2012) Recurring hemizygous deletions in cancer may optimize proliferative potential. *Science* 337:104–109; PMID 22628553.

Genome instability and epigenetic dysregulation in cancer

Darwiche N (2020) Epigenetic mechanisms and the hallmarks of cancer: an intimate affair. *Am J Cancer Res* 10:1954–1978; PMID 32774995.

[Feinberg AP](#) (2016) Epigenetic modulators, modifiers and mediators in cancer aetiology and progression. *Nature Rev Cancer* 17:284–299; PMID 26972587.

Flavahan WA (2017) Epigenetic plasticity and the hallmarks of cancer. *Science* 357:eaal2380; PMID 28729483.

Pena-Diaz J & Jiricny J (2012) Mammalian mismatch repair: error-free or error-prone? *Trends Biochem Sci* 37:206–214; PMID 22475811.

Roy R (2012) BRCA1 and BRCA2: different roles in a common pathway of genome protection. *Nat Rev Cancer* 12:68–78; PMID 22193408.

Schwitalla S (2013) Intestinal tumorigenesis initiated by dedifferentiation and acquisition of stem cell-like properties. *Cell* 152: 25–38; PMID 23273993.

Shen H & Laird PW (2013) Interplay between the cancer genome and epigenome. *Cell* 153:38–55; PMID 23540689.

Cancer genomics and transcriptomics

- Cieslik M & Chinniyam AM (2020) Global cancer genomics project comes to fruition. *Nature* 578:39–40. [Provides a commentary on six key papers published in the same issue from the Pan Cancer Analysis of Whole Genomes consortium.]
- Garraway LA & Lander ES (2013) Lessons from the cancer genome. *Cell* 153:17–37; PMID 23540688.
- Hong M (2020) RNA sequencing—new technologies and applications in cancer research. *J Hematol Oncol* 13:166; PMID 33276803.
- [Kim N](#) (2020) Single-cell RNA sequencing demonstrates the molecular and cellular reprogramming of metastatic lung adenocarcinoma. *Nature Commun* 11:2285; PMID 32385277.
- Kim B (2020) Advancing cancer research and medicine with single-cell genomics. *Cancer Cell* 37:456–470; PMID 32289270.
- Martinez-Jimenez F (2020) A compendium of mutational cancer driver genes. *Nat Rev Cancer* 20:555–572; PMID 32778778.
- [Parsons J & Francavilla C](#) (2020) Omics approaches to explore the breast cancer landscape. *Front Cell Dev Biol* 7:395; PMID 32039208.
- [Paull EO](#) (2021) A modular master regulator landscape controls cancer transcriptional identity. *Cell* 184:334–351; PMID 33434495.
- [Vogelstein B](#) (2013) Cancer genome landscapes. *Science* 339:1546–1558; PMID 23539594.

Cancer therapeutics and monitoring strategies

- Al-Lazikani B (2012) Combinatorial drug therapy for cancer in the post-genomic era. *Nature Biotechnol* 30:1–13; PMID 22781697.
- Crowley (2013) Liquid biopsy: monitoring cancer-genetics in the blood. *Nat Rev Clin Oncol* 10:472–484; PMID 23836314.
- Dancey JE (2012) The genetic basis for cancer treatment decisions. *Cell* 148:409–420; PMID 22304912.
- June CH (2018) CAR T cell immunotherapy for human cancer. *Science* 369:1361–1365; PMID 29567707.
- McDermott U (2011) Genomics and the continuum of cancer care. *N Engl J Med* 364:340–350; PMID 21268726.

Nature Medicine Focus on targeted cancer therapies (2013) *Nature Med* 19:1380–1464. [A collection of various reviews in this area in the November 2013 issue.]

Ward RA (2020) Challenges and opportunities in cancer drug resistance. *Chem Rev* 6:3297–3351; <https://dx.doi.org/10.1021/acs.chemrev.0c00383>

11

Genetic and genomic testing in healthcare

Practical and ethical aspects

DOI: [10.1201/9781003044406-11](https://doi.org/10.1201/9781003044406-11)

CONTENTS

[11.1 AN OVERVIEW OF GENETIC TESTING](#)

[11.2 GENETIC TESTING FOR CHROMOSOME ABNORMALITIES AND PATHOGENIC STRUCTURAL VARIATION](#)

[11.3 GENETIC AND GENOMIC TESTING FOR PATHOGENIC POINT MUTATIONS AND DNA METHYLATION TESTING](#)

[11.4 GENETIC AND GENOMIC TESTING: ORGANIZATION OF SERVICES AND PRACTICAL APPLICATIONS](#)

[11.5 ETHICAL, LEGAL, AND SOCIETAL ISSUES \(ELSI\) IN GENETIC TESTING](#)

[SUMMARY](#)

[QUESTIONS](#)

[FURTHER READING](#)

We end this book by considering how the ever-expanding knowledge of our genome, our genes, and genetics is being used in genetic testing to make a

positive impact on the health of society. Our ability to scrutinize and interpret genetic variation in health and disease has certainly expanded substantially since the last edition of this book. And there have been new developments in treating disease. We discuss how these advances raise ethical considerations for practice, and how further developments in both diagnosis and treatment may raise different versions of the ethical issues.

In some previous chapters we looked at how genetics and genomics are illuminating our understanding of the molecular basis of disease, and how this knowledge has brought about significant—sometimes profound—changes in how we diagnose and treat human disease. For many genetic disorders there remains no adequate treatment; in general, genetic approaches to treatment have advanced much more slowly than our ability to diagnose a genetic cause. Innovative treatments are emerging, however, and it is likely that there will be many more changes over the coming decade.

Genetic testing has been used in the clinic since the late 1960s. The subsequent DNA cloning revolution allowed rapid developments in DNA-based diagnosis. Initially used in just a few medical settings, notably clinical genetics, DNA technologies were then democratized by PCR, an inexpensive DNA technology that was very easy to use and was rapidly taken up. As well as being extensively used in clinical genetics services, PCR-based testing became the standard way of identifying pathogens and so is a key tool used by microbiologists and virologists. And it has come to be increasingly used in other medical specialties, such as hematology and oncology. Of course, there has also been a revolution in DNA sequencing.

As the technologies become exponentially faster and cheaper the entire clinical approach is changing. First, genetic testing is now available, and useful, to all the major divisions of medicine—it is certainly no longer the preserve of specialties labeled “clinical genetics”. Secondly, previous strategies of using a phenotype to determine which bits of a genotype to assess have become inverted: whole genome assays are quite often the first step from which predictions about phenotype might now be made.

Given its speed and affordability, whole genome genetic testing is being offered by more and more private companies direct to the consumer (DTC). DTC genetic testing comes with or without health interpretations, and accompanying healthcare professional input may be totally lacking or minimal. Given that the popular discourse around genetic information is often strikingly optimistic—“your DNA is your blueprint” and so on—there is potential for underestimating the complexity of a DTC output in terms of predictions about future health.

In [Section 11.1](#) we give an overview of genetic testing before describing the technology of genetic testing for detecting chromosome abnormalities and large to moderate-scale DNA copy number variation ([Section 11.2](#)), and testing for point mutations and DNA methylation changes ([Section 11.3](#)). In [Section 11.4](#), we describe how genetic services are organized and the practical applications of genetic testing. (Note that we have previously described applications in pharmacogenetic testing within the context of treatment for genetic disorders; readers interested in this application should consult [Section 9.2](#).)

Finally, in [Section 11.5](#), we consider the range of ethical questions and impacts on society that might arise through the practices of genetic testing and certain applications of genetic technologies to treat disease. We then go on to offer some thoughts and possible directions to turn to so that such issues can be resolved or ameliorated in practice.

11.1 AN OVERVIEW OF GENETIC TESTING

Although the title of this chapter refers to genetic testing in healthcare, it is important to consider that genetic testing requested for other reasons may intersect with healthcare analyses. Genetic testing for crime scene analyses, identity checking, or determining whether biological relationships have been misattributed, might each be conducted for legal reasons, but they may also provide important information for healthcare. Companies offering direct-to-consumer (DTC) genetic testing increasingly sell ancestry tests to pique consumer interest with certificates of ancestral make-up, but often

such data are also analyzed to make healthcare predictions, and so may intersect with healthcare questions. Genetic testing is also important in understanding normal genetic variation in different human populations. And, increasingly, genetic testing is being used to analyze the genetic contribution to common, complex genetic disease rather than being limited to the rare disease diagnoses it was focused on just a few decades ago.

The genetic testing outlined in this chapter is primarily concerned with detecting the relatively small portion of human genetic variation that confers susceptibility to disease. There are different general strategies for carrying out the testing, and different levels and environments at which it is carried out.

The different source materials and different levels of genetic testing

The source material for genetic testing can be cells (usually blood, tumor, skin, embryonic, or fetal cells; see [Table 11.1](#)) or body fluids (blood, urine), stools, and even exhaled breath (increasingly used in assaying certain cancer biomarkers). In addition, testing is sometimes carried out on archived material (often stored blood or tumor samples) from deceased persons to provide information that can be of clinical help to surviving family members.

TABLE 11.1 SOURCES OF MATERIAL FOR GENETIC TESTING	
Source of cells/DNA/RNA	Type of testing or screening
EMBRYONIC/FETAL	
Single cell from a blastomere or a few cells from a blastocyst	preimplantation diagnosis
Fetal DNA in maternal blood	prenatal diagnosis, as early as 6 weeks (testing for paternal alleles). Fetal sexing
Chorionicvillus	prenatal diagnosis at about 9-14 weeks

Source of cells/DNA/RNA	Type of testing or screening
Amniotic fluid	prenatal diagnosis at about 15-20 weeks
Umbilical cord blood	prenatal diagnosis at about 18-24 weeks
ADULT/POSTNATAL	
Peripheral blood	screening for heterozygote carriers. Testing for defined heterozygous carrier genotype. Pre-symptomatic genotype screening or testing. Identity testing (DNA profiling). Testing for chromosome abnormalities
Mouthwash/buccal scrape	
Biopsy of skin/muscle/other tissue	RNA-based testing
Tumor biopsy	cancer-associated genotypes or gene expression patterns
Guthrie card	neonatal screening
ARCHIVED MATERIAL FROM DECEASED PERSONS	
Pathological specimens	genotyping
Guthrie card	possible source of DNA from a deceased individual (not all of the blood spots on the card might have been used up in neonatal screening)

Like other clinical tests, genetic tests may be conducted on individuals, couples, families, communities, or whole populations. They may be conducted at different levels for different purposes, whether informing about genetic risk at the prenatal diagnosis level, at preimplantation, at pre-conception (to inform a decision about future reproduction), at the level of managing an existing genetic condition, and so on. And the tests may be of different types, according to whether the object is to detect a *specific* pre-defined genetic abnormality for some purpose, or to *scan* for a variety of *possible* genetic changes, that then must be further investigated to see if any

of them represent a convincing pathogenic change. See [Figure 11.1](#) for a visual representation with some illustrative examples of the methods used.

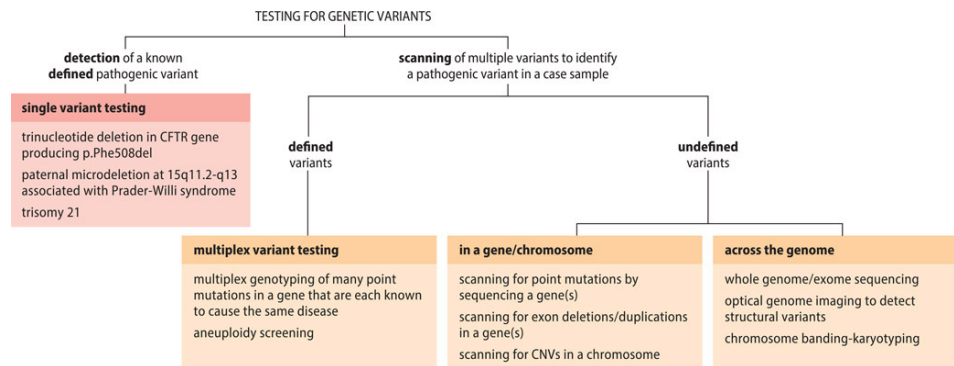


Figure 11.1 Some examples to illustrate how genetic testing is used to confirm or define pathogenic DNA variants. Detection of a single pathogenic variant may be conducted to confirm a previously defined genetic variant or abnormality present in another family member or suggested by the phenotype, or in management of a cancer. Alternatively, genetic testing is conducted to define an unknown causative mutation. If the disease gene locus and many previously identified causative mutations are already known, multiplex genotyping of several, or many, *known* causative mutations may be carried out to scan for the pathogenic variant in an affected individual. Other types of scanning for a causative mutation may be much broader because the pathogenic variants are undefined; here, the object is to define the causative variant. In cases where there is uncertainty over the disease locus, genome-wide scans can be carried out; they may identify many candidate pathogenic mutations that first need to be evaluated. CNVs, copy number variants of large sequences.

Genetic testing started in the 1960s with looking for chromosome abnormalities by examining stained chromosome preparations under the microscope. Linkage analysis and mutation testing for rare highly penetrant mutations within genes underlying monogenic diseases followed in the 1980s with more and more genes being identified, and thus testable, over the next few decades. These services were focused on advising others—existing family members or possible future ones via pregnancy investigations—once a disease or phenotype had come to light.

As techniques improved, the possibility of genetic testing was extended from the prenatal setting (from cells taken at different stages of pregnancy), to also include preimplantation genetic diagnosis (PGD). The latter is conducted in the context of *in vitro* fertilization: tests are carried out to decide which embryos lack the genetic abnormality being investigated and can thus be implanted. More recently, preconception testing for a panel of individually rare autosomal recessive conditions offers improved reproductive choices before any fertilization takes place.

Predictive genetic testing and genetic screening

Predictive genetic testing is increasingly offered to family members as more and more genetic diagnoses are made. Often, such tests are directed at members of a family with a history of a late-onset single-gene disorder, and the aim is to predict whether an asymptomatic member of the family is at risk of the disorder at some *future* point. A positive test result may offer the opportunity to take some medication and/or alter lifestyle factors in the meantime to reduce the disease risk; a negative test result provides reassurance that the predisposing genetic factor has not been inherited.

Genetic testing is also carried out on apparently asymptomatic individuals in communities and populations to identify individuals carrying harmful variants (*genetic screening*). The aim is usually to identify a high-risk subset of the population who can then be offered additional specific testing (such as follow-up prenatal diagnosis after identifying couples who are both carriers of a recessive condition). Unless the diseases are dominated by a few known types of genetic variant, we typically do not know what variants may be present in the individuals. As a result, genetic screening quite often involves assays of gene products or biomarkers associated with the pathogenesis, such as altered metabolites.

Direct versus indirect genetic testing

As previously summarized in Figure 7.1, genetic disorders arise through pathogenic DNA sequence changes or copy number changes. Direct genetic testing means that the test assays for the presence of a causative sequence or copy number mutation. For a very few disorders, exceptional mutational homogeneity allows us to predict the genotype, such as for sickle-cell disease and Huntington disease. But for the vast majority of single-gene disorders, if we do not have prior knowledge of the molecular pathology in a family, we first need to screen candidate genes to identify the causative mutation. After that, we can carry out a direct assay to determine the risks for relatives by seeing whether they carry the disease-associated variant or not.

To identify the causative mutation some type of *gene analysis* is carried out. Sometimes, multiple genes are analyzed, as for certain single-gene disorders (the Lynch syndrome case study profiled in Clinical Box 13 on page 394 gives an example), and for some common cancers. In response to a person enquiring about their family history of breast and ovarian cancer, for example, DNA from an affected relative can be assayed for point mutations/copy number variation in a panel of genes known to cause such familial disease. Candidate pathogenic DNA variants can then be investigated using various approaches described below to assess pathogenicity.

In *indirect genetic testing* the assay does not screen for the pathogenic variant directly; instead, it assays some other factor that is genetically linked to the variant or is a direct consequence of the variant. At the end of the day, a test is sufficient if it lets us know whether or not a particular gene variant is present even if we do not directly assay that variant.

In the past, indirect tests were often carried out using linkage analysis. Polymorphic DNA markers that mapped very close to a disease gene locus would be assayed in order to *infer* the inheritance of disease alleles using the same approach as previously illustrated in Figure 8.2. Such tests had small error rates because of the chance of recombination between marker and disease locus. With the sequencing of the human genome and the

development of rapid gene sequencing techniques, indirect linkage analyses have almost become obsolete.

Some indirect genetic testing assays some consequence of genetic variation, rather than the genetic variation itself. The testing might seek evidence of a gene product, or a characteristic disease-associated biomarker, such as an abnormally elevated metabolite. Sometimes a functional assay can be used. For recessive disorders, a single functional assay might be sufficient to detect a loss of function and can be conveniently carried out in cultured cells in which the gene is expressed. See [Table 11.2](#) for some examples of genetic testing that assay the consequence of genetic variation.

TABLE 11.2 EXAMPLES OF INDIRECT GENETIC TESTS THAT ASSAY SOME CONSEQUENCE OF GENETIC VARIATION

Assay level	Example
Gene product	Proteins: may be assayed by immunohistochemistry (as in Lynch syndrome in Clinical Box 13 on page 394). RNA: as in identifying translocations by targeted sequencing of transcripts of fusion genes
Genome instability	Microsatellite instability in Lynch syndrome. Because of a deficiency in mismatch repair, microsatellites across the genome show aberrant profiles.
Functional assay	Tests for various genes encoding enzymes may be assayed by enzyme assays. DNA testing for the DNA repair disorder Fanconi anemia is often conducted by an <i>in vitro</i> DNA repair assay (DNA detection is complicated because the disease can be caused by a mutation in any one of 15 different genes, some with large numbers of exons). Cultured lymphocytes from an individual are treated with a DNA interstrand cross-linking agent (often diepoxybutane or mitomycin C), and examined to identify chromosomal aberrations resulting from defective repair of the induced DNA cross-linking.

How genetic tests can be evaluated

Given the increasing availability in healthcare practice of genetic tests designed to identify a specific class of DNA change, standards for the assessment of their performance are important. Several frameworks for such evaluation have been proposed. The ACCE model, is one most referenced in the literature, proposes four main criteria:

- **Analytical validity:** how well does the test assay measure what it claims to measure?
- **Clinical validity:** how well does the test predict the projected health outcome?
- **Clinical utility:** how useful is the test result?
- **Ethical validity:** how well does the test meet the expected ethical standards?

The evaluation, particularly for predictive tests or tests for not very penetrant genes, can be more complex than implied by the four categories above. Rapid development and marketing may mean that there is not yet sufficient evidence to answer all these questions, and lack of public health/population level data may mean that ascertainment of particular test results selects for additional familial factors not measured by the test. Performing high-quality randomized controlled trials to demonstrate utility is often difficult, and the lack of evidence of effectiveness of a test may affect an evaluation of cost-effectiveness.

The ACCE process has been used by policy makers to decide about genetic testing for particular disorders using a standard set of 44 targeted questions (<https://www.cdc.gov/genomics/gtesting/acce/index.htm>) that address disorder, testing, and clinical scenarios, as well as analytic and clinical validity, clinical utility, and associated ethical, legal, and social issues.

The analytical validity of the test is determined by two key performance indicators, as follows:

- the **sensitivity**, the proportion of all people with the condition who are correctly identified as such by the test assay
- the **specificity**, the proportion of all people who do not have the condition and who are correctly identified as such by the test assay.

See [Table 11.3](#) for a worked example and for how related measures are defined.

TABLE 11.3 PARAMETERS RELATING TO THE ANALYTICAL VALIDITY OF A TEST

		CONDITION		Sensitivity	$a/a+b$	(90/100 = 90%)
		Present	Absent			
TEST	+ve	a (90)	c (30)	Specificity	$d/c+d$	(1870/1900 = 98.4%)
	-ve	b (10)	d (1870)	False positive rate	$c/a+c$	(30/120 = 25%)
				Positive predictive value	$a/a+c$	(90/120 = 75%)
				False negative rate	$b/b+d$	(10/1880 = 0.5%)
				Negative predictive value	$d/b+d$	(1870/1880 = 99.5%)

Numbers in parentheses are specific values for illustrative purposes only, drawn from 100 people with a hypothetical condition and 1900 lacking the condition. The false positive rate is the proportion of people who test positive for the factor being assayed but do not have the condition. The false negative rate is the proportion of people who test negative for the factor being assayed but who have the condition. The positive predictive value is the proportion of people testing positive who have the condition. The negative predictive value is the proportion of people testing negative who do not have the condition. Note that the sum of the false positive rate and the positive predictive value is always 100%, as is the sum of the false negative rate and the negative predictive value.

The ACCE process produces important, but under-evaluated, by-products. First, it identifies where the gaps in knowledge are in the natural history of a disease (which is important for future research agendas). Secondly, it can identify where the implementation gaps are. After the test findings are given, downstream recommendations can be made, such as screening or interventions. But of course we need to know about, and then reduce, the barriers to implementing these downstream recommendations, as well as effects of the recommended screening or intervention.

11.2 GENETIC TESTING FOR CHROMOSOME ABNORMALITIES AND PATHOGENIC STRUCTURAL VARIATION

In [Section 7.4](#) we detailed the two fundamental classes of chromosome abnormality (large-scale DNA changes that can be detected by standard karyotyping using chromosome banding techniques). They are: numerical abnormalities (in which abnormal chromosome segregation leads to aneuploidy, with fewer or more chromosomes copies than normal); and structural abnormalities (in which standard karyotyping by chromosome banding reveals chromosome rearrangements that produce large-scale deletions, duplications, inversions, or translocations).

Of course, disease can also be caused by structural abnormalities below the limit of detection of standard chromosome banding techniques. They mostly manifest as copy number variants (CNVs), comprising deletions and duplications from over 50 bp to a few Mb of DNA. They usually cause disease by eliminating one or more genes or by inactivating a gene as result of deletion or duplication or one or more exons or gene control regions. Various molecular genetic techniques can be used to screen/detect such DNA changes.

TABLE 11.4 AN OVERVIEW OF (A) MAJOR TECHNIQUES USED TO SCREEN FOR, OR CONFIRM, CHROMOSOME ABNORMALITIES AND PATHOGENIC LARGE COPY NUMBER VARIANTS, AND (B) ONLINE RESOURCES TO ASSIST IN THEIR INTERPRETATION

(A)	TECHNIQUE	APPLICATION
	standard karyotyping (chromosome banding)	General method for screening for chromosome abnormalities. Detailed in Box 7.2 on pp. 204–5. Often now used as a back-up method.
	quantitative fluorescence PCR (QF-PCR)	The front-line method to screen for the common aneuploidies in prenatal diagnosis.

* The older alternative of Southern blot-hybridization is virtually obsolete now, but still used in some labs to detect large deletions in facioscapulohumeral dystrophy – for an example, see Figure 2 in Clinical Box 3 on p. 171. It, and triplet repeat-primed PCR, can be used to detect very large expansion of repeats n disorders such as Fragile X syndrome and myotonic dystrophy.

chromosome SNP microarray analysis	The most commonly used type of chromosome microarray analysis. The method of choice for screening for large deletions and duplications across the genome. Also, a confirmatory method for screening for aneuploidies in prenatal diagnosis.
chromosome FISH (fluorescence in situ hybridization)	Often used to confirm regions of chromosome deletions/duplications identified by chromosome microarray analysis. Also used to screen for amplification of oncogenes (Figure 10.7A on page 377) and in detecting translocations, especially common oncogenic translocations.
RNA fusion screening	RNA fusion panels permit general screening for cancer-causing translocations (using targeted RNA sequencing to identify transcripts of many possible oncogenic fusion genes arising via translocation).
multiplex ligation-dependent probe amplification (MLPA)*	Especially used to test for gene variants where one or more exons are duplicated or deleted. Commercial kits are available for many genes, but this is not an easy method to carry out where kits are not available.
droplet-digital PCR (ddPCR)	A type of quantitative PCR. Offers highly sensitive accurate quantitation of CNVs. Highly versatile method. Needs dedicated PCR machine.

* The older alternative of Southern blot-hybridization is virtually obsolete now, but still used in some labs to detect large deletions in facioscapulohumeral dystrophy – for an example, see Figure 2 in Clinical Box 3 on p. 171. It, and triplet repeat-primed PCR, can be used to detect very large expansion of repeats in disorders such as Fragile X syndrome and myotonic dystrophy.

	genome-wide sequencing	A universal screen that can identify structural variants across the genome (as well as point mutations).
	optical genome mapping	A universal screen for structural variation. This new and different approach can be used to detect large and small structural variants, copy number variations, and complex rearrangements.
(B)	ONLINE RESOURCE TO ASSIST INTERPRETATION	DESCRIPTION
	Decipher database	At http://decipher.sanger.ac.uk . The Database of Chromosomal Imbalance and Phenotype in Humans using Ensembl Resources collects clinical information about chromosomal microdeletions, microduplications, insertions, translocations, and inversions, and displays this information on the human genome map.
	ClinVar	At https://www.ncbi.nlm.nih.gov/clinvar/intro/ . Maps relationships between human variations and phenotypes.
	dbVar	At https://www.ncbi.nlm.nih.gov/dbvar/ . NCBI's database of human genomic structural variation. Documents large variants (>50 bp), including: insertions, deletions, duplications, mobile elements, translocations, and complex variants.

* The older alternative of Southern blot-hybridization is virtually obsolete now, but still used in some labs to detect large deletions in facioscapulohumeral dystrophy – for an example, see Figure 2 in Clinical Box 3 on p. 171. It, and triplet repeat-primed PCR, can be used to detect very large expansion of repeats in disorders such as Fragile X syndrome and myotonic dystrophy.

In [Table 11.4 Part \(A\)](#) we give an overview of the principal techniques used to screen for chromosome abnormalities and large-to moderate-scale CNVs and their main applications. The interpretation of these tests are aided by various databases and online resources ([Table 11.4 Part \(B\)](#)).

Screening for fetal aneuploidies using quantitative fluorescence PCR

The commonest aneuploidies are trisomies 13, 18, and 21 and various types of abnormal sex chromosome number, and prenatal testing has been available for some time. Initial techniques required chromosome culturing so that results took a minimum of two weeks to grow the appropriate cells. More recently, quantitative fluorescence PCR has become the screening method of choice: it offers a much more rapid turnaround, often 24 hours, either by detecting abnormalities on fetal scanning, or increasingly through non-invasive prenatal testing (NIPT) offered to the population of pregnant women, as described below.

Quantitative fluorescence PCR (QF-PCR) is fast, robust, highly accurate, and largely automated. Several pairs of fluorescently labeled primers are used in a *multiplex PCR*—the idea is to simultaneously amplify multiple polymorphic markers on the chromosomes most frequently involved in aneuploidies. For each marker, the amplification products will fall within a characteristic size range of different lengths; as required, two or more markers that have overlapping allele sizes can be distinguished by labeling them using fluorophores that fluoresce at different wavelengths.

Certain polymorphic short tandem repeat polymorphisms are usually selected, often based on tetranucleotide or pentanucleotide repeats to maximize the length difference between alleles. Fluorescently labeled products from the *exponential phase* of the PCR reaction ([Figure 3.4](#)) are separated according to size by electrophoresis through long and extremely thin tubes containing polyacrylamide (capillary electrophoresis). That happens in a commercial DNA analyzer of the type used in capillary DNA sequencing: a detector at a fixed position records the intensity of

fluorescence signals as fragments migrate through the capillary tubes and past the detector ([Box 3.3](#) on page 73 describes the principle of capillary electrophoresis).

Autosomal aneuploidies

To monitor the common autosomal aneuploidies, the QF-PCR screen uses highly polymorphic short tandem repeat markers. An individual marker might not always be informative: in a trisomy, for example, the marker might show identical repeat numbers for all three chromosome copies, just by chance, resulting in an uninformative, single PCR product. The most informative situation occurs when the marker exhibits different numbers of repeats on the three chromosomes. But quite often only two length variants are recorded for a single marker; then quantification becomes important ([Figure 11.2A](#)). Because four or more different markers are used per chromosome, however, there is little difficulty with interpretation (two or more markers are often informative for each chromosome—see [Figure 11.2B,C](#) for a practical example).

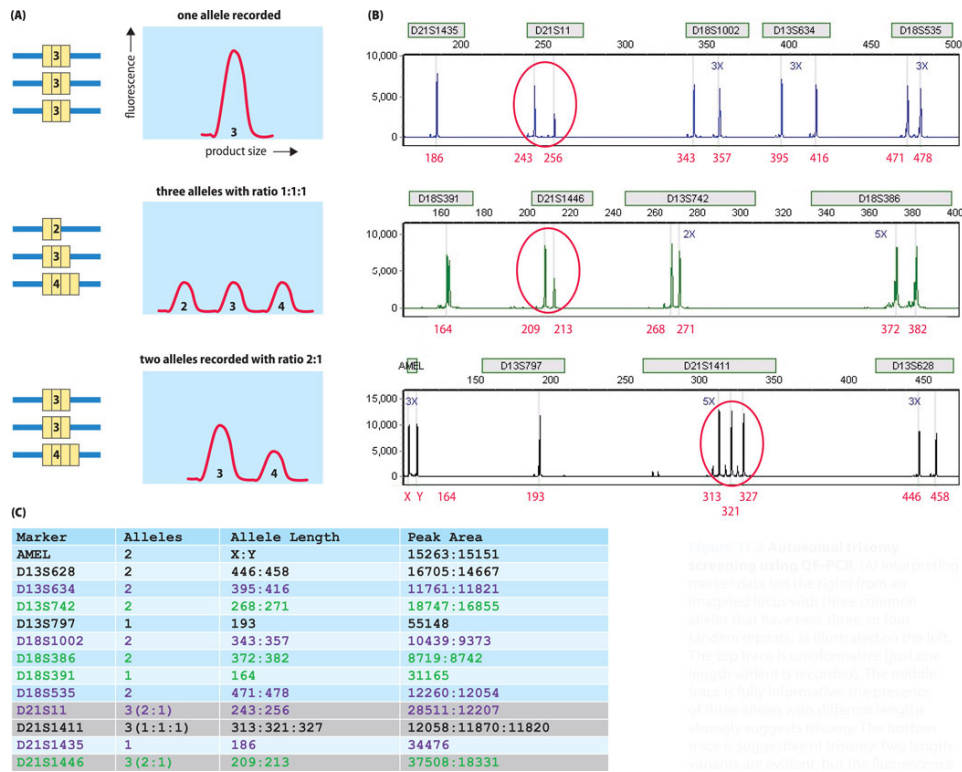
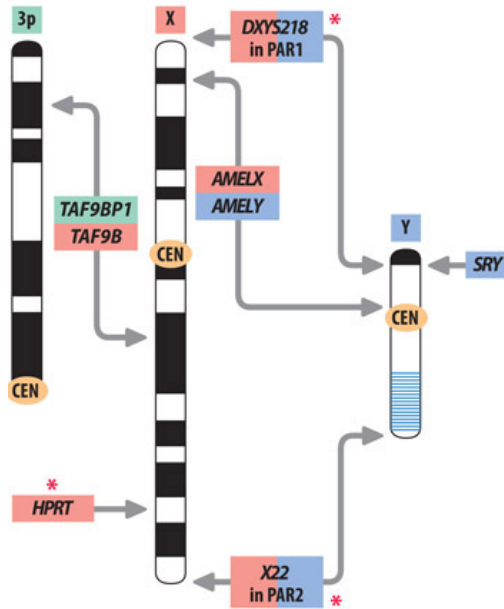


Figure 11.2 Autosomal trisomy screening using QF-PCR. (A) Interpreting marker data (on the right) from an imagined locus with three common alleles that have two, three, or four tandem repeats, as illustrated on the left. The top trace is uninformative (just one length variant is recorded). The middle trace is fully informative: the presence of three alleles with different lengths strongly suggests trisomy. The bottom trace is suggestive of trisomy: two length variants are evident, but the fluorescence associated with allele 3 seems to be approximately twice that associated with allele 4 (the area under the peaks is normally used for quantitation). (B) A practical example. The output shows traces for three sets of markers (shown in blue at top, green in the middle, and black at the bottom) that collectively represent assays for microsatellite markers on chromosomes 13, 18, and 21 (the three autosomes associated with viable trisomies), plus control X and Y markers from the amelogenin genes (see [Figure 11.3](#)). The data highlighted by red ovals strongly suggest trisomy in this individual: three alleles of different sizes for *D21S1411*, and a 2:1 ratio for the two length variants for each of *D21S11* and *D21S1446*. The other chromosome 21 marker, *D21S1435*, is uninformative, showing only one length variant, which is presumably due to three alleles of identical lengths. (C) The calculation of peak areas and interpretation by

SoftGenetics software (rows highlighted in gray are significant). (B,C, data courtesy of Jerome Evans, NHS Northern Genetics Service, Newcastle upon Tyne, UK.)

(A)



(B)

Marker	Alleles	Allele Length	Peak Area
D13S628	2	454.4:461.8	7680:7262
D13S634	2	391.4:406.3	7003:7181
D18S1002	2	341.8:354	6083:5638
D18S386	2	352.2:381.7	7291:5847
D21S1411	2	316.9:333.4	6558:6037
D21S1446	1	212.4	14090
DXYS218	1	239.5	7669
HPRT	1	282.0	6169
SRY	0		
TAF_9	3 (2:1)	3:X	40238:18508
X22	1	218	6680

Figure 11.3 Detecting sex chromosome aneuploidies using QF-PCR. (A) Marker sets. Primer pairs are designed to amplify X-specific markers (*HPRT*), Y-specific markers (*SRY*), markers in the pseudoautosomal regions PAR1 or PAR2 (shared by the X and Y), and highly homologous sequences on the X and Y chromosomes, such as the amelogenin genes *AMELX* and *AMELY* in which a single set of primers can amplify both sequences (which can be differentiated because of small length differences due to insertion or deletion). To gauge the ratio of X chromosomes to autosomes, primers are used to amplify equivalent segments of the *TAF9B* gene on Xq and a highly related

pseudogene *TAF9BP1* on 3p. CEN, centromere. Data from a practical example. The interpretation would be monosomy X (Turner syndrome) on the basis of the absence of the SRY marker and ratio of 2:1 for the length variants from *TAF9BP1* in chromosome 2 and *TAF9* on the X chromosome. (Data courtesy of Jerome Evans, NHS Northern Genetics Service, Newcastle upon Tyne, UK.)

Sex chromosome aneuploidies

The copy number of our sex chromosomes is more variable than that of auto-somes, ranging from monosomy (45,X) to different types of trisomy, tetrasomy, and occasionally pentasomy. Identifying a monosomy using PCR might seem particularly challenging—how can 45,X be distinguished from 46,XX? However, counting the sex chromosomes is possible by using primer sets specific for the X or Y chromosome plus primer sets that simultaneously amplify conserved sequences on both sex chromosomes or on both the X and an autosome (**Figure 11.3**).

Noninvasive fetal aneuploidy screening

Screening for fetal aneuploidies has been made possible by high-throughput sequencing of fetal DNA in maternal plasma, an advanced form of noninvasive prenatal screening. Because there is only a small number of viable human aneuploidies, it is relatively easy to design a series of QF-PCR assays for this purpose. We describe recent major advances in this area in [Section 11.4](#).

Detecting large-scale copy number variants using chromosome SNP microarray analysis

Standard chromosome-banding karyotyping will not detect deletions and duplications less than 6–10 Mb of DNA. However, once the human genome had been sequenced techniques rapidly followed that allow us to detect

smaller copy number variants. Chromosome microarray analysis uses oligonucleotide sequences from well-studied polymorphic loci across the genome to monitor polymorphism at hundreds of loci on each chromosome. We previously explained the principle underlying microarray hybridization in Figure 3.9 on p. 71 and accompanying text in [Section 3.3](#).

The most widely used type of chromosome microarray analysis uses single nucleotide polymorphism (SNP) microarrays. Test DNA samples from individuals are hybridized to microarrays containing oligonucleotides representing the different alleles at each of many thousands of SNP loci across the genome. SNP arrays do not directly compare a patient's test sample with a control sample. Instead, the assay compares the dosage of the individual being tested at any given locus with the equivalent values in a database of SNP array results from control individuals. Readouts of the SNP profiles across individual chromosomes allow us to detect gains and losses of sequences across the genome.

Deletions can be identified simply because of absence of heterozygosity across the deleted area: each SNP in the deleted area should show just a single allele. For duplications, the ratios of alleles will vary: if we imagine a SNP locus with two alleles, say A and B, a normal heterozygote would be scored as AB (with equal representations of alleles A and B); however, in regions of partial trisomy, loci in which both alleles are evident might show skewed allele ratios and might appear as AAB or ABB instead of AB (twice the signal for one allele compared to the other).

As well as identifying large deletions and duplications, SNP microarrays can identify regions of heterodisomy and isodisomy. That is useful for some conditions such as Prader-Willi syndrome (where uniparental disomy can be quite common, even if less frequent than paternal 15q11-q12 deletions), as illustrated in the case study profiled in [Clinical Box 15](#).

CLINICAL BOX 15 CHROMOSOME SNP ARRAY ANALYSIS TO IDENTIFY MIXED

HETERODISOMY/ISODISOMY IN A PRADER-WILLI SYNDROME CASE

David was the second child of unrelated parents Claire (aged 40) and Mike (aged 41). He was born by elective Caesarian section for breech presentation in good condition and weighed 2.8 kg. On Day 2 there was concern because he had never cried and was not interested in feeding, and so he was dependent on nasogastric tube feeds. He was examined by a pediatric neurologist who found him to have severe truncal hypotonia but normal tendon reflexes. She suspected Prader Willi syndrome and requested DNA methylation studies of the PWS critical region at 15q11-q13, which showed absence of the paternal allele, thus confirming the diagnosis. Follow-up chromosome SNP array analysis did not show any clinically significant copy number changes and excluded a 15q11-13 deletion, but revealed two large regions of homozygosity on chromosome 15 (see [Figure 1](#)).

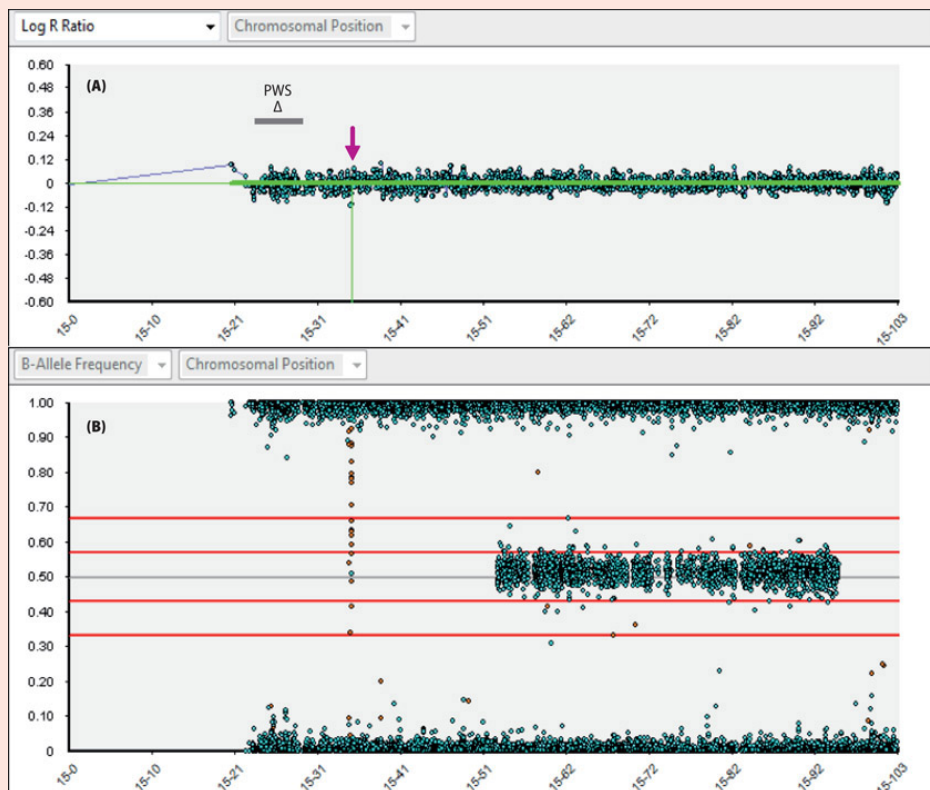


Figure 1 Chromosome SNP array analysis. The scale at the bottom of (A) and (B) represents the 103 Mb chromosome 15 from 15ptel (15-0) to 15qtel (15-103). (A) LogR ratio—the relative fluorescence of SNPs compared to SNPs with a diploid complement. SNPs showing copy number gain would have a positive value; those showing deletions would have a negative score. Almost all SNPs, including over the PWS deletion area (PWS Δ) show a zero logR ratio, indicating two copies (the arrow marks a nullisomic loss where the logR ratio drops off the scale). (B) B-allele frequency—the ratio of alleles calculated by $B/A+B$. In a diploid situation the SNP data points should be arranged in three rows, with homozygous SNPs at top and bottom (B-allele frequency of 1 and 0, respectively) and a central row showing SNPs that are heterozygous. Here, the heterozygote row is confined to just a part of the chromosome (roughly 42 Mb of DNA from 15–52 to 15–94), with two flanking regions of segmental absence of heterozygosity.

The most likely explanation is that this case represents uniparental disomy with a mixture of heterodisomy for the central region of 15q (from the 15–52 to 15–95 positions) and isodisomy (for the flanking regions). The conceptus initially had trisomy 15 because a maternal egg with two different chromosome 15s was fertilized by a normal sperm with a single chromosome 15. Non-disjunction of the two maternal chromosome 15s had occurred after a double crossover so that a central region of 15q was heterodisomic and flanked by isodisomic segments. Trisomy rescue ensued by loss of the paternal chromosome 15.

The family were seen by a clinical geneticist when David was three months old. He was by then able to be bottle fed and was gaining weight on high calorie feeds. He was still floppy but his head control was improving. David was referred to a pediatric endocrinologist and was treated with growth hormone from 10 months of age. At the age of 3 years his height and weight were on the 50th centile and by this stage he had not yet developed any food-seeking behavior. Maternal uniparental disomy for chromosome 15 accounts for 20–25 % of cases of PWS. It occurs due to

rescue of a trisomy 15 conception and the recurrence risk in a future pregnancy is negligible.

Detecting and scanning for oncogenic fusion genes using, respectively, chromosome FISH and targeted RNA sequencing

Balanced translocations and inversions are not detectable by chromosome microarray analyses because there is no appreciable loss or gain of DNA. Both can, however, be detected by traditional chromosome-banding karyotyping, but their contribution to pathogenesis is quite different. Inversions and constitutional translocations are rare, and make a very small contribution to genetic disease. Somatic translocations, by contrast, are common in many kinds of cancer, creating oncogenic *fusion genes* thought to account for ~20 % of human cancer morbidity.

The high overall frequency of somatic translocation in cancer occurs because translocations provide an opportunity for inappropriate oncogene activation. If translocation breakpoints occur in the immediate neighborhood of a protooncogene on one chromosome and close to an actively transcribed gene on the other chromosome, the proto-oncogene can be brought into close proximity to active transcriptional control signals and be inappropriately expressed as part of a fusion gene—Figure 10.8 on page 379 shows the example of *ABL1* activation in chronic myelogenous leukemia after fusing with the *BCR* gene. Oncogenic fusion genes like this are thought to account for ~20 % of human cancer morbidity, but the prevalence of fusion genes shows significant differences in different cancers, and many fusion genes are specific to certain types of cancer.

Traditional chromosome-banding karyotyping in metaphase chromosome preparations takes a minimum of two weeks to deliver results: cells need to be grown in culture and then a spindle poison added during periods of active growth to arrest cells in metaphase. Modern alternative methods, however, can permit rapid detection of fusion genes in interphase cells, such as by using chromosome **fluorescence *in situ* hybridization** (FISH) and targeted RNA sequencing of fusion gene transcripts.

The nature of chromosome FISH and its applications

The essence of chromosome fluorescence *in situ* hybridization (FISH) is to fix either metaphase or interphase chromosome preparations on microscopic slides, treat the slides so as to denature the DNA, and hybridize fluorescently labeled probes of interest to the denatured DNA—see **Figure 11.4A**. Obtaining metaphase chromosome preparations from blood samples takes time (because of the need for cell culturing), and the locations of the fluorescent signals are typically recorded against a background stain that binds to all DNA sequences.

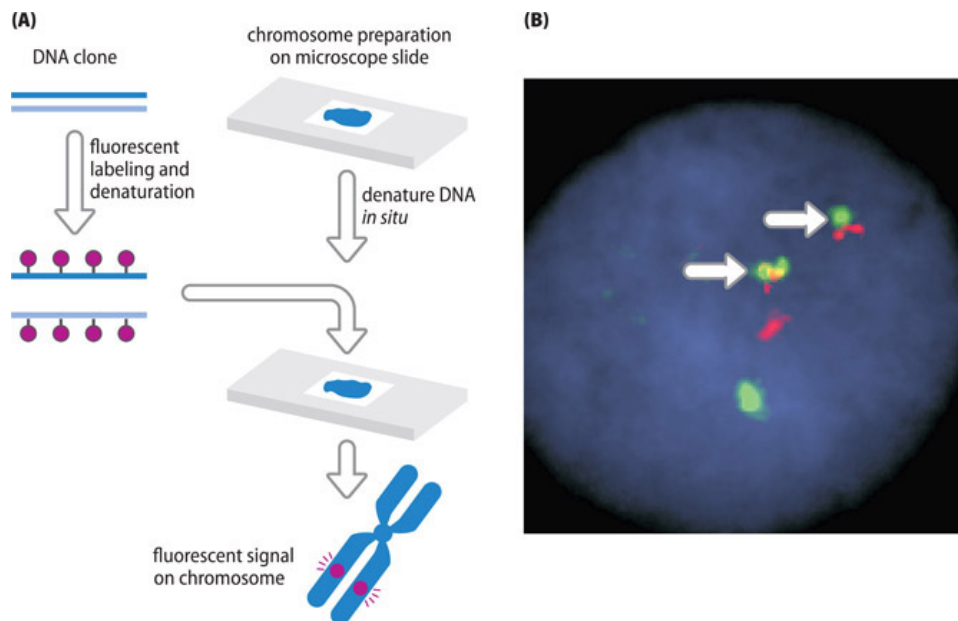


Figure 11.4 Principle of chromosome FISH and use in detecting recurrent translocations producing a fusion gene. (A) Principle of chromosome FISH. A labeled DNA clone of interest is hybridized to a denatured chromosome preparation on a microscope slide (either a metaphase chromosome preparation or an interphase chromosome preparation). When metaphase chromosome preparations are used, as shown here, a double fluorescent signal is often seen, representing hybridization to target sequences on the sister chromatids. (B) Interphase FISH to detect recurrent t(9;22) translocations in chronic myeloid leukemia (CML). Cases with CML often show translocations with breakpoints in the *ABL1* gene (on 9q) and the *BCR* gene (on 22q; see Figure 10.8A). Here, the *ABL1* and *BCR* probes give, respectively, red and green

fluorescent signals. The white arrows indicate characteristic color signals for the fusion genes: very close positioning of red and green signals, with sometimes overlapping orange-yellow signals. By contrast, the red and green signals at bottom are well separated and represent the normal chromosome 9 and normal chromosome 22, respectively. (Courtesy of Fiona Harding, Northern Genetics Service, Newcastle upon Tyne, UK.)

Chromosome FISH is often used to confirm regions of chromosome duplication or deletion that have been suggested by primary screening methods, notably chromosome microarray analysis. It can also be used to rapidly detect in interphase cells two types of oncogene activation in cancer:

- *gene amplification* in particular types of cancer, notably amplification of the *MYCN* gene in neuroblastoma—see Figure 10.7A on page 377)
- *fusion genes* generated by recurrent translocations. Chromosome FISH is notably used for detecting fusion genes in certain types of cancer that are strongly associated with recurrent translocations producing a specific fusion gene. For example, translocations associated with promyelocytic leukemia always produce *PML-RARA* fusion genes, and translocations producing *BCR-ABL1* fusion genes are common in chronic myeloid leukemia.

To detect specific fusion genes, interphase FISH can be carried out using probes representing the two genes that form the fusion gene (selected from regions retained on the translocation chromosomes). In the absence of visible chromosomes, probes from the two genes are labeled with different fluorophores, so that one produces a red fluorescent signal, for example, and the other produces a green fluorescent signal. The translocation chromosomes can readily be identified because here the green and red fluorescent signals are superimposed ([Figure 11.4B](#)).

Targeted RNA sequencing of fusion gene transcripts

The application of chromosome FISH to assaying oncogenic fusion genes arising from translocations is limited because the tests typically assay for the presence of a single fusion gene. They are therefore suited to detecting fusion genes frequently found in certain types of cancer, such as acute promyelocytic leukemia, chronic myeloid leukemia, and promyelocytic leukemia. Using chromosome FISH to carry out general screens for oncogenic fusion genes across a range of cancers would be very time-consuming and costly.

There is a considerable need for methods that can rapidly and simultaneously screen for a wide variety of fusion genes in cancers. And there can be considerable diagnostic value. For example, cells in different types of sarcoma appear very similar (“small blue round cells”), but different types of carcinoma—such as Ewing sarcoma, rhabdomyosarcoma and synovial sarcoma—are associated with particular types of recurrent translocation and specific oncogenic *fusion genes* that can therefore act as diagnostic aids.

In principle, RNA sequencing in cancer cells could allow genome-wide screens for fusion genes (in RNA sequencing RNA transcripts are first converted into double-stranded cDNA then sequenced). However, standard RNA sequencing does not have sufficient sensitivity to detect low expression signals from fusion genes (against the voluminous background expression signals from complex whole transcriptomes). To overcome the difficulty, *targeted RNA sequencing* has been developed using panels of specific biotinylated oligonucleotide probes to selectively enrich for sequences from RNA transcripts of interest. The same principle is used in the *targeted DNA sequencing* that selectively enriches for panels of gene sequences to screen for mutations; we will describe the method in detail in [Section 11.3 \(Box 11.2\)](#) when we consider methods of scanning for point mutations.

The current leader in commercial targeted RNA sequencing of fusion gene transcripts is the Illumina TruSight Pan-Cancer panel. It targets a total

of 1385 cancer genes for gene expression, variant and fusion detection, including detection of gene fusions with both known and novel gene fusion partners.

Detecting pathogenic moderate-to small-scale deletions and duplications at defined loci is often achieved using the MLPA or ddPCR methods

Structural variation includes moderate- to small-scale copy number variants (CNVs) where the sequences are >50 bp in length but below the limit of detection of chromosome banding-karyotyping, and often from hundreds of nucleotides to tens of kb in length. Two important classes of CNVs of this size range are listed below.

- *Inactivating intragenic deletions and duplications.* Large genes are often prone to intragenic deletions and duplications (often resulting from inappropriate pairing of repeat sequences). Pathogenesis may result from loss of important sequence (caused by deletions) or frameshifts (after deletion or duplication of one or more exons). Such copy number variation is typically assayed by multiplex ligation-dependent probe amplification (MLPA) as described below.
- *Unstable oligonucleotide repeat expansion.* PCR assays are often used to amplify the region containing the expanded repeat. The primers are labeled with a 5' fluorophore that gets incorporated into the product that can be detected after capillary electrophoresis. For very large expansions, however, more specialized triplet repeat-primed PCR assays are used— see below.

Various laboratory methods permit analysis of large CNVs. The versatile MLPA method is used widely to scan for exon deletions and duplications, but is limited by availability of commercial kits, as explained below. An alternative method that is rising in importance is a type of quantitative PCR known as droplet digital PCR (ddPCR). We give detailed descriptions of

MLPA and ddPCR later in this section. Two additional methods are used infrequently, and we describe them briefly immediately below.

Southern blot-hybridization is very rarely used now, although it continues to be used for identifying pathogenic large deletions in facioscapulohumeral dystrophy, as shown in Clinical Box 3 on page 170, and sometimes for very large expansions of unstable short tandem repeats. It involves digesting genomic DNA with suitable restriction nucleases, size-fractionating the resulting fragments by agarose gel electrophoresis, denaturing the DNA *in situ* in the gel, then transferring the DNA to a plastic sheet laid over the gel so that the fragments are located on the plastic in a mirror image representation of how they appeared on the gel. Thereafter, the denatured DNA on the filter is hybridized with a suitable labeled gene probe. Readers interested in the details can find an adequate explanation of the technique in the online encyclopedia Wikipedia.

Triplet repeat-primed PCR is a very specialized technique used to detect particularly large-scale expansion of unstable short tandem repeats, as may be seen in disorders such as myotonic dystrophy and Fragile X syndrome. The output of an application in a myotonic dystrophy case study is shown in Clinical Box 5 on page 196. Readers interested in the details of the technique can find them at PMID 9004136.

Multiplex ligation-dependent probe amplification (MLPA)

MLPA is a quick and versatile method that can detect copy number changes over a broad range of DNA lengths. It uses pairs of short single-stranded sequences (called *probes*, but quite distinct from hybridization probes). The probes are designed to bind to specific exons (or to other target sequences whose relative copy number we wish to determine). Each pair of probes is designed to hybridize collectively to a *continuous* target DNA sequence; that is, when they bind to the target DNA, the pair of probes align immediately next to each other. The gap between them can then be sealed

using DNA ligase to give a single probe that is complementary in sequence to the target sequence of interest ([Figure 11.5](#)).

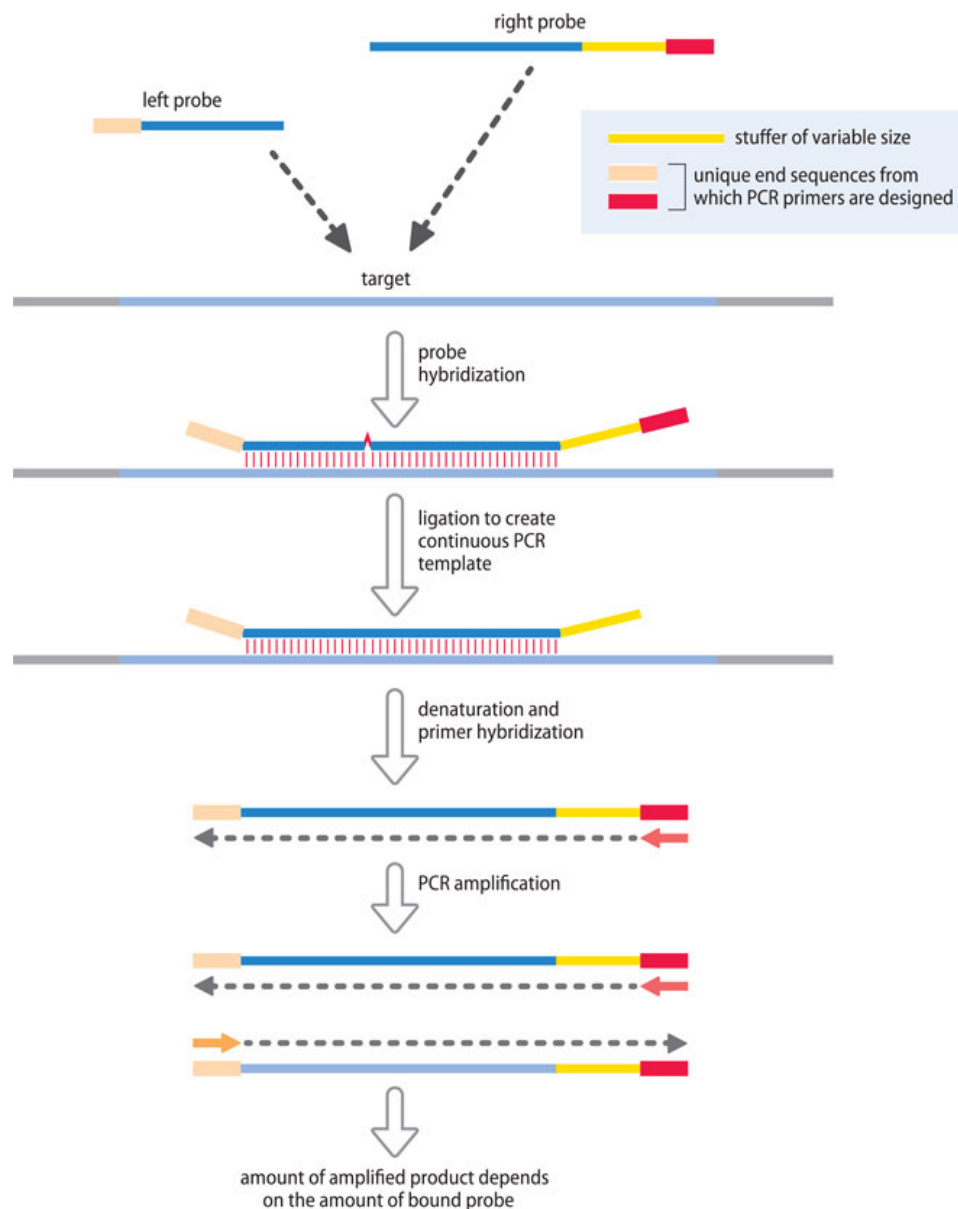


Figure 11.5 The principle of multiplex ligation probe amplification (MLPA). For each target sequence (such as an individual exon), a pair of probes is designed that will hybridize to adjacent sequences within the target and will carry unique end sequences not present in the genome. The aim is to use DNA ligase to seal the left and right probes to give a continuous sequence flanked by the unique end sequences and then to amplify the continuous sequence by using primers complementary in sequence to the unique end sequences. Probe pairs for multiple different target sequences (such as multiple exons

within a gene) are simultaneously hybridized to their target sequences and ligated to form continuous sequences that are then simultaneously amplified in a multiplex reaction. The point of the stuffer fragment is simply to provide a spacer sequence whose length can be varied. This can ensure different sizes of the PCR products from a multiplex reaction (in which multiple probe sets are used simultaneously) so that the products can be readily separated by capillary gel electrophoresis.

The 5' end of one of the probe pair and the 3' end of the other are designed to contain unique sequences not present in our genome. By designing oligonucleotide primers that will bind to regions in the unique end sequences only, the probe sequences are selectively amplified in a PCR reaction.

A key feature of MLPA is that the amount of amplified probe product is proportional to the number of bound copies of the probe, which in turn depends on the number of target sequences the probe has bound to. With a heterozygous deletion, for example, there is one copy of the target sequence instead of two; the amount of bound (and therefore ligated) probe is one-half of the normal amount, and the amount of amplified product is proportionally reduced.

Often, MLPA is designed to be a *multiplex* reaction: up to 55 pairs of probes can be used to bind simultaneously to different target sequences. The left and right probes for each target sequence all have the same set of left and right unique end sequences, and so all ligated probes can be amplified by a common set of primers (specific for the unique end sequences). But the stuffer sequences (described in Figure 11.5) are designed to be of different lengths for different probes, enabling the amplified probes to be physically separated by capillary gel electrophoresis and quantified independently. **Figure 11.6** gives a practical example of how MLPA can be used to screen simultaneously for large numbers of different exons. Useful YouTube video and text summaries of MLPA technology can be accessed from the MRC-Holland website at

<https://www.mrcholland.com/technology/mlpa/technique>

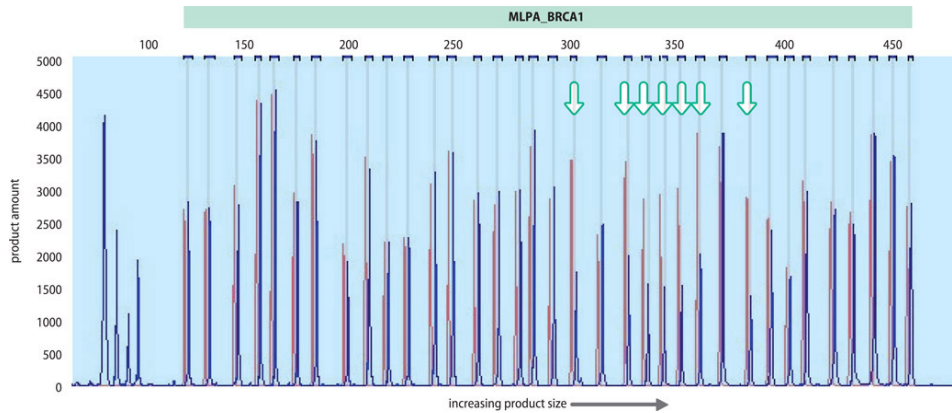


Figure 11.6 Using MLPA to scan for constitutional copy number changes in the exons of the *BRCA1* gene. MLPA scan. The blue peaks at the left from 0 to 110 bp on the horizontal size scale are internal controls. The paired blue and red peaks in the size range 125–475 bp represent comparative MLPA results in a normal control sample (blue) and a test sample (red) for individual exons of the *BRCA1* gene in most cases (however, for some large exons two partly overlapping probes were used). The test sample came from an individual with breast cancer in whom previous DNA sequencing investigations were unable to identify changes in the exons of the *BRCA1* gene. The MLPA analysis shown here identified a deletion that encompassed seven *consecutive* exons (marked by vertical green-outlined arrows)—in each case the blue peak is reduced by roughly one-half, as expected for a heterozygous deletion. Note: the order of the peaks is not the same as the order of the exons within the gene. (Data courtesy of Louise Stanley, NHS Northern Genetics Service, Newcastle upon Tyne, UK).

Droplet digital PCR (ddPCR)

Digital PCR (dPCR), like real-time PCR is a type of quantitative PCR requiring a specialist PCR machine. It offers precise absolute quantification (unlike real-time PCR, which requires reference standards to permit quantitation). A key feature is that individual DNA samples are extensively divided into very many smaller samples, each with limiting amounts of template DNA but with all the components required for PCR amplification of any template DNA. Like real-time PCR, it uses fluorescent probes.

In the popular droplet digital PCR (ddPCR) method, individual DNA in a standard PCR reaction mix are mixed with oil to create an emulsion and microfluidic devices are used to divide a single PCR reaction sample into 20 000 droplets that are individually distributed into microwells containing either zero or one (or at most a very few) template DNA molecules. Each droplet is effectively a PCR mini-reactor where PCR amplification and analysis occur separately from all the other droplets. After the reaction is over, the droplets are individually counted and scored as positive or negative for fluorescence, with application of Poisson statistical data analysis to enable highly accurate DNA quantification. A YouTube video explaining the method is available at https://www.youtube.com/watch?v=Qwma-1Ek-Y4&ab_channel=Bio-RadLaboratories

Two very different routes towards universal genome-wide screens for structural variation: genome-wide sequencing and optical genome mapping

As technology develops rapidly, so too does change. The current profusion of techniques for detecting structural variation may soon be replaced by universal screening systems that detect all types of structural variation. Currently, there are two contenders.

- *Whole genome sequencing (WGS)*. Because WGS is also used to identify point mutations we defer considering this until we cover genetic testing of point mutations in [Section 11.3](#). Suffice it to say that it can also detect structural abnormalities across the genome.
- *Optical genome mapping*. This is a very new and radically different way of scanning structural variation across the genome. Extremely long DNA fragments are first isolated from sample cells, labeled and imaged. Individual color images are then aligned with equivalent patterns from reference genomes to identify structural variation.

A pioneer in optical genome mapping is the Bionano company, whose Saphyr system has recently been released and is being adopted by many

health service laboratories. See [Figure 11.7](#) for an overview of the Saphyr system.

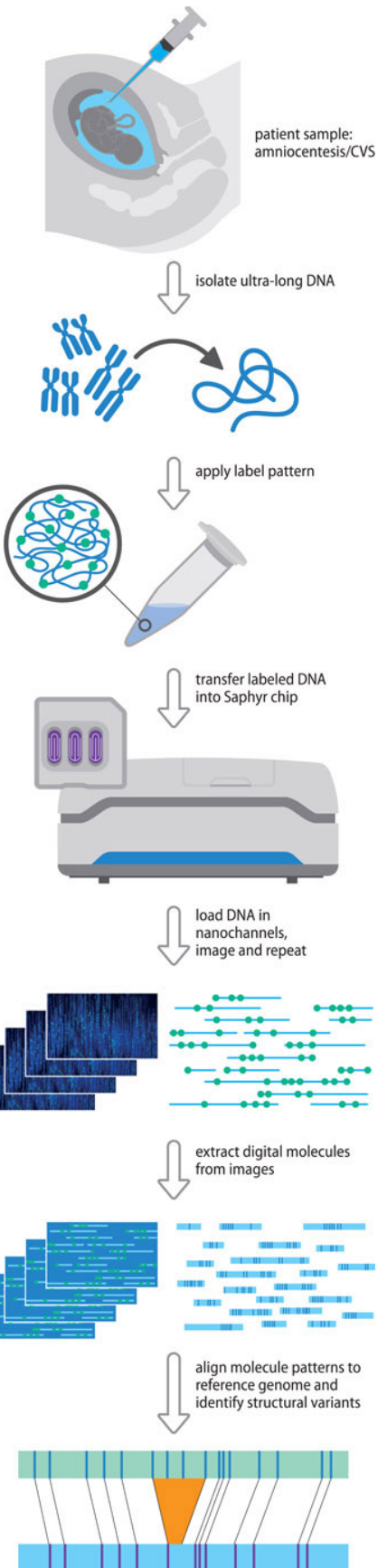


Figure 11.7 Prenatal workflow for optical genome mapping using the Bionano Saphyr system. DNA obtained from cells contained in amniotic fluid or a chorionic villus sample is labeled across the genome at a 6 bp motif by a labeling technology; the resulting label pattern is unique to each individual sample. Labeled ultra-long DNA molecules are loaded on a Saphyr chip, then electrophoresed into nanochannels where they are uniformly linearized for imaging in repeated cycles. The resulting images are processed to extract molecules containing the linear positions of sequence motif labels. Multiple molecules are used to create consensus genome maps representing different alleles from the sample. The sample's unique optical genome map is aligned to the reference genome and differences are automatically called, allowing for detection of structural variations in a genome-wide fashion. (Image modified from: <https://bionanogenomics.com>)

11.3 GENETIC AND GENOMIC TESTING FOR PATHOGENIC POINT MUTATIONS AND DNA METHYLATION TESTING

The most common pathogenic DNA changes are point mutations, mostly single nucleotide changes, or changes to a very small number of nucleotides. These changes can be detected by a range of different methods. Some long-established methods are designed to detect a specific mutation in a defined gene or to scan for any point mutation in a gene or panel of genes associated with a specific disease.

More recently, improvements in the speed and cost of DNA sequencing have brought radical changes to clinical settings. Instead of starting with a pheno-type, or family history of a phenotype, and considering which gene(s) should be analyzed, it is now often easier to perform genome-wide mutation scanning by whole exome sequencing or whole genome sequencing, and then apply the filtering at a later stage. **Next-generation sequencing (NGS)**, which encompasses a variety of sequencing technologies that use massively parallel DNA sequencing has driven that change (see [Box 11.1](#) for an overview). Sanger (dideoxy) sequencing,

previously described in Figure 3.10 on page 72, remains in use, but typically as a confirmatory technique.

BOX 11.1 MASSIVELY PARALLEL (“NEXT-GENERATION”) WHOLE-GENOME SEQUENCING

In standard dideoxy sequencing, individual DNA sequences of interest must first be purified; they are then sequenced, one after another. The sequencing involves DNA synthesis reactions, producing a series of reaction products of different lengths that are then separated by capillary gel electrophoresis ([Figure 3.10](#), page 72). By contrast, massively parallel DNA sequencing (often called next-generation sequencing) is indiscriminate: all of the different DNA fragments in a complex starting DNA sample can be simultaneously sequenced without any need for gel electrophoresis. That allows a vastly greater sequencing output

There are many different types of massively parallel DNA sequencing, but they can be separated into two broad categories: those in which the starting DNA sequences are first amplified by PCR, and those that involve single molecule sequencing (that is, sequencing of unamplified DNA molecules). We give details of run parameters for major commercially available technologies in [Table 3.3](#) on page 75.

Massively parallel DNA sequencing often involves *sequencing-by-synthesis*. That is, the sequencing reaction is monitored as each consecutive nucleotide is inserted during DNA synthesis. [Figure 1](#) shows the workflow that is involved in massively parallel sequencing and gives a simplified illustration of a popular form of sequencing-by-synthesis used by the Illumina company. Alternative methods are used by some other companies.

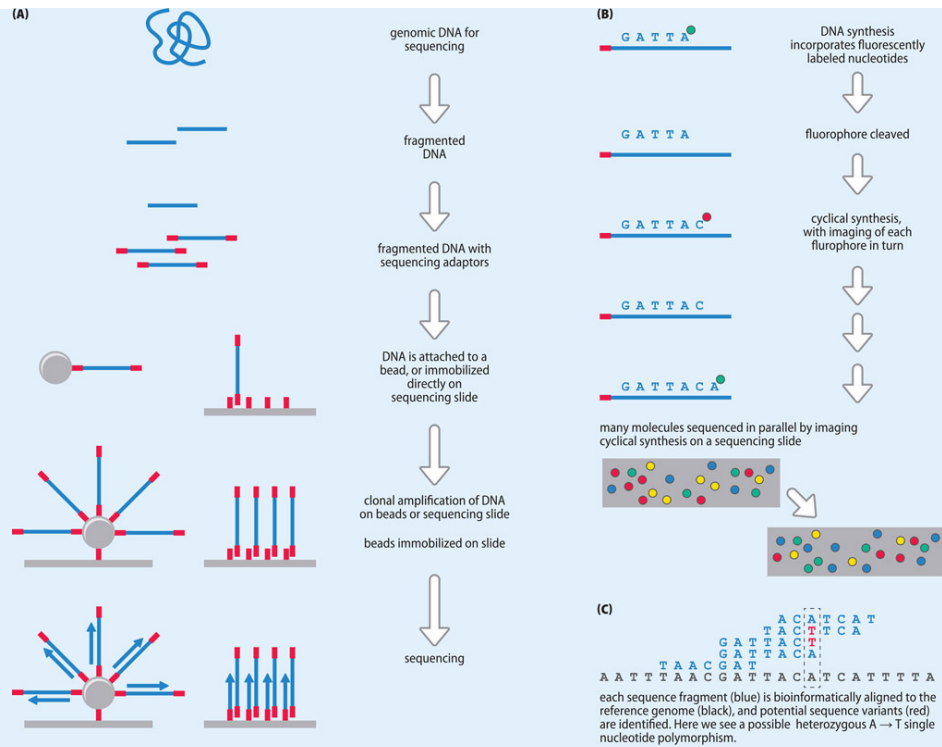


Figure 1 Next-generation sequencing workflow. (A) Genomic DNA is fragmented, and adaptor oligonucleotides are attached. The DNA is then attached either to a bead or directly to the sequencing slide. In either case, the DNA is clonally amplified in this location to provide a cluster of molecules with identical sequences. If beads are used, they are then immobilized on a sequencing slide. (B) The Illumina Genome Analyzer system of sequencing-by-synthesis. The sequence of each fragment is read by decoding the sequence of fluorophores imaged at each physical position on a sequencing slide. Advanced optics permit massively parallel sequencing. (C) Each DNA fragment yields one or two end sequences, depending on whether it is sequenced from one or both ends. These end sequences are computationally aligned with a reference sequence and mismatches are identified. (From Ware JS et al. [2012] *Heart* 98:276–281; PMID 22128206. With permission from the BMJ Publishing Group Ltd.)

We begin with a description of the range of techniques that can be used to identify specific *defined* pathogenic single nucleotide variants (SNVs). Then we consider how one can *scan* for undetermined pathogenic point

mutations, generally by using DNA sequencing at different levels. That started with standard Sanger dideoxy sequencing across exons and exon-intron boundaries plus promoter regions of a specific gene, but often now includes “gene panels”, amplified sequences from multiple genes that are obtained by *targeted DNA sequencing*. At a higher level, genome-wide scanning has begun to be used quite widely, beginning with whole exome sequencing, and more recently whole genome sequencing.

Thereafter, we move on to consider the problem of sequence *interpretation*. Sometimes trying to identify a pathogenic SNV in a single gene can be difficult, but the problem scales up massively when genome-wide sequencing is undertaken: there are huge numbers of candidate pathogenic variants to sift through, as described below. Although a battery of online resources can help us weed out less compelling variants, the task is nevertheless complex, especially for whole genome sequencing. Not only will many variants of uncertain significance arise in the sequence, but there are also ethical issues relating to the discovery of incidental findings that may raise information about additional health issues beyond the health concern for which the test was ordered.

We end this section with testing for DNA methylation changes. Such changes are especially important in cancers, but are also relevant to certain single gene disorders, notably those associated with imprinting defects.

Diverse methods permit rapid genotyping of specific point mutations

Instead of simply scanning for potentially pathogenic mutations, genetic tests may seek to identify a *specific* point mutation at a defined locus. That may be required for different reasons. A member of a family may be found to have a specific pathogenic mutation and there is interest in knowing if other members of the family have the same mutation, or if the mutation has been inherited in early pregnancy. At the population level there may be interest in screening carriers of common mutations such as the sickle cell mutation. And in cancers, tumor biopsies can be tested for the presence of

specific causative mutations to monitor minimal residual disease post-treatment, and to check for mutations that would govern the tumor's response to a targeted drug. DNA sequencing can be used to identify such variants, but it is often much more convenient to use simpler alternative detection methods.

The vast majority of pathogenic point mutations are SNVs, often occurring by substitution, but sometimes by insertion or deletion. Different methods can discriminate between the mutant and normal alleles. For mutations like these, a pair of allele-specific oligonucleotides (ASO) can be designed that represent mutant or normal allele sequences encompassing the sequence containing the point mutation, so that a mutant-specific ASO base pairs perfectly with the mutant sequence and a normal sequence-specific ASO base pairs with the corresponding normal allele. Normal and mutant ASOs that differ at a single base at a central position in the oligonucleotide sequence may allow allele-specific hybridization and that is the basis of SNP-chip hybridization that we described in [Section 11.2](#) and in GWA studies in [Chapter 8](#).

Microarray-based hybridization is very rarely used now for genotyping pathogenic point mutations. Instead, the genotyping quite often depends on designing the mutant ASO and normal ASO to have a single base difference at the 3'-terminal nucleotide, corresponding to the position and identity of the single-nucleotide change to be tested. The mutant ASO base pairs perfectly with the mutant allele; the normal ASO is specific for the normal allele sequence. When base pairing occurs between the normal ASO and mutant DNA, however, there is base mismatch at the 3' end nucleotide of the ASO (see [Figure 11.8A](#)); the same applies in the case of the mutant ASO binding to normal DNA.

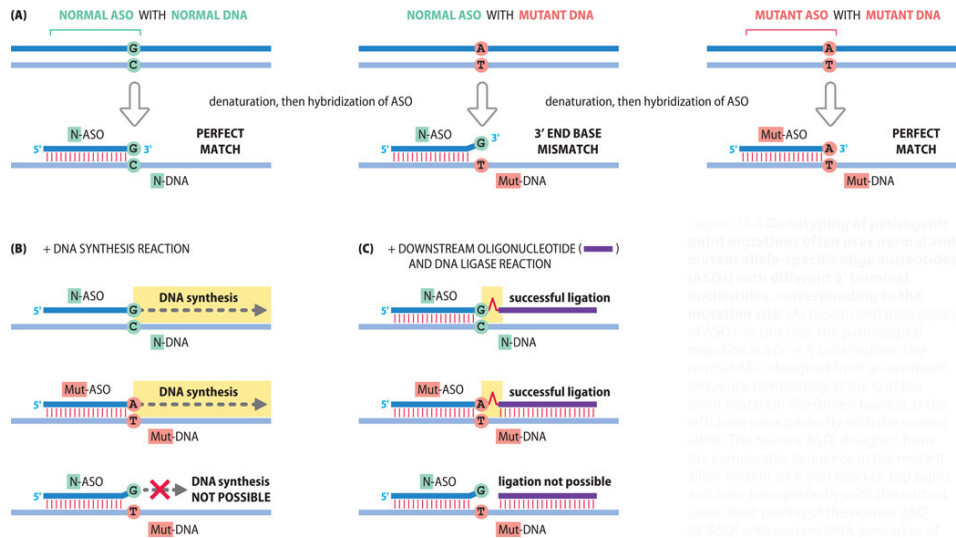


Figure 11.8 Genotyping of pathogenic point mutations often uses normal and mutant allele-specific oligonucleotides (ASOs) with different 3' terminal nucleotides, corresponding to the mutation site. (A) Design and base pairing of ASOs to assess the pathological mutation (A → G substitution). The normal ASO, designed from an upstream sequence terminating at the G of the point mutation site (green bracket at top left), base pairs perfectly with the normal allele. The mutant ASO, designed from the comparable sequence in the mutant allele ends in an A (red bracket, top right) and base pairs perfectly with the mutant allele. (B,C) Exploiting the base mismatch using follow-up DNA synthesis (B) or DNA ligation reaction (C). A normal ASO perfectly base-paired to normal DNA or a mutant ASO perfectly base paired to mutant DNA permits DNA synthesis using the ASO as a primer; but if there a 3' end base mismatch, DNA synthesis cannot be primed by the ASO. Similarly, an ASO perfectly base paired to the DNA can be ligated (indicated by the highlighted red chevron) to a downstream oligonucleotide that is base paired to an immediately adjacent downstream sequence, but an ASO with a 3' end base mismatch cannot be so ligated.

Figure 11.8 Genotyping of pathogenic point mutations often uses normal and mutant allele-specific oligonucleotides (ASOs) with different 3' terminal nucleotides, corresponding to the mutation site.

(A) Design and base pairing of ASOs. In this case the pathological mutation is a G → A substitution. The normal ASO, designed from an upstream sequence terminating at the G of the point mutation site (green bracket at top left) base pairs perfectly with the normal allele. The mutant ASO, designed from the comparable sequence in the mutant allele ends in an A (red bracket, top right), and base pairs perfectly with the mutant allele. Base pairing of the normal ASO (N-ASO) with mutant DNA (center) or of mutant ASO with the normal allele (not shown) results in a 3' end base mismatch. (B,C) Exploiting the base mismatch using follow-up DNA synthesis (B) or DNA ligation reaction (C). A normal ASO perfectly base-paired to normal DNA or a mutant ASO perfectly base paired to mutant DNA permits DNA synthesis using the ASO as a primer; but if there a 3' end base mismatch, DNA synthesis cannot be primed by the ASO. Similarly, an ASO perfectly base paired to the DNA can be ligated (indicated by the highlighted red chevron) to a downstream oligonucleotide that is base paired to an immediately adjacent downstream sequence, but an ASO with a 3' end base mismatch cannot be so ligated.

The difference in base pairing—whether an ASO is perfectly base paired or has a base mispaired—can be exploited in different ways. The *oligonucleotide ligation assay* exploits the base mismatch by following up with a DNA ligation reaction, one that assays the ability to ligate the bound ASO to another oligo-nucleotide designed to base pair to an immediately

adjacent sequence on the template DNA ([Figure 11.8C](#) shows the principle). This type of assay is used less frequently now. Instead, some of the more frequently used methods exploit base mismatching (sometimes at the 3' end and sometimes a central base mismatch) with a subsequent DNA synthesis reaction ([Figure 11.8B](#)). The methods listed below are three of the more commonly used for genotyping SNVs. In addition, mass spectrometry is used for genotyping and we describe that in the section on multiplex genotyping that follows this one.

- *Amplification refractory mutation system (ARMS)*. The ARMS method exploits the inability of an ASO with a 3' end-base mismatch to prime DNA synthesis in a PCR reaction (see [Figure 11.9](#)).

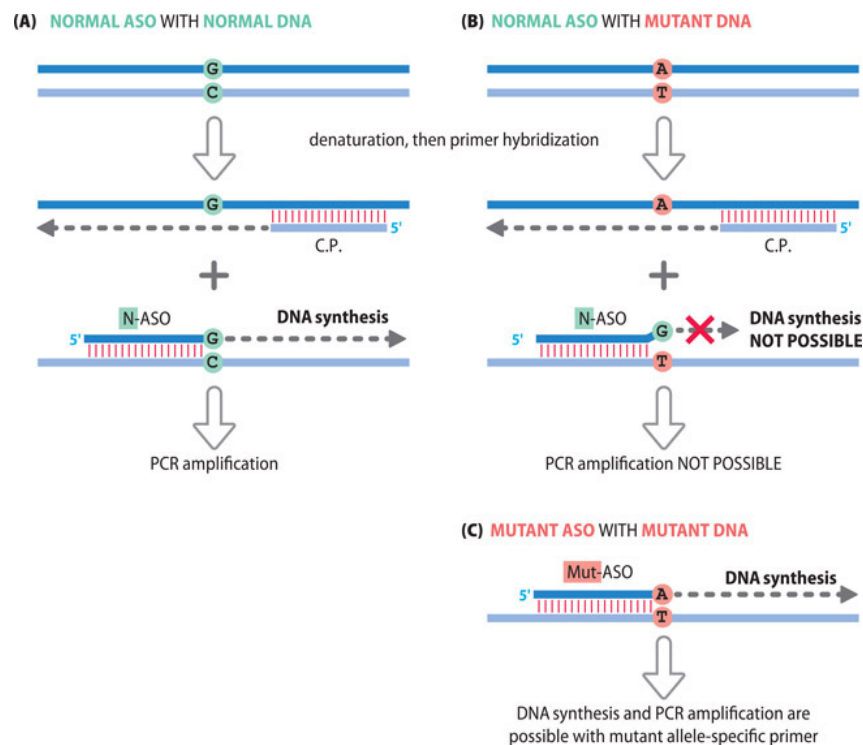


Figure 11.9 The principle underlying the amplification refractory mutation system (ARMS). In a PCR assay, when using a DNA polymerase lacking a proof-reading activity, the nucleotide at the 3' end of each oligonucleotide primer needs to be correctly base paired to the template DNA to allow DNA synthesis. (A) The normal allele-specific oligonucleotide

primer (N-ASO) terminating in a G at its 3' end will allow DNA synthesis (and consequent PCR amplification) because it is correctly base paired to the normal template DNA. (B) The mutant template DNA has a mutation with a single G to A difference. The N-ASO has its 3' terminal nucleotide at the nucleotide position corresponding to the mutation site. The lack of base pairing at the 3' end means that DNA synthesis cannot be primed. The same is true for mutant ASO with normal DNA (not shown). (C) However, if a mutant allele-specific oligonucleotide primer (Mut-ASO), terminating in an A at its 3' end, is used with mutant DNA, DNA synthesis can occur and PCR amplification is possible. Thus, the primer terminating in a G would be specific for normal alleles, and the primer terminating in an A would be specific for the mutant allele. C.P., common primer.

- **Real-time PCR using TaqMan genotyping.** An ASO with a central base mismatch is inadequate in priming DNA synthesis in a quantitative PCR reaction where amplification is tracked using a quantitative fluorescent signal (see [Figure 11.10](#)).

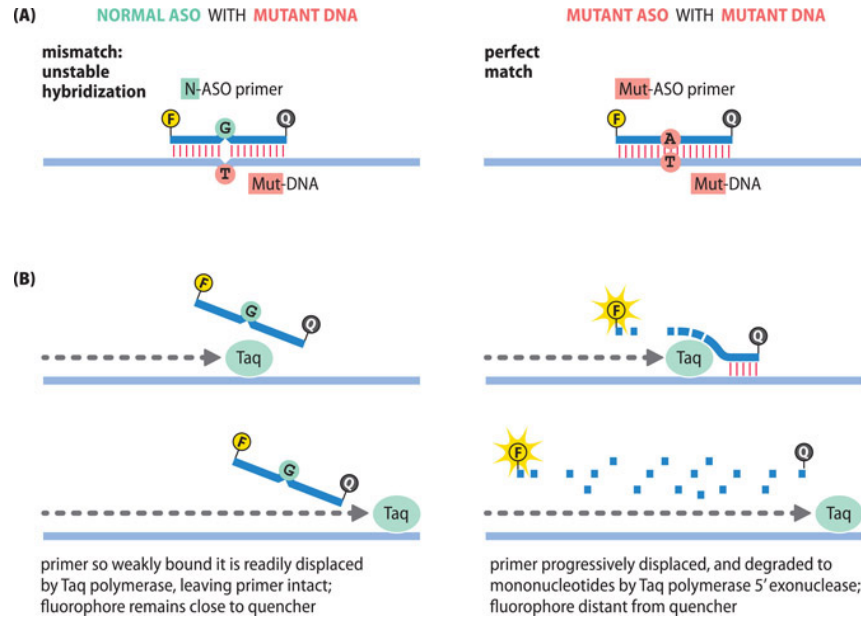


Figure 11.10 Genotyping point mutations using real-time PCR with the TaqMan™ double-dye system. (A) Primer hybridization. A central base mismatch between, for example, a normal allele-specific oligonucleotide (N-ASO) primer and the complementary mutant DNA strand, results in unstable

hybridization, unlike a perfectly base paired ASO. (B) DNA synthesis. In real-time PCR with TaqMan™ genotyping, DNA is synthesized with *Taq* polymerase and monitored continuously by fluorescence. That is possible because the ASO primers carry two different dyes: a 5′ fluorophore (F) and a 3′ quencher (Q). The fluorophore can emit strong fluorescence but when in close range to the quencher its fluorescence signal is suppressed. When a DNA synthesis step occurs during PCR, the *Taq* polymerase effortlessly displaces a weakly bound ASO primer with a central base mismatch, which is simply brushed aside. But a perfectly base paired primer is displaced progressively by *Taq* polymerase whose associated 5′→3′ exonuclease activity sequentially degrades the ASO into mononucleotides that are dispersed in solution, thereby liberating the fluorophore from the inhibitory quencher to cause fluorescence.

- **Pyrosequencing.** Exploits the inability of an ASO with a 3′ end-base mismatch to prime DNA synthesis in a single DNA synthesis step. The method is so-called because, just as in a sequencing reaction, it follows incorporation of an individual nucleotide in a growing DNA chain through the reaction: $dNTP \rightarrow dNMP + PP_i$ where PP_i represents the pyrophosphate released from the $dNTP$ to enable incorporation of a $dNMP$. The released pyrophosphate is used to drive a color reaction (see [Figure 11.11](#)).

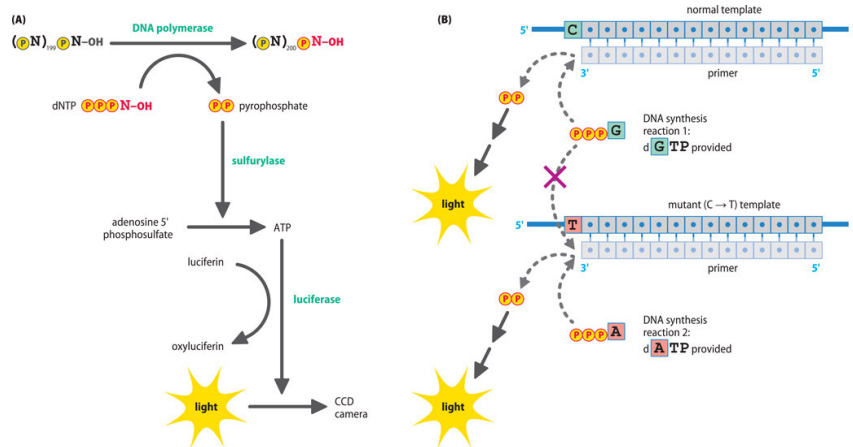


Figure 11.11 The principle underlying pyrosequencing. (A) Incorporating a nucleotide into a growing DNA chain requires cleavage of the $dNTP$

precursor and insertion of a dNMP residue. The remaining pyrophosphate is detected in pyrosequencing by a two-step reaction. First, ATP sulfurylase quantitatively converts pyrophosphate (PPi) to ATP in the presence of adenosine 5' phosphosulfate. Then the released ATP drives a reaction where luciferase converts luciferin to oxyluciferin, thereby generating visible light in amounts proportional to the amount of ATP. Each time a nucleotide is incorporated, a light signal is produced and recorded by a charge-coupled device (CCD) camera. (B) The dNTP precursors for DNA synthesis are provided *individually* and in a set order. If in the first reaction dGTP is the only nucleotide precursor, a G would be incorporated opposite the highlighted C in the normal template at the top, and a PPi residue would be released and trigger light production as shown in (A). But if a mutant DNA template were used (with T replacing C), no light would be produced in the first reaction (a G would not be inserted opposite the T residue). If, however, dATP is provided in a second reaction, an A would be inserted opposite the T in the mutant template, producing PPi and light, but no base incorporation would occur opposite the C in the normal template.

The advantages of multiplex genotyping

Some types of genetic disorder show very limited mutational heterogeneity. Sickle-cell anemia is the outstanding example: its very specific phenotype is always due to a nucleotide substitution that replaces a valine residue at position 6 in the β -globin chain by glutamate. Unstable oligonucleotide repeat disorders such as Huntington disease typically show a very limited range of mutations.

For most single-gene disorders, disease can be caused by any number of different causative point mutations. Nevertheless, certain pathogenic point mutations may be frequent in certain populations and contribute very significantly to disease. The *CFTR* mutation causing the p.Phe508 del cystic fibrosis variant is very common in populations of European origin, for example, as are two hemochromatosis-causing variants in the *HFE* gene that result in the C282Y and H63D amino acid substitutions.

Multiplex genotyping can be performed with the genotyping methods described in the previous section. Generally, that means genotyping dozens of point mutations at a time and so that can be used as a type of *mutation scanning* to see if the pathogenic mutation under investigation is one among a set of *known* pathogenic variants. Commercial kits based on the ARMS method have permitted testing for 50 common cystic fibrosis-associated mutations, covering ~90 % of the pathogenic mutations found in some populations of north European ancestry that have a high frequency of cystic fibrosis. Pyrosequencing and real-time PCR with TaqMan genotyping can also be used in multiplex genotyping.

Multiplex genotyping using mass spectrometry

In mass spectrometry, samples are ionized into charged molecules and the ratio of their mass to charge is measured. In MALDI-TOF mass spectrometry, the ionization occurs by **matrix-assisted laser desorption/ionization** and the mass analyzer is a **time-of-flight** analyzer. Genotyping by MALDI-TOF mass spectrometry typically uses the single-primer extension method to add a single chain-terminating nucleotide that discriminates between mutant and normal alleles. To do this, a primer is annealed to denatured template DNA so that it binds to a sequence terminating one nucleotide upstream of the SNV (single nucleotide variant) site. In the presence of DNA polymerase and chain-terminating dideoxy(dd)NTPs the primer is extended by one nucleotide. Mass analysis then permits genotyping by discriminating between the different reaction products according to their mass (see [Figure 11.12](#)).

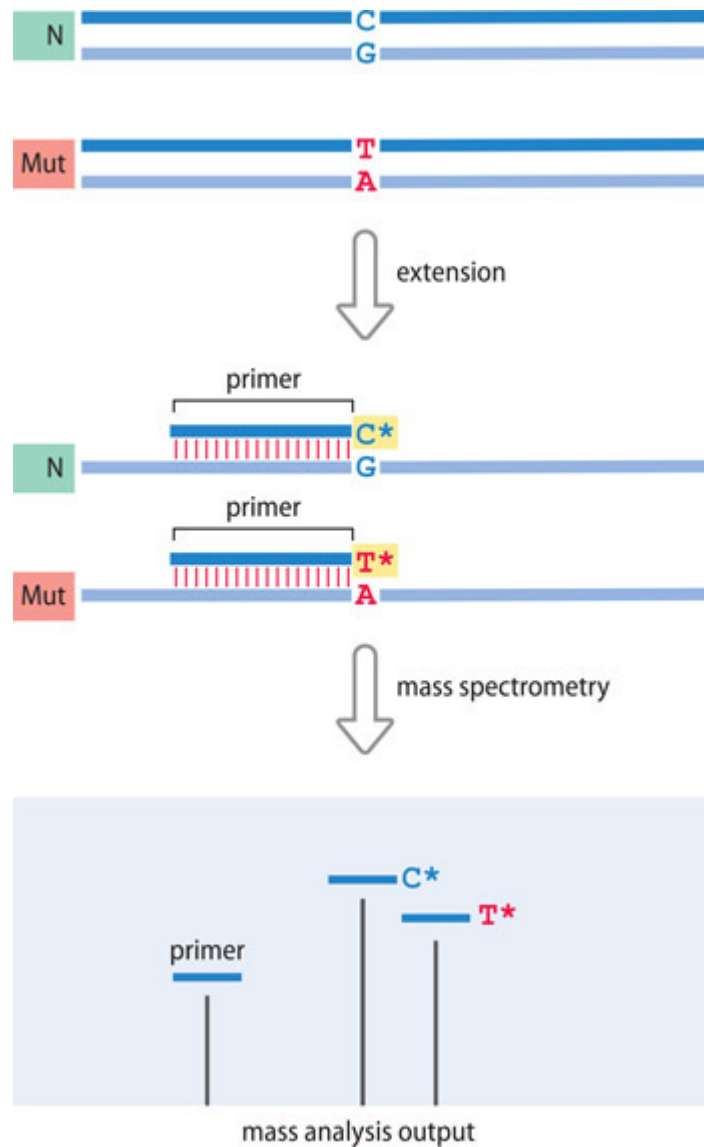


Figure 11.12 Genotyping single nucleotide variants (SNV) using mass spectrometry (MS). Imagine genotyping a heterozygote for a C>T substitution, using a primer designed to bind to the region immediately upstream of the SNV site. For the extension reaction, the DNA is denatured, and the primer, DNA polymerase and chain-terminating dideoxy(dd)NTPs are added, allowing incorporation of a *single* chain-terminating dideoxynucleotide, either ddC for the normal allele (shown as C*), or ddT for the mutant (shown as T*). On the basis of mass alone, the two reaction products, primer extended by C* and primer extended by T*, can be distinguished from each other and from the unextended primer.

Other genotyping methods can determine dozens of genotypes at a time, but MALDI-TOF mass spectrometry using a machine such as the Agena MassArray™ can genotype large numbers of variants. As a result, it has increasingly been used in diagnostic DNA services as an inexpensive way of carrying out genome-wide SNP analyses in *trio testing*, seeking to confirm biological relationships in two parents and affected child prior to carrying out more expensive whole exome or whole genome sequencing.

Mutation scanning: from genes and gene panels to whole exome and whole genome sequencing

As described in the previous section, mutation scanning may be carried out on a limited scale using multiplex genotyping of known pathogenic variants. But for the great majority of mutation scanning the object is to *define* a pathogenic variant whose identity may be difficult to suspect, and may never have been recorded previously. In some cases we may not even know the disease gene locus, and a genome-wide mutation scan may be needed, as described below. Quite often, however, we might wish to scan for mutations in a known disease gene, or in a *gene panel*, a group of genes associated with the same type of disorder. In that case, **targeted DNA sequencing** can be carried out: desired DNA sequences are captured from a genomic DNA sample by a DNA hybridization method and submitted for DNA sequencing. **Box 11.2** gives an overview.

BOX 11.2 TARGETED DNA SEQUENCING FOR MUTATION SCANNING

Targeted DNA sequencing means using a DNA hybridization method to capture desired DNA sequences from a genomic DNA sample so that they can be selectively sequenced, normally by massively parallel (“next-generation”) sequencing. The capture method relies on the extraordinarily high affinity of streptavidin, a bacterial protein, for the vitamin biotin (see [Figure 1](#)).

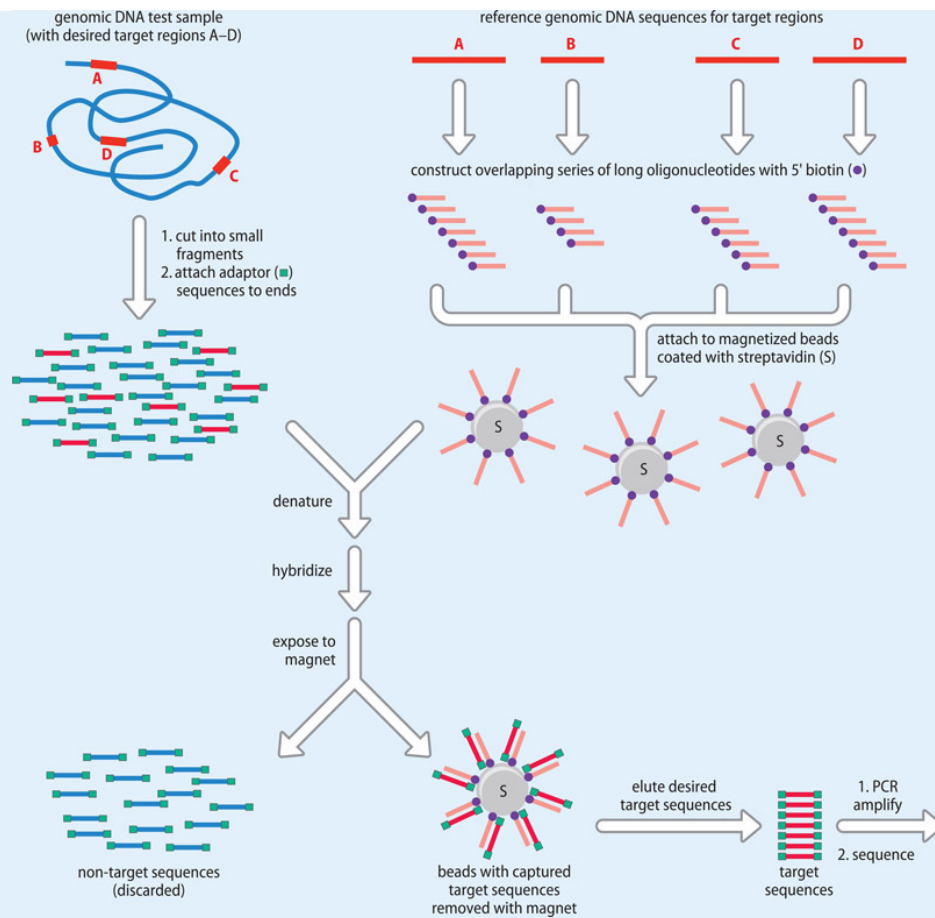


Figure 1 The principle of targeted DNA sequencing. In this example, four DNA regions, A–D, are targeted for sequencing. A series of partially overlapping oligonucleotides are synthesized to represent each of the desired target sequences (top right); each oligonucleotide has a biotin group covalently attached to its 5' end. The biotinylated oligonucleotides are then mixed with magnetized beads coated with streptavidin; the strong biotin-streptavidin affinity means that the oligonucleotides bind strongly to the beads. The genomic DNA sample is fragmented, denatured, and mixed with bead-oligonucleotide complexes. Sequences in the target regions of the genomic DNA sample (shown as red bars) will hybridize to complementary oligonucleotide sequences on the beads. Having fished out the target sequences the beads can be removed with a magnet, and the target sequences can be eluted, amplified, and sequenced.

Note: the target DNA sequencing method shown in Figure 1 can be adapted for **targeted RNA sequencing**. In RNA sequencing, RNA

transcripts isolated from cells are first converted to cDNA then sequenced. Targeted RNA sequencing involves converting total RNA to cDNA then capturing regions of interest using the method described in Figure 1. Targeted RNA sequencing is important for sequencing oncogenic fusion genes, as explained in [Section 11.2](#).

TARGETED DNA SEQUENCING TO SCAN FOR MUTATIONS IN GENES AND GENE PANELS

Targeted DNA sequencing focuses on genes of interest. Much of the sequence of small genes and genes with small introns may be captured as overlapping sequences, but for a gene with many large introns, the focus can be on capturing exons and the immediately flanking intron sequences (to maximize detection of splice-site mutations) plus known major regulatory sequences.

Rather than scan individual genes, *gene panels* are commonly used now, in which sequences are captured from multiple genes that are often implicated in the same disorder, or a collection of similar disorders, and then submitted to massively parallel DNA sequencing. They often represent monogenic disorders and common cancers, and can be narrowly or broadly focused. [Table 1](#) provides some examples; comprehensive lists of curated gene panels are publicly available at sites such as the PanelApp database at <https://panelapp.genomicsengland.co.uk> and the GenCC database at <https://thegencc.org/>.

TABLE 1 EXAMPLES OF HUMAN GENE PANELS USED IN TARGETED DNA SEQUENCING

Gene panel	Gene sequences
Breast cancer (common)	<i>BRCA1, BRCA2, PALPB2</i>
Lynch syndrome	<i>MLH1, PMS2, MSH2, MSH6</i> + 3'UTR of <i>EPCAM</i>
Familial hypercholesterolemia	43 genes*

Gene panel	Gene sequences
Retinal disorders	395 genes*
Illumina TruSight One panels	Up to 6700 genes associated with human disease in two panels

The advantages of gene panels are low costs, often better coverage—when designed well—of the genes of interest than a whole exome panel, and few variants to interpret, so that incidental findings (described below) are less troublesome.

Targeted DNA sequencing has also permitted a type of genome-wide sequencing; see the *whole exome and whole genome sequencing* subsection in the main text. *As listed in the Genomics England PanelApp at <https://panelapp.genomicsengland.co.uk/panels/>

Whole exome and whole genome sequencing

The Illumina TruSight One panels in [Table 1](#) of [Box 11.2](#) effectively represent a “*clinical exome*”. Another, more long-standing, application of targeted DNA sequencing is to use biotin-streptavidin capture and sequencing of a “whole exome” from a genomic DNA sample. Such a captured exome is artificial, designed to be mostly made up of coding DNA sequences (we have more RNA genes than protein-coding genes, and protein-coding DNA accounts for only 1 % of the genome; the bias towards coding DNA is justified on the observation that pathogenic point mutations are concentrated in coding DNA). In addition to coding DNA, however, captured exomes are designed to include some untranslated sequences, notably: short intronic sequences flanking exons (to catch more splice site mutations), many (known) regulatory sequences, and sequences specifying microRNAs.

The whole exome and whole genome sequencing approaches each have their advantages and disadvantages, as listed below.

- *Whole exome sequencing* (WES) is comparatively inexpensive, but it sometimes suffers from inefficient capture; some of the desired coding sequences may be missing.
- *Whole genome sequencing* (WGS) does not suffer from the disadvantage of missing sequences, and picks up structural variants as well as point mutations. It has been used extensively in cancer studies, but it is more expensive; and with so many variants to analyze, interpretation can be much more complex unless the data are filtered.

Virtual gene panels

The curated gene panels described in [Box 11.2](#) need to be periodically updated as new genes are added to the list or sometimes old ones are removed. An alternative approach uses *virtual gene panels*: WGS is carried out and then bioinformatic filters are applied to filter out most of the WGS dataset, retaining just the sequences of the genes of clinical interest. This approach can be expected to become more important as WGS costs fall.

Interpreting and validating sequence variants can be aided by extensive online resources

As described in the previous section, mutation scanning can be carried out at different scales. Currently, the trend is towards developing genome-wide mutation scans. Although comparatively cheap, the capture process in exome sequencing is, however, inefficient, and whole genome sequencing can be expected to replace exome sequencing once sequencing costs become sufficiently low.

Genome-wide scale comes with another cost: the number of validated variants that need to be filtered to arrive at high-probability pathogenic variants is large, about 20 000 for exome sequencing and significantly more for whole genome sequencing (but much less for many gene panels). The

process of interpreting, validating, and filtering sequence variants may consequently be very time-consuming.

To narrow down the choice of variants it may be profitable to focus on three major types of analysis: searching for precedent (by analyzing records of previously confirmed pathogenic variants), assessing sequence conservation (comparing against equivalent sequences in other organisms to draw conclusions on the functional importance of a nucleotide sequence, or often a derived amino acid that is predicted to be substituted), and determining rarity of the pathogenic variant—see [Figure 11.13A](#) for an overview.

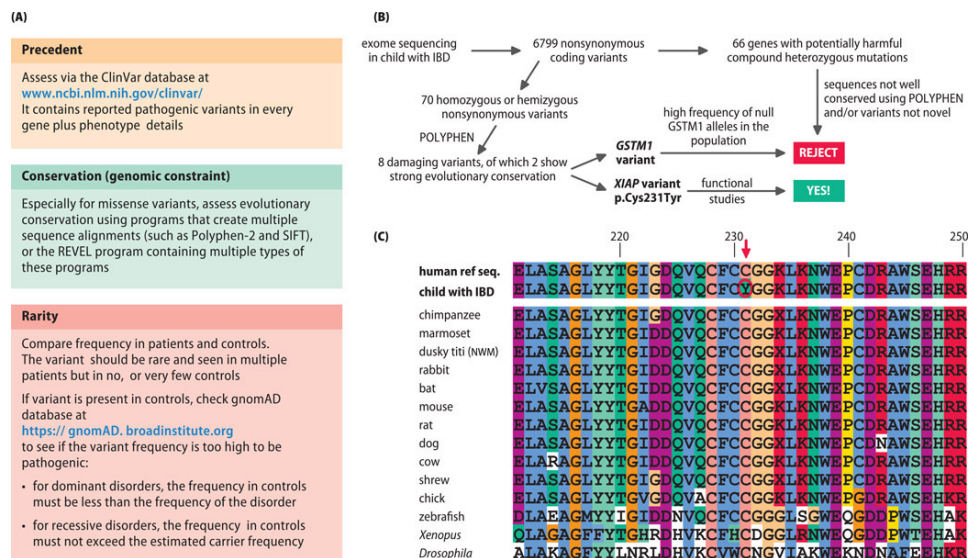


Figure 11.13 Sifting through DNA variants from a mutation scan to permit identification of a pathogenic variant. (A) Three major types of analysis. (B) An example from an early exome sequencing study in a young boy with recessive inflammatory bowel disease (IBD). POLYPHEN screening ultimately led to two strongly conserved novel missense variants. One could be excluded, because null alleles of that gene are frequent. The other, a maternally inherited variant in the X-linked inhibitor of apoptosis (*XIAP*) gene caused a p.Cys231Tyr substitution while the paternal allele carried a deletion. (C) The new *XIAP* variant appeared highly significant due to very strong evolutionary conservation of the Cys-231 amino acid, extending to *Drosophila*. The mutant *XIAP* protein showed loss of normal function in apoptosis and *NOD2* signaling, confirming it as causative. NWM, New World monkey. (Adapted from

Worthey EA et al. [2011] *Genet Med* 13:255–262; PMID 21173700. With permission from Macmillan Publishers Ltd.)

Evidence that a change is pathogenic usually comes from comparing the frequency of the candidate variant in patients and controls. Ancestrally matched control DNA samples may be used (but results are often compared with data stored in general DNA databases—see below). When sifting through variants that are also present in controls, the frequency is very important. Variants at high frequency in controls are very unlikely to be pathogenic, but the type of disorder and penetrance of mutations need to be taken into account. A sequence variant found in healthy male and female controls would usually be eliminated from consideration in a highly penetrant early-onset dominant or X-linked condition, but could contribute to disease in an autosomal recessive or a low-penetrance dominant condition.

Unlike easily identified and interpreted loss-of-function mutations, missense variants can be difficult to evaluate. A non-conservative substitution—replacing an amino acid by another of a different class—is more likely to be pathogenic than a conservative substitution. Conservative substitutions can, however, be pathogenic and nonconservative substitutions can be benign. But this is where evolutionary conservation studies can be very helpful. The concept is simple: if a sequence is functionally important there is pressure from natural selection to maintain that sequence. The sequence is subject to **genomic constraint**. Thus, if the normal amino acid is very highly conserved across a wide range of species, it is likely to be functionally very important, and a mutation producing a nonconservative amino acid change at this position becomes highly significant. Conversely, a substitution is unlikely to be pathogenic if it changes the amino acid to one that is the normal amino acid at an equivalent position in an ortholog from another species.

In Figure 11.13B,C we give an example of how conservation was particularly important in using exome sequencing to identify the genetic cause in a boy who presented aged 15 months with a life-threatening, but

previously unidentified, form of inflammable bowel disease. Because of the severity of the disorder at such an early age, a recessive disorder was expected. Exome sequencing identified 16 124 DNA variants when compared against the human genome reference sequence, with a total of 6799 substitutions that were then analyzed as shown in Figure 11.13B,C to identify a pathogenic missense mutation in the *XIAP* gene. The discovery led to a change in treatment.

If a variant, especially a nonconservative missense variant, arises *de novo* in an affected individual, its candidacy as a contributor to pathogenesis is also increased, especially if it is a nonconservative missense variant. We consider segregation of variants in families below.

Clinical reporting and nomenclature of sequence variants

Diagnostic laboratories generally report DNA variants suspected to be associated with pathogenesis in five categories as follows:

- Pathogenic
- Likely pathogenic
- Uncertain significance
- Likely benign
- Benign.

In reporting nucleotide and protein variants, the HGVS nomenclature is recommended, as described in [Box 11.3](#).

Standards and guidelines for the interpretation of sequence variants

The foundation for current interpretation of sequence variants was recently set by the American College of Medical Genetics and Genomics jointly with the Association for Molecular Pathology ([Richards et al. \[2015\]](#)) under

Further Reading). They proposed 16 criteria supporting pathogenicity (**P**), placing them in four groups according to the strength of the evidence thus: very strong (PVS) with one category, PVS1; strong (PS) with categories PS1 to PS4; moderate (PM) with categories PM1 to PM6; and supporting (PP) with categories PP1–PP5.

BOX 11.3 NOMENCLATURE FOR SEQUENCE VARIANTS

The nomenclature for sequence variants is described in detail in the HGVS (Human Genome Variation Society) website (<http://varnomen.hgvs.org/recommendations/general>). The computer program Mutalyzer (<https://mutalyzer.nl>) will generate the correct name of any sequence variant that a user inputs. The nomenclature for sequence variants requires a reference sequence, which may be one obtained from a database such as RefSeq, the NCBI's reference sequence database at <https://www.ncbi.nlm.nih.gov/refseq>. More recently, the Locus Reference Genomic (LRG) database <https://www.lrg-sequence.org/> has come to be commonly used, providing stable reference sequences for reporting sequence variants with clinical implications. The nomenclature for sequence variants has three main components, as follows:

1. A reference sequence shown by a recognized accession number followed by a symbol describing the type of sequence as follows: **g.** (nuclear genome); **m.** (mitochondrial genome); **c.** (coding DNA); **n.** (non-coding DNA); **r.** (RNA); **p.** (protein)
2. A number, or a range of numbers separated by an underscore, that defines the position(s) changed within the reference sequence
3. A description of the type of change. For nucleotide sequences: a substitution (>), deletion (del), insertion (ins), duplication (dup), a deletion-insertion event (delins), or an inversion (inv). For RNA sequence variants, nucleotides are shown in lower case. For proteins, the symbols fs, *, and ext indicate, respectively, a

frameshift, termination codon, and extension of the protein sequence.

The reference sequence may have an accession number recognized by sequence databases such as NM_000249.4, the *genomic* sequence corresponding to a major *MLH1* gene transcript, or it might be a Locus Reference Genomic database reference number, such as LRG_199t1 which is the genomic sequence corresponding to a primary transcript of the dystrophin gene, the Dp247m isoform (also present in the NCBI database with the accession number NM_004006.2). An example of a full variant sequence would be: LRG_199t1:c.79_80insG (= insertion of a G between nucleotides 79 and 80 of the LRG_199t coding DNA). [Table 1](#) lists some examples of the nomenclature for sequence variants, but omitting the reference sequence for simplicity.

TABLE 1 SOME EXAMPLES OF NOMENCLATURE FOR DESCRIBING SEQUENCE VARIANTS

Example	Interpretation
g.19C>A	substitution of a C by an A at position 19 in the genomic sequence
c.79_80delinsTT	nucleotides 79 and 80 of the coding DNA sequence are replaced by the dinucleotide TT
c.872_875del	deletion of nucleotides 872-875 in the coding sequence
c.*57C>G	replacement of C by G at nucleotide position 57 in the 3' untranslated region
c.178+9A>G	replacement of A by G at the ninth nucleotide within the intron that follows nucleotide number

Note that substitutions are confined to just a single nucleotide; if a CC dinucleotide at positions 79 and 80 were replaced by a TG that would be represented as 79_80 delinsTT, not as two separate substitutions at nucleotides 79 and 80.

Example	Interpretation
	178 in the cDNA (the last nucleotide of the preceding exon)
c.179-3C>T	replacement of C by T at third nucleotide preceding nucleotide 179 in the cDNA (the first nucleotide of the following exon)
p.Asp107His	aspartate at amino acid position 107 is replaced by histidine
p.Gly542*	the codon specifying glycine at amino acid (a.a.) position 542 is replaced by a stop codon
p.Arg123LysfsTer34	a variant with arginine 123 as the first amino acid shifts the reading frame, replacing it with a lysine and terminating after another 33 codons

Note that substitutions are confined to just a single nucleotide; if a CC dinucleotide at positions 79 and 80 were replaced by a TG that would be represented as 79_80 delinsTT, not as two separate substitutions at nucleotides 79 and 80.

As might be expected, the strongest criterion for pathogenicity, PVS1, is evidence of a null variant—such as a nonsense or frameshifting mutation, a change to the canonical GT and AG motifs at splice sites, single- or multi-exon deletions, or a change to the initiation codon—in a gene where loss of function is a known disease mechanism. Other strong criteria for pathogenicity include PS1 and PS2. In PS1, a nucleotide change is reported to give the same missense variant as one previously reported to be pathogenic but via a different nucleotide change. An example: a **GGA** glycine codon known to undergo a pathogenic G>A mutation to give the arginine codon **AGA**, is found in another affected individual to have been replaced by a **CGA** codon, also specifying arginine. PS2 covers a *de novo* mutation in a patient with disease but no family history, and with both paternity and maternity confirmed.

A total of 10 criteria for benign (**B**) variants were placed in three groups: stand-alone with one criterion (BA1); strong (BS) with four criteria (BS1–BS4), and supporting (BP) with five criteria (BP1–BP5). The strongest evidence for a benign variant, occurs when its frequency is exceptionally high: at >5 % in the population for criterion BA1 (as derived from global population variation data), or for BS1, at a higher frequency than could be expected for the disorder. The evidence framework for the criteria above is displayed in [Figure 11.14](#).

	BENIGN (B)			PATHOGENIC (P)		
	STRONG (BS)	SUPPORTING (BP)	SUPPORTING (PP)	MODERATE (PM)	STRONG (PS)	VERY STRONG (PVS)
population data	MAF is too high for disorder (BA1/BS1) OR observation in controls inconsistent with disease penetrance (BS2)			absent in population databases (PM2)	prevalence in affecteds statistically increased over controls (PS4)	
computational and predictive data		multiple lines of computational evidence suggest no impact on gene/gene product (BP4) missense in gene where only truncating cause disease (BP1) silent variant with non predicted splice impact (BP7) in-frame indels in repeat without known function (BP3)	multiple lines of computational evidence support a deleterious effect on the gene/gene product (PP3)	novel missense change at an amino acid residue where a different pathogenic missense change has been seen before (PM5) protein length changing variant (PM4)	same amino acid change as an established pathogenic variant (PS1)	predicted null variant in a gene where LOF is a known mechanism of disease (PVS1)
functional data	well-established functional studies show no deleterious effect (BS3)		missense in gene with low rate of benign missense variant and pathogenic missenses common (PP2)	mutational hot spot or well-studied functional domain without benign variation (PM1)	well-established functional studies show a deleterious effect (PS3)	
segregation data	nonsegregation with disease (BS4)		co-segregation with disease in multiple affected family members (PP1)	increased segregation data →		
de novo data				de novo without paternity & maternity confirmed (PM6)	de novo paternity & maternity confirmed (PS2)	
allelic data		observed in <i>trans</i> with a dominant variant (BP2) observed in <i>cis</i> with a pathogenic variant (BP2)		for recessive disorders, detected in <i>trans</i> with a pathogenic variant (PM3)		
other database		reputable source without shared data = benign (BP6)	reputable source = pathogenic (PP5)			
other data		found in case with an alternate cause (BP5)	patient's phenotype or FH highly specific for gene (PP4)			

Figure 11.14 Evidence framework for criteria classifying pathogenic or benign variants. MAF, minority allele frequency. LOF, loss of function. FH, family history. Path., pathogenic. See text for description of criteria. Note that in 2020 the UK's Association for Clinical Genomic Science (ACGS) released its best practice guidelines for variant classification in rare disease; these can be accessed at <https://www.acgs.uk.com/media/11631/uk-practice-guidelines-for-variant-classification-v4-01-2020.pdf>. Also note that some variants—including some missense mutations, some synonymous mutations with possible splice effects, nonsense and frameshifting mutations where the predicted termination codon lies close to the end of the coding sequence, and some intronic mutations within a few nucleotides of the intron

terminating GT and AG dinucleotides—may not easily be classified and are considered variants of uncertain clinical significance. We consider such variants of unknown significance later in the text. (Adapted from [Richards et al. \(2015\)](#) Genet Med 17:405–424; PMID 25741868 with permission from Springer Nature.)

Note the importance of increasing amounts of segregation data in Figure 11.14, and of *de novo* mutation. For people with a dominant disorder but unaffected parents, a *de novo* mutation has a generally higher likelihood of being pathogenic than an inherited mutation, especially if paternity and maternity are confirmed. If the disorder is familial, the mutation might be checked in other family members. If the mutation does not segregate with disease, it is highly unlikely to be implicated in the disease, assuming a high penetrance. But the reverse is not necessarily true: co-segregation with disease is not evidence that a variant is pathogenic (a nonpathogenic variant at a disease locus has a 50 % chance of residing within the same allele as the true disease-causing mutation, and a 50 % chance of co-segregating with disease). Segregation with disease may need to be studied through multiple meioses in a family.

***In silico* resources for describing and interpreting sequence variants**

We have described a few computational and database resources above, but in [Table 11.5](#) we summarize the range of major internet resources in this area. Note that despite the reliance on *in silico* resources, in some cases it may be necessary to validate an interpretation by laboratory analyses. Functional analyses that show a mutant protein does not carry out the normal functions can provide high confidence in pathogenicity when suitable analyses can be done, as in the example of Figure 11.13. Until recently, non-canonical splice site variants would need to be confirmed by reverse transcriptase-PCR analyses, but powerful new programs such as SpliceAI are providing greater confidence in predicted splice variants.

TABLE 11.5 MAJOR RESOURCES FOR DESCRIBING AND INTERPRETING SEQUENCE VARIANTS AND THEIR CLINICAL RELEVANCE

Resource	Description
ClinGen	At www.clinicalgenome.org . An aid to exploring the clinical relevance of genes and variants. Includes assessment of gene-disease validity; evaluation of gene dosage sensitivity, with the HI haploinsufficiency scores predictive model, based on functional, evolutionary and network properties; clinical actionability; curated variants. Reviewed at PMID 26014595.
ClinVar	At www.clinvar.com . Gives reports of the relationships between human variants and phenotypes, with supporting evidence. Reviewed at PMID 26582918 and 31777943.
gnomAD	At https://gnomad.broadinstitute.org/ . The Genome Aggregation Database (gnomAD), the major resource on human genetic variation, was constructed by aggregating international exome and genome sequence data. It has a gene/missense constraint track with pLI and Z-scores (a measure of constraint at the gene level; for a missense variant, a Z score >3.09 is deemed significant, indicating that it is intolerant to variation). The pLoF program measures a transcript's intolerance to loss-of-function (LoF) variation, with observed/expected (oe) values presented, low values indicating constraint against variation.
HGMD	At http://www.hgmd.cf.ac.uk/ac/index.php . The Human Gene Mutation Database collates published gene lesions responsible for human inherited disease. Available as a free version and as a professional version requiring a subscription. Reviewed at PMID 32596782.
Mastermind Professional	At https://www.genomenon.com/mastermind/ . Powerful publication resource covering genetic variations (and disease-gene associations). Enables graphical and text

Resource	Description
	querying to scan the literature for pathogenic variants, and assists the sensitivity and reproducibility of clinical variant interpretation. Requires a subscription. Reviewed at PMID 33281875.
REVEL	Freely available at https://sites.google.com/site/revelgenomics . Powerful software that predicts pathogenicity of missense variants based on combining scores from PolyPhen-2, SIFR, PROVEAN and 10 other computational tools. Reviewed at PMID 27666373.
Alamut Visual Plus™	At https://www.interactive-biosoftware.com/alamut-visual-plus/ . Proprietary software that curates data from multiple sources (ClinVar, dbSNP, COSMIC, Mastermind, PubMed) plus offers both high-quality missense predictors and splicing predictors (including SpliceSiteFinder-like, MaxEntScan and others) in one place.
SpliceAI	At https://github.com/Illumina/SpliceAI . This deep neural network method can efficiently model mRNA splicing from a genomic sequence and can accurately predict noncoding cryptic splice site mutations in patients with rare genetic diseases—see PMID 30661751.
Coding Constrained Region	Measure of regional missense constraint derived from gnomAD data. Data available at https://github.com . See PMID 30531870 for the map of constrained coding regions (CCR) in the human genome.

Incidental findings and variants of uncertain clinical significance

Genome-wide mutation scanning—whether at clinical exome, whole exome or whole genome levels—is especially prone to some problems that can provide clinical and/or ethical difficulties. One is *incidental findings*, where medically important findings are made that are unrelated to the medical reason for which the genetic test was ordered. Mutation scanning in a patient with a heart defect may disclose, for example, a harmful mutation in a hereditary cancer gene. We consider this when we look at ethical issues related to genetic testing in [Section 11.5](#).

Another issue concerns how to manage *variants of uncertain clinical significance* (VUS) that become a real issue in genome-wide sequencing. Each one of us will have a large number of VUS—see the legend to Figure 11.14 for some examples of common classes of VUS). While most VUS might be expected to be benign, reporting them can cause great anxiety to the patient, leaving both patient and referring physician with unanswerable questions. If they do not get reported, however, the chance of revisiting them at a later stage, when more is known, goes away. That can be undesirable: should there be a pathogenic VUS, responsive action might have been able to be taken to prevent or ameliorate future clinical symptoms.

Detecting aberrant DNA methylation profiles associated with disease

Aberrant epigenetic changes are heavily involved in the development of cancers. As detailed in [Section 6.3](#)), they also make important contributions to several inherited disorders. Often, they occur in response to some genetic change. The inherited disorders that show aberrant DNA methylation include notably imprinting disorders (according to the sex of the transmitting parent, alleles at imprinted gene loci are subject to epigenetic silencing and hypermethylation). In the case of disorders such as Prader-Willi syndrome and Angelman syndrome, for example, testing for aberrant DNA methylation is the front-line molecular genetics test to confirm the diagnosis. That is so because essentially all affected individuals show

aberrant cytosine methylation profiles for the relevant chromosome region (15q11-q12 in both cases), whether there has been a microdeletion on the normal chromosome or uniparental disomy for that chromosome.

Different methods can be used to detect aberrant DNA methylation. Bisulfite PCR methods are inexpensive and often used as a quick way to exclude negative cases for Prader-Willi and Angelman syndromes. MS-MLPA, a methylation-sensitive variant of the MLPA method that was described in [Section 11.2](#), allows simultaneous semi-quantitative detection of the methylation status of genes and their copy number.

Bisulfite sequencing

Methylated and unmethylated cytosines can be distinguished by making the DNA single stranded and treating it with sodium bisulfite (Na_2SO_3). Under controlled conditions, the unmethylated cytosines are deaminated to produce uracils, but 5-methylcytosines remain unchanged. After treatment with sodium bisulfite, the relevant region can be amplified by PCR, during which newly created uracils are read and propagated as thymines. New DNA strands are synthesized without incorporating methyl groups so that any retained methylated cytosines in the template DNA are propagated as unmethylated cytosines. That allows different ways of distinguishing the methylated cytosines from the original unmethylated cytosines.

Figure 11.15 shows how, after treatment with sodium bisulfite, samples can be amplified by PCR and sequenced to distinguish methylated cytosines from unmethylated cytosines. Methylation-specific PCR assays can also be devised by designing alternative PCR primers to have 3¢ nucleotides that are specific for one of the variable nucleotides after treatment with sodium bisulfite (that is, a U or T versus a C).

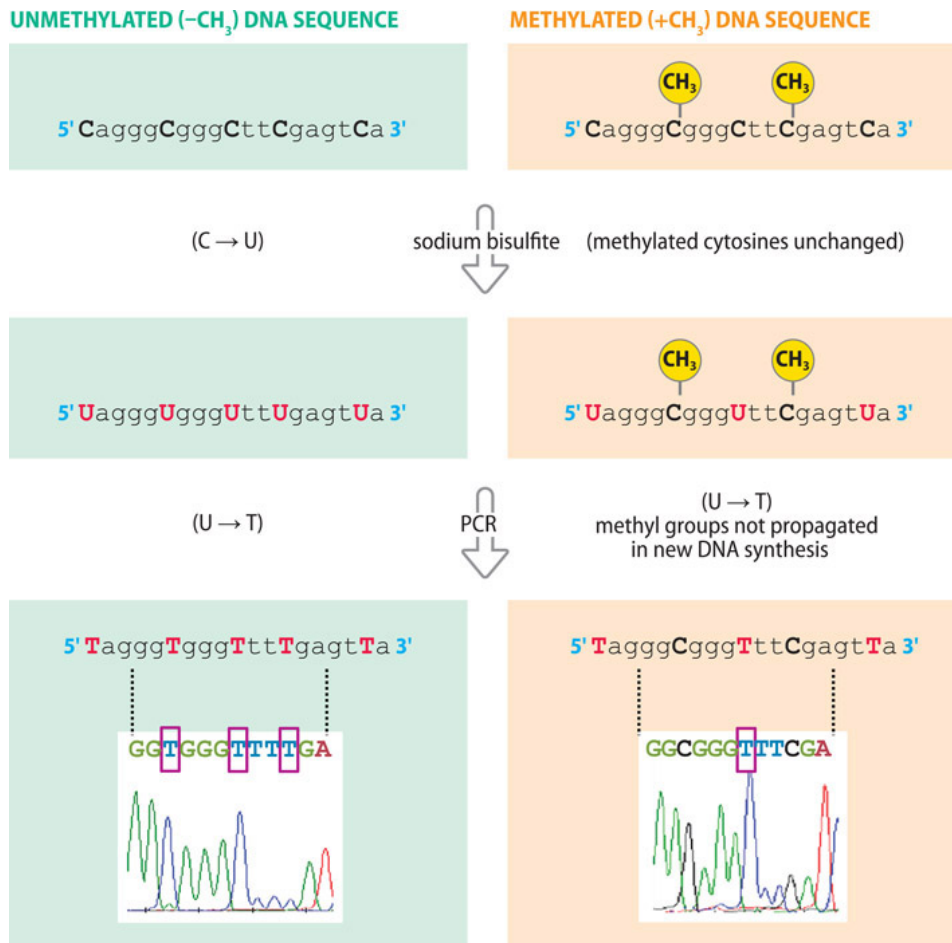


Figure 11.15 Distinguishing methylated cytosines from unmethylated cytosines with the use of sodium bisulfite. Sodium bisulfite converts unmethylated cytosines (left panel) to uracils; after DNA replication in a PCR reaction, they become thymines in newly synthesized DNA. Sodium bisulfite does not react with methylated cytosines, which remain unchanged (right panel). Newly synthesized DNA strands in a PCR reaction are not methylated, and so although the starting DNA is methylated in the right panel, the PCR product is unmethylated. DNA sequencing can identify all unmethylated cytosines because after treatment with sodium bisulfite each unmethylated C becomes a T (shown by boxes in sequencing panels at the bottom); if the cytosines are methylated, the sequence obtained is the same as in DNA that has not been treated with sodium bisulfite. DNA sequencing of PCR products is therefore one way of distinguishing between the two patterns. Alternative assays use methylationspecific PCR by designing primers with a nucleotide at the 3' end that corresponds to a variable site, pairing with either U/T (from an unmethylated cytosine that has been chemically converted by

sodium bisulfite) or C (representing an originally methylated cytosine). Other assays take advantage of methylation-sensitive restriction enzymes (see Text).

Note that very sensitive detection of DNA methylation status can be achieved by bisulfite treatment of sample DNA followed by the *pyrosequencing* method described above, and is often used in cancer studies.

Methylation-sensitive MLPA (MS-MLPA)

This is the front-line test for confirming doubtful cases of Angelman and Prader-Willi syndromes (especially rare mosaics), assessing DNA methylation while simultaneously being able to pick up associated microdeletions. It is also used to detect imprinting center microdeletions, and is useful in cancer studies for rapid assessment of promoter hypermethylation.

MS-MLPA is a slight modification of the MLPA method (described previously in Figure 11.5) that makes use of the methylation-sensitive restriction nuclease *HhaI*. The left MS-MLPA probes are designed to contain a *HhaI* recognition sequence, GCGC, and when denatured and hybridized to a desired target sequence containing clustered CpG dinucleotides, a heteroduplex of MLPA probe and desired target sequence has a recognition sequence for *HhaI*. If the target sequence is methylated, the *HhaI* enzyme cannot cleave the recognition sequence and amplification occurs as in normal MLPA. If it is not methylated, *HhaI* cleaves the heteroduplex and amplification cannot occur (see [Figure 11.16](#)).

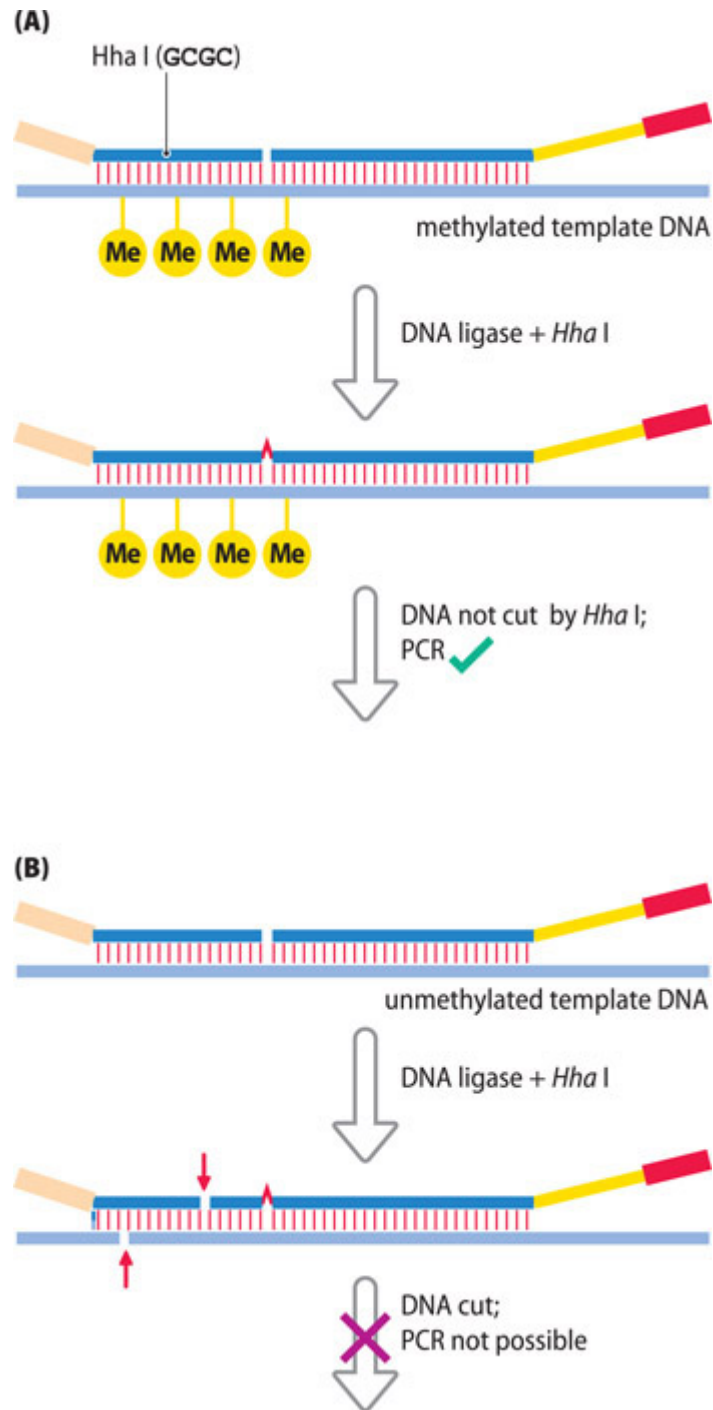


Figure 11.16 The basis of methylation-sensitive MLPA. In this modification of the MLPA method (see Figure 11.5), a *Hha*I restriction site has been engineered into the left MLPA probe. The two MLPA probes are designed to hybridize to a DNA methylation-prone GC-rich sequence on the template DNA, including a CpG within the GCGC sequence complementary to the GCGC of the *Hha*I recognition sequence. If that CpG is

methylated, *HhaI* cannot cut the DNA, and PCR amplification occurs as normal. If it is unmethylated, *HhaI* cuts the DNA and PCR is not possible.

After hybridization the MS-MLPA reaction is split into two parts: one is treated as a normal MLPA reaction to assess copy number; the other treated with *HhaI* to assess methylation status.

11.4 GENETIC AND GENOMIC TESTING: ORGANIZATION OF SERVICES AND PRACTICAL APPLICATIONS

In the previous two sections we covered the technology of genetic and genomic testing, describing how chromosome abnormalities and large-scale copy number and structural variation are detected in [Section 11.2](#), and how testing for point mutations and DNA methylation are carried out in [Section 11.3](#). In those sections we briefly alluded to different clinical settings in which those testing methods are deployed, and the different scales of testing, from individuals and families to communities and populations. Here, we now take a close look at the clinical context of genetic and genomic testing, and the organization and development of these services. We begin with a radical change that is transforming some genetic services into genomic medicine services. Thereafter we focus on different levels at which testing is offered.

The developing transformation of genetic services into mainstream genomic medicine

Genetic services evolved in many countries in the early 1960s to translate new chromosome findings into clinical services. Initially, the services were staffed by cytogeneticists and clinicians; DNA testing was added in the mid-1980s. In some countries services developed around particular genetic conditions, such as the thalasseмииs.

The 1990s saw development of cancer genetic services. Testing for predisposition to breast, bowel, and ovarian cancers served to guide enhanced screening recommendations. A decade later, similar developments took place in cardiogenetic services, as heritable causes of sudden cardiac death came to be explored, including arrhythmias and cardiomyopathies. Diagnostic approaches to the child with developmental delay, possibly with dysmorphic features, have become ever more granular since initial cytogenetic approaches. Of course, the developments over the last decade took place at the same time as our understanding of the molecular basis of genetic disorders increased exponentially. Parallel exponential improvement in available technologies was achieved in terms of both speed and cost.

About 20 years ago, at the turn of the new millennium, genetic testing was available for few disorders; patients were generally seen by clinical geneticists before having a test. More recently, advances in sequencing technology and the development of whole exome and, genome sequencing, have enabled broad genetic testing on an individual patient basis within a clinically useful time-frame. Initially, implementation of such tests was carried out by way of clinical research studies (such as the UK's Deciphering Developmental Disorders project, for example). More recently exome sequencing has been utilized as a clinical diagnostic test.

In some countries, genome sequencing is transitioning to routine healthcare. The UK has been in the vanguard of these developments, but other countries have also invested substantially in establishing national genomic medicine initiatives to address implementation barriers and transition testing from centers of excellence to mainstream medical practice (see [Box 11.4](#)).

BOX 11.4 NATIONAL GENOMIC MEDICINE INITIATIVES AND THE DEVELOPMENT OF MAINSTREAM GENOMIC MEDICINE SERVICES

Within the last decade the governments of at least 14 countries have invested billions of dollars in establishing national genomic medicine

initiatives to address barriers to implementation and transition testing from centers of excellence to mainstream medical practice. In countries such as the UK, France, Australia, Saudi Arabia, and Turkey, necessary resources have been developed enabling genome sequencing of large numbers of patients with rare diseases and cancer. Other countries, such as the US, Estonia, Denmark, Japan, and Qatar, have invested in population-based sequencing projects with return of results to participants.

In the UK, the momentum towards a national genomic medicine service began in 2009 when the House of Lords Science and Technology Committee issued a report calling for a strategic vision for developing genomic medicine. Follow-up workshops involving major stakeholders led to two further reports on *Genetics and Mainstream Medicine* in 2011 and *Genomics in Medicine* in 2012 that set out a route to effective integration of genomics across the NHS. An NHS Long Term Plan then laid out a vision to enable it to harness the power of genomic technology to improve the health of the population and to be the first national healthcare system to offer whole genome sequencing as part of routine care.

In 2013 the UK Government established Genomics England with a mandate to sequence 100 000 genomes from patients with a range of rare diseases and seven common cancers. The objective of sequencing 100 000 whole genome sequences was met in December 2018 (but the communication of results to patients is still ongoing at the time of writing). The follow-up vision now is to sequence half a million genomes by 2023/24 with the aim of improving healthcare for rare diseases and cancers as well as other specific aims such as developing the UK's newborn screening program, developing a pharmacogenomic service and the early detection and treatment of high-risk conditions such as familial hypocholesterolemia, and to link seamlessly with research into genomics so that findings in this setting can be rapidly incorporated into healthcare.

Having previously only been available to citizens via initiatives such as the 100 000 Genomes Project, genomic sequencing is now being offered

therapies. This is no small undertaking. Setting up a consistent and equitable national genomic medicine service is a laudable aim, but one that will require a close focus on implementation issues; attempts to do so in the UK during a global pandemic have, almost inevitably, led to delays to date.

Genetic and genomic testing are being offered now by a much wider range of healthcare practitioners. And an agenda of *mainstreaming genetics* envisages that any branch of healthcare practice should be sufficiently versed in the implications of particular genetic tests to offer such testing directly to their patients. Increasingly, therefore, diagnostic genetic testing will become the responsibility of the clinicians to whom patients are initially referred, and clinical genetic services will take on a molecular pathology role—being involved in multidisciplinary discussions about a patient perhaps, or organizing cascade predictive testing of family members where relevant.

As a result of mainstreaming genetic services, clinicians most skilled in particular groups of disorders will be able to add genetic investigation to the diagnostic tests available to them to personalize their treatments to a greater degree. For example, the discovery that PARP-1 (poly[ADP-ribose] polymerase) inhibitors are particularly effective in treating breast cancer in patients with *BRCA1* and *BRCA2* mutations allows this particular treatment to be initiated at an earlier stage if *BRCA1* and *BRCA2* testing is available promptly at breast cancer diagnosis. In cells with either of these mutations, homologous recombination (one of the two major DNA repair methods) is nonfunctional, but base-excision repair is unaffected. PARP-1 inhibition disables base-excision repair, and thus cells with *BRCA1* or *BRCA2* mutations are no longer able to repair DNA. In the past, tests for *BRCA1* or *BRCA2* mutations were often undertaken largely so that relatives might be offered predictive testing, but the examples above illustrate how rapid genetic testing will increasingly be part of the mainstream specialist

approach to deciding optimal treatment that is personalized to the particular patient's genetic make-up.

An overview of diagnostic and pre-symptomatic or predictive genetic testing

In rare, single-gene conditions, genetic tests are often directed at persons presenting with a clinical problem for *diagnostic testing*. Once the causative (or strongly predisposing) genetic variant has been found, close relatives of that person may be offered some type of predictive genetic testing to see whether or not they have inherited the variant in question. This in turn provides information on whether action is required, and what steps might need to be taken to reduce the likelihood of the clinical problem arising, or to detect it in an early, more treatable stage. Depending on the penetrance, such testing is often described in two slightly different ways:

- *Pre-symptomatic testing* is used when a patient possessing the mutant allele(s) will at some stage develop the condition in question, such as in the case of Huntington disease.
- *Predictive testing* is used if patients are at high *risk* of developing symptoms, such as in the case of carriers of pathogenic *BRCA1* or *BRCA2* variants.

The distinction between diagnostic and predictive testing is relied upon in policy to avoid genetic discrimination (as set out, for example, in the UK government's code on genetic testing and insurance at <https://www.abi.org.uk/globalassets/files/publications/public/genetics/code-on-genetic-testing-and-insurance-final.pdf>). But that same distinction is becoming less clear as whole genome approaches increasingly unearth both diagnostic and predictive findings. Furthermore, the latter will often be significantly less certain in their predictions (with reduced penetrance for example) than the single-gene genetic tests from which services have amassed the most experience to date.

For some diseases the clinical benefit of predictive testing is clear. Take, for example, familial hypercholesterolemia (OMIM 143890), an autosomal dominant disorder commonly caused by pathogenic mutations in the *LDLR* (low-density lipoprotein receptor) gene. Affected individuals normally develop premature cardiovascular disease in the third decade, but early detection of a pathogenic *LDLR* mutation offers the possibility of prevention by lowering LDL-cholesterol through dietary changes and medication. That has led to recommendations of *cascade testing* (testing of relatives after a genetic condition has been identified in a family), either by measuring LDL-cholesterol, or by testing for the familial *LDLR* mutation from as young as the age of 10 years.

Other examples of the balance towards benefits over risks include colorectal cancer syndromes such as Lynch syndrome and familial polyposis coli, which are dominantly inherited. Early detection of a germline mutation that predisposes to these diseases can be followed up by regular colonoscopy surveillance. By identifying and surgically removing polyps before they grow and become dysplastic, the risk of developing late-stage cancer is much reduced. There is a small risk of perforation of the bowel during the required colonoscopy, which needs to be factored into this balance, but this risk is generally lower in the young population who need to be screened before national bowel screening programs kick in, and be in the hands of experienced colonoscopists who run dedicated family history screening programs.

In other familial cancer syndromes, the benefits of predictive testing may be less clear. For example, Li Fraumeni syndrome due to germline *TP53* variants will detect those at high lifetime risks of various cancers including breast and sarcomas. However, the evidence that screening (for example, whole body MRI screening) will make any cancers more amenable to treatment is still quite limited.

Note that the surveillance carried out for some familial cancer syndromes (see [Clinical Box 16](#)) does not prevent cancer—its aim is to identify early cancers while they are still amenable to therapy. To reduce their risk of developing cancer, women from families with a known *BRCA1* or *BRCA2*

variant associated with risk may opt for bilateral mastectomy and/or surgical removal of the ovaries together with the associated fallopian tubes. Predictive testing would be indicated before making this decision, and indeed is not usually offered without a positive test result for the familial mutation.

CLINICAL BOX 16 LYNCH SYNDROME AND FAMILIAL (BRCA1/BRCA2/PALB2) BREAST CANCER: CANCER RISKS AND CANCER SCREENING

LYNCH SYNDROME (HEREDITARY NONPOLYPOSIS CANCER)

The diagnosis is determined on the basis of the pattern of cancers in a family and the age at diagnosis (at least one diagnosis under the age of 50 years), or on the finding of microsatellite instability in tumor tissue, or on immunohistochemistry evidence of abnormal gene expression. Although colorectal cancer is the commonest cancer in the condition, there are several associated cancers ([Table 1](#)).

TABLE 1 CANCER RISKS IN LYNCH SYNDROME COMPARED WITH THE NORMAL POPULATION

CANCER	GENERAL POPULATION RISK (%)	RISK IN LYNCH SYNDROME (%)
Colorectal	5.5	20-80
Endometrial	2.7	20-60
Gastric	<1	1-10
Ovarian	1.6	9-15
Hepatobiliary tract	<1	2-7
Urinary tract	<1	4-5
Small bowel	<1	1-4

CNS, central nervous system.

CANCER	GENERAL POPULATION RISK (%)	RISK IN LYNCH SYNDROME (%)
Brain and CNS	<1	1-3

CNS, central nervous system.

As testing has become more widespread and initiated with weaker family histories, the risk profiles in Lynch syndrome have also widened. This is so because by selecting only the strongest family histories for genetic testing, other familial factors affecting cancer incidence were selected in the process. Screening a general population—unselected for family history—would mean that the associated cancer risks would on average be lower. As yet, there is no widespread population screening for such dominant cancer genes, but the concept is of importance, since many people advocate that their incidental discovery during genome-wide scans should be disclosed [ACMG53/57/73—<https://www.ncbi.nlm.nih.gov/refseq>] or that they be specifically sought as “additional” findings during genomic approaches in the diagnosis of other conditions.

A typical screening protocol for affected and high-risk individuals would be colonoscopy every two years, starting at the age of 25 years, followed by additional endoscopy examination of the esophagus, stomach, and duodenum from 50 years of age. The efficacy of screening for other associated tumors is not proven. Women who have the condition, or who are at high risk, may opt for total hysterectomy and surgical removal of ovaries plus fallopian tubes after completion of their family.

FAMILIAL BREAST CANCER DUE TO *BRCA1/BRCA2/PALB2* VARIANTS

In the general UK population, the lifetime risks of breast cancer and ovarian cancer are roughly 12 % and 2 %, respectively. A person aged 20–25 years with a pathogenic *BRCA1*, *BRCA2*, or *PALB2* variant has a

roughly 70 % risk of going on to develop cancer, notably breast or ovarian cancer ([Table 2](#)).

TABLE 2 CANCER RISKS FOR CARRIERS OF PATHOGENIC BRCA1 /BRCA2/ PALB2 VARIANTS

Cancer type	Lifetime (to age 80 years) risk (%)		
	<i>BRCA1</i>	<i>BRCA2</i>	<i>PALB2</i>
UNAFFECTED CARRIERS			
Breast cancer	60–90	30–85	40–60
Ovarian cancer*	30–60	10–30	2–10
Male breast cancer	0.1–1	5	0.2–6
Prostate cancer	8**	25	7
Other cancers	<5	<5	<5
AFFECTED WOMEN CARRIERS (WITH UNILATERAL BREAST CANCER)			
Cancer in other breast	50% (overall 5-year risk =10%)	50% (overall 5-year risk = 5–10%)	7

* Majority of lifetime risk after 40 years of age.

** Similar to population risk.

When a pedigree indicates a high likelihood of familial breast or ovarian cancer, a typical screening program would commence with annual mammography at 30–40 years of age in women at high risk. Mammography is less sensitive in women aged less than 40 years and so would often be accompanied by MRI breast screening from 30 years of age or earlier. Bilateral mastectomy reduces the risk of developing breast cancer by 95 %, but cancer can still occur in remaining breast tissue on the chest wall.

There is very limited evidence that ovarian screening allows the detection of cancers at a more treatable stage. Women may consider the

surgical removal of ovaries and fallopian tubes—it reduces the risk of ovarian cancer by 95 % (with a small residual risk of primary peritoneal carcinoma) and depending on the age at which it is done, it may also reduce the risk of breast cancer. Ovarian cancer due to these genes is rare below the age of 40, and the risk of bilateral salpingo-oophorectomy at younger ages is significant because of the premature surgical menopause this precipitates.

Cascade testing

Cascade testing means testing of relatives after the identification of a genetic condition in a family. The relatives might be at risk of going on to develop the same single-gene disorder (predictive testing, see above). Unaffected relatives may also be at risk of transmitting a disorder if they carry a harmful allele (heterozygote carriers in recessive disorders, nonpenetrance in dominant disorders) or a balanced translocation.

Different issues need to be considered. How important is it for the relatives to be made aware of the information on the basis of the severity of the condition and the level of risk of a relative developing the condition, or having a child with the condition? What treatment or intervention is available to those who have inherited the factor in question? How might the information change things? How easy will it be for family members to pass information on to relatives? And can, and should, health professionals be involved in such communication?

Take the example of a child with multiple malformations and developmental delay who has inherited unbalanced chromosome translocation products from a parent with a balanced translocation. The translocation will be explained to the parents along with information about their future pregnancies, and also the possibility of other family members carrying the same balanced translocation. In addition to addressing questions from the couple about the risk to future pregnancies and about

their child's future, health professionals need to consider which additional family members should be contacted who might have the same balanced translocation (and be at risk of producing children with unbalanced translocation products), and how to go about this. The same principles apply to cascade testing for carriers of autosomal or X-linked recessive disorders.

Predictive genetic testing in children

Weighing the relative merits and disadvantages of having a predictive test for an adult-onset condition is not easy. Individuals seeking such testing many years before interventions—such as mammography—could be offered, would often, on reflection, delay the predictive test until just before the time it would impact screening recommendations. Alternatively, refining reproductive risks might be a reason to seek predictive testing in conditions where the evidence for interventions is more limited. Predictive testing of children is appropriate when the onset of the condition is usually in childhood, as in the case of multiple endocrine neoplasia (where screening recommendations start from the age of 5). But for later onset conditions it is important to have a conversation with the parents to plan the optimum timing of a test.

For tests, where a significant proportion of adults go on to make a considered decision not to have the test, such as predictive testing for Huntington disease, it is important to preserve the decision until the child is competent enough to do so themselves. A plethora of international guidelines recommend that predictive tests for adult-onset disorders should not be undertaken in children unless a medical intervention applicable to children is both possible and shows clear medical benefit (as in familial hypercholesterolemia). One example of such guidance—from the British Society of Genetic Medicine and currently under revision—gives worked case examples and practical suggestions for consultations with parents requesting such testing (<https://www.bsgm.org.uk/about/our-history/>).

Pre-symptomatic genetic testing for conditions whose course cannot be altered by medical intervention

There are no current interventions to delay the onset of most of the late-onset neurological disorders. Nevertheless, Huntington disease, a devastating neurological disorder that often does not manifest itself until later stages in life, was one of the first conditions for which predictive testing was offered. After initial concerns that individuals testing positive for this disorder might take their own lives, predictive testing was introduced with caution. Several sessions are usually scheduled with a genetic service to explore the pros and cons of such testing; in these settings experience has shown that predictive testing precipitates a catastrophic event—suicide, suicide attempt, or psychiatric hospitalization—in less than 1 % of cases.

The uptake of testing in people at 50 % risk of Huntington disease is about 10–20 %, interestingly much lower than the uptake imagined prior to the test being available. Some evidence also suggests that initial thoughts about testing are more enthusiastic than considered decisions taken after a discussion about the pros and cons. Young adults who undergo testing generally do so to assist in making career and family choices. Another group opting for testing are those who have reached the age by which signs and symptoms would usually have presented; they wish to be tested so that they can reassure their children and grandchildren that the condition has not been passed down their branch of a family.

Whilst pre-symptomatic testing for highly penetrant conditions such as Lynch syndrome or Huntington disease is often pitched as a “yes” or “no” result, for any condition with a penetrance of <100 %, individuals who test positive for the gene variant may never develop the condition. And as population genetic testing becomes more widespread (and is not selected on the basis of pheno-type), lower penetrance variants of these conditions will be found, and communication around risk, and the interventions available to manage it, is becoming more complex.

The different ways in which diagnosis of genetic conditions is carried out in the prenatal period

Couples who have a family history of a serious genetic disorder usually want to know whether they are at risk of having an affected child. If they are at risk, they might choose not to have children at all, or to adopt an unrelated child. Other times the genetic condition in question can be avoided through the use of egg or sperm donation, or by preimplantation diagnosis so that only healthy embryos are selected, as described below. Since the latter often involves the financial burden of private fertility services, some will choose natural conception and prenatal diagnosis in which the fetus is tested to see if it has inherited the genetic condition. Given that the opportunities for therapies *in utero* remain very limited, such prenatal diagnosis is usually offered on the understanding that a pregnancy will be terminated if the fetus is affected, though clearly this will always ultimately be a pregnant woman's decision.

Careful risk assessment and communication of the options is key in prenatal diagnosis (Box 11.5). Accurate predictive genetic testing is possible for single-gene disorders in which the major genetic variant contributing to disease has been identified in an affected family member. Prenatal diagnosis may also be carried out in situations in which there is an increased risk of transmitting a chromosomal aneuploidy (advanced maternal age is an important risk factor). Or one parent might have been identified as a carrier of a balanced translocation, and there is a risk that a fetus with unbalanced translocation products might be viable but have severe problems.

Traditionally, prenatal diagnosis has involved collecting a sample of fetal tissue recovered by an invasive procedure. A sample may be taken from the chorion (the outermost extra-embryonic membrane), and fetal DNA can be isolated from the cells obtained ([Figure 11.17A](#)); there is a roughly 1 % excess risk of miscarriage. The sample can be taken any time in the pregnancy from 11 weeks onward, but typically in the first trimester (to allow the possibility of early termination of pregnancy).

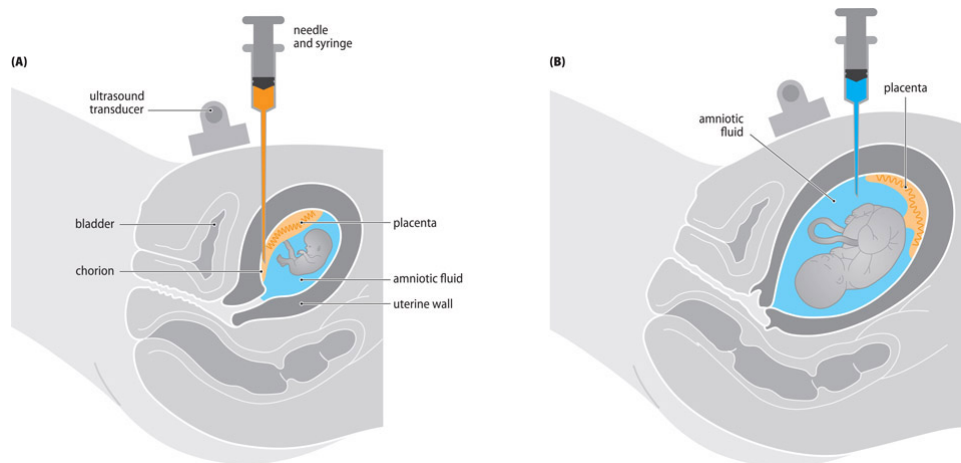


Figure 11.17 Invasive prenatal diagnosis using chorionic villus sampling or

amniocentesis. (A) Chorionic villus sampling. As shown here, this is usually carried out by a transabdominal approach guided by ultrasound under local anesthetic. (B) Amniocentesis.

Amniocentesis is the other major alternative sampling method, and it also has a small risk of miscarriage. A sample of amniotic fluid is taken at, or close to, 16 weeks of gestation ([Figure 11.17B](#)); it provides fetal cells that are processed to give either chromosome preparations to check for chromosome abnormalities, or fetal DNA samples for analysis.

Pre-implantation genetic testing can also be carried out to prevent the transmission of a harmful genetic defect by using *in vitro* fertilization (IVF). We describe that separately below.

CLINICAL BOX 17 GENETIC CONSULTATIONS AND GENETIC COUNSELING

Genetic consultations

Genetic consultations often start with a child or adult with a genetic disorder—at other times consultations are initiated by a person’s concern about other family members. Rather than focus on his or her own health issues, the person may bring up the subject of a family history of a medical problem such as cancer. Together, the patient and clinician

construct the pedigree, the basic tool in the genetics clinic. Diagnoses are then confirmed by using, for example, cancer registries, or death certificates in the case of deceased individuals, or by requesting consent to access medical information of living relatives.

Confirming family history diagnoses can be really important in risk assessment. For example, a purported family history of bowel cancer may in fact be diffuse gastric cancer, or ovarian cancer. In that case different genes should be scrutinized, and the geneticist should consider whether, and which, genetic tests are appropriate, and whom it is most appropriate to test first. If there is a relative who has the disorder, it would often be more appropriate to test the relative first to establish the causative mutation, which would then be the basis of a predictive test.

Genetic counseling

Parents at risk of having a child with a genetic disorder, and affected individuals and relatives of an affected family member, may benefit from **genetic counseling**, the process by which they are informed of the consequences and nature of the disorder, the probability of developing or transmitting it, and the options open to them. Genetic counseling may be provided by doctors or by professionals specifically trained as genetic counselors who may have a science or nursing background.

The counselor aims to provide the necessary information to help family members to make a decision based on a patient's values and circumstances, rather than direct them toward a particular decision. Such non-directive approaches have been important in difficult decisions about termination of pregnancy, or predictive testing for untreatable conditions. Recently, however, the evidence basis of certain interventions based on genetic testing has been improving (for example, regular surveillance improves the life expectancy of people with familial polyposis coli), and so a certain directiveness on the part of the counselor may be more appropriate.

As well as offering general support, the counseling process has at its core the determination of risks of a condition. That may be relatively simple for single-gene disorders based on Mendelian principles, but as detailed in [Section 5.3](#) there are often complications, such as lack of penetrance or variable expressivity.

The risk estimate may be determined by a Bayesian calculation in which a prior probability (such as the risk predicted from Mendelian principles alone, for a single-gene disorder) is modified by some other relevant information. For an X-linked recessive disorder, for example, the daughter of an obligate carrier would have a 50 % risk of herself being a carrier. However, the carrier risk for a woman whose maternal grandmother is an obligate carrier but whose own mother's status is unknown (a 50 % chance of being a carrier) can be modified by circumstance. Such Bayesian calculations are now often computerized (for example, see Cambridge University's CanRisk program at <https://canrisk.org/>). But an understanding of the basic principles remains key to sense-checking the outputs of such programs in case of incomplete/inaccurate data entry.

In [Figure 1](#), individual I-2 is an obligate carrier of the X-linked recessive condition because she has two affected boys. III-3 is concerned that her mother, II-3, might be a carrier. Because we do not know her status, II-3 has a 50 % chance of being a carrier; if she were a carrier, she would have a 50 % chance of transmitting the mutant allele to III-3. That is, the probability that III-3 is a carrier, based on this information alone, would be $50\% \times 50\% = 25\%$. On drawing the pedigree, however, III-3 is found to have four brothers, none of whom are affected—this additional conditional information alters the risk.

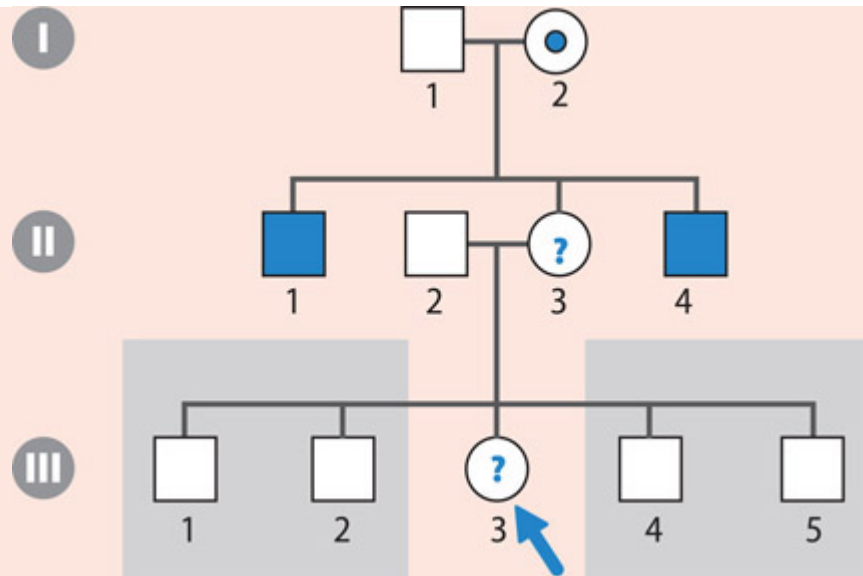


Figure 1 Genetic risk in a pedigree with a childhood-onset X-linked recessive condition. If III-3 were an only child, her risk of being a carrier would be 1 in 4 (her mother, II-3, is the daughter of obligate carrier I-2 and has a 1 in 2 chance of being a carrier; if so, III-3 also has a 1 in 2 risk of inheriting the mutant allele). However, III-3 subsequently mentions that she has four grown-up brothers, none of whom are affected. That additional information suggests that the probability that II-3 is a carrier is much less than 0.5, and that means the chance that III-3 is a carrier is greatly reduced—but by how much?

Bayesian analysis to account for conditional information

If II-3 were a carrier, it would be possible, but unusual, that she would have had four unaffected sons. The new conditional information suggests that she is more likely not to be a carrier, and the risk that III-3 would be a carrier should therefore be much reduced.

The question is: by how much? To answer that question and to give a new risk estimate based on all the information, Bayesian analysis is used. Four steps are involved, as listed below.

1. Identify all the different scenarios that can explain the observations.

2. For each scenario, calculate the prior probability and conditional probability.
3. Multiply the prior probability by the conditional probability to obtain a joint probability for each scenario.
4. Determine what fraction of the total joint probability is represented by each individual scenario to get a posterior probability for each of the three scenarios.

If we discount fresh mutation, there are three possible scenarios in this case: (A) II-3 is not a carrier, and so III-3 is also not a carrier; (B) II-3 is a carrier, but III-3 is not a carrier (because she did not inherit the mutant allele); (C) II-3 is a carrier and III-3 is also a carrier (because she inherited the mutant allele). As detailed in [Figure 2](#), Bayesian analysis suggests that scenario A is by far the most likely—the ratio of the probability for the three scenarios is 32:1:1 for A:B:C. Coming back to the original question, the probability that III-3 is a carrier is given by the posterior probability for scenario C, which is 1/34 (or close to 3 %), substantially less than the prior probability of 25 %.

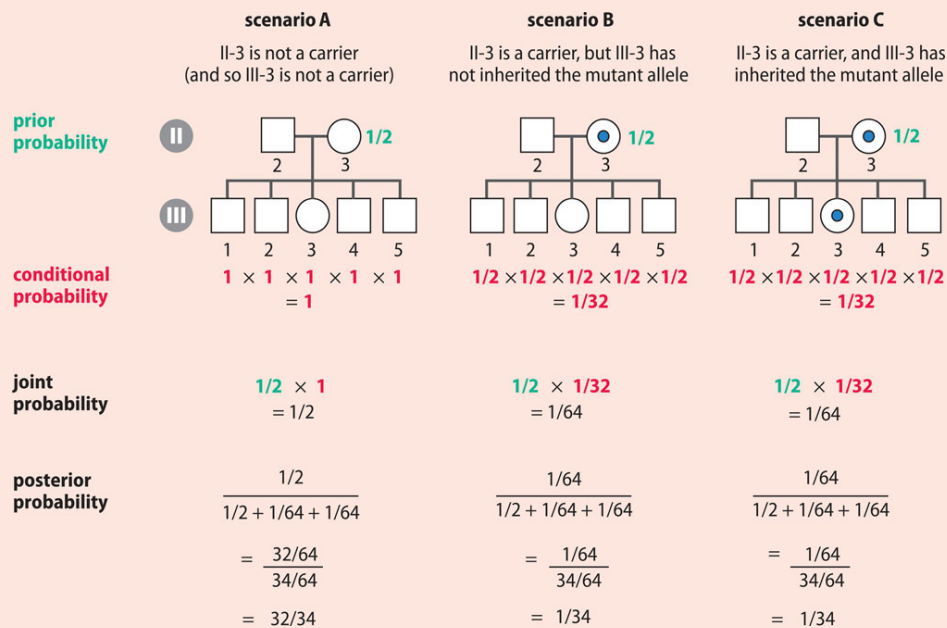


Figure 2 According to Bayesian calculations, the risk that III-3 in Figure 1 is a carrier is only about 3 %. Here, the prior probability is the standard risk due to

Mendelian segregation, and the conditional probability is the *multiplicative product* of the individual probabilities that individuals in generation III have the status that is attributed to them. The probability that an individual male in generation III is unaffected is 1 in 2 if II-3 is a carrier (scenarios B and C), or 1 if II-3 is not a carrier (scenario A). The probability that III-3 is not a carrier is 1 in scenario A (because her mother is not a carrier), or 1 in 2 when her mother is a carrier (scenarios B and C). The joint probability is the product of the prior probability and conditional probability, and the posterior probability is the fraction of the total joint probabilities (for all scenarios) that is attributable to one scenario.

Preimplantation genetic testing is carried out to prevent the transmission of a harmful genetic defect using *in vitro* fertilization

Preimplantation genetic testing is a technique used to identify or screen for genetic defects in embryos created through *in vitro* fertilization before pregnancy so that an apparently healthy embryo can be implanted into the uterus. The procedure is technically challenging because it typically involves analyzing a single cell (as a way of monitoring the genotype of the oocyte or of the early embryo), and is not widely available.

To infer the genotype of an oocyte, polar bodies are sometimes analyzed. More commonly, a single cell (blastomere) is removed from the very early embryo for testing ([Figure 11.18](#)). For technical reasons, some centers prefer to analyze a few cells taken from the outer trophoctoderm at the later blastocyst stage (the trophoctoderm will give rise to extra-embryonic membranes). In either case, the remaining embryo can be implanted successfully and is viable.

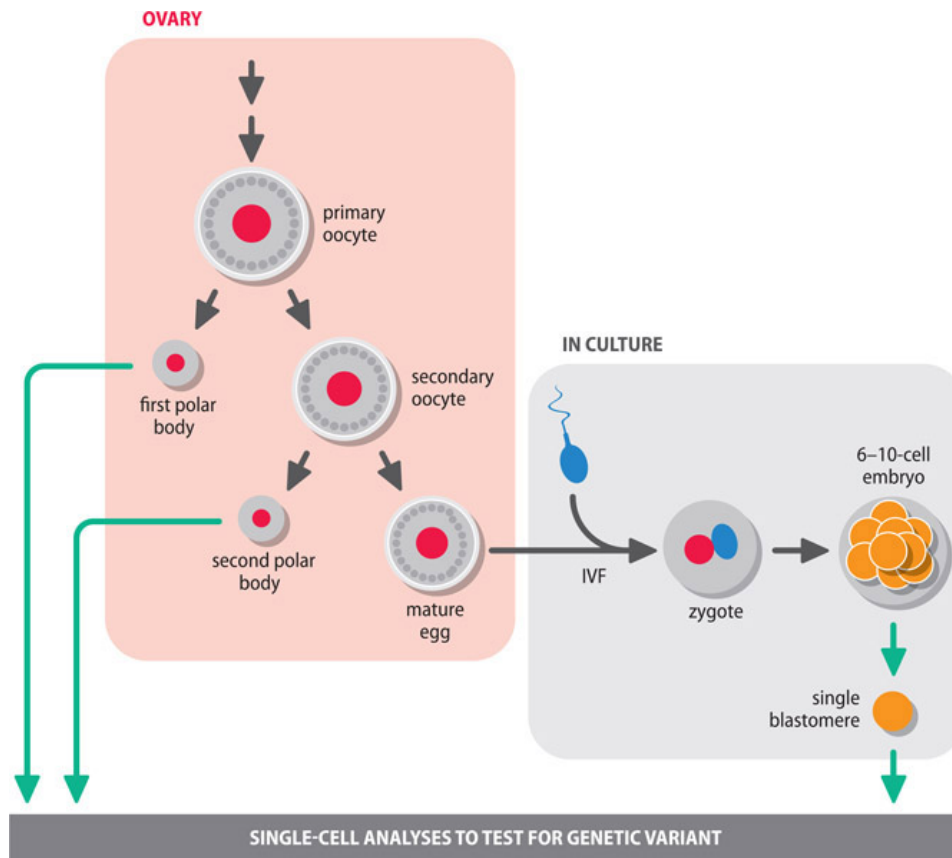


Figure 11.18 Preimplantation genetic diagnosis often involves analyzing single cells. Unlike for sperm cells, the meiotic divisions giving rise to an egg cell are asymmetric: the primary oocyte divides to give a secondary oocyte and a polar body, and the secondary oocyte divides to give the mature egg cell and a second polar body. The polar bodies are disposable and can be analyzed to infer whether the egg cell is carrying a specific harmful genetic variant or a chromosomal aneuploidy. If not, IVF proceeds with what appears to be a normal egg cell. More commonly, a single cell is sampled from the early embryo and tested for the presence of the harmful genetic variant. If the test result is negative, the remaining embryo is implanted in the uterus, and development can proceed normally. Because it can be challenging to obtain data from a single cell, some centers prefer to allow the embryo to develop further and remove a few cells from the blastocyst for testing.

Standard assisted reproduction techniques are used to obtain embryos for testing: ovarian stimulation (to produce eggs that are then collected under sedation), addition of sperm, and assessment of the *in vitro* fertilization

(IVF) and of the embryos produced. In the case of single blastomere analyses, individual embryos are grown in culture to reach the 6–10-cell stage. At this stage a small hole is made in the zona pellucida and a single cell is removed through the hole for testing. Despite the loss of one cell for analysis, the embryo will go on to develop normally.

There are two broad categories of preimplantation genetic testing (PGT), as listed below.

- *Diagnosis*. This applies to couples who are at risk of transmitting a specific genetic abnormality: one or both parents have previously been shown to carry a pathogenic variant or chromosome abnormality that the test is designed to identify.
- *Screening* is performed on couples who may have difficulty conceiving but have no *known* genetic abnormality. Here, the embryo is screened for the presence of any chromosomal aneuploidy.

In both cases, the object is to implant normal embryos only, to avoid the birth of an affected child (in diagnostic cases) or to improve the pregnancy success rate (in screening cases).

For preimplantation genetic diagnosis, prior identification of mutant alleles in one or both parents allows a test in which one or more relevant DNA regions in the DNA from the biopsy are PCR-amplified and sequenced.

If there has been difficulty in identifying a parental mutation, indirect genetic linkage tests can be conducted using a well-established set of polymorphic markers that span the disease gene locus. Occasionally, the test seeks to identify the transmission of a chromosomal abnormality and involves interphase FISH.

The process of achieving a pregnancy becomes medicalized (with potential side effects associated with ovarian hyperstimulation). Additionally, the likelihood of a successful pregnancy outcome is quite low: it is only about 1 in 5 at the start of an IVF treatment cycle (sometimes no

embryos are suitable for transfer, depending on the number of eggs fertilized, and the number and quality of unaffected embryos), but increases to 1 in 3 after embryo transfer.

Noninvasive prenatal testing (NIPT) and whole genome testing of the fetus

Short fragments of cell-free DNA, both fetal and maternal, are present in maternal blood. The fetal DNA fragments arise from placental cells undergoing apoptosis; the maternal DNA fragments originate from the occasional degradation of the mother’s cells through apoptosis and necrosis. The fetal DNA fragments are in the minority, accounting for around 10–15 % of the total cell-free DNA in the maternal circulation between 10 and 20 weeks of gestation. Analysis of cell-free DNA in a maternal plasma sample can therefore be sufficient to investigate the genetic composition of the fetus. Obtaining maternal blood is of course somewhat invasive; the non-invasive terminology is used comparatively, given the much higher degree of invasiveness—and risk—that is linked to amniocentesis or chorionic villus sampling.

Because the cell-free DNA in maternal plasma is dominated by maternal DNA, the easiest fetal DNA sequences to identify are those inherited exclusively from the father (they can be readily amplified and detected). That includes Y-chromosome DNA sequences, and noninvasive fetal sexing is now routinely available with a sensitivity of about 90 % and a specificity of 98 %. Testing has also been possible for other exclusively paternal sequences in certain situations (see [Table 11.6](#) for some applications).

TABLE 11.6 APPLICATIONS OF NONINVASIVE TESTING FOR VARIOUS GENETIC CONDITIONS	
Genetic condition	Noninvasive testing/diagnosis
Serious X-linked recessive disorders	Fetal sexing test: identifying a female fetus avoids need for subsequent invasive prenatal diagnosis with associated miscarriage risk, but not for a male fetus

Genetic condition	Noninvasive testing/diagnosis
Congenital adrenal hyperplasia (21-OH deficiency)	Fetal sexing test: abnormal androgen production in affected female fetuses results in virilization of the external genitalia. After early identification of a female fetus, the fetal adrenals can be suppressed by the oral administration of dexamethasone to the mother. <i>CYP21A2</i> haplotype testing can now be offered from 8 weeks via NIPD if pre-pregnancy work-up (from both parents and previously affected child) suggests informative <i>CYP21A2</i> haplotype testing
Hemolytic disease of the newborn	Testing for paternal rhesus D blood group: rhesus D-negative women may be at increased risk of hemolytic disease of the newborn (because of a previous affected pregnancy or raised antibody titer). If a paternal rhesus D is identified, the pregnancy needs to be monitored closely because of the risk of fetal anemia
Cystic fibrosis	Haplotype testing can be offered if DNA is available from both parents with confirmed mutation and DNA from previously affected child or confirmed non-carrier child. Mutation testing for <i>CFTR</i> is also possible if father is known to be carrier of particular <i>CFTR</i> variant
Various craniosynostosis syndromes	Such as Aperts syndrome, Crouzon syndrome, and achondroplasia where a parent is known to have a pathogenic <i>FGFR2/3</i> variant
Duchenne/Becker muscular dystrophy	Dystrophin haplotype testing where familial mutation is known
Spinal muscular atrophy	Where both parents are known to be carriers

Technological breakthroughs

It is technically easy to test cell-free DNA in maternal plasma for the presence of exclusively paternal DNA sequences. More comprehensive testing has been difficult as fetal markers are not readily distinguished from maternal homologs, and there is a large background of circulating maternal DNA in maternal plasma. Technological advances in noninvasive prenatal testing and screening over the last decade have dramatically opened up this field, offering an exciting new window on fetal diagnosis and fetal screening, with the list of current possibilities likely to expand rapidly.

A major breakthrough came from overcoming technical obstacles to what is a very simple principle: counting the parental haplotypes. For any very short genome region, three haplotypes exist in the freely circulating DNA in maternal plasma: the maternal haplotype that is transmitted to the fetus (M_t), the maternal haplotype that is not transmitted to the fetus (M_u), and the paternally transmitted haplotype (P_t).

Because the DNA in maternal plasma will be a mix of DNA originating from degraded maternal cells plus a relatively small amount of fetal DNA, typing for individual DNA markers will show a small excess of alleles from M_t haplotypes over alleles from M_u haplotypes ([Table 11.7](#)). To detect such a small difference reliably, a very specific test would be needed; however, with massively parallel DNA sequencing it is comparatively easy to count millions (or even billions) of DNA molecules—permitting very specific testing.

TABLE 11.7 DISTINGUISHING BETWEEN THE TWO MATERNAL HAPLOTYPES BY COUNTING ALLELES AT MATERNALLY HETEROZYGOUS MARKER LOCI IN DNA FROM MATERNAL PLASMA

Contribution made by DNA from:	Expected count of alleles at marker loci on:	
	Transmitted maternal haplotype, M_t	Untransmitted maternal haplotype, M_u

Contribution made by DNA from:	Expected count of alleles at marker loci on:	
	Transmitted maternal haplotype, M_t	Untransmitted maternal haplotype, M_u
Mother (M_t+M_u)	$N(1-\epsilon)$	$N(1-\epsilon)$
Fetus (M_t+P_t)	$N\epsilon$	0
Total	N	$N(1-\epsilon)$

ϵ is the fraction of the DNA in maternal plasma that originates from fetal cells. N is the number of haplo-types with the allele of interest that have been analyzed.

NIPT can be used with targeted massively parallel DNA sequencing: certain genome regions of interest are captured from the genomic DNA (see [Box 11.2](#) above for the principle) and sequenced.

One application of NIPT is fetal aneuploidy screening. Previously, the problem here had been distinguishing fetal autosomes from the maternal equivalents. When fetal DNA accounts for 10 % of the DNA in maternal plasma, the amount of chromosome 21 DNA increases by just 5 % if the fetus has trisomy 21. Because of massively parallel DNA sequencing, that small difference can be readily detected (trisomy 21 can be detected with a sensitivity of 99 % and a specificity of 99 % using this method). Current evidence suggests that NIPT is more cost-effective as a screening tool (to define a high-risk group that can then be offered confirmatory amniocentesis) than as a diagnostic procedure.

An overview of the different types of genetic screening

The genetic testing described above is reactive: it is carried out in response to individuals seeking medical help or advice about the risk of developing or transmitting a genetic disorder. That is to say, it comes to the attention of medical services as a disease phenotype in a family member. A causative genotype is then sought so that other family members at risk can be tested to see if they have the pathogenic variant.

In *genetic screening*, the genetic tests are carried out in communities and populations and genotypes are used to predict phenotypes that may manifest

at some future time. In population screening, a particular population is screened with a targeted enquiry.

With the advent of whole genome sequencing, there is an added dimension to the screen; the whole genome assay is in effect a screen, to which targeted enquiries are then made depending on the clinical question.

Genetic screening of populations can be carried out using biochemical and physiological markers (as products of genetic variants) as well as looking directly at genetic make-up. In [Section 8.3](#) we considered a type of longitudinal population screening, exemplified by the UK Biobank project, in which comprehensive testing is carried out on people at regular intervals over decades. That is a research-led type of screening without any (or only very exceptional) feedback of findings to participants. In contrast, the three types of genetic screening listed below are primarily directed at providing clinical benefit to the subjects tested.

- *Pregnancy screening.* The object is to identify whether or not the pregnancy is at a very high risk of leading to the birth of a child with a serious genetic disorder. The motivation for the test is usually to prevent the birth of an affected child. (Less commonly, the test might be requested to allow psychological preparation and medical management planning for such a birth, while also offering psychological benefit, should the test indicate that the fetus is unaffected.) It may entail screening for a serious single-gene disorder in communities where that disorder is prevalent, or for aneuploidies. As described below, technological advances now permit comprehensive genetic profiles to be obtained for a fetus. That might lead to new ways of treating disease *in utero*.
- *Newborn screening.* This is carried out in many countries, but to variable extents. A major motivation has been to target early treatment in serious disorders for which early intervention can make a substantial difference and may lead to disease prevention. Genetic screening of newborns began with certain metabolic disorders, and this class of disorder is still a major focus.

- *Carrier screening.* This is also carried out in many countries, often targeted at particular ancestral groups (for example, Tay-Sachs or sickle cell screening) with an aim of identifying carrier couples of a mutant allele for a range of severe autosomal recessive disorders. More recently, advances in sequencing techniques have led to expanded carrier screening approaches where many different carrier states are screened for simultaneously.

Pregnancy screening for fetal abnormalities

Specific maternal screening programs have been undertaken in the first trimester to identify fetuses at high risk of common and serious single gene disorders—such as sickle-cell disease and thalassemia—that are prevalent in certain communities. However, the focus for most prenatal screening is maternal screening for fetal aneuploidy, notably the commonest chromosomal abnormality, trisomy 21 (causing Down syndrome).

As described above, massively parallel DNA sequencing is increasingly used to screen DNA in maternal plasma (which includes small amounts of DNA from fetal cells) for evidence of aneuploidies such as trisomy 21. This type of NIPT is still largely offered after a “high risk combined” screen result based on three parameters. One is *nuchal translucency*, the skin thickness at the back of the neck, as measured by ultrasound scanning between 11 and 14 weeks of gestation; it is determined by the amount of fluid that collects here (which is often greater in Down syndrome babies). A second factor is the mother’s age (the risk increases 16-fold as the maternal age increases from 35 to 45 years). The third factor is based on altered levels of certain maternal serum proteins, such as an increased level of free b-HCG (human chorionic gonadotropin) and a decrease in PAPP-A (pregnancy-associated plasma protein A).

Private clinics are increasingly offering NIPT to a general pregnant population but here the pre-test probability of an aneuploidy is usually less than 1 %. Whilst this may still be a risk some couples would not wish to tolerate, a quick look at Bayes theorem (Box 11.5 on page 458) tells us that

the lower the population frequency, the higher is the chance that a positive result will be a false positive. For example, if a 30-year-old woman has a chance of an aneuploidy of 0.1 % and the NIPT has a sensitivity and specificity of around 99 % (commonly advertised as such), then the chance that her positive test is a true positive is <10 %. For a 1 % figure the probability of a false positive is 50 %. This therefore has the potential to lead to a much higher rate of inappropriate invasive follow-up testing than people commonly realize.

On the basis of combined screening, approximately 2 % of women will have a greater than 1 in 150 risk (compared with an overall population risk of about 1 in 670); they will be offered chorion biopsy for definitive aneuploidy testing. There will be an adverse outcome in 20 % of these women (which includes trisomy 13 and trisomy 18 in addition to trisomy 21); that still means there is no chromosome abnormality in 80 % of women who take up chorion biopsy—the development of a reliable test based on cell-free fetal DNA in the maternal serum is therefore a major advance. The combined screening detects 90 % of all affected pregnancies.

First-trimester ultrasound is important in estimating the date of delivery and for nuchal measurement, and is essential for accurate estimates of gestational age needed for risk calculations based on the levels of the maternal serum proteins described above. Additional ultrasound is routinely offered in pregnancy at around 20 weeks of gestation to look for structural anomalies (a significant proportion of which are due to chromosomal or Mendelian disorders).

Newborn screening allows the possibility of early medical intervention

Newborn screening was pioneered in the late 1960s. Screening for phenylketonuria (PMID 20301677) used dried blood spots collected on a filter-paper card (the Guthrie card) at 5 days of age. Assays for congenital hypothyroidism (which has a number of causes, few of which are genetic), were added shortly afterward. For both conditions the rationale was

prevention of the developmental delay that would inevitably ensue in the absence of medical intervention (which involves dietary changes for phenylketonuria and hormone replacement for congenital hypothyroidism—see [Table 11.8](#)).

TABLE 11.8 NEWBORN SCREENING PROGRAMS FOR SELECTED AUTOSOMAL RECESSIVE DISORDERS AND CONGENITAL HYPOTHYROIDISM

Genetic disorder	Prevalence	Type of screening	Treatment of affected individuals
Congenital hypothyroidism	1 in 5000	assay of free thyroxine or thyroid-stimulating hormone in serum	hormone replacement
Cystic fibrosis	1 in 2500 in European populations	screen for immunoreactive trypsinogen, then confirm by scan for <i>CFTR</i> mutations	antibiotics, chest physiotherapy, pancreatic enzyme replacement for those with pancreatic insufficiency
Galactosemia	1 in 75000	assay of levels of erythrocyte galactose-1-phosphate and galactose-1 -	change of diet to reduce intake of galactose

Data from guidelines proposed by the American College of Medical Genetics and Genomics, which makes information on screening programs for individual disorders available through PubMed (PMID 21938795) and the NCBI bookshelf (<http://www.ncbi.nlm.nih.gov/books/NBK55827/>).

CFTR, cystic fibrosis trans-membrane conductance regulator gene; HbF, fetal hemoglobin; HPLC, high-performance liquid chromatography; IEF, isoelectric focusing.

Genetic disorder	Prevalence	Type of screening	Treatment of affected individuals
		phosphate uridylyltransferases	
Phenylketonuria	~1 in 12000	plasma amino acid analysis to show increased phenylalanine: tyrosine ratio	change of diet to reduce intake of phenylalanine (Clinical Box 9 on page 234)
Sickle-cell disease	~1 in 500 with African ancestry	hemoglobin separation by electrophoresis, IEF, or HPLC. DNA studies may be used to confirm genotype	hydroxyurea (increases HbF in red blood cells, reducing transfusion requirement and decreasing frequency and severity of vaso-occlusive events; prophylactic penicillin

Data from guidelines proposed by the American College of Medical Genetics and Genomics, which makes information on screening programs for individual disorders available through PubMed (PMID 21938795) and the NCBI bookshelf (<http://www.ncbi.nlm.nih.gov/books/NBK55827/>). *CFTR*, cystic fibrosis trans-membrane conductance regulator gene; HbF, fetal hemoglobin; HPLC, high-performance liquid chromatography; IEF, isoelectric focusing.

Inborn errors of metabolism have been a major focus of newborn screening for two reasons. First, they have been studied for decades, and there is a highly developed understanding of the molecular basis of disease, allowing useful early medical interventions in some cases.

The second advantage is that inborn errors of metabolism are typically amenable to easy-to-use screening systems that work at the gene-product or metabolite level, and are applicable to easy-to-access patient samples, such as blood or urine. A disease allele may have any one of a potentially very large number of different mutations; if the gene has many exons, the screening can be laborious. However, all that heterogeneity at the DNA level often has a rather uniform effect at the gene-product level: a single assay can often detect abnormalities in the product or characteristic changes in certain metabolites.

As a result, it is usual to use assays at the gene-product level, or assays for disease-associated metabolites (tandem mass spectrometry—which allows the parallel testing of multiple metabolites in blood and urine samples—can efficiently screen for a range of metabolic disorders at low cost).

Benefits versus disadvantages of newborn screening

More recently, other disorders have been added to screening lists, and the huge advances in massively parallel DNA sequencing have led to proposals to greatly increase the number of disorders that are screened for.

In addition to the large costs of implementing national screening programs, any screening program will include false positives. Anxiety can be generated in families who receive a positive screen result but whose child is unaffected on second testing (and as mentioned above, the more tests that are taken, the greater is the chance of receiving a false positive result). Accordingly, some countries have taken a quite conservative approach. In the UK, for example, national newborn screening is restricted to nine rare but serious conditions: phenylketonuria, congenital hypothyroidism, medium-chain acyl-CoA dehydrogenase deficiency (MCAD), sickle-cell disease, cystic fibrosis, maple syrup urine disease, isovaleric acidaemia, homocystinuria, and glutaric aciduria type 1.

By contrast, the American College of Medical Genetics and Genomics (ACMG) has recommended screening for 54 conditions (including hemoglobin abnormalities, various inborn errors of amino acid, fatty acid, or organic acid metabolism, biotinidase deficiency, congenital adrenal hyperplasia, galactosemia, and cystic fibrosis).

Early treatment might not be of clinical benefit in all of the conditions screened, but there can be other benefits. One benefit might be a greater awareness of the disorders and a greater sharing of information, increasing knowledge of the natural history of these very rare disorders. Another is that parents will be informed about the condition and recurrence risks before they have further children. Some countries have piloted newborn screening for Duchenne muscular dystrophy, not because of therapeutic benefit but because if a child does not present until 4 years of age, couples may already have a second affected child at the time of diagnosis.

Newborn screening using whole genome sequencing (WGS)

The UK government has recently announced plans for newborn screening using whole genome sequencing (<https://www.gov.uk/government/publications/genome-uk-the-future-of-healthcare>). At the time of writing, Genomics England are leading a public consultation on the possibility of introducing whole genome sequencing as the primary technology through which to offer newborn screening, which offers the potential of diagnosing many more conditions, as well as indicating sensitivities to future pharmacological interventions, in a newborn population. Whilst that may sound appealing, the disadvantages will need careful attention before existing screening programs expand in this way.

As alluded to above, WGS here is a new type of screen in itself; it is the assay upon which the newborn screen is performed, and this can be predetermined by determining which filters are applied to the data. This is similar to the filtering of tandem mass spectrometry outputs used in current

screening programs. That is to say, only certain conditions will be looked for in the data that have the potential to reveal more. Questions arise about who then might access the remaining data for predictions, and when they might do so. For example, will parents have a right to obtain or access the data? Can they—or the data analysts they instruct—inspect it for adult-onset conditions, uncertain findings and so on? These questions about such future predictions often seem qualitatively different when talking about existing data from WGS than data not brought into existence through traditional screening techniques. That said, current newborn screening primarily relies on interpreting the peaks from tandem mass spectroscopy; many more abnormal gene products could be analyzed this way than is currently routine.

Patient support organizations such as the Genetic Alliance UK are cautiously supportive of whole genome approaches to newborn screening, but have called for careful scrutiny of the ethical issues involved.

Different types of carrier screening can be carried out for autosomal recessive conditions

Carrier screening can be carried out at the pre-conception level or antenatally, but it can also sometimes be a “side-effect” of newborn screening (in that newborn screening does not aim to detect carrier status that is not relevant to an individual until of reproductive age). Ideally carrier screening is done when couples are planning to have children, to see whether they are both carriers of the same condition, but often the window of opportunity for this may be small or opaque.

The example of β -thalassemia screening

Approximately 70 000 babies are born each year with this disorder, the incidence being highest in Mediterranean countries, India, Africa, Central America, the Middle East, and Southeast Asia. To treat the resulting anemia, affected individuals require repeated blood transfusions; however,

that causes iron overload, which in turn leads to liver damage and cardiomyopathy. Iron chelation therapy is then used to increase iron excretion, prolonging life expectancy well into the fourth decade of life and usually beyond that.

Carrier screening can be undertaken using mean corpuscular volume and mean corpuscular hemoglobin levels in the standard full blood examination; various methods are used to confirm the diagnosis. In 1973, carrier testing was introduced across Greece and Cyprus after educational programs at schools, in the armed forces, maternity clinics, through the mass media, and, in Cyprus, through the Orthodox Church. Sardinia introduced screening a few years later.

Subsequently, many countries have developed screening programs. In Iran, many provinces of Turkey, the Gaza Strip, and Saudi Arabia the testing is mandatory for couples registering for marriage. In the Gaza Strip, couples have to sign a declaration that they are aware both are carriers if they continue with the marriage. These countries have opted for screening before pregnancy; in other countries, screening occurs in antenatal clinics—if the woman is found to be a carrier, testing is offered to her partner. Although consent is an intended prerequisite, screening evaluations have indicated that patient awareness and understanding of the program are very variable.

There has been a significant reduction in affected births in countries with screening programs, partly due to altered marriage plans but mainly due to the uptake of prenatal diagnosis and termination of pregnancy. For example, the incidence of b-thalassemia in Sardinia when screening was introduced in 1975 was 1 in 250 births; by 1995 it was 1 in 4000 births. In Cyprus the number of affected births in 1974 was 51, in 1979 it was 8, and there were no affected births between 2002 and 2007. Similar marked reductions have been reported after the introduction of antenatal screening programs in Taiwan and Guangdong China.

The example of Tay-Sachs disease screening

Carrier screening programs have also sometimes been directed to particular population groups with a high incidence of a serious disorder. For example, the recessive disorder Tay-Sachs disease (PMID 20301397) is a progressive neurodegenerative disorder that is rare in most populations (with a carrier frequency of about 1 in 300 in Europe and America), but is especially common in Ashkenazi Jews (about 1 in 27 is a carrier). This inborn error of metabolism presents with progressive weakness and loss of motor skills at between 3 and 6 months, followed by seizures, blindness, spasticity, and death usually before 5 years of age. It is caused by failure to produce the enzyme hexosaminidase A, as a result of genetic mutation in the *HEXA* gene. As a consequence, a fatty substance, GM2 ganglioside, accumulates in brain cells and nerves, damaging and eventually destroying them.

Carrier testing based on assaying serum hexosaminidase A began in 1970, when it was recognized that carriers may be distinguished from non-carriers by this assay. Testing is available through health services in many countries; the Dor Yeshorim organization also offers genetic screening to Ashkenazi Jews worldwide through orthodox Jewish High Schools and in community testing sessions. When testing is undertaken in orthodox schools, the results may not be given directly but instead be available at a later stage for couples considering marriage. This screening program has led to a significant reduction in the number of children born with Tay-Sachs disease in this community.

Preconception couple carrier screening

The object of preconception couple carrier screening is to identify couples who are each carriers of a pathogenic variant for the same severe autosomal recessive disorder (which can be any from a range of such disorders). Such screening programs have been in place over the past several decades in communities where a particular disease has a high prevalence (for example, screening for Tay-Sachs or sickle cell disease). With the advent of faster, cheaper, whole genome approaches (and the realization that everyone is a

carrier of roughly 1–10 autosomal recessive conditions), “*expanded pre-conception screening*” (ECS) approaches can offer simultaneous screening for say, 100, serious recessive conditions, regardless of population prevalence.

The chance of any one recessive condition may be extremely low, but the combined chance of finding a couple where both members are carriers for the same recessive condition is around 1 % in unselected populations. Innovative approaches that have been piloted in the general population of the Netherlands, and elsewhere, have disclosed only couple results; that is to say, prospective parents are told in cases where they both test positive for the same disease carrier state, and individual carrier results are not disclosed on the basis they will have no medical consequences. Such expanded screening can also be done in early pregnancy but at this point reproductive options are more limited (termination or not) than if preconception screening is carried out.

New genomic technologies are being exploited in cancer diagnostics

As sequencing technologies have improved in depth as well as breadth, they have played a crucial role in elucidating cancer mechanisms. Single-cell sequencing is helping to define the evolution of cancers, and the complex relationships between different cancer subclones is being defined over space and time, demonstrating the enormous heterogeneity of cancers and the difficulty of successfully treating them. Genetic and genomic technologies have also driven improvements in testing for cancer in different ways.

Diverse cancer biomarkers

Different genes associated with certain types of cancer can provide **biomarkers** of those cancers that when detected provide clinically useful information about the cancer. Different types of nucleic acid biomarker can

be found, including alleles with a specific pathogenic point mutation, oncogenic fusion genes, and specific gene expression signatures. Information from a detected biomarker can be used in different ways: to diagnose a cancer phenotype; to predict the likely response of the cancer to defined drugs; to indicate the likely clinical course of the cancer; and finally, to monitor the cancer (assessing the presence of mutant clones, and so on). [Table 11.9](#) gives examples of some of the very many different biomarkers used in cancer testing.

TABLE 11.9 EXAMPLES OF DIFFERENT ROLES FOR DNA AND GENE EXPRESSION BIOMARKERS IN CANCER TESTING

Role	Gene/expression biomarker	Cancer type (comment)
Diagnostic	<i>BCR-ABL1</i>	chronic myeloid leukemia (see Figure 10.8A)
	<i>JAK2</i>	myeloproliferative disease (specific mutations confirm diagnosis of clonal disease)
	<i>EWS-FLI1</i>	Ewing sarcoma
Predictive	<i>HER2</i>	breast cancer (amplification predicts response to anti-HER2 antibodies)
	<i>BRAF</i>	melanoma (specific point mutations predict response to specific BRAF inhibitors)
	<i>KIT,PDGFRA</i>	gastrointestinal stromal tumors (specific point mutations predict response to c-KIT/PDGFR inhibitors)

MRD, minimal residual disease.

* Multi-gene expression signatures.

Role	Gene/expression biomarker	Cancer type (comment)
Prognostic	<i>TP53</i>	chronic lymphocytic leukemia (specific point mutations are indicative of poor outcome)
	<i>BRAF</i>	metastatic colorectal cancer (specific point mutations are indicative of poor outcome)
	* <i>MammaPrint</i> (70-gene)	breast cancer (risk stratification)
	* <i>OncotypeDx</i> (21-gene)	breast cancer (risk stratification)
Disease monitoring	<i>BCR-ABL1</i>	chronic myeloid leukemia (detection of MRD)
	<i>PML-RARA</i>	acute promyelocytic leukemia (detection of MRD)

MRD, minimal residual disease.

* Multi-gene expression signatures.

Multiplex testing using targeted DNA sequencing

As described in [Box 11.2](#) on page 441, targeted DNA sequencing allows DNA sequences from any genome region of interest to be selectively captured and sequenced. Multiplex testing for panels of cancer susceptibility genes have now been adopted by many diagnostic services. There has been close liaison between oncology and genetic services to determine which cancer diagnoses have a significant germline predisposition, how this might affect treatments, and subsequent cascade testing of family members.

In the UK, the NHS genome medicine service has more recently delineated the aim to provide a uniform cancer testing service for all

cancers, and current panels for cancer (and for rare and inherited disease) are listed in the NHS National Genomic Test Directory at: <https://www.england.nhs.uk/publication/national-genomic-test-directories/>

Noninvasive cancer testing uses “liquid biopsies”

Another promising recent development is noninvasive cancer testing. Instead of taking a tumor biopsy (which can be difficult, according to the type of cancer), different approaches allow the analysis of freely circulating tumor DNA in plasma, as described in Clinical Box 14 on page 412. (They were stimulated by the application of high-throughput DNA sequencing in noninvasive prenatal testing of the fetal genome, as described above.)

The freely circulating DNA originates from cells undergoing apoptosis or necrosis, which includes originally healthy cells as well as inflamed cells and diseased cells, such as cancer cells. This means that tumor-specific variants need to be detected against a background of circulating DNA from non-tumor cells in the same individual.

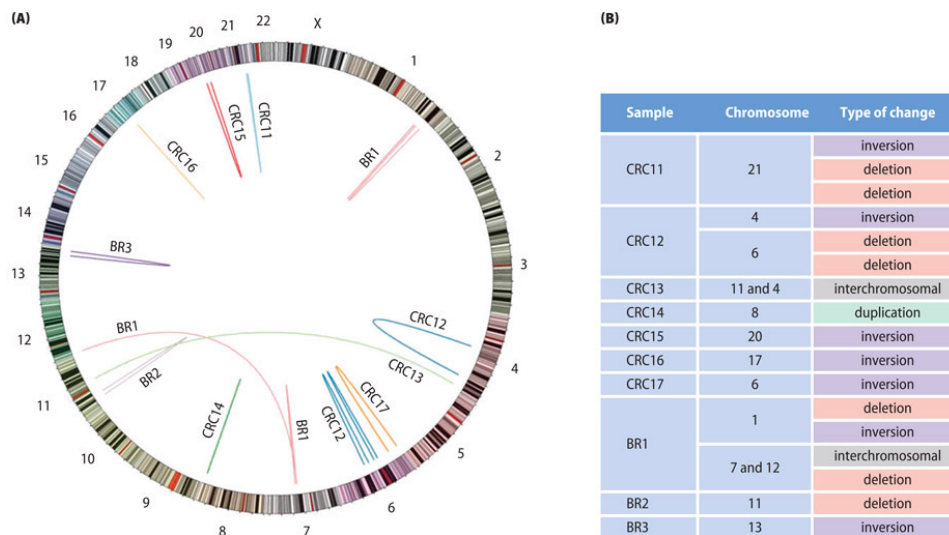


Figure 11.19 Detection of tumor-specific rearrangements by massively parallel DNA sequencing in plasma samples. (A) A Circos plot indicating rearrangements identified in tumor-cell DNA present within plasma samples from 10 cancer patients, 7 with colorectal cancer (CRC11 to CRC17), and 3 with breast cancer (BR1 to BR3). No

rearrangements were identified in DNA obtained from plasma samples from 10 unaffected controls. (B) Observed DNA rearrangements. A table with details of the breakpoint coordinates is found in the original paper. Note that droplet-digital PCR methods, described in [Section 11.2](#), can permit very sensitive quantitation of copy number variation (Adapted from Leary RJ et al. [2012] *Sci Transl Med* 4:162ra154; PMID 23197571. With permission from the AAAS.)

Massively parallel DNA sequencing can be applied to the analysis of plasma DNA. As well as detecting sequence variants it can identify tumor-specific chromosome alterations with comparative ease ([Figure 11.19](#)). Whilst this represents another technological breakthrough, the use in clinical practice will depend on the ability to detect cancers accurately at a more treatable stage. Different methods can be used to quantitate mutant alleles and copy number variants, but droplet digital PCR (described near the end of [Section 11.2](#)) can be particularly useful. Because of its high sensitivity, it can rapidly detect *minimal residual disease* (the small number of cancer cells that remain in the body after treatment that may become active, start to multiply and cause a relapse).

Bypassing healthcare services: the rise of direct-to-consumer (DTC) genetic testing

Up to this point we have considered genetic testing offered through healthcare services. That was the only option until recently, when commercial genetic testing began to be offered directly to consumers. Two stimuli in particular have led to the growth of DTC genetic services: the recent rapid decrease in the cost of genetic testing, and increasing identification of genetic variants conferring susceptibility to diseases that are common in populations.

The purpose of DTC genetic testing is quite different to that of healthcare-led genetic testing. Genetic testing organized through healthcare services is targeted to people at high risk of developing specific genetic conditions and serves to explain and/or manage health problems. By contrast, DTC testing targets healthy people and the rationale is to facilitate

life planning. The development of DTC testing is taking place in the context of a public discourse about personalized/precision medicine and genetics that tend to enthusiastically promote it in a very optimistic light, rarely dwelling on potential concerns and limitations. Such an over-optimistic perspective potentially raises inappropriate expectations of our ability to interpret what common genetic variants mean for our health.

Most DTC genetic tests rely on inexpensive SNP-chip genotyping, which checks for the presence or absence of single nucleotide polymorphisms (SNPs), or small insertions or deletions throughout a genome. SNP-chip genotyping detects common genetic variants well, but detection of very rare variants is poor: “calls” of rare cancer predisposing variants, such as pathogenic variants in *BRCA1* and *BRCA2* and those causing Lynch syndrome, are often false positives.

Genome sequencing is another method becoming more widely used in DTC genetic tests. These tests are not so vulnerable to miscalling as SNP arrays are. However, detecting variants is not the same as knowing their clinical effects—as described above, interpretation of genetic variants can be challenging and often difficult in the absence of a (familial) phenotype.

DTC tests currently sit outside much of the regulation governing clinical genetic testing, but claim to provide insight into issues as diverse as ancestry, nutrition, athletic ability, and child talent. Many testing providers also claim to help provide insight on health. The range of health information that might *potentially* be provided could include the items listed below.

- *Polygenic risk scores*—Combining many different common variants across the genome may serve to place someone in a broad risk category, such as: “your genes predispose you to weigh about 3 % more than average”. The validity and utility of these risk scores for predictive clinical purposes is hotly debated. In our opinion, although polygenic scores may be useful in researching the causes of disease, or stratifying populations into higher and lower risks,

they are rarely able to usefully predict disease—see the end of Section of 8.2 for more detail.

- *Genotyping at specific points*—looks at specific variants that influence the chance of developing particular diseases, such as: “you have two copies of the e4 variant in the *APOE* gene. People with this result have an increased risk of developing late onset Alzheimer disease.” This type of testing can also be used to identify variants that affect drug metabolism.
- *Carrier screening*—looks at specific variants to identify people who are carriers for particular recessive genetic conditions, such as: “one variant detected in the *CFTR* gene. If you and your partner are both carriers, each child may have a 25 % chance of having this condition.” Many carrier tests are ancestry specific: they test for specific carrier variants common in a particular ancestral group. If someone with a different ancestry were a carrier, this would probably not be detected because it would likely be due to a different variant (which the test would not check).
- *Uninterpreted “raw” genetic data*—See below, under DIY genetics, some DTC genetic test companies provide access to uninterpreted genetic data. Customers can download their data and seek an interpretation using third party services. These usually work by cross-referencing the data against freely available genetic databases and by constructing a report based on interpretations in these databases (which may not be up to date). They may report variants and disease risks that were not reported, or referred to, by the original DTC genetic test company, and might repurpose raw data from tests designed to answer other questions, such as ancestry, to try to provide health information.

In reality, the health information provided by many DTC companies is far from comprehensive. For example, a recent analysis of 15 DTC genetic testing companies advertising to UK consumers found that none of them complied with all the UK Human Genetics Commission’s principles for

good practice regarding consumer information. The “personalized medicine” that genetic testing promises is often portrayed in an optimistic light by the mainstream media, and genetic technology is generally presented as highly accurate. As a result, people may perceive genetic testing as clearly predictive, and expect that the results will help them plan for the future.

DIY genetics

Do-it-yourself (DIY) genetics has also risen in popularity. Here, people ask for raw data from DTC companies and process this themselves via third-party interpretation services. But many variants that are called in the raw data of a DTC test and sent for clinical confirmation are false positives. This limitation is often not appreciated by DTC customers, or the health professionals they may subsequently visit, leading to anxiety and inappropriate medical interventions.

The limitations of DTC genetic tests

The predictive value of these tests is often low when there is no family history of disease. A person with no medical or family history of disease X is informed that “you have a disease X-causing (or ‘disease X-predisposing’) genetic variant.” It may be that there are currently unmeasurable protective genetic (or other) factors in that person’s family that mean that the variant is less likely to lead to disease X in that person.

Even if a person does have a family history, identifying a “high genetic risk” via DTC genetic testing does not mean that they will definitely develop the condition. A study of people with a genetic form of diabetes found that up to 75 % of those who carry a particular missense variant in the *HNF4A* gene—specifying an R114W substitution—developed diabetes by age 40. However, a recent study looking at the same variant in UK

Biobank participants who were not pre-selected as having diabetes showed that only 10 % developed diabetes by age 40.

False positives are common, especially when SNP-chip genotyping and third-party interpretation services are used. “Miscalls” due to inaccurate genotyping of rare variants when using SNP-chips can be common: a recent study using SNP-chips to genotype very rare pathogenic *BRCA1* and *BRCA2* variants in UK Biobank participants found that 96 % were false positives. Additionally, the databases used by third party interpretation services to interpret the data may not be up to date, and variants may be classified incorrectly due to outdated evidence.

False negatives are another significant concern. This can happen because DTC genetic tests tend to prioritize breadth over detail. At the time of writing, for example, the 23andMe “genetic health risk” report for *BRCA1* and *BRCA2* currently only checks for just three disease-causing variants, which are mainly relevant for people with Ashkenazi Jewish ancestry. But there are thousands of different pathogenic *BRCA* variants that the test does not check for. As a result, about 80 % of people with disease-causing *BRCA* variants in the general population might be expected to be given false reassurance that their *BRCA1* and *BRCA2* testing was negative.

The downsides of improved sensitivity through whole genome sequencing: increased uncertainty about what variants mean

The prior probability of any one variant identified via whole genome sequencing (WGS) being causative for a patient’s rare disease is extremely low. Attempts to catalogue human genetic variation, for example via the 1000 Genomes Project, show that a typical human genome differs from the reference human genome at between 4 and 5 million sites. Most of these variations will be benign, some may subtly impact on risk of various common diseases, and a very small number will have the potential to cause serious disease.

Careful filtering of WGS datasets is therefore crucial to produce a meaningful output. This in turn requires a significant change in mindset

from the one that prevailed in the days when most pathogenic variants were identified after phenotype-driven single-gene sequencing. In those cases, identified variants would have a much higher prior probability of being causative. Now, in the complex world of WGS datasets, variants should be “innocent until proven guilty”. Translating this principle into clinical practice, however, is difficult in the context of a prevailing view that sees a genome sequence as a “blueprint” that speaks for itself. In practice, extensive filters are often applied by bioinformatic pipelines, to create virtual gene panels (as described in [Section 11.3](#)), which means that only certain variants stand a chance of becoming a communicated result.

Even when variants are consistently picked out through a pipeline, there remains considerable discrepancy in how different laboratories interpret the same variant. One study showed how laboratories using the same guidelines agreed on their classification in just one-third of cases; about a quarter were classified so differently that different medical interventions would be recommended. International guidelines for variant interpretation are helpful but it is arguably very unlikely that they will ever reach a stage where the outputs from WGS can be easily filtered to produce clinically meaningful results for a person without the need for additional expertise that links (familial) phenotype with genotype.

Improving knowledge of variant interpretation also leaves us with a difficult legacy, that some patients will have been diagnosed incorrectly with genetic conditions, yet healthcare services to date have no systematic attempts to revise the data and recontact patients seen in the past. Furthermore, there is often no threshold for communicating genetic variation of uncertain significance (VUS). There is some evidence that people misinterpret these as being definitely pathogenic or definitely benign. However, there is also evidence that many people are uncomfortable with the idea that decisions about non-disclosure might be made without involving them.

Genomic data repositories linked to phenotypes are expanding, but, as of 2022, there is a significant skewing of data from people of European ancestry. Lack of diversity in populations that contribute to genomic data

has been linked to both missed, as well as false positive, diagnoses of disease in populations of non-European ancestry. Although there have been many calls to improve the diversity of genomic data sets, and indeed reference genomes utilised, distributing the benefits of genomic research to all populations needs to address the needs of hitherto underserved communities who may already feel disenfranchised by developments in genomics that are of little apparent utility to them.

11.5 ETHICAL, LEGAL, AND SOCIETAL ISSUES (ELSI) IN GENETIC TESTING

We have described the medical advantages of genetic testing and changing landscape through rapidly evolving genetic technologies. In this section we consider the ethical, legal, and societal issues raised by these developments, often abbreviated as ELSI.

For monogenic disorders, or disorders where there is a single genetic factor that explains most of the risk, close relatives and potential offspring of persons who have inherited (and therefore “carry”) the relevant mutant alleles (or genetic variants) may be at high risk of developing the disorder. Positive identification of harmful genetic variants in one person therefore raises the stakes for unaffected relatives who may subsequently be found to have inherited the same genetic variants. As we enter an age where genetic testing—even for single gene disorders—creates a data resource on an individual’s entire genome, considerations of how such data should be stored and accessed to utilize their potential to predict current or future health issues (for both the person tested, and their biological relatives) are urgently needed.

Genetic information as family information

The familial nature of genetic information often generates discussion on confidentiality issues. The confidentiality of individual patients should be respected. But we also need to ensure that their close relatives have access

to information on possible inheritances that may be relevant for their own health and life choices. Clinical guidance in this area has increasingly taken the stance that genetic information should be seen as confidential to families, not individuals—although for a given individual the personal consequences of having a genetic change, that is, the clinical symptoms that arise from it, should be confidential to them alone. While patient confidentiality is important in genomics, as in other areas of medicine, the duty to maintain this confidentiality is not absolute: it must be balanced against others, such as the duty to prevent serious harm.

Utilizing genetic information obtained for one person to provide accurate testing for a relative is different from informing a family of the particular details of an individual patient’s medical problems. Research indicates that patients often see genetic information as belonging to their family rather than exclusively to them. Healthcare professionals, however, are often reticent about taking a family-centered approach to the confidentiality of genetic information in practice. They worry that this stance could disrupt family dynamics or erode patient trust in the health service (see [Dheensa et al. \[2017\]](#) under Further Reading).

Health professionals need to recognize the nuance above. In many cases, a default approach of not disclosing any information without written consent from specified people, is not “playing it safe” from a legal perspective; it is contrary to professional guidance and vulnerable to legal challenge. When faced with uncertainty about whether to disclose information, health professionals should undertake (and document) a balancing act, considering whether in this instance the duty to preserve individual patient confidentiality is, or is not, outweighed by the potential benefits of disclosure to a patient’s wider family. **Box 11.6** reports a recent UK court case centered around the personal versus familial nature of genetic information.

CLINICAL BOX 18 CASE STUDY: CONSENT VERSUS DUTY OF CARE TO FAMILY MEMBERS

The ABC vs St George's Case (<https://www.bailii.org/ew/cases/EWHC/QB/2020/455.html>) involved a patient at St George's Hospital who, on the grounds of diminished responsibility, had been convicted of killing his wife, and been detained in a secure unit. While there, he developed signs and symptoms consistent with Huntington disease (HD), but he did not want any of his family members to know this. The doctors noted that the man's daughters each had a 50 % chance of inheriting the condition from him, and when informed that one daughter was in the early stages of an unplanned pregnancy, they discussed whether to tell her of her risks. Although they thought she had a right to know, they felt they could not tell her this without her father's consent, which he steadfastly withheld.

After the daughter had given birth, one of her father's doctors told her, by mistake, about the HD diagnosis. She subsequently had a genetic test and found she had inherited the condition. She claimed her father's doctors were negligent in withholding this information from her. She argued that, had she known she would develop the condition herself, and be unable to raise her child, she would have had a termination of pregnancy (regardless of whether the baby had inherited the condition or not).

Although the case for negligence was dismissed, the ruling made it clear that doctors in this situation had a duty to weigh the daughter's interests in the balance, and that using the absence of her father's consent as a veto was misplaced. The ruling confirmed the professional guidance that was already in place (which recommended a breach of confidence where the harms to identifiable others were significant, and which used familial genetic examples to illustrate this). Professional guidance also urges doctors to distinguish a breach of medical confidentiality (the fact that he had a diagnosis of HD) from disclosure about the familial risk (the fact that signs, symptoms, and family history could suggest a genetic condition). Whilst the daughter might then infer she inherited HD from her father, this inference would not equate to a breach of his confidence.

Genetic testing in one person may thus raise a series of ethical (and legal) questions about current and future family members (see [Figure 11.20](#)). They include:

- Whose responsibility is it to alert relatives of their risks, and how might this best be facilitated?
- How can these familial aspects be appropriately covered in any consent process?
- What does it mean to respect confidentiality when a result might indicate others are at risk?
- When is the risk of a future condition a legitimate reason to terminate a pregnancy?
- Is it appropriate to test children for genetic conditions unlikely to manifest until adulthood?

<p>Consent</p> <ul style="list-style-type: none"> • A result in one person can reveal results in others who have not consented to such testing • Difficulty of adequately informing someone about complex genetic testing, with many possible outcomes • Testing those who lack capacity to consent for benefit of their relatives, or interest of their parents • Is individual consent necessary to share familial findings? • How to incorporate a “right not to know” genetic information 	<p>Perception that genetics is clear-cut</p> <ul style="list-style-type: none"> • Technologies now highly accurate but predicting health consequences is often more difficult • Direct-to-consumer tests can have high false positive and negative results • VUS result on a laboratory report may suggest it is pathogenic • Diversity issues: Significant skewing of ancestry in worldwide genomic databases means that interpretation of variants in non-European populations is less robust and sometimes wrong 	<p>Confidentiality</p> <ul style="list-style-type: none"> • Should a person’s confidentiality be breached to alert at-risk relatives that they too could be tested, and if so, when? • Does dealing with genetic information create a duty of care to family members not known to the clinician? • Does availability of an evidence-based intervention make a difference to familial disclosure? • Does familial genetic testing – such as trio testing – create clinical responsibilities to family members?
<p>Incidental findings</p> <ul style="list-style-type: none"> • What to do when a test finds genetic predispositions not being looked for? • Should such predispositions be sought opportunistically? • Does availability of an evidence-based intervention make a difference to above? • Is the possibility of changing reproductive options an actionable finding? • Is there a meaningful way to cover the possibility of incidental findings in a consent process? • What to do on discovery that biological relationships are different to expected 	<p>The scope of genetic testing</p> <ul style="list-style-type: none"> • Large amount of information possible from small sample (such as saliva) throughout life-course (or before and after) • Uncertain predictive results in pregnancy, yet often require a binary decision • Difficulty of anonymizing DNA/biobank databases that include whole genome analyses, with possibility of discrimination • Possible duty to reanalyze DNA or recontact past patients when evidence may have changed. How could medical systems adapt to incorporate? 	<p>Clinical-research boundary</p> <ul style="list-style-type: none"> • Clinical interpretation of variants relies on access to research databases on an iterative basis — boundary often unclear • Many genetic test results provide uncertain answers until researched in large populations • Need for system where healthcare learns directly from the [research] data of patients. Currently much data not utilized because of consent/ privacy concerns

Figure 11.20 Some ethical issues in genetic testing. VUS, variant of uncertain clinical significance.

These issues arise to some extent in any genetic testing, but given that the diagnostic and predictive power of genetic testing in multifactorial disorders is weaker they are less stark in common conditions.

Genetic health professionals may find themselves in the rather uncomfortable position of meeting, and having access to information on,

different family members who do not know about each other. Through genetic testing they may also discover that the biological relationships between some family members are different from the assumed relationships (as a result of misattributed paternity, unsuspected adoption, or sperm donation—a case study will be presented below, in the section that includes incidental findings). Routine use of *trio testing*—in which samples of both parents’ DNA determine whether suspect DNA variation in the child is *de novo* or not—significantly increases the chances of discovering misattributed biological relationships.

Consent issues in genetic testing

Consent to healthcare testing—including genetic testing—is an important aspect of respect for a person’s autonomy. The consent process is meant to ensure that a person understands the nature and purpose of giving a sample, or of undergoing the medical intervention—see [Figure 11.21](#) for the aspects to discuss during consent for genetic testing, as recommended by the UK’s Joint Committee on Genomic Medicine, and see [Figure 11.22](#) for a record of discussion form for clinical genetic or genomic testing proposed by the same organization.

(A) Points for practitioners to consider when offering genomic tests

- genomic information is often complex, and much genomic variation has unclear, uncertain or limited clinical significance
- genetic tests may be predictive or diagnostic (or have elements of both) and may accordingly have different impacts on clinical management
- some genetic variants may only predict disease well if found in the context of a medical or family history of the relevant disease
- information found in one person may inform the healthcare of their family members; approaches that focus solely on individual patients neglect this aspect and risk creating problems later on
- genomic analysis may reveal unexpected information, for example, about family relationships
- large international data collections – that transcend healthcare boundaries – will often be necessary to gather the evidence of clinical manifestations of a particular variant
- there is a clinical – research continuum; some testing and interpretation occurs at this boundary and cannot clearly be categorized as one or the other

(B) Elements for a consent discussion in genomic clinical practice

- test results may predict future health as well as diagnose current problems
- results may be relevant to other family members (consider outlining how communication with family members might be facilitated, and by whom)
- genomic tests may take longer than other medical tests: patients should be given likely timescales for availability of tests results, or components of the results.
- the scope and limits of the proposed testing (that is, what will and will not be tested for and communicated, as well as when and how)
- genomic tests may generate additional, unexpected or incidental findings (consider giving examples or outlining how these might be dealt with)
- outcomes from genomic testing may be uncertain or unclear
- interpretation of genomic results may be updated in the future and may need periodic re-evaluation
- DNA samples are routinely stored (in contrast to most other biosamples which are discarded after the tests are complete)
- stored DNA samples from one family member are routinely used as quality assurance for clinical testing in other family members
- it is sometimes necessary to share this data more widely across the NHS, or occasionally outside it, to gather evidence to inform variant interpretation or evaluation of family history; absolute anonymization may not be possible, and might compromise the utility of sharing.

Figure 11.21 Recommendations for the consent process from the UK’s Joint Committee on Genomic Medicine.

RECORD OF DISCUSSIONS regarding testing and/or storage of genetic material

I have discussed genomic/genetic testing with my health professional and I understand that:

Family implications

1. The results of my test *may* have implications for other members of my family. I acknowledge that my results may sometimes be used to inform the appropriate healthcare of others. This could be done in discussion with me, or in such a way that I am not personally identified in this process.

Uncertainty

2. The results of my test *may* reveal genetic variation whose significance is not yet known. Deciding whether such variation is significant may require sharing of information about me including (inter)national comparisons with variation in others. I acknowledge that interpretation of my results may change over time as such evidence is gathered.

Unexpected information

3. The results of my test *may* reveal a chance of a disease in the future, and nothing to do with why I am having this test. This may be found by chance, while focusing on the reason for my test, and I may then need further tests to understand what this means for me. If these additional findings are to be looked for, I will be given more information about this.

DNA storage

4. Normal laboratory practice is to store the DNA extracted from my sample even after the current testing is complete. My sample might be used as a 'quality control' for other testing, for example, that of family members.

Data storage

5. Data from my test will be stored to allow for possible future interpretations.

Health records

6. Results from my test and my test report will be part of my patient health record.

Note of other specific issues discussed (*eg referral to particular research programmes, insurance*):

I agree to genetic/genomic investigations*	DATE ____/____/____
----- Patient/parent signature	Discussion undertaken by: (clinician's name and signature) -----

Affix sticky label or fill in details

Patient name: _____

Date of birth ____/____/____

Patient address: _____

Genetics ref. _____ 1 COPY for notes, 1 COPY for patient to retain

**insert details here, eg to investigate the cause of my child's developmental delay / family history of cancer / heart disease etc*

Figure 11.22 Record of Discussion form following the consenting process, as recommended by the UK's Joint Committee on Genomic Medicine.

As a rule, the process of seeking consent ensures that a person understands the nature and purpose of the procedure, or intervention in question, thereby asserting their right to self-determination. This therefore

applies to individuals who have capacity (or competence) and is not possible for children or adults who lack capacity.

There are three essential criteria for legally valid consent:

1. The person providing consent must have sufficient, appropriate information to be able to make a decision.
2. They must be competent to make a decision.
3. The decision must be voluntarily given (free from coercion).

Qualifiers around consent can confuse this definition. For example, consent is not consent unless it is informed, so it is not always clear what the mantra of “informed consent” means. It might be more helpfully understood as “information that needs to be provided in order for a person to give their consent”.

Verifying past consent, or updating consent, so that family members can benefit, may not be possible because contact may have been lost, or it may not be clinically appropriate because the family member seeking information may be concerned about his or her confidentiality being compromised. For example, a pregnant woman who wishes to undergo prenatal diagnosis may not want anyone to know about the pregnancy until the test results are available. The necessity to seek consent from another family member for release of the information could lead to a breach of confidentiality for the pregnant woman.

In much of medical practice, an intervention or treatment may be proposed because it will directly address a health problem or its treatment. Sometimes it is so integral to a patient’s care that their consent might be assumed or inferred. For example, it would be unusual to seek specific, separate consent to check a person’s blood pressure or cholesterol level. Inferring consent to a genetic or genomic test may be more problematic because such a test may reveal many different types of results, for example, current or future health problems; predispositions to conditions that may never manifest; information that requires more research before it is of clinical utility; or information that is of (greater) relevance to relatives.

Consent to genomic testing therefore needs to incorporate the complexity, uncertainty, and open-endedness of these many types of “results”.

For adults who lack capacity, and are therefore unable to provide consent, a genetic test can be undertaken if it is believed to be in the best interests of the adult concerned. Those close to the adult, for example someone appointed as power of attorney for health and welfare, might help to inform what is in their best interest. It is important to remember, however, that capacity is decision-specific. A person may lack the capacity for certain decisions, but not others. For very young children (as in newborn screening) a person with parental responsibility may give consent for genetic testing. Generally speaking, adults are presumed to have capacity to consent, but may not have, whilst children (under 16–18 in most countries) are presumed not to have capacity but may demonstrate it depending on the age and maturity of the child.

A different consent lens for genomic testing?

Consent for medical investigations, including blood tests, is traditionally anchored at one point in time. It may need to evolve in the case of genomics, however, so that broad consent is obtained for the creation of a personal genomic resource that might be obtained, for example, at birth; subsequent clinical encounters seeking consent would be required to interrogate that resource with different questions at different stages of a person’s life.

On a practical level, returning to test this resource multiple times (rather than repeating a blood test) will require a significant change in data storage and access within a health service and need to address as well as medical record-keeping so that, for example, a whole genome sequence can be accessed two decades after birth to answer questions about adult-onset predispositions. Such a new approach would also need to be able to revise data interpretations as new knowledge is brought to bear on them. Traditional notions of informed consent are difficult to apply to situations

where the possible outcomes are so unknown, both by virtue of the individuality of the genomic data, but also due to the complexity of navigating through that data to a “result”.

The generation of genetic data is outstripping the ability to provide clinical interpretation

The threshold for initiating genetic testing as an investigation is being lowered. It may now be done simply out of interest (for example, ancestry testing) rather than through suspicion of a particular heritable factor. Consequently, the risk of overinterpreting genetic variation to predict disease increases. As mentioned in [Section 11.4](#), direct to consumer (DTC) genetic tests are sold as providing answers; people buying them may understandably expect their results to be clearly predictive of future health.

One common pitfall is to compare a genetic test result to a “zero risk,” rather than to a population risk. For example, after his polygenic risk score identified a 15 % risk of developing prostate cancer by age 75, a former British Health Minister declared that having a genetic test “may have saved my life”. Experts immediately disputed the usefulness of this result, and his interpretation of it: in the UK the average lifetime risk of a man developing prostate cancer is 18 %. More men die *with* prostate cancer rather than *from* prostate cancer.

Careful framing of results (for example, comparing them with population risks) may mitigate the risk of over-interpretation. However, this relies on information being provided in an accessible manner. Users need to know how important it is to read the information carefully, which may not be obvious in the context of a societal discourse that tends to present genetic results as strongly predictive. The assumption that DTC genetic testing empowers people to reduce their future disease risk is undermined by evidence suggesting that learning about genetic predisposition to particular diseases rarely leads to sustained lifestyle change.

The ability to generate genomic data has substantially outstripped our ability to interpret its significance for an individual (see [Figure 11.23](#) for an

analogy), and while improvements in genomic technology are in many cases driving improvements in healthcare, interpretation of what such data means in a clinical setting, and what sort of intervention should be offered as a result, lags behind. The Global Alliance for Genomics and Health (GA4GH) predicts that by 2025, over 60 million people will have had their genome sequenced in a health-care context, but as suggested above, pathways for managing the output from genome sequencing are still in their infancy.



Figure 11.23 Improvements in genomic technologies can be likened to improved efficiency in the fishing industry. Single-gene approaches are like fishing for a particular fish, that one wants for dinner and knows how to cook. Whole genome approaches, by contrast, are like trawling the ocean bed. In such a case, one may not want to use the entire yield in one go, or be able to use it, or even to recognize what is in the net. One might often be better off throwing some of the catch back into the ocean to pick them out when they are matured. (Left, From Just [Dance/Shutterstock.com](https://www.shutterstock.com/user/Dance/). Right, From Susi [Nodding/Shutterstock.com](https://www.shutterstock.com/user/Nodding/))

The detailed but unfocused approach of genomic tests gives opportunities to answer questions that go beyond the problems that led to a patient having a test (see incidental findings) but how well such questions can really be answered, and at what cost, is as yet unclear. At any given time, deciding which of the multitude of possible outputs from genomic tests should be considered a “result”—anticipated or otherwise—is challenging, not least because the links between many genetic variants and diseases are often unproven or poorly understood. Multidisciplinary input and collaboration will be key to interpreting the significance of genomic results.

New disease gene discovery and changing concepts of diagnosis

Exome and genome sequencing are powerful diagnostic tools. Take, for example, the Deciphering Developmental Disorders Project in the UK. It recruited patients with severe undiagnosed disorders (who generally already had had any currently available diagnostic genetic testing). Thereafter, exome sequencing was carried out in family trios (two parents plus affected child) to achieve a 40 % diagnosis rate for the first one thousand or so family trios in the study.

The search for a diagnosis has often been described as a journey, with parents of children with rare genetic disorders anticipating that a diagnosis may guide treatment, prognosis, acceptance and social support. However, identification of new rare disease genes may be changing the impact of receiving a diagnosis, and in many cases very little is known about the long-term effects of newly identified genetic conditions. Health professionals may find themselves in the position of learning about the effects of possible disease-causing variation(s) in a gene through meeting the patients in whom such genetic changes have been discovered.

Often the pathogenic variant will be in a gene newly thought to be linked to developmental disorders; there will be little, if any, published literature to draw on. Health professionals then have to speculate on whether the detected genetic change is the cause of the patient’s health problems, and how it will impact the patient or their family in the longer term. This has

often led to patient support and awareness groups taking on an increasingly important role, as families gather to share their lived experience of newly diagnosed rare genetic conditions, in turn informing clinical services.

The agnostic approach of exome and genome sequencing is also challenging our previous concepts of existing genetic diagnoses. Genome-wide trawls often find apparently pathogenic variants in well-described disease genes in patients whose clinical phenotypes fall outside the boundaries of the phenotype expected. For example, loss-of-function variants in *SOX2* are known to cause anophthalmia and microphthalmia in addition to other phenotypes such as developmental delay and structural brain anomalies. Eye abnormalities were thought to be a key feature of *SOX2*-related disorders, and so *SOX2* would be requested as a genetic test only in patients who had absent or small eyes. Recently, via “genotype-first” approaches, loss-of-function *SOX2* variants have been found in people with developmental delay but without anophthalmia or microphthalmia, broadening the phenotypic spectrum associated with this gene.

Complications in diagnosing mitochondrial disease

Until recently, as noted above, a diagnosis of mitochondrial disease was often made rather late in a patient’s investigative journey. With the advent of WGS plus the ability to accurately determine levels of mtDNA variants, such diagnoses might be made well before a patient exhibits the hitherto classical symptoms of mitochondrial disease. But earlier diagnosis may also mean that the range and severity of subsequent disease is difficult to predict. One reason for this is that mutant mtDNA level in blood invariably underestimates the levels present in less accessible, clinically manifesting, post-mitotic tissues, such as the brain.

Take the example of the m.3243A>G mutation in *MT-TL1*. It causes a relatively mild phenotypes (diabetes and deafness) at low mutant levels. But at higher levels, it causes complex disease presentations, including MELAS (mitochondrial encephalopathy, lactic acidosis, and stroke-like episodes). As part of mainstreaming genetics and genomics, a diabetologist may

suggest that a young person with familial diabetes should undergo WGS testing, with the application of a monogenic diabetes virtual gene panel.

Identification of a m.3243A>G variant would not only constitute a primary finding and provide a diagnosis, but also imply a risk of developing additional future phenotypes. For some of these—such as hearing loss, cardiac involvement, and renal dysfunction—screening may alter the course of the disease. But for others, such as stroke-like episodes (SLEs), there is as yet no early intervention to alter the clinical course of the disease. As mentioned, the predictive value of extrapolating m.3243A>G levels in blood to brain tissue is limited, and this raises important questions about when and how to communicate uncertain findings, especially where there is no clinical action to offer as a result.

mtDNA single-nucleotide variants are passed on by females to their offspring. However, as a random small sample of wild-type and mutant mtDNA are “bottlenecked” into each individual ovum (as illustrated in Figure 7.17 on page 214), the resultant mutant load in the child can be considerably different to that in its mother, including a much more severe phenotype in the child. A woman found to carry the m.3243A>G variant may choose to proceed with a pregnancy, have prenatal testing with the option of terminating a pregnancy with high variant levels, or (in some jurisdictions) undergo mitochondrial donation *in vitro* fertilization (see below). Each option may result in considerable anxiety, and the woman may already have children who have therefore been *de facto* tested for what is typically an adult-onset condition.

Complications arising from incidental, additional, secondary, or unexpected information

The potential for discovering “other” information depends on several factors. It might depend on what question led to genetic testing (was it a diagnostic or screening test, for example), how broad or targeted the genomic analysis is, and what level of interrogation of the genomic sequence takes place. For example, if someone has a genomic test to

investigate a familial tendency to cardiomyopathy, finding a disease-associated *BRCA1* variant may be entirely unexpected information, and incidental to the question at hand. Whether this is possible will depend on the filters applied to the bioinformatic pipeline.

When a test is requested for a particular reason, there has been much recent discussion about how far findings that may indicate future disease should be routinely sought. Such findings are usually termed additional or secondary findings when a routine search for them has been done. Those discovered whilst looking for something else are reported as incidental, or unexpected findings. A gene panel approach looking for recessive conditions in childhood may, for example, find heterozygous gene variants conferring adult-onset cancer or neurological conditions.

The ability to find genetic variants unrelated to the clinical problem that a patient presents with are an inevitable consequence of the increased sensitivity of genomic testing. This is of course not so different from other types of clinical tests: a whole-body MRI scan done to investigate one symptom may reveal a quite unsuspected tumor or aortic aneurysm, for example. But there are at least two subtle differences for incidental findings in genetic tests, as listed below. First, they may predict clinical manifestations many decades from the point of their discovery. Secondly, they may also predict clinical manifestations for close relatives.

Opportunistic finding of (other) health risks could be considered helpful, of course; but working out how to handle this information raises difficult questions. In 2013, the American College of Medical Genetics and Genomics (ACMG) suggested that when performing clinical sequencing, laboratories should automatically seek and report pathogenic variants in 56 genes associated with “medically actionable” conditions (revised in 2021 to 73 genes). The main rationale was the potential benefit of diagnosing disorders where preventative measures and/or treatments were available, with the aim of improving health.

The ACMG recommendations above proved controversial. The debate centered around whether patients should have a right to choose not to know such information. Other questions are yet to be fully addressed. They

include the following: (a) What constitutes a “medically actionable” finding? (b) What is the predictive value of such findings in the absence of a phenotype or family history of the relevant disorder? and (c) How do we reconcile this with the statement that this search is not validated for population screening?

Analysis of data from the 1000 Genomes cohort demonstrated that approximately 1 % of “healthy” people will have a “medically actionable” finding in one of the 56 ACMG-listed genes. What this might mean on an individual basis, however, is often unclear. Most of our knowledge of the effects of variation in gene *X* has been gathered by studying people identified as having a gene *X* variant, and they have been tested because of a personal or family history of gene-*X*-associated disease. That inevitably biases the sample from which our conclusions are drawn.

It is less clear what it might mean to find, for example, an apparently pathogenic variant in a gene linked to cardiomyopathy in a person with no personal or family history of heart problems (see **Box 11.7** for a case history). This has important implications for cascade testing of relatives. To what extent should testing and subsequent screening be offered in a family based on an incidental finding of a genetic variant thought to be predictive of a particular condition, if there is no clinical evidence that anyone in the family, including the person in whom the genetic variant in question was first identified, is actually affected by it?

CLINICAL BOX 19 CASE STUDY: POOR PREDICTIVE VALUE GENETIC TESTING IN ABSENCE OF CLEAR CLINICAL PHENOTYPE

A two-year-old boy was investigated for “absence spells”. He had no loss of consciousness, and after being investigated in detail for epilepsy, no abnormalities were found. The community pediatrician attempted to reassure the parents that this is a normal feature in some children, and that he would likely grow out of these spells. As a precaution, however, the boy was referred to a pediatric cardiologist, who also found no

abnormalities: his baseline ECG was defined as within normal limits, and he had no family history (to 3rd-degree relatives) of any cardiac problems.

The cardiologist had been to a presentation about mainstreaming genetics and realized that long QT syndrome (leading to increased chance of sudden cardiac death) can be difficult to diagnose in childhood. He therefore requested screening of a gene panel “to exclude long QT syndrome”. A *KCNQ1* variant associated with long QT was identified, and described on the laboratory report as “likely to be pathogenic”. A reveal device was inserted but no abnormalities in the boy’s QT interval were recorded during subsequent absence spells. As a precaution he was treated with beta blockers, and cascade genetic testing of his family was initiated. This revealed that his three-year-old sister, father, paternal aunt (and her two children, aged 4 and 8) and paternal grandfather all carried the same variant. Cardiac investigations of their phenotype, at rest, with exercise, and pharmacological challenge were all normal or equivocal. All carriers in the family were prescribed beta blockade and two members of the family were referred for possible implantable cardiac defibrillator insertion.

LEARNING POINTS

The significance of genomic variants found in the absence of a phenotype can be very unclear. It is easy to see why investigations and treatments were requested as a precaution here, but also quite possible that a significant health resource has gone into investigating and treating many members of a family when no-one is at increased risk of sudden cardiac death. Such psychological and financial costs are significant, and have the potential to be burdensome to mainstreaming agendas. It is important that the data from these sorts of examples are collected systematically, and learned from, to improve future practice.

Broad genomic testing also has the potential to detect carrier status for recessive and X-linked conditions. On a disorder-by-disorder basis, being a

carrier for a genetic condition is very rare (with notable exceptions such as hemochromatosis and cystic fibrosis). But it is very common, and “normal”, to be a carrier for a genetic condition. A population study simultaneously testing carrier status for 100 or so recessive disorders in nearly 25 000 people found that 25 % were carriers for at least one of the disorders, and 5 % were carriers for multiple disorders.

For most people, being a carrier will have no impact on their life at all. However, if their partner happens to be a carrier for the same condition, the implications can be profound: each of their children would have a 25 % chance of being affected by the genetic condition. This is particularly relevant for couples known to be biologically related, and couples with common ancestry (who will have a higher chance of both being carriers for the same recessive condition).

Because of the increased scope of carrier screening and because being a carrier for one or more recessive genetic conditions is very common, carrier results for recessive genetic conditions are increasingly conveyed on a couple basis. That is, carrier status is only communicated if relevant in the context of a particular relationship, where both individuals in a couple are carriers for the same condition (see [Section 11.4](#) above). Making the status of “being a carrier” part of normal variation would be a welcome development, but the notion of a couple’s result rather than an individual result needs careful consideration, not least in terms of recording this information in medical records.

Interestingly in the UK’s 100 000 genomes project, although participants were offered a subset of variants on the ACMG list as “additional, looked for findings” no additional findings have yet been communicated from the project (at the time of writing). That is so although consent for such extra investigations was obtained up to six years previously; many research participants no longer recall what they consented to. What is urgently needed are long term implementation projects that assess the penetrance of these variants in a general population as well as an evidence basis for interventions offered (see **Box 11.8** for a case history)

Consent issues in testing children

Genetic testing of children raises additional consent issues. Should children at risk be tested for adult-onset conditions? Or screened for carrier status for serious recessive disorders? The answer to both questions is usually no; unless there is clear medical benefit at that time, testing should be delayed until the child has the capacity to make the choice.

Sometimes parents will request such testing because they consider that such knowledge would be helpful. However, if there is no chance of a childhood onset of the condition, and no interventions or actions that can be taken now to alter the course of the condition, then a plethora of international guidelines recommend deferring such testing until the child is old enough to consent themselves. This also respects the child's "right to an open future" whereby decisions that can be delayed are, so that options for the child are not curtailed. This applies especially to conditions in which adults might sometimes choose not to be tested; testing during childhood would then deny the child the right not to give consent that he or she could exercise as an adult.

Current professional guidelines in many countries therefore stipulate that children should not normally be tested in this way unless there is clear medical benefit in early testing. Testing for familial hypercholesterolemia is one such example: early detection of a pathogenic *LDLR* mutation offers the possibility of prevention by lowering LDL-cholesterol through dietary changes and medication, and testing in *LDLR* mutations families is recommended from the age of 10 years.

CLINICAL BOX 20 CASE STUDY: MISATTRIBUTED GENETIC PARENTAGE AS AN EXAMPLE OF AN INCIDENTAL FINDING

Meena and Joe are seen in the genetics clinic after their daughter Ana is born with serious health problems. Whole exome sequencing (WES) on a sample from Ana finds two pathogenic variants in a gene associated with a

severe autosomal recessive condition. Further testing is needed to ensure that these variants were inherited on separate chromosomes, one from each parent. If so, Ann has no working copy of the gene; the true cause of her health problems has been found. Tests on parental samples show that Meena has one of the variants; Joe has neither. Further testing shows that Joe is not the biological parent of Ana.

Meena and Joe had previously been told that, as a couple, the chance of their future babies being affected by the condition was likely to be 1 in 4 (25 %). They were told that if a genetic cause for Ana's health problems were found, they could have prenatal genetic testing in future pregnancies. However, as Joe is not the genetic parent of Ana, the chance of Meena and Joe having a baby with the autosomal recessive condition would be very low. Prenatal genetic testing, with its associated miscarriage risk, would not be indicated.

KEY POINTS

- This scenario is most likely to be one of misattributed genetic paternity. However, gamete donation may result in other misattributed genetic relationships.
- Genetic testing can reveal unexpected social information as well as medical information; ideally this possibility should be made clear during the consent process, although the presence or absence of consent will not necessarily help to determine whether, when or how such a finding should be disclosed.
- While a clinician may feel uncomfortable introducing this type of “social” information into discussions, it can have medical relevance, for example, in predicting recurrence risk for future pregnancies.
- Different professional duties may arise when responding to existing information than considering whether potential information should be sought.

- Trio testing (analyzing the genome of both parents together with the child under investigation to improve the diagnostic yield) is now standard practice in the investigation of rare diseases.
- The interpretation of variation discovered through WES or WGS (whole genome sequencing) is challenging; there is an enormous amount of variation across the genome even within genes that can cause severe diseases, so that demonstrating inheritance from both parents (or in the case of a dominant condition demonstrating the finding in Ann is *de novo* and not inherited from a healthy parent) is important for diagnostic interpretation.
- The possibility of discovering such findings increases with trio testing. Some services have a policy of not disclosing such findings and/or labelling a sample as “failed” if it is not genetically related to the child. However, this may lead to repeat sampling if the reasons for the trio failure is not made explicit.

Ethical and societal issues in prenatal diagnosis and testing

Prenatal diagnosis for serious genetic disorders has long been available in many developed societies. As noninvasive prenatal screening technology develops and is standardized, the miscarriage risk of invasive procedures (chorion biopsy or amniocentesis) is becoming less of an obstacle to prenatal diagnosis; it also allows for a diagnosis at an early stage of pregnancy. After a diagnosis of a harmful genetic variant, terminating a pregnancy is accepted in many societies, although support is often far from universal. First trimester, or early second trimester, terminations are usually less traumatic medically and socially for all involved.

For those who consider that terminating a pregnancy can be justified for serious genetic disorders, another issue remains: where do we draw the line that divides serious disorders from non-serious disorders? Some couples might wish to contemplate termination for what many other people might consider mild disorders, such as congenital deafness.

And as we move towards being able to routinely analyze a fetal genome through non-invasive tests, further questions arise. In the past, prenatal genetic testing was usually only offered when a particular fetal phenotype was noted or suspected, meaning that filtering and interpretation of genetic variants identified could be anchored in attempts to explain an existing health concern. Advanced genomic testing is now increasingly used in pregnancies where there is no prior suspicion of genetic abnormality, producing information on genotype without the phenotypic data required to give it meaning. This increases the difficulty in predicting whether, and how, particular genetic variants might affect future development and health (see [page 464](#) for a worked example on prior probability).

A challenge to healthcare scientists, clinicians, and parents, therefore, is deciding what qualities prenatal genotypic variation should have in order to be constructed as a “result”. At the same time, such tests are often re-requested in order to make binary decisions about whether to continue a pregnancy or not. A range of professional organizations are developing guidelines on the use of advanced genomic testing during pregnancy. However, the discovery of ambiguous findings—such as variants with uncertain clinical significance, susceptibility loci for neurodevelopmental problems, and susceptibility to adult-onset diseases—remains a difficult management problem. Any decision to terminate—or not—will need to be made well before we know whether any of these will manifest.

Preimplantation genetic diagnosis (PGD)

For any single gene disorder, a proportion of embryos produced will be unaffected: 50 % in the case of autosomal dominant transmission where one parent carries the pathogenic variant, and 75 % in the case of autosomal recessive condition (although two-thirds of the unaffected group will be carriers). After genetic testing of embryos, PGD offers the option of implanting just those embryos expected to have a normal phenotype from

genetic testing (see Figure 11.18 on page 460 for the practicalities). It thus avoids the difficult choice to terminate a pregnancy.

In the UK, any PGD is tightly regulated by the Human Fertilisation and Embryology Authority (HFEA), and practitioners are asked to consider the welfare of the future child in their decisions around which embryos to implant. Difficult questions can arise concerning the purpose of PGD, and are hotly debated. Instead of using PGD to avoid transmission of a genetic condition, for example, might it be used instead, to select an embryo purely on the basis of a predicted HLA antigen profile that is a close match to that of an existing child in the family who needs a tissue transplant?

Newborn genome sequencing

In the preface to the previous edition of this book, although the prospect of whole genome screening of newborns (neonates) might have seemed on the distant horizon then, in 2014, we did ask this question: *might we soon live in societies in which genome sequencing of citizens becomes the norm?* Well, the time when this happens is looking much closer now. In the Americas and Europe, extrapolations of recent genome sequencing suggest that by 2030, at least 60 million citizens will have had their genomes sequenced. And should the goals be met of China's 15-year Precision Medicine Initiative—at \$9.2 billion, the largest of its kind—another 100 million genomes may be delivered by 2030.

In the UK, following the 100 000 genome project, there are plans to analyze the genomes of five million people (Our Future Health—<https://ourfuturehealth.org.uk/>). And the vision to offer newborn screening via whole genome sequencing (**Figure 11.24**) is much closer to a realistic endeavor.

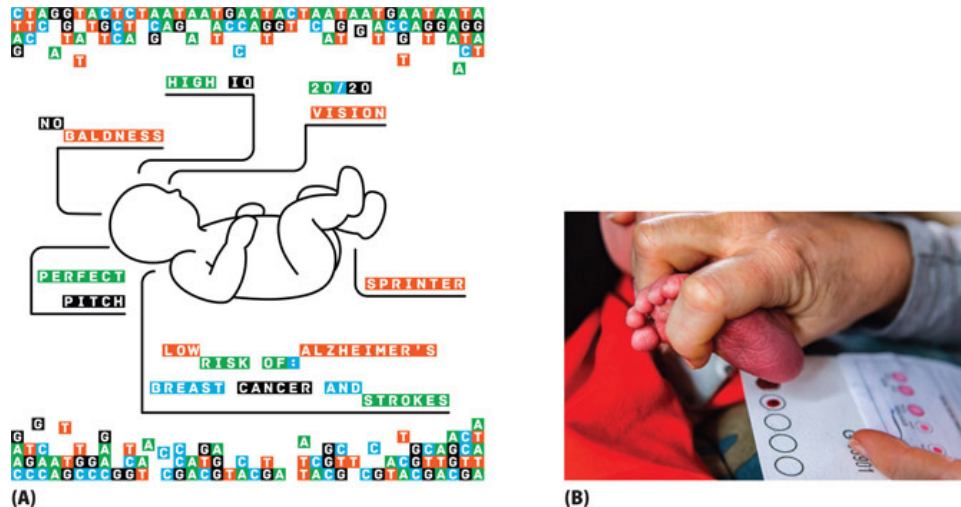


Figure 11.24 Will genome sequencing of neonates be the future norm?(A) Prenatal whole genome sequencing. From Olga Boat / [Shutterstock.com](https://www.shutterstock.com), permission. (B) From Valmedia / Science Photo Library, with permission.

At the time of writing, the UK government has just published the results of a public consultation exercise on the idea of newborn screening by whole genome sequencing (WGS), available at: <https://www.gov.uk/government/news/public-dialogue-on-the-use-of-whole-genome-sequencing-in-newborn-screening>. Its conclusions were as follows:

1. It would be acceptable to identify a wider set of conditions than the current newborn screening program if they impact the infant in early childhood and there are treatments and interventions to cure, prevent, slow progression, or personalize treatments.
2. A comprehensive genetic database should be established so that people from ethnic minority backgrounds are not disadvantaged by receiving more uncertain, or less accurate, diagnoses than the rest of the population.
3. The full complexities of whole genome sequencing must be recognized within any consent processes including:
 - a. its implications for the wider family
 - b. that 21st-century families come in many forms

- c. while parents give consent on behalf of the newborn, the child may have a different view as they grow up, including on whether their genomic data are used for research
- d. that the screening test has potential to look for many more conditions than current newborn screening tests, and that some of these may not appear for many years, or be poorly predicted by genetic variation alone.

The dialogue participants confirmed that in many ways, sequencing and analyzing genomes is the easy part. The really difficult questions revolve around how predictive the results are, what conditions it would be acceptable to look for, what information to give to whom and when, and how to help parents make informed choices about tests that could have important implications for their child, for themselves, and maybe for others in their family over many years.

Ethical and social issues in some emerging treatments for genetic disorders

Rapid developments in diagnosing and delineating molecular disease mechanisms have advanced treatment prospects for an increasing number of genetic conditions. They include the use of therapeutic monoclonal antibodies and other proteins produced by genetic engineering ([Section 9.2](#)), and various gene and RNA therapies ([Sections 9.3](#) and [9.4](#)).

The vast majority of gene therapies involve genetic modification of the somatic cells of patients, and have no consequences for future generations. Germline gene therapy has potential consequences for future generations and is widely banned. However, a recent proven treatment for certain severe mitochondrial diseases is effectively a type of germline gene therapy in which donor mtDNA becomes incorporated into the germ line. That happens by a type of *in vitro* fertilization in which mitochondria in the early embryo or egg cell are replaced by mitochondria from an oocyte donor (see

Figure 9.26 on page 356.). This type of therapy is legally permissible in the UK.

If the technology of genetic modification using CRISPR-Cas or similar genome editing method advances in the future so that it becomes highly efficient, and safe to carry out, the prospect is raised of germline genome editing. That may open the door to *genetic enhancement*, the prospect of modifying the genome to select for some quality perceived to be desirable.

We give an overview of some of the ethical and social issues raised by treatments for genetic disease and genetic enhancement in [Table 11.10](#) and enlarge on two areas immediately below this: inequality of treatment provision; and treatment for mitochondrial disease by mitochondrial replacement. And we conclude with a section on the prospects and ethics of germline modification of nuclear DNA and genetic enhancement.

TABLE 11.10 SOME ETHICAL AND SOCIAL ISSUES IN TREATING GENETIC DISEASE AND GERMLINE GENOME MODIFICATION

Treatment/genome modification	Ethical and/or social issues
Drugs provided by the pharmaceutical industry	Inequality of availability because of moderate to sometimes large costs
Invasive procedures required by new treatments	Concerns about invasive procedures needed in tiny infants, such as intubation and tracheostomy to administer Nusinersen in infants <1 year of age with severe spinal muscular atrophy. Invasive treatment can be hard to stop, once begun, even when futility becomes clear
Genetically engineered mAbs and other “recombinant” proteins	Gross inequality of availability, as a result of huge annual manufacture expenses

mAbs, monoclonal antibodies.

Treatment/genome modification	Ethical and/or social issues
Licensed somatic gene therapy and RNA therapeutics	Gross inequality of availability through huge expense of treatment
Mitochondrial replacement therapy	Some ethical concerns about germline alteration, even although the donated mtDNA is natural, as opposed to artificially altered
Germline genomic editing	Major ethical concerns about alteration of the germ line having unforeseen consequences for future generations

mAbs, monoclonal antibodies.

Inequality of treatment availability

The press typically reports advances in treatment of genetic disorders with great enthusiasm. There is often little mention of the downsides, which include the huge inequality of availability of many of these treatments because of costs that can sometimes be staggering. That can also apply to treatments used for decades, allowing refinement of the production process, not just major advances that have recently burst on the scene that might be expected to be initially expensive.

Take the example of hemophilia. An estimated 20 000 people in the US are living with this inherited bleeding disorder, and more than 60 % of them have moderate or severe hemophilia requiring lifelong treatment with expensive drugs and clotting factors. During the 1960s the average life expectancy for a patient with hemophilia was ~12 years. Recombinant factor VIII was made in 1984 and approved for medical use in the US in 1992; now people diagnosed with hemophilia can anticipate a near-normal life expectancy if treated with recombinant factors VIII or IX. In 2020 the *American Society of Hematology Clinical News* reported that treating an

adult patient by replacing factor VIII or IX with the genetically engineered recombinant protein, or by using the newer bispecific antibody emicizumab, costs somewhere in the region of \$300 000 to \$500 000 *per year*.

Of course, experimental gene therapies and RNA therapeutics are also very expensive; few have been licensed thus far. Conventional drugs produced by the pharmaceutical industry to treat or prevent genetic disease can vary in cost and availability. For example, statins and beta blockers are not so expensive and are widely available, but new drugs can be very expensive: lifetime treatment of cystic fibrosis using Vertex Pharmaceuticals' effective Trikafta drug (a combination of the elexacaftor, tezacaftor, and ivacaftor drugs described in [Section 9.2](#)) costs more than \$6 million dollars per patient.

The ethics of treating mtDNA disorders by mitochondrial donation

Recall that mitochondrial DNA (mtDNA) is present in hundreds to thousands of copies per cell and is strictly maternally inherited. Disorders in which a person is homoplasmic (100 % mutant mtDNA), or has a high percentage of mutant mtDNA, can result in a clinically severe disorder (mitochondria are the batteries of a cell; defective mitochondria especially affect organs needing the most energy: brain, muscles, and heart). Disorders like these are incurable, and reproductive choices have been mainly limited to egg donation or preimplantation genetic diagnosis to select embryos with the lowest percentage of variant mtDNA.

A woman with a *heteroplasmic* disease-causing mtDNA variant has a mix of mutant and normal mtDNA. She might have very few symptoms or be unaffected, but because of the *mitochondrial genetic bottleneck* ([Figure 7.17](#) on page 214) only a very few of the available mtDNAs pass from early primordial germ cell precursors into the egg, but in an unpredictable fashion. As a result, a heteroplasmic woman might quite often produce eggs with a high load of mutant mtDNA.

Mitochondrial replacement therapies (also known as *mitochondrial donation*) were detailed at the end of [Section 9.4](#). Two different *in vitro* fertilization methods can be used for this purpose (described in Figure 11.18 on page 460). The essential point is that in the case of a woman with a heteroplasmic mtDNA variant, the normal nuclear DNA present in the unfertilized (or fertilized) egg is removed, then injected into an enucleated donor egg containing healthy mitochondria (before being fertilized *in vitro*), or into the already fertilized enucleated donor egg. Fertilization occurs using sperm provided by the prospective father. The resulting “three-parent babies”, as sensationally reported by the world’s press, might give the erroneous impression of three equivalent genomes passed to the child; of course, the donor contributes a mitochondrial genome only, just 0.0005 % of the size of each of the two parental genomes.

The UK was the first country to regulate the use of this approach, after the Human Fertilisation and Embryology Authority (HFEA) conducted a scientific review and public consultation that informed a parliamentary debate to approve their use in a clinical setting. Interested readers can find a recent review of regulation of the method in different countries at PMID 31961722 and a neat summary of the ethical debate in a Nuffield Council of Bioethics [report](https://www.nuffieldbioethics.org/assets/pdfs/Novel_techniques_for_the_prevention_of_mitochondrial_DNA_disorders.pdf) at https://www.nuffieldbioethics.org/assets/pdfs/Novel_techniques_for_the_prevention_of_mitochondrial_DNA_disorders.pdf.

The ethics of germline gene modification for gene therapy and genetic enhancement

Mitochondrial donation therapy, as described immediately above, may be a type of germline gene therapy, but it has a clear benefit with arguably few ethical concerns. Consider that it is simply a question of replacing a small amount of damaged genetic material, and, importantly, the replaced genetic material has not been artificially edited or designed in any way. Instead, the procedures essentially involve replacing damaged mitochondria by healthy mitochondria containing natural mtDNA. The treatment is new and there

has, as yet, been limited experience, but it will be important to carefully monitor the safety of these treatments.

Germline modification of the nuclear genome, by contrast, would involve making artificial genetic changes to germline cells that might be transmitted down the generations. The current genetic technologies for modifying the nuclear genome are led by CRISPR-Cas genome editing, but impressive though this technique might be, the technology is currently imperfect. When the CRISPR/Cas9 system was used by a group in China recently to correct pathogenic variants in the *HBB* and *G6PD* genes in human zygotes, the efficiency and accuracy of the correction procedure was variable. Errors, notably “off-target” effects can be introduced unknowingly that might have harmful consequences for generations to come.

Even if the technology were refined and matured to the stage where it was efficient on every occasion with no off-target effects, there might still be unintended consequences because of our imperfect knowledge of the complex nuclear genome. At this stage, we might step back and ask why we would ever want to carry out germline modification of the nuclear genome. Studying genetically modified germline cells in culture might be considered desirable for our basic understanding of these cells, but could germline genetic modification that may be transmitted through reproduction ever be ethically acceptable? We look at two scenarios below.

Germline (nuclear) gene therapy

A review published in *Nature* in 1998 reported that panelists at a symposium on “Engineering the human germ line” held in 1998 at the University of California almost unanimously argued in favor of implementing germline gene therapy, once techniques for altering the germ line could be conducted safely and effectively in human embryos, and “regardless of the concern that its use might lead to an ethical morass” (PMID 9537311). James D Watson was reported as telling the symposium that “scientists should proceed unhindered towards germline engineering”

and advocating that such therapy must be spared excessive regulation, adding: “if there is a terrible misuse and people are dying, then we can pass regulation”. The European view was different: after a bioethics convention produced by the Council of Europe in 1997 a total of 22 European states supported the argument that such genetic manipulation should not be carried out if the aim was to introduce a permanent modification in the genome (that might be transferred down the germ line).

A possible argument in favor of germline gene therapy might be the desire to eliminate the risk of an inherited disease to future generations at the population level. However, for recessive conditions, only a very few of the disease alleles are carried by affected people (the great majority are in healthy heterozygotes), and most serious dominant or X-linked diseases are largely maintained in the population by recurrent mutation.

And, in a correspondence letter to *Nature* in 1998, in response to coverage of the “Engineering the human germ line” symposium, Anne Maclaren pointed out that there was simply no need for germline gene therapy (PMID 9565021). Rather than seek to “correct” harmful sequences in embryos, one could test cells taken from the early embryo using preimplantation genetic diagnosis, and then just implant the normal cells. After all, for nuclear genes, the highest risk for transmitting disease comes from variants associated with single-gene disorders, with at its highest, a 50 % risk in the case of dominant disorders, leaving 50 % normal embryos.

Genetic enhancement and “designer babies”

To some people, the enthusiasm for germline nuclear gene therapy might have concealed an ulterior motive. In her 1998 correspondence letter to *Nature*, Anne Maclaren publicly asked James Watson whether he had simply forgotten about the possibility of preimplantation genetic diagnosis (which would make germ-line gene therapy pointless), or whether it was germline engineering for *genetic enhancement* that he wished to proceed unhindered. No reply was disclosed.

Caught up with this selection of desired traits through genetic enhancement is the prospect that people might wish to use *in vitro* fertilization and preimplantation diagnosis simply to detect and select embryos that offer certain desired qualities, and reject the rest even though they do not harbor a genetic condition.

A demand for “designer babies” with multiple desired qualities might conceivably become a reality in the future if we had a much higher level of information about which genes to modify and if genetic manipulations on the germ line were to be so efficient that the technology became extremely safe. Some people argue that there would be a moral imperative to undertake human germline editing once the techniques are sufficient advanced, but in the real world this is mitigated by the fact that it is not usually possible to ensure a better life.

The moral arguments above also tend to rely on an overly deterministic view of a genome sequence, and the role of variation within in it, in the etiology of the disease or traits. Certainly, most common diseases cannot simply be attributed to specific genetic variants that we could edit away. Multiple, poorly understood genetic and environmental factors interact to influence the expression of diseases with a genetic component, even well understood “monogenic” disorders. As mentioned above, population-level genome analyses are now demonstrating that many genetic “mutations” are much less predictive than previously thought. Furthermore, human genome editing might introduce new risks just as it reduces old ones; or remove protections not yet clearly delineated. Similarly, the genetic basis of character traits, or particular talents, is so complex and multifactorial that acting on any such moral imperative, even if this was uniformly agreed, remains the terrain of science fiction for the foreseeable future.

SUMMARY

- The analytical validity of a test evaluates how well the assay measures what it claims to measure. Many genetic tests with

high analytical validity have low or absent clinical validity.

- A genetic test assay is said to have a high sensitivity if a high proportion of all people with the condition are correctly identified as such, and a high specificity if a high proportion of all people who do not have the condition are correctly identified as such.
- Healthcare service-led genetic tests may be used to confirm a clinical diagnosis, predict the likelihood of developing or transmitting a genetic disorder, predict the clinical course, or help monitor disease. (Some additional tests assay for drug responses, as described in [Section 9.1.](#))
- Most genetic tests are designed to detect chromo-some abnormalities or pathogenic DNA variants. They may involve scanning for an undefined abnormal DNA variant, or testing for a one or more defined pathogenic variants.
- Rather than detect causative genetic variants, some genetic tests indirectly assay a convenient disease-associated characteristic, either altered expression products, altered gene function, or a characteristic disease biomarker such as an abnormally elevated metabolite.
- Genotyping specific single-nucleotide variants often makes use of pairs of allele-specific oligonucleotides that hybridize specifically to template DNA with either the normal or variant sequence.
- Targeted DNA sequencing means using biotin-streptavidin capture of desired genome sequences so that they can be selectively analyzed by DNA sequencing. It may be used to capture multiple genes associated with a specific disease or group of diseases (gene panels), or a whole exome (containing all coding sequences and some untranslated sequences).
- Whole genome sequencing (WGS) may be used in identifying rare disease genes but is comparatively expensive.

Identification of a pathogenic variant is not eased by the sheer number of background genetic variants, and difficulties in interpreting some variants.

- In broad genome scans additional pathogenic variants may incidentally be found that are associated with phenotypes other than those for which the test was ordered (incidental or secondary findings).
- As the costs of WGS fall, virtual gene panels are increasingly used: bioinformatic filters are applied to screen out most of the genome sequence of patients, leaving genes of interest, such as all genes associated with heart disease or mitochondrial disease.
- Assessing the pathogenicity of sequence variants can be difficult. Identifying precedence (previous occurrences of the variant), genomic constraint (strong evolutionary conservation of the sequence at the mutation site), and rarity of the sequence variant is often helpful.
- Genetic screening means carrying out proactive assays to identify individuals at increased risk of carrying harmful genetic variants. Past approaches to target particular communities or populations with an elevated incidence of, for example, a particular autosomal recessive disorder, are rapidly being replaced by screening of entire populations.
- Traditional prenatal diagnosis has used invasive procedures to recover and analyze fetal cells from early pregnancy. In preimplantation diagnosis, genetic testing occurs on embryos produced by *in vitro* fertilization in the context of assisted reproduction.
- In preconception screening, couples can be screened for many different autosomal recessive conditions. Carrier couples then have more reproductive options than when this is first discovered in pregnancy, or after birth.

- In noninvasive prenatal testing, samples of freely circulating DNA recovered from maternal plasma are analyzed. The plasma DNA is a mixture of fetal and maternal DNA issuing from degraded cells. It can be analyzed to infer fetal DNA variants and the fetal genome sequence.
- Cascade testing means testing of relatives after identifying a person with a pathogenic mutation. Relatives will be at a higher risk (than the general population) of being carriers of a recessive disorder or chromosome translocation, or of developing a childhood-onset or late-onset dominant disorder.
- Pre-symptomatic diagnosis can be carried out on asymptomatic individuals at risk of developing a genetic disorder later in life. If a person is identified as carrying the mutant allele, follow-up screening can be carried out, and in some cases treatment regimes can be followed to reduce disease risk.
- In direct-to-consumer genetic testing, commercial companies carry out genetic tests and feed back results without involving healthcare professionals. The main focus may be on genetic ancestry, but predictions about future health risks are often given.
- Genetic testing for susceptibility to common diseases can identify individuals at increased disease risk; because the disease susceptibility here is multifactorial, even the best polygenic risk tests can measure just the genetic component to disease risk.
- Mainstreaming genetics envisages incorporation of genetic testing into mainstream medicine—much as, say, radiology or hematology has been incorporated in the past. We still need radiologists to interpret complex imaging; it seems likely that genetic professionals will be asked to fulfil similar advisory roles.

- Clinical genome sequencing can identify pathogenic variants, but if done without reference to a phenotype (for example prenatally), predictions are often much less clear than people expect. Furthermore, each person has a large number of variants whose clinical significance is weak or uncertain.
- Clinical genome sequencing is being incorporated into existing healthcare systems of many economically advanced societies. Significant bioinformatic and electronic networking challenges persist, as do ethical concerns about releasing data when we currently have imperfect knowledge of the clinical significance of many variants.
- Genetic testing is unusual in that the results often have potential implications for close relatives, as well as for the person tested. Professionals may sometimes need to balance preserving the confidences of one person with the prevention of harm to relatives (by alerting relatives to particular screening, for example).
- Clinical information about the person tested should be held in confidence, but the genetic factor that led to a diagnosis might be considered confidential to several family members. This means that relatives can sometimes be alerted to their risk without breaking the confidences of others.
- Consent for genetic testing should address the complexities of genetics including implications for family members, the fact that uncertain information may be found, or that interpretation of findings may change over time.
- Genetic testing of children should only be done if it benefits them as children. If the test predicts adult-onset conditions for which there is no beneficial intervention in childhood, testing should usually be delayed until the child can be involved in the decision-making process. Disclosure of data that would enable such predictions to be made—from whole exome

sequencing at birth, for example—should not be done until it benefits the child.

- Treatment of genetic disease normally has direct consequences for just the person treated. Genetically modifying the germ line would potentially have consequences for future descendants and is widely banned.
- Mitochondrial replacement (also called donation), a way of avoiding transmission of severe mitochondrial disorders, is a type of germ line modification that simply replaces the mitochondria of a heteroplasmic woman by intact mitochondria from a donor egg cell.

QUESTIONS

Questions can be downloaded by visiting the following link, under Support Materials: www.routledge.com/9780367490812.

FURTHER READING

Genetic testing overviews and resources

Genetic Testing Registry [An electronic resource established by the US National Institutes of Health to serve as a central repository of genetic tests]. Available at: <https://www.ncbi.nlm.nih.gov/gtr/>

Katsanis SH & Katsanis N (2013) Molecular genetic testing and the future of clinical genomics. *Nat Rev Genet* 14:415–426; PMID 23681062.

Korf BR & Rehm HL (2013) New approaches to molecular diagnosis. *J Am Med Assoc* 309:1511–1521; PMID 23571590.

Identifying chromosome abnormalities and large-scale DNA changes

- Cusenza VY (2021) Copy number variation and rearrangements assessment in cancer: comparison of droplet digital PCR with the current approaches. *Int J Mol Sci* 22, 4732; PMID 33946969.
- Mann K & Ogilvie CM (2012) QF-PCR: application, overview and review of the literature. *Prenatal Diagn* 32:309–314; PMID 22467160.
- Schaffer LG (2013) Microarray-based cytogenetics. In Gersen SL & Keagle MB (Eds), *The Principles of Clinical Cytogenetics*, 3rd ed., pp. 441–450. Springer.
- Wapner RJ (2012) Chromosomal microarray versus karyotyping for prenatal diagnosis. *N Engl J Med* 367:2175–2184; PMID 23215555.
- Willis AS (2012) Multiplex ligation-dependent probe amplification (MLPA) and *prenatal diagnosis*. *Prenatal Diagn* 32:315–320; PMID 22467161.

Genotyping point mutations and DNA methylation profiling

- Heyn H & Esteller M (2012) DNA methylation profiling in the clinic: applications and challenges. *Nat Rev Genet* 13:679–692; PMID 22945394.
- Kwok PY (2001) Methods for genotyping single nucleotide polymorphisms. *Annu Rev Genomics Hum Genet* 2:235–258; PMID 11701650.
- Syvanen A-C (2001) Accessing genetic variation: genotyping single nucleotide polymorphisms. *Nat Rev Genet* 2:930–942; PMID 11733746.
- von Kanel T & Huber AR (2013) DNA methylation analysis. *Swiss Med Wkly* 143:w13799; PMID 23740463.

Genome-wide and disease-targeted sequencing in mutation scanning

Choi M (2009) Genetic diagnosis by whole exome capture and massively parallel DNA sequencing. *Proc Natl Acad Sci USA* 106:19096–19101; PMID 19861545.

Rehm HL (2013) Disease-targeted sequencing: a cornerstone in the clinic. *Nat Rev Genet* 14:295–299; PMID 23478348.

Yang Y (2013) Clinical whole exome sequencing for the diagnosis of mendelian disorders. *N Engl J Med* 369:1502–1511; PMID 24088041.

Interpreting and classifying sequence variants (see also [Table 11.5 on page 447](#))

Findlay GM (2021) Linking genome variants to disease: scalable approaches to test the functional impact of human mutations. *Hum Mol Genet* 30:R187–R197; PMID 34338757.

Hanna RE (2021) Massively parallel assessment of human variants with base editor screens. *Cell* 184, 1064–1080; PMID 33606977.

[Richards S](#) (2015) Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. *Genet Med* 17:405–424; PMID 25741868.

Wong AK (2021) Decoding disease: from genomes to networks to phenotypes. *Nat Rev Genet* 22:774–790; PMID 34341555.

Wright CF (2019) Assessing the pathogenicity, penetrance, and expressivity of putative disease-causing variants in a population setting. *Am J Hum Genet* 104(2):275–286; PMID 30665703.

Genetic counseling and prenatal genetic testing

- Brezina PR (2012) Preimplantation genetic testing. *Br Med J* 345:e5908; PMID 22990995.
- Clarke A (2019) *Harper's Practical Genetic Counselling*, 8th ed. Hodder-Arnold.
- Hui L & Bianchi DW (2013) Recent advances in the prenatal interrogation of the human fetal genome. *Trends Genet* 29:84–91; PMID 23158400.
- Lo YMD & Chiu RWK (2012) Genomic analysis of fetal nucleic acids in maternal blood. *Annu Rev Genom Hum Genet* 13:285–306; PMID 22657389.

Predictive testing and genetic screening

- Umbarger MA (2014) Next-generation carrier screening. *Genet Med* 16:132–140; PMID 23765052.
- Wilcken B (2011) Newborn screening: how are we travelling, and where should we be going? *J Inher Metab Dis* 34:569–574; PMID 21499716.

Genomic medicine

- Manolio TA (2019) Opportunities, resources, and techniques for implementing genomics in clinical care. *Lancet* 394:511–520; PMID 31395439.
- Shendure J (2019) Genomic medicine—progress, pitfalls and promise. *Cell* 177:45–67; PMID 30901547.
- Snape K (2019) The new genomic medicine service and implications for patients. *Clin Med* 19(4):273–277; PMID 31308102.
- Stark Z (2019) Integrating genomics into healthcare: a global responsibility. *Am J Hum Genet* 104;13–20; PMID 30609404.
- Williams MS (2019) Early lessons from the implementation of genomic medicine programs. *Annu Rev Genom Hum Genet* 20:389–411; PMID 30811224.

Wise AL (2019) Genomic medicine for undiagnosed diseases. *Lancet* 394:533–540; PMID 31395441.

Genetic testing of cancers and common genetic disease

Berger MF & Mardis ER (2018) The emerging clinical relevance of genomics in cancer medicine. *Nat Rev Clin Oncol* 15:353–365; PMID 29599476.

Gonzalez de Castro D (2013) Personalized cancer medicine: molecular diagnostics, predictive biomarkers and drug resistance. *Clin Pharmacol Ther* 93:252–259; PMID 23361103.

Mars N (2020) Polygenic and clinical risk scores and their impact on age at onset and prediction of cardiometabolic diseases and common cancers. *Nat Med* 26:549–557; PMID 32273609.

Sud A (2021). Will polygenic risk scores for cancer ever be clinically useful? *NPJ Precis Oncol* 5:40; PMID 34021222.

Wald NJ & Old R (2019) The illusion of polygenic disease risk prediction. *Genet Med* 21:1705–1707; PMID 30635622.

Wang L & Wheeler DA (2014) Genome sequencing for cancer diagnosis and therapy. *Annu Rev Med* 65, 33–48; PMID 24274147.

Ethical issues in genetic testing

Caulfield T & McGuire AL (2012) Direct-to-consumer genetic testing: perceptions, problems and policy responses. *Annu Rev Med* 63:23–33; PMID 21888511.

de Jong A (2011) Advances in prenatal screening: the ethical dimension. *Nat Rev Genet* 12:657–663; PMID 21850045.

[Dheensa S](#) (2017) Approaching confidentiality at a familial level in genomic medicine: a focus group study with healthcare professionals. *BMJ Open* 7:e012443; PMID 28159847.

- Lucassen A & Gilbar R (2018) Alerting relatives about heritable risks: the limits of confidentiality. *BMJ* 361:k1409; PMID 29622529.
- Hollands GJ (2016) The impact of communicating genetic risks of disease on risk reducing health behaviour: systematic review with meta-analysis. *BMJ* 352:i1102; PMID 26979548.
- Manrai AK (2016) Genetic misdiagnoses and the potential for health disparities. *N Engl J Med* 375:655–665; PMID 27532831.
- Marcon AR (2018) Representing a “revolution”: how the popular press has portrayed personalized medicine. *Genet Med* 20:950–956; PMID 29300377.
- Ross LF (2013) Technical report: ethical and policy issues in genetic testing and screening of children. *Genet Med* 15:234–245; PMID 23429433.

Clinical and public health genomics: challenges and ethics

- Ackerman JP (2016) The promise and peril of precision medicine: phenotyping still matters most. *Mayo Clin Proc* 91:1606–1616; PMID 27810088.
- Bertier G (2018) Is it research or is it clinical? Revisiting an old frontier through the lens of next-generation sequencing technologies. *Eur J Med Genet* 61:634–641; PMID 29704685.
- Dheensa S (2018) Towards a national genomics medicine service: the challenges facing clinical-research hybrid practices and the case of the 100 000 genomes project. *J Med Ethics* 44(6):397–403.
- Faden RR (2013) An ethics framework for a learning health care system: a departure from traditional research ethics and clinical ethics. *Hastings Cent Rep* 43:S16–S27; PMID 23315888.
- Jackson L (2021) Use of SNP chips to detect rare pathogenic variants: retrospective, population based diagnostic evaluation *BMJ* 372:n214; PMID 33589468.

- Johnson SB (2020). Rethinking the ethical principles of genomic medicine services. *Eur J Hum Genet* 28:147–154; PMID 31534213.
- McEwen JE (2013) Evolving approaches to the ethical management of genome data. *Trends Genet* 29:375–382; PMID 23453621.
- Miller DT (2021) *Recommendations for reporting of secondary findings in clinical exome and genome sequencing, 2021 update: a policy statement of the American College of Medical Genetics and Genomics* *Genet Med* 23(8):1391–1398; PMID 34012069.
- Milne R (2021) Demonstrating trustworthiness when collecting and sharing genomic data: public views across 22 countries. *Genome Med* 13:92; PMID 34034801.
- Nuffield Council on Bioethics (2018) *Genome Editing and Human Reproduction: Social and Ethical Issues*. Nuffield Council on Bioethics, London.
- Popejoy AB (2018) The clinical imperative for inclusivity: race, ethnicity, and ancestry (REA) in genomics. *Hum Mutat* 39:1713–1720; PMID 30311373.
- Wright CF (2019) Genomic variant sharing: a position statement. *Wellcome Open Res* 4:22; PMID 31886409.
- Wright CF (2019) When genomic medicine reveals misattributed genetic relationships-the debate about disclosure revisited. *Genet Med* 21:97–101; PMID 29904162.

Glossary

3'end

The end of a DNA or RNA strand that is linked to the rest of the chain only by carbon 5' of the sugar, not carbon 3' ([Box 1.1](#), [Figure 1](#)).

5'end

The end of a DNA or RNA strand that is linked to the rest of the chain only by carbon 3' of the sugar, not carbon 5' ([Box 1.1](#), [Figure 1](#)).

adaptive immunity/adaptive immune system

Specific immune responses that rely on the recognition of foreign antigen by antibodies and T-cell receptors.

allele frequency

The frequency of an allele in a population; that is, the proportion of all alleles at a locus that are the allele in question (often inaccurately represented as gene frequency).

allele

Individual version of a gene or DNA sequence at a locus on a single chromosome; often also used loosely to describe genetic variants at the protein level.

allogeneic

Describing cell and organ transplantation (or the transplanted cells) in which the donor cells are genetically different from that of the recipient. Compare *autologous*.

amino acid

The fundamental repeating unit of a polypeptide; a building block for a protein ([Figure 2.2](#) and [Table 7.3](#)).

amplification

1. An artificial increase in DNA sequence copy number as a result of *cloning* or *PCR* ([Section 3.1](#)). 2. A natural increase in gene copy number in response to natural selection in organisms ([Figure 4.8](#)) or tumors ([Figure 10.7](#)).

anaphase lag

Loss of a chromosome because it moves too slowly at anaphase to get incorporated into a daughter nucleus.

aneuploidy

A chromosome constitution with one or more chromosomes extra or missing from a full (euploid) set – see pp. 211–2.

angiogenesis

Process whereby new blood vessels are formed by sprouting from existing vessels.

annealing

Process whereby two single-stranded nucleic acids form a stable double-stranded nucleic acid by *base pairing*. The reverse of denaturation.

anticipation

The tendency for the severity of a condition to increase in successive generations (p. 129). Commonly due to bias of ascertainment, but a genuine outcome in the case of some *dynamic mutations*.

antigen

A molecule that can induce an adaptive immune response or that can bind to an antibody or T-cell receptor.

antigen presentation

The process by which antigen is presented in combination with an MHC (HLA) protein on the surface of certain cells so that it can be recognized by receptors on lymphocytes ([Section 4.4](#) and [Box 8.3Figure 1](#)).

antisense RNA

An RNA transcript that has a *complementary sequence* to a mRNA (or some functional noncoding RNA). Naturally occurring antisense RNAs, made using the non-template strand of a gene, are important regulators of gene expression.

antisense (or template) strand

The DNA strand of a gene that, during transcription, is used as a template by RNA polymerase for the synthesis of mRNA ([Figure 2.1](#)).

apoptosis

A natural way of getting rid of unwanted or diseased cells in which the cell is targeted for destruction by various stimuli. Rapid fragmentation of the cell follows, after which the resulting cell fragments are phagocytosed by neighboring cells.

association

A tendency of two *characters* (such as diseases or marker alleles) to occur together at nonrandom frequencies. Association is a simple statistical observation, not a genetic phenomenon, but can be caused by *linkage disequilibrium* ([Section 8.2](#)).

autoimmune disorders

Diseases that arise because the distinction between self and nonself fails so that the body mounts an abnormal immune response against one or more self molecules.

autologous

Describing cells or tissues that were obtained from or pertain to the same individual.

autosome

Any chromosome other than the sex chromosomes, X and Y.

autozygosity

In an inbred person, homozygosity for alleles identical by descent.

balancing selection

Selection working simultaneously in opposite directions on the same variant; can result in heterozygotes for a harmful mutation having a higher biological *fitness* than normal homozygotes (p. 136).

base complementarity

The relationship between bases on opposite strands of a double-stranded nucleic acid: A always occurs opposite T (or U in RNA) and G always occurs opposite C in DNA (but in RNA, G sometimes base pairs with U).

base pair/base pairing

The outcome/process of stable hydrogen bonding between two complementary bases, a purine and a pyrimidine. The bases may reside on opposing strands of a duplex nucleic acid ([Figure 1.4](#)), or on the same RNA strand ([Figure 2.4A](#)). Transient DNA-RNA and RNA-RNA base pairing can allow functional interaction between different molecules.

benign tumor

An abnormal cell growth that is confined to a specific site within a tissue and shows no evidence of invading adjacent tissue.

biologic

Any biological drug, such as a therapeutic monoclonal antibody or recombinant protein.

biomarker

Any characteristic that is objectively measured and evaluated as an indicator of normal biological processes, pathogenic processes, or pharmacologic responses to a therapeutic intervention.

biotin–streptavidin system

A tool for isolating labeled molecules. The bacterial protein streptavidin happens to bind biotin (vitamin B7) with exceptionally high affinity. Biotinylated molecules can be isolated by using streptavidin-coated magnetic beads ([Figure 8.7](#)).

blastocyst

An embryo at a very early stage of development when it consists of a hollow ball of cells with a fluid-filled internal compartment ([Figure 2](#) in [Box 9.2](#)).

blastomere

One of the multiple cells formed when the fertilized egg undergoes cleavage divisions.

boundary elements

Regulatory DNA sequences that define the boundary between differentially regulated loci by limiting or opposing the action of enhancer elements

capping

A stage in RNA processing. A special nucleotide, 7-methylguanosine triphosphate, is joined by a 5'–5' phosphodiester bond to the 5' end of a primary transcript. Capping is important for the stability of the RNA.

cancer

1. One of a heterogeneous group of disorders whose common features are uncontrolled cell growth and cell spreading. 2. A tumor that has become *malignant*.

carrier

A person, usually asymptomatic, who carries a genetic variant that can cause disease after being transmitted to the next generation, or that can contribute to disease in later life.

case-control study

A study in which samples from affected individuals (cases) are analyzed and compared with equivalent samples from unaffected control individuals.

cDNA (complementary DNA)

DNA synthesized by the enzyme reverse transcriptase using RNA (often mRNA) as a template.

centromere

The primary constriction of a chromosome, separating the short arm from the long arm, and the point at which spindle fibers attach to pull chromatids apart during cell division.

character (or trait)

An observable property of an individual, such as eye color or ABO blood group type.

chimera

An organism derived from more than one zygote.

chromatid

One of a pair of sister chromatids that form when a chromosome replicates and persist until the anaphase stage of mitosis (see [Figure 1.10](#)).

chromatin

The nucleoprotein material of a chromosome.

chromatin remodeling

Movement, dissociation, or reconstitution of nucleosomes in chromatin, as part of the systems controlling chromatin conformation.

chromosomal microarray analysis

Clinical application of *microarray hybridization*. The usual object is to scan a genomic DNA sample for changes in copy number (deletions or duplications) of large DNA segments.

chromosome

In eukaryotes, a nucleoprotein structure formed when a nuclear DNA molecule is complexed with various types of proteins and occasionally some RNAs. The complexing helps compact the immensely long DNA molecules.

cis-acting

(of gene regulation by short sequence elements) Term used to describe any gene regulation in which a regulatory DNA or RNA sequence controls the expression of some other sequence present on the *same* nucleic acid molecule ([Figures 6.1, 6.2](#)).

cis-acting RNA

a type of regulatory long noncoding antisense RNA that remains attached to the DNA strand from which it is transcribed but can base pair with the sense strand transcribed from the opposing DNA strand of the same DNA molecule, and thereby regulate its expression.

clones/cloning

Identical copies (of a DNA sequence, a cell, or an organism)/process of making the same. In genetic research, this often means cells containing identical recombinant DNA molecules.

CNV

See *copy number variation*.

coding DNA

A segment of DNA whose sequence is used directly to specify a polypeptide (via a mRNA).

co-dominant

Term used to describe a heterozygous state in which both alleles are fully expressed.

codon

A sequence of three nucleotides (strictly in mRNA, but by extension, in genomic coding DNA) that specifies an amino acid or a translation stop signal.

coefficient of inbreeding

The proportion of loci at which a person is homozygous by virtue of the consanguinity of their parents ([Section 5.2](#)).

coefficient of relationship

Of two people, the proportion of loci at which they share alleles identical by descent ([Box 5.2](#)).

complementary sequences (or strands)

Nucleic acid sequences (or strands) that can form a stable double-stranded nucleic acid by *base pairing*.

complementary DNA

See *cDNA*.

compound heterozygote

A person with two different mutant alleles at a locus.

conformation

Of a complex molecule, the three-dimensional shape—the result of the combined effects of many weak noncovalent bonds.

consanguineous

Description of persons who are closely related because they have descended from a very recent common ancestor (often within the previous three or four generations), usually as a result of a marriage between cousins.

conservative substitution

A nucleotide substitution that changes a codon so that it makes a different, but chemically similar, amino acid.

conserved sequence

DNA or amino acid sequence that is identical or recognizably similar across a range of organisms, suggestive of an important function.

constitutional

(of genetic variation, mutation, chromosome abnormality) Present in the genetic material of the zygote, and therefore present in every nucleated cell of a person.

constitutive heterochromatin

Heterochromatin that remains condensed throughout the cell cycle. Found at centromeres plus some other regions. See [Box 2.3](#), [Figure 2.8](#).

copy number variation (CNV)

Variation between individuals in the number of copies in their genomes of a specific, moderately long to large DNA sequence (from hundreds of base pairs to many megabases). The term CNV is also used to denote a rare copy number variant (frequency less than 1%); if the frequency is above 1%, copy number polymorphism (CNP) is often used (pp. 91–2).

CpG island

Short stretch of DNA, often less than 1 kb long, containing frequent unmethylated CpG dinucleotides. CpG islands tend to mark the 5' ends of genes ([Box 6.1](#)).

CRISPR-Cas

A type of natural prokaryotic adaptive immunity. Adapted as a *genome editing* technique that uses artificial RNA *guide sequences*. See [Figures 9.22, 9.24](#).

cryptic splice site

A sequence in pre-mRNA with significant homology to a splice site. Cryptic splice sites may be used as splice sites when splicing is disturbed or after a base substitution mutation that increases the resemblance to a normal splice site ([Figure 7.4](#)).

cross-linking

(in DNA) Abnormal occurrence of covalent bonds directly linking two bases. The cross-linked bases may be on the same strand or on opposite strands ([Figure 4.1](#)). In proteins, the disulfide bond is a natural form of cross-linking ([Figure 2.5](#)).

crossover

An act of meiotic *recombination*, or the physical manifestation of that (as seen under the microscope) ([Figures 1.13](#) and [1.14](#)).

cytokines

Extracellular signaling proteins or peptides that act as local mediators in cell–cell communication.

dedifferentiation

Epigenetic reprogramming of a *differentiated cell* so that the cell becomes less specialized ([Box 9.1](#)).

denaturation

1. Dissociation of double-stranded nucleic acid to give single strands.
2. Destruction of the three-dimensional structure of a protein by heat or high pH.

derivative chromosome

A chromosome that has been structurally rearranged, for example by translocation, but retains a centromere.

differentiation

(of a cell) Natural process of epigenetic modification that causes a cell to become more specialized.

diploid

Having two copies of each type of chromosome; the normal constitution of most human somatic cells.

direct repeats

Two or more copies of a sequence that occur in the same 5' → 3' direction on a single DNA strand. Usually used to mean repeats that are separated on the DNA; repeats that are directly adjacent to one another are normally described as *tandem repeats*.

distal

(of chromosome) Comparatively distant from the centromere ([Box 7.2](#)).

DNA libraries

The result of cloning random DNA fragments or molecules to produce a collection of cells containing different recombinant DNAs (which must then be screened to find any desired sequence).

dominant

In human genetics, any trait that is expressed in a heterozygote.

dominant-negative effect

The situation in which a mutant protein antagonizes the function of its normal counterpart in a heterozygous person ([Figure 7.19](#)).

dosage-sensitive gene

A chromosomal gene that, when present in one copy instead of the normal two copies is associated with disease (*haploinsufficiency*), or that can cause disease when overexpressed ([Box 7.3](#)).

driver mutations

In cancer, mutations that assist tumor development, being subject to positive selection during tumorigenesis as opposed to passenger mutations in tumorigenesis (which are not positively selected or causally implicated in cancer development).

duplex

A double-stranded nucleic acid.

dynamic mutation

An unstable expanded repeat that changes in size between parent and child ([Section 7.3](#)).

embryonic stem (ES) cell line

Embryonic stem cells that have continued to proliferate after subculturing for a period of 6 months or longer and that are judged to be *pluripotent* and genetically normal.

endonuclease

An enzyme that cuts DNA or RNA at an internal position in the chain.

enhancer

A set of clustered short sequence elements that stimulate the transcription of a gene and whose function is not critically dependent on their precise position or orientation ([Section 6.1](#)).

epigenetic

Heritable (from mother cell to daughter cell, or sometimes from parent to offspring), but not produced by a change in DNA sequence.

epigenetic marks (or settings)

Patterns of epigenetic modification, such as DNA methylation, histone modification, and nucleosome spacing patterns that permit chromatin to switch between open (transcriptionally active) and condensed (transcriptionally inactive) forms.

epigenome

The totality of epigenetic marks in a cell.

epimutation

A change in chromatin organization causing a change in expression of one or more genes *without any change to the DNA sequence* ([Figure 6.21](#)). Can be induced by mutation at a distant gene locus regulating chromatin modification, by environmental factors (resulting in metabolic changes or inflammation, for example), and certain chromosome abnormalities (*position effects*).

episome

Any DNA sequence that can exist in an autonomous (self-replicating) extrachromosomal form in the cell.

epistasis

Literally 'standing above'. Gene A is epistatic to gene B if A functions upstream of B in a common pathway. Loss of function of A will cause all the effects of loss of function of B, and maybe other effects as well.

epitope

The part of an immunogenic molecule to which an antibody responds.

euchromatin

The fraction of the nuclear genome that contains transcriptionally active DNA and that, unlike heterochromatin, adopts a relatively extended conformation.

exon

Originally, any segment of an RNA transcript that is retained during RNA *splicing*, but now used widely to mean the corresponding sequence in genomic DNA. Individual exons may contain coding sequences that are translated and/or noncoding sequences ([Figure 2.1](#)).

exome

The totality of exons in a genome.

exon shuffling

An evolutionary process in which exons from one gene are copied and inserted into a different gene ([Figure 2.16](#)).

exon skipping

Occasional failure to include an exon within an RNA transcript ([Figure 6.7](#)).

exonuclease

An enzyme that digests a DNA or RNA strand from one end. It may be a 3' or 5' exonuclease.

facultative heterochromatin

Heterochromatin that may reversibly decondense to form euchromatin, depending on the requirements of the cell. See [Box 2.3](#).

FISH

See *fluorescence in situ hybridization*.

fitness (*f*)

In population genetics, a measure of the success in transmitting genotypes to the next generation, relative to the most successful genotype. Also called biological or reproductive fitness. *f* always lies between 0 and 1.

fluorescence *in situ* hybridization (FISH)

Hybridization of a fluorescently labeled probe to the denatured DNA of chromosome preparations that have been immobilized on a solid surface ([Figures 11.4](#)), or to the RNA of similarly immobilized cells.

fluorophore (or fluorochrome)

A fluorescent chemical group, used for labeling nucleic acids or proteins ([Box 3.2](#)).

founder effect

High frequency of a particular allele in a population because the population is derived from a small number of founders, one or more of whom carried that allele.

fragile site

Location on a chromosome where the chromatin of metaphase chromosomes appears condensed under certain culture conditions. Most examples do not cause disease.

frameshift

A change in the base sequence of coding DNA that removes or adds nucleotides so as to change the translational reading frame ([Box 2.1](#)).

G-banding

The standard method of identifying chromosomes under the microscope. See [Figure 2.8](#) for an example.

gain-of-function mutations

Mutations that cause the gene product to do something abnormal, rather than simply to lose function. Usually the gain is a change in the timing or level of expression ([Sections 7.7](#) and [10.2](#)).

gamete

Sperm or egg; a haploid cell formed when a germ cell precursor undergoes meiosis.

gene

1. A functional DNA that is used to make a valuable RNA or protein end-product. 2. A factor that controls a phenotype and segregates in pedigrees according to Mendel's laws.

gene conversion

A naturally occurring nonreciprocal genetic exchange in which a short sequence of one DNA strand is altered so as to become identical to the sequence of another DNA strand ([Figure 7.9](#)).

gene dosage

The copy number of a gene. Alteration of the normal gene copy number causes reduced expression (too little gene product) or overexpression (too much product). For *dosage-sensitive genes* ([Box 7.3](#)), the amount of gene product made is critically important.

gene editing/genome editing

Making desired changes to a *specific* gene (or other target) sequence in the genome of *intact* cultured cells. See also *gene targeting* and *CRISPR-Cas*.

gene family

A set of related genes that arose by some type of gene duplication ([Section 2.5](#)).

gene frequency

see *allele frequency*.

gene knockout

The targeted inactivation of a predetermined gene within intact cells so as to artificially create a *null allele*.

gene pool

All the genes (in the whole genome or at a specified locus) in a particular population.

gene silencing/gene suppression

Gross or significant reduction in gene expression occurring naturally by altered epigenetic settings, and that can occur both naturally and artificially through *RNA interference*.

gene targeting

A type of gene editing using homologous recombination to specifically alter a pre-determined gene of interest within intact cells ([Box 9.2](#), [Figure 2](#)).

gene therapy

Treating disease by genetically modifying the cells of a patient. May involve adding a functional copy of a gene that has lost its function, inhibiting a gene showing a pathological gain of function, or, more generally, replacing a defective gene.

genetic background

The genotypes at all loci other than one locus under active investigation. Variations in genetic background (*modifier genes*) are a major reason for imperfect genotype–phenotype correlations ([Section 7.9](#)).

genetic code

The relationship between a codon and the amino acid it specifies ([Figure 7.2](#)).

genetic counseling

The process in which one or more members of a family that have, or are at risk of developing or transmitting, an inherited disease are informed by health professionals of the consequences and nature of the disorder, the probability of developing or transmitting it, and the options open to them.

genetic drift

Random changes in allele frequencies over generations because of random fluctuations in the proportions of the alleles in the parental population that are transmitted to offspring. Only significant in small populations.

genetic redundancy

Partly or completely overlapping function of genes at more than one locus, so that *loss-of-function mutations* at one locus do not cause overall loss of function.

genome/genomics

The total set of different DNA molecules of a cell, organelle, or organism/study of the same. The human genome consists of 24 different chromosomal DNA molecules and one mitochondrial DNA molecule.

genomic constraint

The constraint that *natural selection* imposes on variation at functionally important DNA sequences

genome browser

A computer program that provides a graphical interface for interrogating genome databases.

genome editing

Artificial manipulation of an intact cell that is designed to make a double-strand break at just one locus and subsequently to make a desired change to the base sequence at that locus. See [Figure 9.24](#) for an example.

genome (or gene) imprinting

See *imprinting*.

genomewide association (study) or GWA(S)

The standard approach to identifying factors governing susceptibility to complex disease ([Figure 8.15](#)).

genotype

The genetic constitution of an individual, either overall or at a specific locus.

germ line

The germ cells (gametes) and those cells that give rise to them; other cells of the body constitute the soma.

germ-line (or gonadal) mosaic

An individual who has a subset of germ-line cells carrying a mutation that is not found in other germ-line cells.

guide RNA

A short RNA that can base pair with a specific target sequence in order to guide some DNA-targeting (or RNA-targeting) enzyme to recognize the target sequence.

haploid

Term used to describe a cell (typically a gamete) that has only a single copy of each chromosome (for example the 23 chromosomes in a human sperm or egg).

haploinsufficiency

A locus shows haploinsufficiency if producing a normal phenotype requires more gene product than the amount produced by a single functional allele ([Box 7.3](#)).

haplotype

A series of alleles found at linked loci on a single chromosome ([Box 4.3](#) and [Figure 8.2](#)).

haplotype block

A region of DNA showing limited haplotype diversity ([Box 8.4](#)).

Hardy–Weinberg law (or equilibrium)

The simple relationship between allele frequencies and genotype frequencies that is found in a population under ideal conditions ([Section 5.4](#)).

hemizygous

Having only one copy of a gene or DNA sequence in diploid cells. Males are hemizygous for most genes on the sex chromosomes. Deletions occurring on one autosome produce hemizyosity in males and in females.

heritability

The proportion of the causation of a character that is due to genetic causes ([Section 8.2](#)).

heterochromatin

Highly condensed chromatin showing little or no evidence of active gene expression. Facultative heterochromatin may reversibly decondense to form *euchromatin*, depending on the requirements of the cell, but *constitutive heterochromatin* remains condensed throughout the cell cycle.

heteroduplex

Double-stranded DNA in which there is some mismatch between the two strands.

heteroplasmy

Mosaicism, usually within a single cell, for mitochondrial DNA variants ([Section 7.6](#)).

heterozygous/heterozygote

Having two different alleles at a particular locus/an individual with this property.

heterozygote advantage

The situation when a person heterozygous for a mutation has a reproductive advantage over both homozygotes for this mutation and also normal homozygotes. Sometimes called overdominance. Heterozygote advantage is one reason why severe recessive diseases may remain common ([Section 5.4](#)).

homologs (homologous chromosomes)

The two copies of a chromosome in a diploid cell. Unlike sister chromatids, homologous chromosomes are not copies of each other: one was inherited from the father and the other from the mother.

homologs (genes)

Two or more genes whose sequences are significantly related because of a close evolutionary relationship. They include *orthologs*, equivalent genes in two or more species that evolved from a single gene present in a common evolutionary ancestor, and *paralogs* that evolved by gene duplication such as the two α -globin genes present in humans.

homoplasmy

Of a cell or organism, having all copies of the mitochondrial DNA identical, as opposed to *heteroplasmy*.

homozygous/homozygote

Having two identical alleles at a particular locus/a person with this property. For clinical purposes a person is often described as homozygous *AA* if they have two normally functioning alleles, or homozygous *aa* if they have two pathogenic alleles at a locus, regardless of whether the alleles are in fact completely identical at the DNA sequence level. See also *autozygosity*.

hybridization

(of nucleic acids and oligonucleotides) Process in which complementary single strands are allowed to base pair (*anneal*) to form duplexes.

hybridization stringency

The degree to which the conditions (temperature, salt concentration, and so on) during a hybridization assay permit sequences with some mismatches to hybridize. High stringency conditions allow perfect matches only ([Figure 3.7](#)).

immunotherapy

Traditionally, type of therapy that uses substances to stimulate or suppress the immune system so as to help the body to fight cancer or other diseases, but now including the use of genetically engineered antibodies and T cells.

imprinting

(of certain mammalian genes) An epigenetic phenomenon in which the expression of the gene is determined by its parental origin (pp.160–3).

indels

Insertion/deletion variants, often involving a single nucleotide, but sometimes involving more nucleotides. (The definition is a little imprecise; in practice it usually includes variants that differ by possessing/lacking a sequence of up to 50 nucleotides.)

induced pluripotent stem (iPS) cells

Somatic cells that have been treated with specific genes, gene products, or other agents to reprogram them to resemble pluripotent stem cells. They can then be induced to differentiate into desired cell types ([Box 9.1](#)).

innate immunity/innate immune system

System of nonspecific response to a pathogen using the natural defenses of the body, as opposed to the *adaptive immunity/adaptive immune system*.

inner cell mass (ICM)

A group of cells located internally within the blastocyst which will give rise to the embryo proper ([Box 9.1](#), [Figure 2](#)).

insulator

DNA element that act as a barrier to the spread of chromatin changes or the influence of *cis*-acting elements.

interphase

All the time in the cell cycle when a cell is not dividing.

intron

Originally any segment of a transcript that is cut out and discarded during RNA *splicing*, but now widely used to mean the corresponding sequence in genomic DNA ([Figure 2.1](#)).

isochromosome

An abnormal symmetrical chromosome consisting of two identical arms, either the short arm or the long arm of a normal chromosome.

isoform

Alternative form of a protein as a result of differential expression of the same gene or through the production of different but highly related proteins from two or more loci.

karyotype

A summary of the chromosome constitution of a cell or person, such as 46,XY, but widely used loosely to mean an image showing the chromosomes of a cell sorted in order and arranged in pairs.

ligand

Any molecule that binds specifically to a receptor or other molecule, such as the trimeric FASLG ligand that binds to the FAS receptor in [Figure 10.13](#)..

ligase

DNA ligase is an enzyme that can seal single-strand *nicks* in double-stranded DNA or covalently join two oligonucleotides that are hybridized at adjacent positions on a DNA strand.

lineage

(of cells) In development, the ancestry and descendants of a cell, as traced backward or forward through successive cell divisions.

linkage analysis

Any statistical method that aims to identify chromosomal regions that co-segregate with a disease gene, or other gene of interest.

linkage disequilibrium

A statistical association between particular alleles at separate but linked loci, normally the result of a particular ancestral haplotype being common in the population studied. An important tool for high-resolution mapping ([Section 8.2](#)).

liposome

A synthetic lipid vesicle that can be used to introduce DNA into cells.

liquid biopsy

A test done on a blood sample to look for cancer cells from a tumor that are circulating in the blood, or for pieces of DNA from tumor cells in the blood.

locus (plural: loci)

A unique chromosomal location defining the position of an individual gene or DNA sequence.

lod score (Z)

A measure of the likelihood of genetic linkage between loci. The log (base 10) of the odds that the loci are linked (with recombination fraction q) rather than unlinked. For Mendelian characters a lod score greater than +3 provides minimal evidence of linkage; one that is less than -2 is evidence against linkage ([Box 8.1](#)).

loss-of-function mutations

Mutations that cause a gene product to lose its function, partly or totally ([Section 7.7](#)).

loss of heterozygosity (LOH)

Homozygosity or hemizygosity in a tumor or other somatic cell when the constitutional genotype is heterozygous. Evidence of a somatic genetic change ([Section 10.2](#) and [Figure 10.11](#)).

major histocompatibility complex (MHC)

A large gene cluster containing multiple genes including, notably, genes that function in antigen recognition by binding fragments of antigens and presenting them on the surface of T cells. The human MHC is known as the HLA complex (see Boxes 4.4 and 8.3). *See also* MHC restriction.

malignant tumor

A tumor whose cells show evidence of spreading (invading adjacent tissue, disseminating through the bloodstream and/or lymphatic system).

marker

(molecular) A chemical group or molecule that can be assayed in some way.

meiosis

The specialized reductive form of cell division used exclusively to produce gametes ([Figures 1.13](#) and [1.14](#)).

Mendelian

Description for a character whose pattern of inheritance suggests it is caused by variation at a single chromosomal locus.

mesenchyme

Connective tissues.

messenger RNA (mRNA)

A processed gene transcript that carries protein-coding information to cytoplasmic ribosomes.

meta-analysis

A statistical analysis of combined data from a number of independent studies of the same topic.

metastasis

The process whereby cells from a primary malignant tumor are disseminated via the blood stream or lymphatic system to establish secondary tumors at distant sites in the body.

MHC restriction

The requirement that when a T cell is confronted with a complex of a self-MHC molecule and a foreign peptide antigen bound to it, it will only respond to the antigen when it is bound to a *particular* MHC molecule ([Box 8.3](#), [Figure 1](#)).

microarray hybridization

A nucleic acid hybridization assay in which thousands to millions of different oligonucleotide (or DNA) probes are fixed at specific grid coordinates on a miniature solid surface and allowed to hybridize to complementary sequences within a solution containing a heterogeneous test sample population of labeled DNA or RNA molecules ([Figure 3.9](#)).

microbiome (or microbiota)

The aggregate of microorganisms that share our body space; most of them are found in the gastrointestinal tract.

microRNAs (miRNAs)

Short (21–22-nucleotide) RNA molecules encoded within normal genomes that have a major role in the regulation of gene expression ([Figure 6.10](#)).

microsatellite

Small array of *tandemrepeats* of a very simple DNA sequence, usually 1–4 bp, for example (CA)_n. The total length of the array is usually less than 0.1 kb. A polymorphic microsatellite is alternatively known as a short tandem repeat polymorphism ([Figure 4.6](#)).

mismatch repair

A form of DNA repair in which very simple DNA replication errors (nucleotide substitutions and deletions/insertions of one or two nucleotides) are repaired ([Figure 10.16](#)).

missense mutations

Changes in a coding sequence that cause one amino acid in the gene product to be replaced by a different one ([Section 7.2](#)).

mitosis

The normal process of cell division, which usually produces daughter cells genetically identical to the parent cell ([Figure 1.12](#)).

modifier (gene)

A gene whose expression can influence a phenotype resulting from a mutation at another locus ([Section 7.9](#)).

monozygotic

Originating from a single zygote, as in identical twins (other twins are dizygotic, having originated from different zygotes).

mosaic

An individual who has two or more genetically different cell lines derived from a single zygote. The difference may be point mutations, large-scale mutations or chromosomal abnormalities ([Box 5.3](#)).

mRNA

See *messenger RNA*.

mtDNA

Mitochondrial DNA ([Figure 2.12](#)).

multifactorial

A character that is determined by some unspecified combination of genetic and environmental factors.

mutagen

An agent that results in an increased mutation frequency.

mutation

1. A localized change in the base sequence of a DNA molecule. 2. The process that creates it.

mutation scanning/screening

Testing for any *undefined* change in the base sequence of a genome or genome component (notably exon, gene, or exome) in the hope of identifying abnormal variants correlating with disease. As opposed to testing for a *specific* DNA variant.

natural selection

Process whereby the population frequencies of alleles change by causing a change in the biological *fitness* of the individuals who carry them. Many alleles cause reduced biological fitness (*purifying* or *negative selection*); a few alleles cause increased biological fitness of the

individuals who carry them (*positive selection*). See also *balancing selection*.

ncRNA

See *noncoding RNA*.

next generation sequencing

(also called *massively parallel sequencing*) Any method that permits very high-throughput DNA sequencing by sequencing many molecules in parallel. See [Box 11.1](#).

nick (in DNA)

Cleavage of a single phosphodiester bond on one DNA strand only.

non-allelic homologous recombination (NAHR)

Recombination between misaligned DNA repeats, either on the same chromosome, on sister chromatids or on homologous chromosomes. NAHR generates recurrent deletions, duplications, or inversions ([Section 7.4](#)).

noncoding RNA (ncRNA)

mature RNA transcript that is not translated to make a polypeptide ([Figure 2.7](#)).

nondisjunction

Failure of chromosomes (sister chromatids in mitosis or meiosis II; paired homologs in meiosis I) to separate (disjoin) at anaphase ([Figure 7.16](#)). The major cause of numerical chromosome abnormalities.

nonhomologous end joining

Form of repair of double-strand breaks in DNA that involves the fusion of broken ends without copying from a DNA template.

non-penetrance

The situation when somebody carrying an allele that normally causes a phenotype to be expressed does not show that phenotype, as a result of interaction with alleles of other genes (*modifier genes*) or with non-genetic factors ([Figure 5.12](#)).

nonsense mutation

A nucleotide substitution that changes a codon specifying an amino acid so that it becomes a premature termination codon ([Section 7.2](#)).

nonsense-mediated mRNA decay

A cellular mechanism that degrades mRNA molecules containing a premature termination codon more than 50 nucleotides upstream of the last splice junction ([Box 7.1](#)). A stop codon less than 50 nucleotides from the last splice junction may often be harmless, but sometimes a short toxic polypeptide may be produced.

nonsynonymous substitution (or mutation)

A change in the sequence of a codon that results in a different codon interpretation. [Table 7.2](#) gives the different classes.

nucleosome

The basic structural unit of chromatin, comprising 146 bp of DNA wound around an octamer of histone molecules ([Figures 1.7](#) and [6.13](#)).

nucleotide

The fundamental repeating unit of a nucleic acid, consisting of a sugar to which is covalently attached a base and a phosphate group ([Figure 1.2](#)).

null allele

Any mutant allele where the normal gene product is not made or is completely non-functional.

odds ratio

In case-control studies, the relative odds of a person with or without a factor under study being a case ([Table 8.6](#)).

OMIM

Online Mendelian Inheritance in Man database at <https://www.ncbi.nlm.nih.gov/omim>

oncogene

A gene that when activated in some way (often by a change that stimulates its expression) can help to transform a normal cell into a tumor cell. Originally the word was reserved for activated forms of the gene (while the normal unactivated cellular gene was called a proto-oncogene), but this distinction is now widely ignored.

open reading frame

A continuous sequence of *coding DNA*.

origin of replication

A site on a DNA molecule where replication can be initiated.

orthologs

Homologous genes present in different organisms having descended from a common ancestral gene.

PCR (polymerase chain reaction)

The standard technique used to amplify short DNA sequences ([Figure 3.3](#)).

penetrance

The frequency with which a genotype manifests itself in a given phenotype.

pedigree

A limited family tree; a more extensive family tree is a kindred.

personalized medicine

A model of health care in which medical decisions and practice are tailored to the individual patient. Knowledge of a person's genome, for example, can allow more informed decisions about the suitability of prescribing certain drugs, and knowledge of cancer mutations may allow suitably targeted therapies.

pharmacodynamics

The study of the response of a target organ or cell to a drug.

pharmacogenetics

The study of the influence of individual genes or alleles on the metabolism or function of drugs.

pharmacokinetics

The study of the absorption, activation, catabolism, and elimination of a drug.

phasing

Converting genotypes into haplotypes in genome wide association studies..

phenocopy

A person or organism that has a phenotype normally caused by a certain genotype but does not have that genotype. Phenocopies may be the result of a different genetic variant, or of an environmental factor.

phenome

The totality of phenotypes of an individual organism.

phenotype

The observable characteristics of a cell or organism, including the result of any test that is not a direct test of the genotype.

phosphodiester bond

The link between adjacent nucleotides in DNA or RNA.

plasmid

A small circular DNA molecule that can replicate independently in a cell. Modified plasmids are widely used as cloning vectors ([Section 3.1](#)).

pleiotropy

The common situation in which variation in one gene affects several different aspects of the phenotype.

ploidy

The number of complete sets of chromosomes in a cell. Gametes are *haploid* and most normal cells are *diploid*, but some of our cells naturally have multiple chromosome sets (polyploidy) or none at all (nulliploidy).

pluripotent (of a mammalian stem cell)

Capable of giving rise to descendant cells that participate in the formation of all of the tissues of an embryo except the extraembryonic membranes.

PMID

PubMed identifier, a seven-digit or eight-digit number that, when typed into the query box at the NCBI PubMed database (<http://www.ncbi.nlm.nih.gov/pubmed/>), allows electronic access to a specific article in a biomedical journal.

point mutation

A mutation causing a small alteration in the DNA sequence at a locus, often changing just a single nucleotide.

polyadenylation/poly(A) tail

Addition of 200 or so adenosines to the 3' end of a mRNA. The resulting poly(A) tail is important for stabilizing mRNA ([Section 2.1](#)).

polygenic

Description of a character determined by the combined action of a number of genetic loci. Polygenic theory ([Box 8.2](#)) assumes that there are very many loci, each with a small effect.

polygenic risk score

Assessment of the risk of a specific condition based on the collective influence of very many genetic variants including variants not known to be associated with genes relevant to the condition.

polymorphism

The existence of two or more variants (alleles, phenotypes, sequence variants, chromosome structure variants) at significant frequencies in the population. Often also used more loosely to mean any sequence variant present at a frequency of more than 1% in a population.

polypeptide

A string of amino acids linked by peptide bonds. Proteins may contain one or more polypeptide chains.

positional cloning

Identifying a disease gene using knowledge of its chromosomal location.

position effect

Complete or partial silencing of a gene after some chromosome rearrangement that results in the gene becoming heterochromatinized – see [Figure 6.22](#).

positive selection

Selection in favor of a particular genotype that confers increased biological *fitness* ([Section 4.3](#)).

potency

Of a cell, its potential for dividing into different cell types. Cells can be totipotent, *pluripotent*, multipotent, or committed to one fate.

premutation allele

Among diseases caused by *dynamic mutations*, a repeat expansion that is large enough to be unstable on transmission but not large enough to cause disease.

primary structure

Of a polypeptide or nucleic acid, the linear sequence of amino acids or nucleotides in the molecule.

primary transcript

The RNA product of transcription of a gene by RNA polymerase, before splicing. The primary transcript of a gene contains all the exons and introns.

primer

A short oligonucleotide, often 16–25 bases long, which base pairs specifically to a target sequence to allow a polymerase to initiate the synthesis of a complementary strand.

primordial germ cells

Cells in the embryo and fetus that will ultimately give rise to germ-line cells.

probe

Known DNA or RNA fragments used in a hybridization assay to identify closely related target sequences within a complex, poorly understood population of nucleic acid molecules (the test sample)—see [Section 3.3](#).

prodrug

An inactive precursor to a therapeutic drug that is administered to a patient and activated within the body after natural conversion by a drug-metabolizing enzyme or other component ([Section 9.2](#)).

promoter

A combination of short sequence elements, usually just upstream of a gene, to which RNA polymerase binds so as to initiate transcription of the gene ([Figure 6.1](#)).

protective factor

A variant that reduces susceptibility to disease ([Table 8.11](#)).

proofreading

An enzymatic mechanism by which DNA replication errors are identified and corrected.

proteome/proteomics

All the different proteins in a cell or organism/study of the same.

proximal (of a chromosomal location)

Comparatively close to the centromere.

pseudoautosomal regions (or sequences) (PAR)

Regions with identical genes at the tip of the short arms and, separately, at the tips of the long arms of the X and Y chromosomes ([Figure 5.7](#)). Because of X–Y recombination, these genes move between the X and the Y ([Figure 5.8](#)), behaving as alleles that show an apparently autosomal mode of inheritance.

pseudogene

A DNA sequence that shows a high degree of sequence homology to a non-allelic functional gene but is itself nonfunctional or does not make a protein like its closely related homolog (but it may, however, make a functional non-coding RNA) ([Box 2.4](#)).

purifying (negative) selection

A form of natural selection in which harmful mutations that wreck or disturb the function of an important DNA sequence tend to be removed from the population.

purine

A double-ringed organic nitrogenous base that is a constituent of a nucleic acid, notably adenine (A) and guanine (G)—see [Figure 1.3](#).

pyrimidine

A single-ringed organic nitrogenous base that is a constituent of a nucleic acid, notably cytosine (C), thymine (T), and uracil (U)—see [Figure 1.3](#).

quantitative character

A character such as height, which everybody has but to differing degrees (in contrast with a dichotomous character such as polydactyly, which some people have and others do not).

quantitative PCR (qPCR)

PCR methods that allow accurate estimation of the amount of template present ([Section 3.2](#)). See also *real-time PCR*.

quantitative trait locus (QTL)

A locus that contributes to determining the phenotype of a continuous character.

reactive oxygen species (ROS)

Chemically reactive molecules or atom containing oxygen, such as oxygen ions, oxygen radicals, and peroxides. Formed within cells as a natural by-product of normal oxygen metabolism, they have important roles in cell signaling and homeostasis but cause DNA damage (see [Section 4.1](#)).

reading frame

During translation, the way in which the continuous sequence of the mRNA is read as a series of triplet codons. There are three possible forward reading frames for any mRNA, and the correct reading frame is set by correct recognition of the AUG initiation codon (see [Box 2.1](#)).

real-time PCR

A form of quantitative PCR in which the accumulation of product is followed in real time, allowing accurate quantitation of the amount of template present ([Section 3.2](#)).

recessive

Referring to a character that is manifested in the homozygote but not in the heterozygous state.

recombinant

In linkage analysis, a gamete that contains a haplotype with a combination of alleles that is different from the combination that the parent had inherited ([Figure 8.5](#)).

recombinant DNA

An artificially constructed hybrid DNA containing covalently linked sequences ([Figure 2](#) in [Box 3.1](#)).

recombination (or crossover)

Exchange of DNA sequences between paired homologous chromosomes at meiosis ([Figures 1.13](#) and [1.14](#)).

regenerative medicine

Using stem cell cultures to provide replacement cells for cells lost through disease (or injury).

relative risk

In epidemiology, the relative risks of developing a condition in people with and without a susceptibility factor ([Table 8.3](#)).

replication fork

In DNA replication, the point along a DNA strand where the replication machinery is currently at work ([Figure 1.5](#)).

replication origin

See *origin of replication*.

replication slippage

A mistake in replication of a short tandemly repeated DNA sequence that results in newly synthesized DNA strands with more or fewer copies of the tandem repeats than in the template DNA ([Figure 4.6](#)).

reprogramming (cellular, nuclear or epigenetic)

Large-scale epigenetic changes to convert the pattern of gene expression in a cell to that typical of another cell type or cell state. Often occurs in cancers ([Section 10.3](#)) and can be artificially induced in cells ([Box 9.1](#)).

restriction endonuclease

A bacterial enzyme that cuts double-stranded DNA at a short (normally 4, 6, or 8 bp long) recognition sequence ([Box 3.1](#)).

restriction fragment length polymorphism (RFLP)

A DNA polymorphism that creates or abolishes a recognition sequence for a restriction endonuclease. When DNA is digested with the relevant enzyme, the sizes of the fragments will differ, depending on the presence or absence of the restriction site ([Figure 4.4](#)).

restriction site

A site on a DNA molecule that is cleaved by a restriction endonuclease.

retrogene

A functional gene that appears to be derived from a reverse-transcribed RNA ([Box 2.4](#)).

retroposon (or retrotransposon)

A member of a family of mobile DNA elements that transpose by making an RNA that is copied into a cDNA which integrates elsewhere in the genome ([Section 2.5](#)).

retrovirus

An RNA virus with a reverse transcriptase function, enabling the RNA genome to be copied into cDNA before integration into the chromosomes of a host cell.

reverse transcriptase

An enzyme that makes a DNA copy of an RNA template; an RNA-dependent DNA polymerase ([Table 1.1](#)).

ribozyme

A natural or synthetic catalytic RNA molecule.

risk ratio

In family studies, the relative risk of disease in a relative of an affected person compared with that of a member of the general population ([Section 8.2](#)).

RNA gene

A gene that makes a functional noncoding RNA ([Figure 2.7](#)).

RNA interference (RNAi)

A cellular defense system activated by the presence of long double-stranded RNA sequences and designed to protect against viruses and excessive transposon activity within cells ([Figure 9.20](#)). Its discovery allowed specific *gene silencing/suppression* using *siRNAs*.

RNA polymerase

An enzyme that can add ribonucleotides to the 3' end of an RNA chain. Most RNA polymerases use a DNA template to make an RNA transcript.

RNA processing

The processes required to convert a primary transcript into a mature messenger RNA, notably capping, splicing, and polyadenylation.

RNA sequencing/RNA-Seq.

Sequencing cDNA as an indirect method of sequencing RNA. Some new technologies in principle allow direct sequencing of RNA.

RNA splicing

See *splicing*.

secondary structure

The path of the backbone of a folded polypeptide or single-stranded nucleic acid, determined by weak interactions between residues in different parts of the sequence ([Box 2.2](#)).

segmental duplication

The existence of very large, highly related DNA sequence blocks on different chromosomes, or at more than one location within a chromosome.

segregation

1. The distribution of allelic sequences between daughter cells at meiosis. Allelic sequences are said to segregate, non-allelic sequences to assort. 2. In pedigree analysis, the probability of a child's inheriting a phenotype from a parent.

selection

See *natural selection*.

selective sweep

Process whereby *positive selection* for a favorable DNA variant causes a reduction in variation in the population at the immediately neighboring nucleotide sequences ([Box 4.2](#)).

sense strand

The DNA strand of a gene that is complementary in sequence to the template (antisense) strand and identical to the transcribed RNA sequence

(except that DNA contains T where RNA has U). Quoted gene sequences always give the sense strand, in the 5' → 3' direction ([Figure 2.1](#)).

sensitivity (of a test)

The proportion of all true positives that the test is able to detect ([Table 11.3](#)).

sib

Brother or sister.

silencer

Combination of short DNA sequence elements that suppress the transcription of a gene.

silent mutation

Has the same meaning as *synonymous substitution*, which is the preferred term because sometimes this type of change can result in altered gene expression and disease ([Figure 7.4B](#)).

single nucleotide polymorphism/variant

See *SNP/SNV*.

siRNA (short interfering RNA)

Double-stranded RNA molecules 21–22 nucleotides long that can dramatically shut down the expression of genes through RNA interference ([Figure 9.21](#)).

sister chromatid

One of the two paired chromatids of a single chromosome that form after DNA replication and remain joined at the centromere until the anaphase stage of mitosis. Non-sister chromatids are present on different but homologous chromosomes ([Figure 1.10](#)).

SNP (single nucleotide polymorphism)

A nucleotide position in the genome where two or occasionally three alternative nucleotides are common in the population. The dbSNP database lists human SNPs but includes some rare pathogenic variants and some variants that involve two or more contiguous nucleotides.

SNV (single nucleotide variant)

A rare DNA variant (frequency less than 0.01) that can be seen to differ at a single nucleotide position from the consensus sequence in the population.

somatic cell

Any cell in the body that is not part of the germ line.

specificity (of a test for a condition)

A measure of the performance of a test that assesses the proportion of all people who do not have the condition who are correctly identified as such by the test assay. Specificity = (1 – false positive rate) ([Table 11.3](#)).

splice acceptor site

The site that defines the junction between the end of an intron in RNA and the start of the following exon. The junction sequence often conforms to the consensus sequence yyyyyyyyyynyagR, where y is a pyrimidine, n is any nucleotide, and R is a purine that is the first nucleotide of the exon.

splice donor site

The site that defines the junction between the end of an exon in RNA and the start of the following intron. The junction sequence often conforms to the consensus sequence (C/A)AGguragu, where r is a purine and capital letters denote the end nucleotides of the exon.

splicing

The process whereby some precursor RNA transcripts are cleaved into sequences, some of which (exons) are retained and fused (spliced) to give

the mature RNA whereas others are discarded (introns).

stem cell

A cell that can act as a precursor to differentiated cells but retains the capacity for self-renewal. Can be a tissue stem cell that gives rise to a limited number of cell types ([Figure 9.17](#) and Box and 10.2) or a *pluripotent* stem cell ([Box 9.1](#)).

stop (termination) codon

An in-frame codon that does not specify an amino acid but instead acts as a signal for the ribosome to dissociate from the mRNA and release the nascent polypeptide. See [Figure 2.3](#) for the principle and [Figure 7.2](#) for the different types of stop codon.

stratification

A population is stratified if it consists of several subpopulations that do not interbreed freely. Stratification is a source of error in association studies and risk estimation.

stratified medicine

A model of health care in which different medical treatments are targeted to subsets of the same disease according to which disease-associated genetic variants a person possesses.

stringency (of hybridization)

The choice of conditions that will allow either imperfectly matched sequences or only perfectly matched sequences to hybridize ([Figure 3.7](#)).

stroma

Supportive tissue of an epithelial organ, tumor and so on, consisting of connective tissues and blood vessels.

structural variation

Large-scale DNA variation that involves moving or changing the copy number of moderately long to very long DNA sequences, by one of various mechanisms: translocation, inversion, insertion, deletion, or duplication ([Section 4.2](#)).

supplementation therapy

(also called augmentation therapy) Therapy intended to supplement some deficiency, as opposed to the great majority of drug therapies that are designed to inhibit some disease process.

susceptibility factor

A variant that provides increased risk of developing a specific disease.

synonymous substitution (or silent mutation)

A nucleotide substitution that changes the sequence of a codon without any change in the amino acid that it specifies, but some may cause altered splicing and disease ([Figure 7.4B](#)).

tandem repeats

Any pattern in which a sequence of one or more nucleotides in DNA is repeated and the repetitions are directly adjacent to each other. See [Figure 2.12A](#) for an example.

targeted DNA sequencing

The process in which a defined subset of a genome (containing target sequences of interest) is captured then submitted for DNA sequencing ([Box 11.2](#)).

telomere

Specialized structure that stabilizes the ends of linear chromosomes. See [Figure 1.9](#) for telomeric DNA structure.

termination codon

See *stop codon*.

terminal differentiation

The state of a cell that has ceased dividing and has become irreversibly committed to some specialized function.

therapeutic window

The range of plasma drug concentrations that are of therapeutic benefit without causing extra safety risks due to drug toxicity.

tissue

A set of contiguous functionally related cells.

trait

See *character*.

trans-acting

The term used to describe any gene regulation in which the expression of some sequence on a DNA or RNA molecule is regulated by a different molecule or molecular assembly (in practice, a different RNA or a protein that is usually expressed from a remote gene and needs to diffuse to its site of action) ([Figures 6.1, 6.2](#)).

transcription factor

DNA-binding protein that promotes the transcription of genes. Some are ubiquitous, promoting transcription in all cells, but many are tissue-specific.

transcription unit

A segment of DNA that is used to make a primary RNA transcript (see [Figure 2.1](#)). May occasionally span multiple genes, as in the transcription of mtDNA ([Figure 2.12](#)) and in the transcription of adjacent 28S, 5.8S, and 18S rRNA genes.

transcriptome/transcriptomics

All the different RNA transcripts in a cell or tissue/the study of the same.

transdifferentiation

Epigenetic reprogramming of the nucleus of a cell, causing it to change from one cell type to another, such as from a skin cell to a neuron.

transduction

1. Relaying a signal from a cell surface receptor to a target within a cell.
2. Using recombinant viruses to introduce foreign DNA into a cell.

transfection

Direct introduction of an exogenous DNA molecule into a cell without using a vector.

transformation (of a cell)

1. Uptake by a competent microbial cell of naked high-molecular-weight DNA from the environment.
2. Alteration of the growth properties of a normal eukaryotic cell as a step toward evolving into a tumor cell.

transgene

An exogenous gene that has been transfected into cells of an animal or plant. It may be present in some tissues (as in human gene therapies) or in all tissues (as in germ-line engineering, for example in the mouse—see [Box 9.2](#)). Introduced transgenes may integrate into host cell chromosomes or replicate extrachromosomally and be transiently expressed.

transgenic animal

An animal in which artificially introduced foreign DNA (a transgene) becomes stably incorporated into the germ line ([Box 9.2](#)).

transit amplifying cells

The immediate progeny by which stem cells give rise to differentiated cells. Transit amplifying cells go through many cycles of division, but they eventually differentiate (Boxes 9.12 and 10.2).

translocation

Transfer of chromosomal regions between nonhomologous chromosomes ([Figure 7.13](#)).

transposon/transposon repeat

A mobile genetic element/a member of a repetitive DNA family containing some members that are able to transpose but also many inactivated copies of transposons ([Table 2.6](#), [Figure 2.15](#)).

trophoblast

Outer layer of polarized cells in the blastocyst that will go on to form the chorion, the embryonic component of the placenta ([Figure 2](#) of [Box 9.1](#), [Figure 2](#)).

tropism

The specificity of a virus for a particular cell type, determined in part by the interaction of viral surface structures with receptors present on the surface of the cell.

tumor suppressor gene

A gene that is commonly inactivated in tumors (by an inactivating mutation, by deletion as a result of abnormal chromosome segregation/recombination, or by epigenetic silencing). Classical tumor suppressor genes normally work to inhibit or control cell division.

unequal crossover

Recombination between chromatids that have paired up slightly out of alignment. See [Figure 7.8](#).

unequal sister chromatid exchange

The same process as *unequal crossover* but involves sister chromatids. See [Figure 7.8](#).

uniparental diploidy

A 46,XX diploid conceptus in which both genomes derive from the same parent. Such conceptuses never develop normally ([Figure 6.19](#)).

uniparental disomy

A cell or organism in which both copies of one particular chromosome pair are derived from one parent. Depending on the chromosome involved, this may or may not cause disease ([Figure 6.24](#)).

unrelated

Ultimately everybody is related; the word is used in this book to mean people who do not have an identified common ancestor in the last four or so generations.

untranslated region (5' UTR, 3' UTR)

Regions at the 5' end of mRNA before the AUG translation start codon, or at the 3' end after the stop codon ([Figures 2.1](#), and [2.3](#)).

variant (in relation to DNA)

A sequence that is different from the majority sequence but exists at a low frequency (<0.01 ; that is, less than 1%) in the population.

vector

A nucleic acid that is able to replicate and maintain itself within a host cell and that can be used to confer similar properties on any sequence covalently linked to it.

X-chromosome inactivation (or X-inactivation)

The *epigenetic* inactivation of all except one of the X chromosomes in the cells of humans and other mammals that have more than one X ([Figure 6.20](#)).

zygote

The fertilized egg cell.

Index

Note: The index covers the main text but not the summaries, questions or glossary.

Prefixes have been ignored for filing unless integral to a topic (so ‘ β -sheets’ at ‘beta’ but ‘ β -thalassemia’ at ‘thalassemia’). The same applies to numeric prefixes but unavoidable numerics have been sorted as though spelled out (so ‘5-methylcytosine’ will be at ‘methyl’ but ‘7SL RNA’ at ‘seven’).

Page numbers ending with ‘B’, ‘F’ or ‘T’ indicate that the listed topic is dealt with on that page *only* in a box, figure or table. ff. following a page number means that the item appears not only in the given page number but also in multiple consecutive pages.

1000 Genomes Project [44](#), [88B](#), [92](#), [272](#), [472](#), [481](#)
100 000 Genomes Project [88B](#), [451B](#), [482](#), [484](#)
11p15 imprinted gene cluster [34](#), [173F](#)
15q11 imprinted gene cluster [34](#), [174F](#)
18S rRNA [50](#), [142](#), [208](#), [472](#)
21-hydroxylase deficiency, see [steroid 21-hydroxylase deficiency](#)
28S rRNA [33](#), [50](#), [142](#), [208](#)
45,X see [Turner syndrome](#)
47,XXY and 47,XYY [225](#)
49, XXXXY [117F](#)
5' and 3' untranslated regions [28](#)

5' → 3' exonuclease [28](#)
5S rRNA [142](#)
5.8S rRNA [50](#), [142](#), [208](#)
5,6-dihydrouridine [29F](#)
5-methylcytosine [84](#), [85F](#), [150T](#), [156](#), [167](#), [295–6](#), [448](#)
6-mercaptopurine [318](#), [320T](#)
7-methylguanosine [28](#)
7S DNA [45F](#)
7SK RNA, [34](#)
7SL RNA [34](#), [34F](#), [49B](#), [52](#)

A

α -thalassemia X-linked mental retardation [167T](#)
AAVs (adeno-associated viruses) [336T](#), [337](#), [343–4](#), [352](#)
abasic site, in DNA [81–2](#)
ABL1 oncogene *see* [BCR-ABL1 fusion gene](#)
ABO antigens/blood group [93](#), [111](#)
ABO gene [111](#), [221B](#)
ACCE framework [422](#)
acetylation
 of histone tails [153F](#)
 of proteins [30T](#)
N-acetyltransferases (NAT1 and NAT2) [317](#)
achondroplasia [113](#), [130](#), [190](#), [190T](#), [219F](#), [462T](#)
ACMG *see* [American College of Medical Genetics and Genomics](#)
acrocentric chromosomes
 definition [205](#)
 human [36F](#), [51](#), [208](#), [209F](#)
acute myeloid leukemia (AML) [378T](#), [398](#), [403](#)
acute promyelocytic leukemia [378T](#), [429](#), [468T](#)

ADA *see* [adenosine deaminase](#)

adaptations/adaptive evolution [95](#), [95T](#), [96](#)

thrifty phenotype (thrifty gene) hypothesis [296](#)

adaptive immune system [229CB](#), [283B](#)

adenine

structure of [4F](#)

base pairing with thymine [4F](#)

deamination by hydrolytic attack to give hypoxanthine [231](#),
[232](#)

deamination in RNA editing [84–85](#)

mispairing [392](#)

structure [3F](#)

adeno-associated viruses (AAVs) [336–337](#), [343–344](#), [352](#), [357](#)

adenomas [370F](#), [392](#), [396](#)

adenomatous polyposis coli *see* [APC](#)

gene [263](#), [265B](#), [370](#), [385T](#), [387B](#), [395–6](#), [401](#), [402F](#), [407F](#)

adenosine deaminase (ADA), deficiency [341](#)

adenoviruses, in gene therapy [336T](#)

adeno-associated viruses, in gene therapy [336T](#)

adenovirus vectors [342](#)

adoption studies

Danish schizophrenia study [258](#)

ADRB1 and ADRB2 receptors [318](#), [319T](#)

adverse drug reactions [311–2](#), [314](#), [319](#), [319–20B](#)

affected sib-pairs

use for linkage analysis [260–1](#), [260F](#), [261T](#)

age/aging

apparently accelerated in some disorders [85B](#)

cancer incidence and aging [366](#), [371](#)

cell senescence [367–8B](#), [388](#)

epigenetic changes during [295](#)

maternal age effects in Down syndrome [212](#)

paternal-age-effect disorders [190](#), [190T](#)

see also [progeria](#)

age-related macular degeneration [288](#)

and complement gene factors [288](#)

Alamut Visual Plus program [447T](#)

alanine

chemical class [185T](#)

pathogenic polyalanine expansion [193](#), [194T](#)

structure [25F](#)

ALDP, peroxisomal membrane protein [242](#)

ALFRED database [92T](#)

alkaptonuria [305](#)

allele frequency, definition [130](#)

allele frequencies

disease allele frequencies differ in populations [130](#)

factors affecting [132–3](#)

influence of purifying selection [132](#)

allele-specific oligonucleotides (ASO) [67](#)

for rapid genotyping of point mutations [436–7](#), [437F](#), [438](#),
[438F](#)

alleles [77](#), definition [110](#)

allelic associations

an explanation for [263–4](#)

compared to genetic linkage [263F](#)

nature of [263](#), [263F](#)

allelic exclusion [102](#)

allelic heterogeneity [125–6](#)

allogeneic bone marrow transplantation [340](#)

allogeneic stem cells [353](#)

α 1-antitrypsin deficiency [228](#), [308T](#)

inclusion bodies and protein aggregation, [228](#), [228F](#)

α 1-antitrypsin, Pittsburgh variant [219–20](#)

α -helices [31B](#), [144F](#)

All of Us project [88B](#)

allele frequencies, mutation vs. selection [135](#)

alphoid DNAs [51](#)

ALT (alternative lengthening of telomeres) pathway [368B](#)

alternative splicing [93T](#), [147](#)

causing altered tau mRNA location [147](#)

classes of [146F](#)

evolutionary conservation [147](#)

and two different reading frames in *CDKN2A* gene [146F](#)

Alu repeats, evolutionary origin from 7SL RNA [49B](#)

see also [repetitive sequences](#), [human](#)

Alzheimer disease [470](#)

% concordance in MZ and DZ twins [257T](#)
amyloid- β , central role [286](#), [286F](#)
APOE* ϵ 4 risk factor [284](#), [285F](#)
biological pathways in pathogenesis [287F](#)
common susceptibility factors [286–7](#)
dominantly inherited pedigree [259](#), [259F](#)
genes involved in Mendelian subsets [284](#), [284T](#)
and genes in inflammation pathways [288](#)
presenilin genes and [260](#)
prionoid disease [230B](#)
protective variants for [289T](#)
rare associated variants [286](#)
shared pathways for susceptibility factors in common
Mendelian subsets [287F](#)

American College of Medical Genetics (ACMG) [465](#), [480–2](#)

American College of Medical Genetics and Genomics [445](#)

amino acids, classical and general

binding by specific tRNAs [26](#)
C- and N-terminal ends [25F](#)
chemical classes [185T](#)
covalently linked to tRNA [27](#), [29F](#)
N-terminal methionine [28F](#)
occasional cleavage of [28F](#)
repetitive -NH-CH-CO- motif [25F](#)
structures of the 20 common amino acids [24](#), [25F](#)

amino acids, rare

citrulline [309F](#)

selenocysteine, 21st amino acid [184F](#)

AML *see* [acute myeloid leukemia](#)

ammonia, in the urea cycle [309F](#)

amniocentesis [434F](#), [457](#), [461–2](#), [483](#)

amplification of DNA

by DNA cloning in cells, *see* [DNA cloning](#)

cell-free, *see* [PCR](#)

forming double minute chromosomes [377F](#)

see also [gene amplification](#)

amplification refractory mutation system (ARMS) [437](#), [438F](#), [439](#)

AMY1A gene, copy number changes [95T](#), [97](#), [98F](#), [221B](#)

α -amylase, salivary [97](#), [98F](#)

amyloid- β (A β), production of [287F](#)

amyloid family proteins

aggregation of [230](#)

diseases associated with [230B](#)

amyloid fibril [230B](#)

amyotrophic lateral sclerosis (motor neuron disease)

cytoplasmic aggregation of SOD1 protein in [230B](#)

anaerobic glycolysis in cancer [389](#)

anaphase [13](#), [14F](#), [15F](#), [211](#), [390](#)

anaphase lag [211](#)

ancestral chromosome segments, sharing of [267](#), [268F](#)

see also [human chromosomes](#)

ancestry testing [470](#)

androgen-insensitivity syndrome (testicular feminization syndrome)

[224](#)

androgen production, abnormal [201B](#), [224](#), [307F](#), [462T](#)

androgen receptor gene [224](#)

androgenetic embryo [163F](#), [171](#)

aneuploidy/ies

- in cancer cells [390](#)
- as chromosome abnormalities [211–2](#)
- fetal aneuploidy screening [424](#), [425F](#)
- noninvasive [425](#)
- gene dosage problems [163–4](#)
- prenatal genetic testing of [423T](#), [424–5](#), [426F](#), [463–4](#)
- quantitative fluorescence PCR [424](#), [425](#), [425F](#)
- and regulatory gene mutation [224](#)
- segmental [207T](#), [224–5](#)
- of sex chromosomes [424](#), [426F](#)
- whole chromosome [224–5](#)

Angelman syndrome (AS) [161](#), [168B](#), [173T](#), [174](#), [174B](#), [175B](#), [203T](#)

UBE3A mutation in [175B](#), [203T](#), [207T](#), [220](#)

angiogenesis [369](#), [369T](#)

and cancer cells [369](#), [369T](#)

angiomyolipomas [324](#)

angiotensin-converting enzyme (ACE) [319](#), [319t](#)

animal disease models *see* [disease models](#)

aniridia type 1 [143F](#)

ankylosing spondylitis [266B](#), [279B](#)

HLA-B27 association [263F](#), [266B](#)

annealing *see* [hybridization](#)

anonymity and confidentiality [473–4](#), [475B](#)

anophthalmia [479](#)

ANRIL (CDKN2B-AS1) antisense RNA [159T](#)

anti-cancer defense systems [366](#)

anti-proliferative agents [324](#)

antibodies

genetically engineered [326–7](#), [327T](#), [328](#), [328T](#)

intrabodies [327–8](#)

scFv antibodies [327F](#), [328](#)

see also [monoclonal antibodies](#); [therapeutic antibodies](#)

anticipation, genetic disorders [129](#), [129F](#), [194](#)

antigen-presenting cells, professional

main cell types [103](#)

producing co-stimulatory molecules [103](#)

antigen presentation [99](#), [103](#), [265B](#)

antisense oligonucleotides

use in gene silencing [345](#)

antisense (template) DNA strand [23F](#)

antisense RNA(s) [34F](#), [35](#), [150T](#), [159T](#), [172](#), [173F](#), [278T](#)

antisera, panels of [105B](#), [264](#), [266B](#)

antibody diversification [400](#)

APAF1 [385F](#)

APC gene (adenomatous polyposis coli) [395](#)

and cancer susceptibility [370](#), [385T](#), [401](#), [402F](#), [407F](#)

epithelial cancer evolution [370F](#)

in familial adenomatous polyposis [385T](#)

Wnt pathway and [370](#), [385T](#), [396](#)

APOBEC cytidine deaminases

in antibody diversification [102](#), [400](#)

C → U RNA editing [400](#)

in hypermutation (kataegis) in cancers [400](#)
APOE (apolipoprotein E) gene [285](#), [470](#)
*APOE** ϵ 2, *APOE** ϵ 3, *APOE** ϵ 4 alleles [285–6](#), [285F](#)
apolipoprotein B mRNA, RNA editing [147](#)
apoptosis

after severe DNA damage [391](#)
avoidance of, by tumor cells [369T](#)
inhibited by proto-oncogenes [375–6](#)
mitochondrial pathway [385F](#)
promoted by tumor suppressor genes [384](#), [385F](#)
response to DNA damage [384](#)
triggered by double-strand DNA breaks [83](#)
role of p53 [384](#), [385F](#)

apoptosis pathways [384](#)

regulation by p53 [384](#), [385F](#)
regulation by some oncogenes [384](#)

APP (amyloid precursor protein) gene [260](#), [284T](#)

AR (androgen receptor) gene [224](#)

Arginine

chemical class [185T](#)
methylation of [153](#)
structure [25F](#)

ARMS (amplification refractory mutation system) [437](#), [438F](#)

ascertainment bias [122](#)

Ashkenazi Jews [472](#)

and study of founder effects [133](#)

asparagine

chemical class [185T](#)
structure [25F](#)
aspartate/aspartic acid

chemical class [185T](#)
structure [25F](#)

assisted reproduction *see* [in vitro fertilization](#)

association analyses [261](#)

basic principles [263–4](#)
confounding sample structure [272–3](#)
explanations for association of alleles in a population [263–4](#)
HLA and candidate gene studies [264](#), [265–5B](#)
see also [genomewide association studies \(GWA/GWAS\)](#)

Association for Molecular Pathology [445](#)

assortative mating [132](#)

asymmetric cell division [331B](#), [374B](#)

ataxia telangiectasia [85–6B](#)

atherosclerosis, as amyloid disease [230B](#)

ATM (ataxia-telangiectasia mutated) protein kinase [391](#), [391F](#)

ATRX gene [167T](#)

AUC (area under the curve) [278](#), [278–9B](#)

augmentation therapy *see* [supplementation therapy](#)

autism spectrum disorder

frequent immune pathway dysfunction [288](#)

high frequency of de novo CNVs [277](#), [278T](#)

autoantibodies [266B](#)

autoimmune diseases/autoimmunity

HLA variants strongest risk factors [265–6B](#)

importance of complement factors [287–9](#)
PTPN22 R620W variant as modifier of risk [289](#)
autoimmune responses [103](#)

co-stimulatory molecules in [103](#)

autologous cells

in cancer therapy [411](#)
genetically modified in ex vivo gene therapy

autologous cells [340–1](#), [354](#), [411](#)
autologous versus allogeneic transplantation [340](#)
autophagy [282B](#)
autosomal aneuploidies [424](#), [425F](#)
autosomal dominant disorders

parent-of-origin effects [129](#), [129F](#)
variable expressivity [128](#), [128F](#)

autosomal dominant inheritance

fitness of affected individuals [135](#), [135F](#)
mutant allele transmission [135](#), [135F](#)
patterns of [112–3](#), [113F](#)

autosomal recessive disorders

consanguinity [114](#), [114–5B](#)
disease-related phenotypes in carriers [115–6](#)
newborn screening for [464–5](#), [465T](#)

autosomal recessive inheritance patterns [113–4](#), [114F](#), [115](#)

fitness of affected individuals [135F](#), [135F](#)

mutant allele transmission [135F](#), [135F](#)
patterns of [113–4](#), [114F](#), [115](#)
autozygosity/autozygous [247](#)
autozygosity mapping, *see* [linkage mapping](#)
avastin [409](#), [409T](#)
azoospermia [203T](#), [225](#)

B

β 2-microglobulin and *B2M* gene [354](#)
B cells, activation-induced cytidine deaminase [102](#)

cell-specific Ig production [100–1](#), [101F](#), [102](#)

balanced chromosome translocations
balancing selection [97](#), [136](#)

see also [overdominant selection](#)

Bardet-Biedl syndrome [126](#), [126F](#)
Barr bodies [116](#), [117F](#), [164](#)
barrier elements [143](#), [168](#), [169F](#), [173F](#)
base cross-linking (between DNA bases) [82](#)

pyrimidine dimers induced by UV light [81F](#)
repair of interstrand crosslinks [85–6B](#)

base excision repair (BER) [82–3](#), [85B](#), [102](#), [392](#)
base pairing

A-T and G-C base pairs, structure [4](#), [4F](#)
A-T and G-C base pairs, relative strengths [4](#), [4F](#)
in double helix [31B](#)
prevalence of [5B](#)
in single-stranded RNAs [29F](#)

base wobble [184](#)

bases *see* [nucleic acid bases](#)

Bayesian analysis [458–9B](#), [468T](#)

BCL2 oncogene [378T](#), [409T](#)

inhibitor of mitochondrial apoptosis pathway [384](#)

BCR-ABL1 fusion gene [378](#), [378F](#), [408](#), [412](#)

amplification of [412](#)

see also [chronic myelogenous leukemia](#); [Philadelphia chromosome](#)

BCR-ABL1 fusion gene [378](#), [379F](#), [429](#), [468T](#)

BCR-ABL1 fusion protein [409T](#), [412](#)

Becker muscular dystrophy (BMD) [126](#), [346B](#)

Beckwith-Wiedemann syndrome [129](#), [129F](#) [172](#), [173T](#)

Benzo(*a*)pyrene [4F](#)

beta-adrenergic receptors [318](#), [319T](#)

β -globin genes/gene clusters [47F](#), [47T](#)

β -(pleated)-sheets

in protein aggregation [229–30B](#)

structure of, [31B](#)

β -turns [31B](#)

biologics [301](#), [310–1](#), [325](#)

biomarker(s) [281](#), [420–1](#)

in breath [419](#)

in cancer [389](#), [408](#), [413B](#), [419](#), [468](#)

in phenylketonuria [235F](#)

biopsies, invasive versus liquid (from tumors) [412–3](#)

Bionano Saphyr system [433](#), [434F](#)

biotin-streptavidin capture system [69T](#), [250F](#), [441B](#), [442](#)
biotinylated probes [430](#)
biotinidase deficiency [465](#)
bisulfite sequencing [448–9](#), [449F](#)
bivalents [15F](#), [16–17](#)
BLAST computer programs [38](#), [39T](#), [40–1](#)
blastocysts [332](#), [339B](#)
blastomeres [419](#), [460](#), [460F](#)
BLAT program [41](#)
blepharimosis
blood-brain barrier [286](#)
blood cells, origin of [341F](#)
Bloom syndrome [85–6B](#)
bone dysplasia [219](#)
bone marrow transplantation [340](#)

treating blood cell disorders [306](#)

see also [hematopoietic stem cells](#)

bottleneck *see* [mitochondrial genetic bottleneck hypothesis](#); [population bottleneck](#)

boundary elements [144](#)

separating euchromatin and heterochromatin [168](#), [169F](#)

see also [barriers](#); [insulators](#)

BRAF oncogene [468T](#)

brain,

immune privileged organ [288](#)

imprinting of *UBE3A* gene in neurons but not glial cells [161](#)

main target of prion toxicity [229B](#)

branch site, *see* [splice sites](#)

BRCA1 and *BRCA2* genes [262](#), [385T](#), [386](#), [391](#), [396](#), [401](#), [409T](#), [470–1](#)

BRCA1, *BRCA2* gene panels for mutation screening [442B](#)

driver mutations in primary breast cancer [402](#), [402F](#)

familial breast cancer association [452](#), [453B](#)

genome instability after mutation [386](#)

maintaining constitutive heterochromatin [169](#)

MLPA scanning for exon copy number changes in *BRCA1*
[432F](#)

roles in DNA repair [381](#), [391F](#), [452](#)

somatic nonsynonymous mutations per tumor

targeting PARP-1 inhibitors at [452](#)

see also [breast cancer](#)

BRCA1 and 2 proteins [391F](#)

breast cancer

CD 44+ CD24- cells in cancer initiation [375B](#)

driver mutations in primary cancers [402F](#)

familial [453](#), [454B](#)

gene panel [442B](#)

λ S relative and lifetime risks [255T](#)

mutation scanning [423F](#), [442B](#)

mutational processes of, dissected [400](#), [400F](#)

monoclonal antibody therapy [328T](#)

new molecular subtyping [406](#)

somatic nonsynonymous mutations per tumor [399F](#)

sporadic [386](#)

targeted therapy for [409T](#), [412T](#), [452](#)

breath, cancer biomarkers in [419](#)

brittle bone disease see [osteogenesis imperfecta](#)

C

C → T mutations, very frequent in vertebrates, [84](#), [85F](#)

C-terminal ends (polypeptides) [25F](#)

Caenorhabditis elegans [297](#)

CAG repeats, see [polyglutamine repeats, expansion of](#)
calico cat, X-inactivation in [164](#)

cancer(s)

anti-cancer defense systems, natural selection [366](#)

age of onset [371](#)

as diseases of stem cells [372](#), [374–5B](#)

cancer databases and browsers [398T](#)

cancer genomics [397ff.](#)

chromosomal instability in [389–90](#), [390F](#)

chromothripsis and chromoplexy [391](#)

definition and terminology [362](#)

different from other genetic diseases [364–5](#)

and disease gene identification [401](#)

driver and passenger mutations [369–70](#)

discriminating between [401](#)

epigenetic dysregulation in [389](#)

genome-epigenome interactions [396–7](#), [397T](#)

genomewide RNA sequencing and link with biology [405–6](#)

in childhood [371](#)

immunosurveillance to kill cancer cells [366](#)

intratumor heterogeneity, different levels of [372](#), [373F](#), [373T](#)

metabolism-epigenome link in cancer [403](#), [404F](#)

long noncoding RNA and miRNA involvement [388–9](#)

number of mutations in different cancers [398–9](#), [399F](#)

somatic mutations play a major role in [364](#)

tricarboxylic acid cycle genes involved in cancer [403](#), [404F](#)

viruses causing human cancers [367](#)

why not everybody succumbs [365](#), [366](#)
cancer cells

acquired biological capabilities [369T](#)
altered metabolism [366](#)
biological characteristics distinguishing [366](#), [367](#)
clonal expansion [370](#)
defense against cytotoxic T lymphocytes [369T](#)
DNA methylation profiles [396](#)
energy surprisingly from glycolysis [366–7](#)
epigenetic reprogramming of [403](#), [404F](#)
immortality, selection pressure to achieve

via telomerase activation [368B](#)
via ALT pathway activation [368B](#)

metabolic changes [367](#)
mutational processes and signatures [400](#)
unregulated proliferation [366](#)
telomerase expression [367–8B](#), [392](#), [403](#)
Warburg effect and [367](#)
see also [cancer stem cells](#)

cancer evolution

biological pathways in [406](#), [407F](#)
by accelerating mutation [371](#)
clonal evolution [370F](#), [372](#)
genome destabilization [371](#)
multi-stage nature [369–72](#)
tracing the mutational history of cancers [404](#), [405F](#)

cancer genes

Cancer Gene Consensus (Cosmic) [403](#)
classified by function [395](#)
driver genes vs. passenger genes [395](#)
epigenetic mediators [395–6](#)
epigenetic modifiers [395–6](#)
epigenetic modulators [396](#)
methods to identify [397](#), [401](#)
nonclassical [403](#)
two fundamental classes of [375](#)
see also [oncogenes](#); [tumor suppressor genes](#)

cancer genetic testing and detection [468–9](#)

different roles for DNA biomarkers [468](#), [468T](#)
different roles for gene expression biomarkers [468](#), [468T](#)
imaging via increased glucose uptake [366](#)
multiplex testing using panels of cancer susceptibility genes
[469](#)
via noninvasive liquid biopsies [412–3B](#), [469](#)

Cancer Genome Project [398](#)

cancer stem cells

cancers as diseases of stem cells [372](#), [374–5B](#)
explanation for intratumor heterogeneity [372](#), [373F](#)
and resistance to therapy [411](#)
single-cell analyses of [405F](#)
target cells in cancer development [374–5B](#)

cancer therapies

CAR-T cell therapy [410–11](#), [410F](#)

cytokine storms in [411](#)

combinatorial therapies, promise of [413](#)
drug resistance evolves [412](#)
immune checkpoint therapy [410](#)
immunotherapy with monoclonal antibodies [409](#), [409T](#)
targeted therapies, the need for [408](#)
targeted therapies using small molecule drugs [408–9](#), [409T](#)
imatinib, the first successful drug therapy [408](#), [409T](#), [411](#)
tumor recurrence [411–2](#)

capillary electrophoresis [73–4B](#)
cap, at 5' end of mRNAs [28](#)
capsid [336](#)
carcinogenesis [362](#)
CAR-T cell therapy *see* [cancer therapies](#)
carcinoma [370F](#)
cardiac QT interval [320B](#)
cardiomyopathy [126T](#), [346B](#), [348](#), [466](#), [481](#)
carriers

autosomal recessive disorders [113](#)
genetic screening [463](#)
preconception screening [467](#)
risk assessment [458–9B](#)
see also [heterozygosity/heterozygous](#)

cascade testing [455](#)
cassette exon [146F](#)
exon duplication [24](#)
case-control studies [264](#), [264T](#)
cas nuclease [345](#)
caspase [385F](#)
CDK (cyclin-dependent kinases) [383](#)
CDK2-cyclin E complex

regulation of [383](#)
CDKN1C gene [159T](#), [172](#), [173T](#), [174](#)
CDKN2A gene [396](#)

alternative splicing [146F](#), [147](#)
p14 and p16 isoforms via alternative splicing [146F](#), [147](#), [384](#),
[385T](#)
pivotal in cell cycle control [384](#)

CDKN2B gene, transcription repressed by an antisense RNA, ANRIL
[159T](#), [396](#)
Celiac disease

HLA association [266B](#)

cell cycle [11](#), [12F](#)

arrest of [391](#), [391F](#)
checkpoints [383](#), [391F](#)
G₀ phase [11](#), [13](#)
G₁ and G₂ phases [11](#)
M phase [11–13](#), [12F](#)
interphase [11](#)
S phase [11–13](#), [12F](#)
rapid cell division and [364](#)

cell cycle—apoptosis pathway [406](#)
cell death

balance with cell proliferation [365](#), [365F](#)
see also [apoptosis](#); [cell senescence](#)

cell differentiation

in cancer [372B](#), [374B](#)

dedifferentiation [333B](#), [369T](#), [397](#)
epigenetic mechanisms [151T](#)
regulation of [396](#)
reversibility of [404F](#)
and stem cells [331–2B](#), [353](#), [355](#)
terminally differentiated cells [11](#), [331](#), [332B](#)
transdifferentiation [333B](#), [353](#)

cell division(s)

asymmetric versus symmetric [331B](#)
number of mitotic in a human lifetime [124B](#)
total number required to form gametes [189](#), [190F](#)
see also [mitosis](#); [meiosis](#)

cell-free DNA [412–3B](#), [461](#), [464](#)

cell-mediated immunity [265B](#)

cell plasticity, single-cell analyses [405F](#)

cell proliferation/growth

balance with cell death [365](#), [365F](#)
contact inhibition [367B](#)
dysplastic and hyperplastic [362](#)
regulation of [365](#), [365F](#)

cell senescence [367–8B](#), [388](#)

cell signalling

12 key pathways in cancer [406](#), [407F](#)
RAS-PI(3)IK pathway in cancer [406](#), [407F](#)

cellular disease models *see* [disease models](#)

cellular memory [150](#)

CENP-A [151T](#)

histone H3 variant [154T](#)
centimorgan (cM) unit [244](#)
central dogma (of molecular biology) [7–8](#)
centric fusion, see [Robertsonian translocation](#)
centromeres [10](#), [12F](#), [13](#), [14F](#), [15F](#)

chromosome banding nomenclature and [204–5B](#)
establishment by epigenetic mechanism [151T](#)
function of [10](#)
heterochromatin at [10](#)
instability if DNA is poorly methylated [167T](#)
poor sequence conservation [10](#)
highly methylated satellite DNAs at [51](#), [90](#)
specific histone H3 (CENP-A) [154T](#)
structure of [10](#), [10F](#)

CFH (complement factor H) gene [261](#)
CFTR gene (cystic fibrosis transmembrane regulator) [41](#), [41F](#)
CG dinucleotide see [CpG dinucleotide](#)
chaperone molecules [30](#), [226](#)
characters

continuous vs discrete [252](#)

Charcot-Marie-Tooth disease [126T](#), [203T](#), [207](#), [221B](#)
chemical drugs see [small molecule drugs](#)
CHEK2 protein [391F](#)
chiasma(ta) [15F](#)
childhood
consent issues [482](#)
testing guidelines [455–6](#)
childhood cancers [318](#), [364](#), [371–2](#), [385T](#)
chimeras [212](#), [339B](#)
chimeric antibodies [327F](#)

chimeric antigen receptor (in CAR-T cell therapy) [410–11](#)
chimeric genes [220](#), [337–8](#), [379F](#), [408](#)
cholesterol [318](#), [322](#), [349F](#)

low-density lipoprotein cholesterol (LDL) [453](#), [482](#)
high-density lipoprotein cholesterol (HDL-C) [291F](#)
metabolism [286](#)
see also [familial hypercholesterolemia](#)

chorionic villus sampling [419T](#), [434F](#), [457F](#), [461](#)
chromatids

mispaired/misaligned, *see* [unequal crossover](#)
sister chromatids [12F](#), [13](#), [14F](#), [15F](#), [16–7](#)

chromatin

chromatin fiber [9](#), [9F](#)
DNA compaction, effects [155F](#)
general structure [9](#)
looped domains of chromatin fiber [9F](#)
modifications and gene expression [150](#), [150T](#), [151T](#), [152](#), [152F](#),
[154](#), [154F](#), [154T](#), [155F](#)
see [euchromatin](#); [heterochromatin](#); [histones](#)

chromatin diseases [167](#), [167T](#)
chromatin effector proteins [155F](#)
chromatin erasers [167](#), [167T](#)
chromatin modifier genes [167](#), [167T](#)
chromatin readers [167](#), [167T](#)
chromatin remodeling [152](#)
chromatin states/structure

changes of, affecting gene expression [151](#)

open vs condensed [151–2](#), [152F](#)
chromatin writers [167](#), [167T](#)
chromodomain [154](#)
chromosome analysis

 Giemsa (G-) banding [36F](#)
 spectral karyotyping [390](#), [390F](#)

chromosomal instability (CIN) [389–91](#)
chromosome analysis

 chromosome FISH, principle of [428](#), [429F](#)
 chromosome SNP microarray analysis
 Giemsa (G-) banding [36F](#)
 optical genome mapping [433](#), [434F](#)
 spectral karyotyping [390](#), [390F](#)

chromosomes [2](#)

 acrocentric, definition [205B](#)
 in the cell cycle [10–12](#), [12F](#), [13](#)
 chromatin structure [9](#)

function [9](#), [10](#)

 homologous (homologs) [16F](#), [16–7](#)
 17F
 metacentric, definition [225B](#)
 in mitosis and meiosis [13ff](#)
 ploidy [10](#)
 structure and function [8–9](#), [9F](#), [10](#), [10F](#)
 submetacentric, definition [225B](#)
 see also [chromosome abnormalities](#); [euchromatin](#);
 [heterochromatin](#); [human chromosomes](#);

chromosome abnormalities

- acentric chromosomes [207](#), [209F](#)
- chromosome instability in tumor cells [371](#), [390–2](#)
- chromosome microdeletions/duplications [202](#), [203F](#), [203T](#), [207](#), [207T](#)
- chromoplexy [391](#)
- chromothripsis [391](#)
- constitutional [205](#)
- derivative chromosomes [208](#)
- dicentric chromosomes [207](#), [209](#), [209F](#)
- disease gene identification via [248](#), [248T](#)
- double minute chromosomes [377F](#)
- genetic testing for [423ff.](#)
- interstitial deletions [121](#), [206T](#), [208F](#)
- inversion [207](#), [208F](#)
- isochromosomes [209](#), [209F](#)
- large-scale deletions and duplications [182](#), [199](#), [207](#), [208F](#), [215](#), [224–5](#), [248T](#)
- nomenclature [210T](#)
- numerical abnormalities [206T](#), [206–9](#)
- ring chromosomes [206T](#), [208F](#)
- structural abnormalities [206T](#), [206–9](#)
- see also* [aneuploidies](#); [translocations](#)

- chromosome-banding karyotyping [423T](#)
- chromosome break mapping [240](#), [248](#)
- chromosome engineering [339B](#)
- chromosome instability [167T](#), [371](#), [390](#), [392](#), [396–7](#), [405F](#)
- chromosome/chromatin remodelling [150T](#), [152](#), [154–5](#), [395](#)
- chromosome recombination, *see* [recombination](#)
- chromosome segregation errors [80](#)
- chromosome SNP microarray analysis [423T](#), [425–6](#), [427B](#)

chromosome translocations, see [translocations](#)

chromothripsis [361](#), [391](#)

chronic granulomatous disease [207T](#), [326T](#)

chronic lymphocytic leukemia (CLL) [388](#), [399F](#), [400](#), [409T](#), [486T](#)

inferring mutational history of [405F](#)

chronic myelogenous leukemia (CML) [374–5B](#), [378](#), [412](#), [429](#), [468T](#)

treatment with imatinib [408](#), [412](#)

Circos plots [473F](#)

circular RNAs [34T](#)

as miRNA sponges [149F](#)

cirrhosis

of the liver [288](#)

primary biliary cirrhosis [256](#)

cis-acting (elements)

boundary elements as *cis*-acting elements [43](#)

definition [140](#)

in gene regulation [140–1](#), [141F](#)

enhancers and silencers as *cis*-acting elements [143](#)

long noncoding RNAs as *cis*-acting elements [140](#), [159](#), [159T](#),
[160F](#)

promoters as *cis*-acting elements [140](#)

working at the DNA level [141F](#)

working at the RNA level [140](#), [141F](#)

cisplatin [81F](#)

citrulline [309F](#)

clade, of related mtDNAs [292B](#)
Claes-Jensen syndromic X-linked mental retardation [167T](#)
clinical exome [175B](#), [442](#), [442B](#)
ClinGen database [447T](#)
ClinVar database [42](#), [423T](#), [443F](#)
CLL, *see* [chronic lymphocytic leukemia](#)
clonal expansion/evolution in cancer, [405F](#)
clonal expansion, mtDNA [213](#), [370](#)
clones

- cell clones [58F](#), [59F](#)
- DNA clones *see* [DNA clones](#)

cloning vectors [58](#), [58F](#), [59–60](#)

- bacteriophage vectors [60](#)
- plasmid vectors [59–60](#), [61B](#)
- see also* [viral vectors](#)

cloud computing [398](#)
CLU (clusterin) gene [287F](#), [288](#)
CML, *see* [chronic myelogenous leukemia](#)
co-dominant phenotypes [111](#), [115](#)
co-stimulatory molecules [410F](#)

- made by professional antigen presenting cells [103](#)

codeine [313](#), [315](#), [322](#)
coding DNA

- principle of [22](#)
- proportion of human genome [43](#)
- translational reading frame [26–7B](#)

Coding Constrained Region data [447T](#)

codons

64 possible codons [26](#)
and anticodon, in tRNA [27](#), [29F](#)
function of [27](#)
genetic code [184F](#)
initiation codon [126](#)
stop codons [27](#)

coefficient of inbreeding [115B](#)
coefficient of relationship [114–5B](#)
cohesins [12F](#), [12–3](#), [14F](#)
COL1A1 and *COL1A2* genes [222](#), [223F](#)

dominant-negative effects in osteogenesis imperfecta [222](#),
[223F](#)

collagens

glycine and proline in [185](#)
Gly-X-Y tripeptide repeat [222](#)
triple helical structure [185](#), [222](#)

colorectal/colon cancer

familial adenomatous polyposis (FAP) [392](#)
genetic screening [454–5B](#)
frequent mutations in *APC*, *TP53*, *KRAS* genes [401](#)
hereditary nonpolyposis cancer (HPNC) see [Lynch syndrome](#)
mismatch repair deficiency
monoclonal antibody treatment [328T](#)
multi-stage evolution [370F](#)
somatic mutation number and age [371](#)
see also [mismatch repair deficiency](#).

colorectal tumors

frequency of chromosome instability [389](#)

frequency of genomic instability [389](#)

common ancestor, of human and mouse [43](#)

common ancestors, of evolutionary

common ancestors, family [114–5B](#), [123B](#)

complement C4 genes

copy number important in lupus [289](#), [290F](#)

excess activity in schizophrenia [288–9](#)

role in synaptic pruning [289](#)

complement genes

in age-related macular degeneration

C3 [288](#)

C4A/B [47F](#), [277T](#), [288–9](#), [290F](#)

C2 and CFB [88](#)

CFH [281](#), [288](#)

CR1 [287F](#), [288](#)

complementary DNA (cDNA)

libraries of [62](#)

preparation of [62](#)

complementary sequences / strands (in nucleic acids) [5](#), [5B](#), [6F](#), [66](#)

complex (common) genetic disease

assessment and prediction of disease risk [278–9B](#)

cancers, *see under* [cancer headings](#)

common vs. rare variants [276–7](#), [276T](#), [281](#), [281F](#)

disease risk prediction [254–5](#), [255T](#)

environmental factors [284](#), [290T](#), [291](#), [291F](#), [292](#), [294–6](#)
epigenetic factors [294–5](#)
genetic architecture of disease [280ff](#)
identifying susceptibility genes [261–2](#)
importance of immune system pathways [287–9](#)
lack of penetrance [255](#)
phenotype classification difficulties [256](#)
protective factors [281](#), [285F](#), [289–90](#), [290F](#)
roles of genetic factors in determining phenotypes [281](#), [281F](#)
strong genetic contribution for some diseases [258](#)
susceptibility factor concept [255](#)
compound heterozygotes [113](#), [114F](#), [221](#), [228](#)
computer programs

Alamut Visual Plus [447T](#)
BLAST [38](#), [38F](#), [39T](#)
BLASTP [40](#)
BLAT [39T](#)
ENSEMBL [39T](#)
HCOP [39T](#), [41F](#)
HomoloGene [39T](#), [41](#)
Mutalyzer [444B](#)
PolyPhen-2 [274](#), [443F](#), [447T](#)
pLoF [447T](#)
PROVEAN [274](#), [447T](#)
REVEL [443F](#), [447T](#)
SIFT [274](#)
SpliceAI [447T](#)
TBLASTN [41](#)

concordance rates of disease, MZ and DZ twins [257](#), [257T](#)
confidentiality, and genetic testing [473–4](#), [475F](#)
congenital adrenal hyperplasia [462T](#), [465](#)

congenital contractual arachnodactyly [240](#)

congenital hypothyroidism [464](#), [465T](#)

treatment [306](#), [308T](#)

consanguineous/consanguinity [112](#), [114](#), [114–5B](#)

coefficient of relationship [114B](#)

coefficient of inbreeding [115B](#)

fraction of genes in common with relatives [115B](#)

consent issues

form for clinical practice [477F](#)

genetic testing [474–7](#)

germ-line therapy [487](#)

conservative substitution [184](#)

conserved genes *see* [evolutionary conservation](#)

contact inhibition [366](#), [367B](#)

contiguous gene syndromes [207T](#)

copy number polymorphisms (CNPs) [277](#)

associated with complex disease [277T](#)

copy number variant/variation (CNV) [90](#), [92](#), [92F](#), [277](#)

detection of large-scale CNVs [425–6](#), [427B](#)

detection by whole genome sequencing in cancers [398](#)

coronary artery disease, protective variants for [289T](#)

COSMIC database [378T](#), [398](#), [398T](#), [403](#)

cousin marriages, counselling [458B](#)

CpG (CG) dinucleotide, *see also* [DNA methylation](#)

CpG islands [156B](#)

CREBBP gene [167T](#)

Creutzfeldt-Jakob disease, vCJD [229B](#)

CRISPR-Cas genome/gene editing [344F](#), [350](#), [350F](#), [351–2](#), [351F](#)

base editing [344F](#)

homology-directed DNA repair [351](#), [351F](#)

for making disease models [338B](#)

natural function of CRISPR-Cas [350](#), [350F](#)

prime editing [344F](#)

repair of mutant gene [345](#)

therapeutic applications [351–2](#), [351F](#)

crizotinib [409T](#)

Crohn's disease [277T](#)

% concordance in MZ and DZ twins [257T](#)

CFH (complement factor H) gene and [261](#)

FUT2 variants and [290](#)

high genetic contribution [257](#)

NOD2 susceptibility factor, [192–3](#), [193F](#), [216–2](#), [262F](#), [263](#),
[274](#)

protective variants for [289T](#)

susceptibility genes involved in autophagy [282B](#)

see also [inflammatory bowel disease](#)

cross-linking (DNA bases), *see* [base cross-linking](#)

cross-linking (polypeptides), *see* [disulfide bonds/bridges](#)

crossovers *see* [recombination](#)

crossing over, in meiosis I [15F](#)

Crouzon/Pfeiffer syndrome [190T](#)

cryptic splice sites [187](#), [188F](#)

identifying with SpliceAI [447T](#)

CTCF gene [395](#)
CTCF protein [173F](#)
CTLA4 / *CTLA4* [409T](#), [410](#)
CYP3A4 [315–6](#)
CYP3A5, and *CYP3A7* enzymes [316](#)
CYP21A2 gene [201–2B](#)
CYP21A1P pseudogene [201–2B](#)
CYP2C9 enzyme [316](#), [321](#)
CYP2C19 gene [315–6](#), [317T](#)
CYP2D6/*CYP2D6* [315–6](#), [316F](#), [317T](#)
CYP4F2 enzyme [321](#), [321F](#)
Cysteine

chemical class [185T](#)
cross-linking by disulphide bridges [31](#), [32F](#), [99F](#), [185T](#)
protein folding role [185](#)
selenocysteine [31](#)
structure [25F](#)

cystic fibrosis [136](#), [482](#)

ARMS mutation scanning [439](#)
gene therapy impractical [343](#)
lifetime disease risk [252](#)
locus-specific databases [192-T](#)
newborn screening [465T](#)
novel drug therapies [323–5](#)
protein misfolding [226](#)

CYT1 and *CYT2* (phosphatidylinositol-3-kinase) [147](#)
cytidine deaminases

in antibody diversification /RNA editing [352](#), [400](#)
excess production causing hypermutation [400](#)

cytochrome P450

- genetic variation [315–6](#)
- non-invasive testing [462T](#)
- phase I drug metabolism and [315](#)

cytochrome P450 gene superfamily [315](#)

cytokine storms [411](#)

cytokines [341](#)

cytokinesis [12–3](#), [14F](#), [15F](#)

cytosine(s)

- structure of [4F](#)

- distinguishing methylated and unmethylated [448–9](#), [4489F](#), [450F](#)

cytotoxic T lymphocytes (CTL) [99](#), [366](#)

- in immunosurveillance [366](#)

- interaction with MHC/HLA [103](#), [265F](#), [354](#), [410](#)

- suppression of, by cancer cells [369T](#)

D

D4Z4 array [170F](#), [171B](#)

Darier-White disease [247F](#)

Darwinian selection see [natural selection](#)

databases, generic

- clinical [42](#)

- human gene disorders and underlying genes [111B](#)

- human genetic variation [92T](#)

- human pathogenic mutation [192T](#)

databases, specific

ALFRED [92T](#)
[Clinicaltrials.gov359](#)
ClinVar [42](#), [423T](#), [443F](#), [447T](#)
COSMIC (Catalog of Somatic Mutations in Cancer) [192T](#),
[378T](#), [398T](#), [403](#), [447T](#)
dbSNP [92T](#), [224](#)
dbVar [92T](#), [423T](#)
DECIPHER [423T](#)
DGV database [92T](#)
Genecards [111B](#)
GeneReviews [11B](#)
Human Gene Mutation Database [192T](#)
Human Gene Nomenclature Committee (HGNC) database [39](#)
IMGT/HLA [104T](#)
IARC's *TP53* [397](#)
LRG (Locus Reference Genomic) database [444B](#)
MITOMAP [45F](#), [192T](#), [216](#), [292B](#)
OMIM (Online Mendelian Inheritance in Man) [42](#), [111B](#)
RefSeq database [39T](#)
RefSeqGene database [39T](#)
SpliceDisease [192T](#)
Wiley database of gene therapy trials [340](#)

de novo mutations

assessing pathogenicity [446](#)
frequency [189](#)
and mosaicism [123](#)

deafness

autosomal [191T](#)
congenital [483](#)
recessively inherited [125](#), [125T](#)

DECIPHER database [423T](#)

Deciphering Developmental Disorders study (DDD,UK) [451](#), [479](#)

dedifferentiation [333F](#), [369T](#), [397](#)

deletion

frameshifting in coding DNA [26–7B](#)

in-frame in coding DNA [26–7B](#)

in mitochondrial DNA [215](#), [216T](#)

see also [base deletion](#); [chromosome abnormalities](#); [indels](#)

denaturation [65](#), [65F](#)

dentatorubropallidolusian atrophy [194T](#)

dendritic cells, origin of [341F](#)

depurination [81](#)

derivative chromosomes [206T](#), [208](#), [209F](#)

designer babies [488–9](#)

developmental origins of adult health and disease [296–7](#)

dexamethasone [307F](#)

DGV database [92T](#)

diabetes, transient neonatal [173T](#)

diabetes, type 1

% concordance in MZ and DZ twins [257T](#)

HLA association [266B](#)

diabetes, type 2

% concordance in MZ and DZ twins [257T](#)

as amyloid disease [230B](#)

diet and [259](#), [290T](#)

thrifty gene hypothesis and [296](#)

variable heritability [259](#)

diamniotic twins [295](#)

dicentric chromosomes [207–8](#), [209F](#)
dicer (endo)ribonuclease [149](#), [344F](#), [345](#), [347](#), [348F](#), [349F](#)
dideoxy DNA sequencing and

Next Generation Sequencing compared [74–5](#), [435B](#)

dideoxynucleotides 722–3, [440](#), [440F](#)

diepoxybutane, inducing interstand crosslinks in DNA [86B](#)

diet

amylase, lactase gene variants selected after change of diet [95](#),
[95F](#), [97](#), [98F](#)

dietary fat intake and *LIPC* genotypes [291](#), [291F](#)

low in phenylalanine to treat phenylketonuria [235B](#)

thrifty phenotype and [296](#)

type 2 diabetes and unbalanced diets [259](#), [290T](#)

differentially methylated regions (DMRs) [161](#)

differentiation *see* [cell differentiation](#)

DiGeorge syndrome [203T](#)

digital PCR, principle of [432](#)

see also [droplet digital PCR](#)

digoxigenin labeling system [69B](#)

diploid cells vs. haploid cells [10–11](#)

direct repeats [202](#), [203F](#)

direct to consumer tests (DTC) [470–2](#), [478](#)

disease gene identification

cancer susceptibility genes [397](#), [401](#)

candidate gene approach [240](#)

via chromosome abnormalities [248](#)

via exome sequencing [249](#), [250F](#) [251F](#), [251T](#)

mutation screening, the final step [241](#)
positional candidate approaches [241](#)
positional cloning [240–1](#)
disease haplotypes, principle of [243F](#)
disease models

animal [337](#), [338–9B](#), [339–40](#)
cellular [337](#)
non-rodent [339–40](#)

disease prevention *see* [prevention of disease](#)
disease risk

calculating in single-gene disorders [113](#), [117F](#), [118](#)
calculating odd ratios [264](#), [264T](#)
complexity for common genetic disease [254–5](#)
empiric risks [255](#)
Hardy-Weinberg applications [131–2B](#)
lifetime risks, contrasting values for mendelian and
multifactorial disorders [254](#), [255T](#)
protective factors (genetic) [285F](#), [289](#), [289T](#), [290](#), [290F](#)
relative risk (risk ratio) [254](#), [255T](#)

distal locations, on chromosomes [205B](#)
disulfide bonds/bridges [31](#), [32F](#), [99F](#), [185T](#)
DMPK gene [196–7B](#)
DNA cloning [58T](#), [58–62](#)

in bacterial cells [59–60](#). [59F](#)

DNA damage [81–3](#). [81F](#)

alkylating agents and [82](#)
base deletion [81](#)

base modification [81](#), [81F](#)
base-base crosslinking [81–2](#), [81F](#)
causing cell cycle arrest or apoptosis [82](#)
deamination [81](#)
depurination [81](#)
DNA adducts [82](#)
damage responses [85B](#)
p53 role in protecting against [384](#)
pyrimidine dimers [81F](#)
responses/sensors [82](#), [390–1](#), [391F](#)

inherited disorders of [85B](#)

simple reversal of [82](#)
single- and double-strand breaks [83](#)
see also [DNA repair](#)

DNA double-strand breaks

common in cancer cells [384](#)
failure to repair [391](#), [391F](#)

DNA duplex

sense strand [7F](#)
template (antisense) strand [7F](#)

DNA duplication *see* [repetitive DNA](#)

DNA helicases [6F](#), [142](#)

promotor TFIIB and D [142F](#)

DNA libraries.

cDNA [62](#)
genomic DNA [62](#)

DNA ligases

in DNA cloning [62B](#)

DNA ligase IV [83](#)

DNA looping [143](#), [143F](#)

DNA methylation

5-me CpG binding proteins [156](#)

across length of a gene [156B](#)

across the genome [156B](#)

as brake on transposon proliferation [156](#)

in cancer cells [396](#)

changes during aging [295](#)

changes in early development [158](#), [158F](#)

CpG islands and [156B](#)

CpG target sequence [156](#), [157F](#)

and DNA demethylation [152F](#), [157F](#)

de novo methylation [157](#), [157F](#), [158F](#)

detecting aberrant [448–9](#), [449F](#), [450F](#)

as epigenetic mechanism [150T](#), [151](#)

function in mammalian cells [155–6](#)

hemimethylated DNA [157](#)

hypomethylation [150T](#), [155F](#), [172](#), [295](#), [396](#), [397T](#)

hypermethylation of pericentromeric DNA [156B](#)

maintenance of [157F](#), [158F](#)

mechanisms [156–7](#), [157F](#)

in open and condensed chromatin [155F](#)

S-adenosylmethionine, as methyl donor [296](#)

satellite DNA extensively methylated [155](#)

symmetric CG methylation [157](#)

DNA methyltransferases (DNMTs) [157](#)

DMNT1 [157](#), [157F](#)

DNMT3A/3B [157](#), [157F](#)

DNA mismatch repair deficiency, [293–4](#), [394F](#), [295–6B](#)

why especially associated with colon cancer [394](#)

DNA nanoparticles [335](#)

DNA nicks [392](#)

DNA nickases [351](#), [351F](#)

DNA polymerases

3'–5' exonuclease activity [80](#)

alpha, beta, gamma and delta [7T](#)

high and low fidelity [7T](#)

nonclassical vs. classical DNA-dependent [7T](#)

RNA-dependent [7T](#)

see also [reverse transcriptases](#)

DNA repair

of base cross-links [85B](#), [99–100](#)

DNA polymerases involved in [7T](#)

of double-strand breaks (DSBs) [83](#), [84F](#), [85B](#), [391](#), [391F](#)

inherited disorders of [85B](#)

PARP1 targeting as cancer therapy [409T](#)

of single strand breaks (SSBs) [83](#)

of mitochondrial DNA [7T](#)

DNA repair mechanisms [82–4](#)

base excision repair (BER) [7T](#), [82–3](#), [85T](#), [102](#), [392](#), [452](#)

homologous recombination (HR-mediated) [83](#), [84F](#), [85B](#), [391](#),
[391F](#)

nucleotide excision repair (NER) [84](#), [85T](#), [392](#)

see also [mismatch repair](#); [nonhomologous end joining.\(NHEJ\)](#).

DNA repair genes [396](#)

DNA replication

DNA synthesis [5](#), [6F](#)

5' to 3' direction [3B](#)

Okazaki fragments [6](#), [6F](#), [7](#), [7T](#)

lagging and leading strands [6](#), [6F](#)

pyrophosphate produced [6F](#)

mitochondrial DNA [45F](#)

replication errors [80](#), [82](#)

correction of [83](#)

replication fork [6F](#)

replication slippage [91](#), [91F](#)

semi-conservative nature of [6](#)

semi-discontinuous nature of [6F](#)

telomeric end-replication problem [367–8B](#)

see also [replication origins](#); [DNA polymerase](#)

DNA sequence variants

assessing pathogenicity [442–4](#), [443F](#), [445–6](#), [446F](#), [446–7](#),
[447T](#), [448](#)

clinical reporting of [444](#)

criteria for classifying variants [445–6](#), [446F](#), [447](#)

genomic constraint [443F](#), [444](#)

nomenclature for [444–5B](#)

sifting through [443–4](#), [443F](#)

triad of precedent, conservation and rarity [443–4](#), [443F](#)

variants of uncertain clinical significance (VUS) [436](#), [464F](#),
[447–8](#), [472](#), [475F](#)

DNA sequencing

commercially available platforms [75T](#)

Human Genome Project (HGP) [22](#), [36–7](#), [87](#)

principles of [71–5](#)

Sanger (dideoxy) [72F](#), [72–3](#)

single-molecule sequencing [75](#), [75T](#)

see also [massively parallel DNA sequencing \(Next Generation Sequencing\)](#)

DNA structure

antiparallel nature [3B](#)

antisense (template) strand [5](#)

complementary sequences [5](#)

strand asymmetry [3](#), [3B](#)

DNA variant types

advantageous variants [79](#), [94](#), [96B](#), [286](#)

damaging variants, average number inherited by a person [189](#)

see also [DNA sequence variants](#)

DNMT1 methyltransferase [157](#), [157F](#)

DNMT3A gene [395](#)

in *de novo* DNA methylation [157F](#)

in ICF syndrome [167T](#)

DNMT3A methyltransferase [157](#)

DNMT3B gene [167T](#), [396](#)

DNMT3B methyltransferase [157](#)

Dolly the sheep [151](#), [332](#)

dominant and recessive phenotypes [110–1](#)

definition of co-dominant [111](#)

definition of dominant [111](#)

definition of recessive [111](#)

dominant disorders

loss-of-function and gain-of-function mutations [220–1](#), [221B](#)

see also [autosomal dominant](#)

Dominant megacolon (Dom) mouse phenotype [240](#)

dominant-negative effects

producing severe loss-of-function [222](#), [223F](#)

p53 mutants and [386](#), [386–8B](#)

Dor Yeshorim organization [467](#)

dosage-sensitivity

aneuploidies and monosomies [163](#), [224](#)

genes expressing [220–1](#), [221B](#), [222](#), [225](#)

haploinsufficiency and [220–1](#), [221B](#)

see also [copy number](#); [gene dosage](#)

double helix, DNA

base pairing and anti-parallel strands [5B](#)

double minute chromosomes [377](#), [377F](#)

double-strand breaks in DNA (DSBs), [154T](#), [169T](#), [208](#), [351](#), [377–8](#)

repair of [83](#), [84F](#), [85B](#), [391](#), [391F](#)

Down syndrome [463–4](#)

combined screening for [464](#)
maternal age effects [212](#)
droplet digital PCR (ddPCR) [423T](#), [432–3](#)
drug activation [313](#)
drug development [311](#)

major stages [311F](#)

drug-handling enzymes [311–2](#)
drug interactions Flockhart table [315](#)
drug metabolized by one or multiple cytochrome P450 enzymes [315](#)
drug metabolism

genetic variation in Phase II metabolism enzymes [317](#)
Phase I and Phase II reactions [312–3](#), [313F](#)
metabolic ratio [315F](#)
slow and fast metabolizers [314–5](#), [314F](#)
stages affected by genetic variation [312](#)

drug resistance, cancer [411–2](#)
drug responses, adverse drug reactions [319–20](#)
drug screening, and disease models [337](#)
drug targets [318](#)

genetic variation in [318–9](#), [319T](#)

drug therapy see [adverse drug reactions](#); [cancer therapy](#); [small molecule drugs](#)
drug types, see [small molecule drugs](#); [therapeutic proteins](#); [monoclonal antibodies](#); [CAR-T cells](#)
drug toxicity testing [337](#)
Duchenne muscular dystrophy (DMD) [126](#), [199](#), [462T](#), [465–6](#)

exon skipping therapy [346B](#), [349T](#)

positional cloning of *DMD* gene [248T](#)
duplex [4](#), see also [heteroduplex](#); [homoduplex](#)
duplications, see [chromosome abnormalities](#); [gene duplication](#); [whole-genome duplication](#)
Dutch *Hongerwinter* [296](#)
DUX4 retrogene/gene [169–70](#), [170F](#), [170–1B](#)
dyskeratosis congenita [85](#), [191T](#)
dynamic mutations [129](#), [194–5](#)
dysplastic cell proliferation [362](#)
dystrophin gene [346B](#)
dystrophin protein, Dp40 isoform [93T](#)

E

*Eco*RI restriction nuclease and methyltransferase [60–1B](#)
Edwards syndrome [211](#)
egg cells

haploid [11](#)
huge numbers of mtDNA11
number of cell divisions to make a [189](#), [190F](#)

elastase [219–20](#), [228](#)
electromyography [197B](#)
electrophoresis

capillary electrophoresis [73–4B](#)
slab gel electrophoresis [73–4B](#)

Ellis-van Creveld syndrome [134](#), founder effect in Amish families
[134T](#)
ELP4 gene [143F](#)
embryofetopathy, phenylketonuria as [235B](#)
embryonic development

DNA methylation in [158](#), [158F](#)
effects on adult health [296](#)
embryonic stem cells (ES cells) [337](#), [338–9B](#), [353](#)
mouse [337](#), [338–9B](#)
human [353](#)

embryonic stem (ES) cell line [339B](#)
embryos

androgenetic [163F](#), [171](#)
gynogenetic/pathenogenetic [163F](#), [171](#)

emicizumab [487](#)

emphysema [228](#)

ENCODE (Encyclopedia of DNA Elements) project [43](#)

end-replication problem, at telomeres [367–8B](#)

endometrial cancer [454B](#)

endoplasmic reticulum aminopeptidases (ERAPs)

endosymbiont hypothesis [212](#)

enhancers

competition for [160T](#)

histone modification of [154](#), [154T](#)

lens and retina-specific [143F](#)

rapidly evolving, [43](#)

roles in splicing, see [splicing enhancers](#)

roles in transcription [143](#), [143F](#)

versus silencers [143](#)

ENSEMBL program [39T](#)

environmental effects

adaptations to [95](#), [95T](#), [96](#)

affecting phenotype [233–4](#)
genotype-phenotype correlation and [233–4](#), [234F](#)
liability thresholds [253–4B](#)
environmental factors

in complex diseases [290–2](#), [290T](#)
in embryonic development [296–7](#)
see also [gene-environment interactions](#)

EP300 gene [167T](#)

EPCAM gene [394](#)

epidermolysis bullosa [222](#)

epigenetic dysregulation

in cancer cells [389](#), [395](#)

in complex diseases and aging [294–6](#)

in Mendelian disorders [167ff.](#)

principles of [165–6](#)

rationale [395](#)

epigenetics [2](#), [140](#), [150](#)

DNA and chromatin modelling [151T](#)

in monozygotic twins [295–6](#)

primary and second epimutations [165–6](#), [166F](#)

transgenerational effects [296–7](#)

epigenetic gene regulation [140](#)

epigenetic marks/settings [150](#)

creating with chromatin ‘writers’ [152](#)

heritability of [150](#)

interpreting with chromatin ‘readers’ [152](#)

removing with chromatin ‘erasers’ [152](#)

resetting of [151](#)
stability of [151](#)
epigenetic mechanisms [140](#)

amyloid and prion mutant protein mechanisms [228](#), [229–30B](#)
five mechanisms affecting chromatin structure [150T](#)
long noncoding RNA effectors [158–9](#), [159T](#), [160F](#)

epigenetic reprogramming

artificial reprogramming of pluripotent stem cells [332B](#)
in cancer cells [395](#), [397T](#), [403](#), [404F](#)
in the early embryo [151T](#)
in germ cell development [151T](#)

epigenome(s) [361](#)

definition [294](#)
dysregulation in cancer cells [389](#)
epigenome-metabolism linkages [403](#), [404F](#)
genome-epigenome interactions in cancer [396–7](#), [397T](#)
high variability of [295](#)
how environmental factors interact with [295–6](#)
transgenerational epigenetic inheritance [296–7](#)

episome [234](#)

epistasis [232](#), [263](#)

Epstein-Barr virus [376](#)

ERAP (endoplasmic reticulum aminopeptidase) [265B](#)

ERBB2 (*HER-2*) oncogene [377](#), [403](#), [407F](#)

ERBB4 protein [146F](#), [147](#)

ethical considerations

animal models [337](#)

genetic enhancement and designer babies [488–9](#)
genetic testing [482–4](#)
germ line gene therapy [329](#), [488](#)
mitochondrial DNA replacement [355](#), [356F](#), [487–8](#)
newborn genome sequencing [484–5](#)
stem cells [332B](#)

euchromatin [37](#), [37F](#), [154T](#)

% in human genome [45F](#)
barrier elements [143](#)
boundary elements [143](#)
heterochromatin [143](#)
insulators [143](#)

eukaryotes

endosymbiont hypothesis to explain origins [212](#)

EVC gene [134](#)

evolutionary conservation

of alternative splicing patterns [147](#)
contrasting degrees for centromeric and telomeric DNA [10](#)
functional constraint and [43–44](#), [94](#)
gene identification through [240](#)
(genomic) constraint [443F](#), [444](#), [447T](#)
heterochromatin DNA [36F](#), [37B](#)
sequence conservation due to purifying selection [44](#)
p53 protein, human vs mouse [38](#)

evolutionary mechanisms

exon duplication (tandem) [47F](#)
exon shuffling by retrotransposition [53](#), [53F](#)

gene birth and loss [43](#)
gene duplication [47](#), [48F](#)
gene amplification through natural selection [97](#), [98F](#)
genome duplication [42–3](#)
retrogene formation by retrotransposition [49B](#)
evorlimus, mTOR inhibitor [324](#)
Ewing sarcoma [468T](#)
exclusion mapping [246B](#)
EXO1 exonuclease [393F](#)
exome [123](#), [249](#)

clinical exome gene panel [175B](#), [442B](#)

Exome Aggregation Consortium (ExAc) [88B](#)
exome sequencing [248–50](#)

in cancer studies [398](#), [400](#), [402](#)
identifying genes underlying recessive monogenic disorders
[249](#), [251T](#), [251F](#)
identifying a gene for recessive inflammatory bowel disease
[443F](#)
see also [clinical exome](#); [whole-exome sequencing.\(WES\)](#)

exon inclusion therapy for spinal muscular atrophy [346–7B](#)
exon-junction complexes [186B](#)
exon shuffling [53](#), [53F](#)
exon skipping

caused by splice site mutations [187](#), [187F](#)
therapeutic for Duchenne muscular dystrophy [346B](#)

exons

absent in some genes [22](#)

average size in human genes [22T](#), [23](#), [23F](#)
extension or truncation of, by abnormal splicing [187](#), [187F](#)
tandem duplication of [47F](#)
exon deletions/duplications

scanning for using MLPA [430](#), [432F](#)

exonic splice enhancer (ESE) [347B](#)
exonic splice suppressor (ESS) [347B](#)
expanded preconception screening (ECS) [467–8](#)
extravasulation [363F](#)
EZH2 gene [396](#)

F

F8 (factor VIII) gene [203](#), [204F](#)

intrachromatid recombination in [202](#), [203F](#)
obtained by functional cloning [240](#)

F9 (factor IX) gene [343](#)

facioscapulohumeral dystrophy (FSHD) [169–70](#), [170F](#), [170–1B](#)

facultative heterochromatin [37B](#), [159](#)

associated histone modifications [154T](#)

FADD adaptor [385F](#)

familial adenomatous polyposis (FAP) [385T](#), [392](#)

familial cancers

breast/ovarian cancer [385T](#)

germline mutations in TS genes [385T](#)

hereditary nonpolyposis cancer [454B](#)

heritable oncogene mutations [384](#), [385T](#)

Li-Fraumeni syndrome [387B](#), [453](#)

and sporadic [381](#)

two-hit paradigm [381–2](#), [382F](#)

familial hypercholesterolemia [113](#), [283](#), [308T](#), [322](#), [451B](#), [453](#), [456](#),
[482](#)

gene panel for [442B](#)

familial melanoma [385T](#)

family members and genetic testing [455](#), [458B](#), [463](#), [474B](#), [475T](#),
[476T](#), [481](#)

family studies

Amish families [134](#), [134F](#)

inter- and intrafamilial variation [121F](#), [128F](#)

recording pedigrees [111–12](#), [112F](#)

Fanconi anemia [85–6B](#), [421T](#).

functional assay [421F](#)

FAS ligand (FASLG) [385F](#)

FAS receptors [385F](#)

Fatal familial insomnia [229B](#)

FBN1 fibrillin gene [240](#)

FBN2 fibrillin gene [240](#)

FDA (US Food and Drug Administration) [326–7](#), [408](#)

ferritin

translational regulation [147](#), [148F](#)

heavy chain gene family [47C](#)

fetal aneuploidy screening [463](#)

fetal ‘combined screening’ (Down syndrome) [464](#)

fetal tissue sampling [419T](#)

FGFR1 (fibroblast growth factor receptor type 1) gene [402](#)

FGFR2 (fibroblast growth factor receptor type 2) gene [190T](#)

FGFR3 (fibroblast growth factor receptor type 3) gene [190T](#), [219](#), [219F](#)

FGFR3 protein [130](#), [190T](#)

fibrillins

fibrillin (*FBN1* and *FBN2*) genes [240](#)
in Marfan syndrome [222](#)

Finland [133](#)

Finnish population, study of founder effects [133](#)

FISH (fluorescence *in situ* hybridization) [377F](#), [390](#), [390F](#), [401](#), [423T](#)

FITC (fluorescein isothiocyanate labeling) [69B](#)

fitness, and purifying selection [135](#), [135F](#)

fluorochrome [390F](#)

fluorophores/fluorochromes [69B](#), [73](#), [390F](#), [424](#), [430](#), [435F](#), [438F](#)

FMR1 gene [197T](#), [198](#)

folate [298](#)

and neural tube defects [290T](#)
and S-adenosylmethionine [296](#)

founder effects [133](#), [134F](#), [137](#), [154](#)

fragile sites [194T](#), [195](#), [197T](#)

fragile X-associated primary ovarian insufficiency (FXPOI) [198](#)

fragile X mental retardation syndrome [169](#), [194T](#), [195](#), [197T](#)

cause of [198](#)
premutations [198](#)

fragile X tremor-ataxia syndrome (FXTAS) [198](#)

frameshift mutations [181F](#), [186](#), [187F](#), [188F](#), [189](#), [193](#), [23](#), [345](#), [346B](#), [394](#)

see also [translational reading frame](#)

Francis Galton [252](#)
Friedreich ataxia [169](#), [194T](#)
frontal dementia and/or amyotrophic lateral sclerosis [194T](#)
frontotemporal lobar degeneration [230B](#), [256](#)

cytoplasmic aggregation of tau protein in [230B](#)

α (1,2)-fucosyltransferase [290](#)
fumarylacetoacetate hydrolase (FAH) [307F](#), [308](#)
functional cloning [240](#)
functional constraint on human DNA sequences, estimating
fusion (onco)gene, detection [428–9](#), [429F](#), [430](#)
FUT2 gene variant, susceptibility to Crohn's/type I diabetes [290](#)

G

G-banding [204–5B](#)
G₀ and G₁ phase [11](#), [12F](#), [13](#)
G₂ phase [11](#), [12F](#)
G₁-S transition [383](#)
G-U base pairing, in RNA [33](#), [33F](#)
gain-of-function mutations [218–9](#), [219F](#), [220](#)

and loss-of-function in one gene [223–4](#)
oncogene activation [378–9](#), [390F](#)
in some tumor suppressor genes [386](#)
see also [missense mutations](#)

galactosemia [465](#)
gametes [13](#)

differential methylation of sperm and eggs [163](#)
DNA methylation in gametogenesis [158](#), [158F](#)
number of cell divisions needed to make [189](#), [190F](#)
why each one is genetically unique [16–7](#), [17F](#)

gametogenesis

epigenetic reprogramming [158](#), [158F](#)
sex differences [14](#)

gammaretroviruses, [336](#), [337T](#), [342](#)

Garlin syndrome [385T](#)

gastric cancer [397](#), [404](#), [454B](#), [458B](#)

gastrointestinal stroma tumours (GIST) [468T](#)

gel electrophoresis

pulsed field [73B](#), [171B](#)

slab gels and capillary gels [73–4B](#)

GenCC (gene curation coalition) database [442B](#)

GENCODE database [39T](#), [45](#), [46T](#)

gene amplification [97](#), [181](#), [181F](#), [315](#), [316F](#), [407F](#)

AMY1A gene [97](#), [98F](#)

CYP2D6 gene and ultrafast metabolizers [315](#), [316F](#)

in oncogene activation [376F](#), [377](#), [377F](#), [407F](#), [423T](#), [428](#), [468T](#)

gene augmentation therapy *see* [gene supplementation therapy](#).

gene birth and loss, during evolution [43](#)

gene bodies, histone modification of [154T](#)

gene conversion [182T](#), [200](#), [200F](#), [201–2B](#), [382](#)

gene dosage [50](#), [116](#), [118](#), [163–5](#), [289](#), [447](#)

and aneuploidy [210](#)

see also [copy number](#); [dosage-sensitivity](#).

gene duplication [93T](#)

genetic drift [133](#), [133F](#), [135](#), [137](#)

evolutionary advantage [50](#)

and olfactory receptors [98](#)
producing protein diversity [98](#)
and segmental duplication [46](#)
tandem repeats [46](#), [47F](#), [48F](#)
gene-environment interactions (GxE)

case-control studies [293](#)
GWA studies [293](#)
importance of [291](#), [291F](#)
prospective cohort studies [293–4](#), [294T](#)

gene expression

basics of transcription [7](#), [7F](#), [8](#)
effect of condensed versus open chromatin structure [151–2](#),
[152F](#), [155F](#)
position effects [151](#), [168](#), [169F](#)
naturally monoallelic for imprinted genes [160–1](#), [160T](#)
naturally monoallelic for many X-linked genes [163–4](#), [164F](#)
see also [gene regulation](#)

gene families

clustered [47](#), [47F](#)
examples of human [47–8](#), [47F](#), [47T](#)

gene knockouts [338B](#)

gene panels

obtained by targeted DNA sequencing [436](#), [440](#), [441–2B](#)
Pan-Cancer gene panel [430](#)
Syndromic Intellectual Disability gene panel [175B](#)
see also [clinical exome](#); [virtual gene panels](#)

gene pool [130](#)

gene regulation

2 fundamental types [139–40](#)

cis- and *trans*-acting effects, DNA level [140](#), [141F](#)

cis- and *trans*-acting effects, RNA level [140](#), [141F](#)

gene silencing/suppression [344–5](#)

artificial, by using RNA interference [347–8](#), [348F](#), [349F](#), [349T](#)

via DNA methylation [155](#)

via genomic imprinting [151T](#)

via heterochromatin spreading [169F](#)

via position effect [151](#)

gene targeting [349](#)

via homologous recombination [338–9B](#)

gene therapy

broad strategies [329–30](#), [330F](#)

delivery problems [330–1](#)

disease models, importance of [337](#), [338–9B](#), [339–40](#)

efficiency and safety aspects [333–4](#)

first successful [341–2](#), [342F](#)

gene suppression /silencing therapy [344](#), [344F](#), [345](#), [345F](#), [347–8](#), [349F](#), [349T](#)

germ-line gene therapy [329](#), [485–6](#), *see also* [mitochondrial replacement therapy](#).

in vivo and *ex vivo* [334–5](#), [335F](#)

in vivo gene transfer [342–4](#)

non-viral delivery [334](#), [336](#)

plasmid and liposome vectors [335](#)

RNA and oligonucleotide therapeutics [344ff.](#)

- safety versus efficiency [335–6](#)
- somatic vs germline [329](#)
- splicing modulation therapy [344F](#), [345](#), [346–7B](#)
- therapeutic genome/gene editing [349–50](#), [350F](#), [351](#), [351F](#), [352](#), [352T](#)
- viral delivery systems [334](#), [336–7](#)
- gene therapy clinical trials [340](#)
- gene therapy clinical trials worldwide database [340](#)
- GeneCards database [111B](#)
- GeneReviews* resource [111B](#)
- genes [2](#), [7](#)

- mitochondrial [44](#), [45F](#)
- number in human genome (GENCODE) [46T](#)
- protein-coding vs. RNA genes [22](#), [22T](#), [32](#)
- single-exon/intronless [186B](#)
- see also* [pseudogenes](#), [retrogenes](#)

- generic drugs [325F](#)
- Genetic Alliance UK [466](#)
- Genetic association and linkage compared [263F](#), [264](#)
- genetic code

- nuclear and mitochondrial DNA [184F](#)
- redundancy in [26](#), [183–4](#), [184F](#)

- genetic counseling [122](#), [128](#), [131–2B](#), [458–9B](#)
- genetic drift [133](#), [133F](#), [135](#), [137](#)
- genetic enhancement [486–9](#)
- genetic heterogeneity *see* [allelic heterogeneity](#); [locus heterogeneity](#).
- genetic mapping

- DECIPHER database [423T](#)
- Human Genome Project (HGP) [22](#), [36–7](#)

(International) HapMap Project [268](#), [269–70B](#), [270](#), [272–3](#)
see also [linkage analysis](#); [association analyses](#)
genetic screening [420](#)

neonatal [258](#)

genetic test parameters

false negative rate [422T](#)
false positive rate [422T](#)
negative predictive value [422T](#)
positive predictive value [422T](#)
sensitivity of a test [422T](#)
specificity of a test [422T](#)

genetic testing, laboratory services

an overview [418–422](#)
chromosome abnormalities and structural variation [423ff.](#)
detection vs. scanning [420F](#)
direct versus indirect testing [421](#), [421T](#)
DNA methylation testing [448](#), [449F](#), [449](#), [450F](#)
evaluating genetic tests [422](#), [422T](#)
pathogenic point mutations [433ff.](#)

genetic testing, clinical, population and ethical aspects

carrier cascade testing [455](#)
confidentiality concerns [474–7](#)
consent issues [473–4](#), [475F](#)
diagnostic and presymptomatic/predictive testing [453–4](#)
direct-to-consumer (DTC) testing [418](#)
incidental findings [479](#)
newborn screening (standard) [464–5](#), [465T](#), [466](#)

- newborn screening via whole genome sequencing [466](#)
- noninvasive prenatal testing [461–2](#), [462F](#)
- predictive genetic testing and genetic screening [420](#)
- pregnancy screening for fetal aneuploidies [463–4](#)
- preimplantation testing [460–1](#)
- sources of materials for [419](#), [419T](#)

genetic testing, service provision and development

- commercial service provision [418](#)
- mainstreaming of genetic and genomic medicine [452B](#), [479](#), [481B](#)
- national genomic medicine initiatives [451](#), [451B](#)

genetic treatment of disease

- an overview of, [303–5](#)

genetic variation

- classes of [78T](#)
- constitutional [78](#)
- functional (effects on the phenotype) [93pp.](#)
- functions of genes showing highest [79](#)
- heteroplasmy (mtDNA), *see* [mitochondrial DNA](#), [heteroplasmy](#)
- pathogenic, broad classes of [180–1](#)
- post-zygotic vs. somatic [78](#), [78T](#)
- in proteins, origin of [93](#), [93T](#)
- see also* [DNA variants](#); [mutations](#)

genetic variation in humans

- advantageous variants [96–7B](#)
- causing adverse drug reactions [312](#), [314](#), [319](#), [319T](#), [319–20B](#)
- chimpanzee comparison [133](#)

comparatively high in Africans [270B](#)
databases of [92T](#), [192T](#)
extreme for HLA, Ig and T-cell receptors [99](#), [99F](#)
limited due to recent population bottleneck [269B](#)

genome

defined [2](#)
size and organism complexity [43](#)
see also [human genome](#)

genome-epigenome interactions in cancer [396–7](#), [397T](#)
Genome Aggregation Database Consortium *see* [gnomAD](#)
genome browsers [39T](#)
genome editing [330F](#), [349](#)

in making disease models [338B](#)
using zinc finger nucleases [352T](#)
see also [CRISPR-Cas genome/gene editing](#)

genome evolution [7T](#), [24](#), [42–4](#)
genome instability

due to reduced methylation in pericentromeric heterochromatin
[156B](#)
due to reduction of centromeric heterochromatin [169](#)
in cancer [366](#), [368B](#), [369T](#), [389–94](#)

genome organization programs [39T](#)
genome sequencing

see [Human Genome Project](#); [human population genomics](#);
[personal genome sequencing](#); [whole genome sequencing](#)

genomewide association (GWA) studies (GWAS) [268](#)

carrying out GWA with SNP chips [270](#), [270F](#)
difficulty in moving from associated SNP to causal variant [273](#)
identifying casual variants [274](#)
limitations of [274–5](#)
Manhattan plots [271F](#)
meta-analyses enabled by genotype imputing [272–3](#)
‘missing heritability’ problem [275](#)
phasing and genotype imputation [271–2](#), [272F](#)
stringent P values [271](#)
Visscher polygenic statistical approach [275–6](#)
Wellcome Trust Case Control Consortium [273](#)
genomewide cancer studies [408–18](#)
genomewide DNA sequencing [423T](#)

cancer studies [399](#), [416–17](#)
identifying novel cancer genes [402](#)
incidental findings [479](#)

genomewide linkage analysis
genomewide RNA sequencing [405–6](#)
genomic constraint [444](#), [447T](#)
genomic imprinting [151T](#), [160](#), [160T](#)

evolutionary conflict between mothers and fathers [161](#)
extent of in mammals [161](#)
imprinting control region [161](#)
imprints, reversibility of [161](#), [162F](#)
occasionally tissue-specific [161](#)
origin of in mammals [161](#)
reversal of imprinting pattern between generations [162F](#)

genomic instability and cancer
Genomic Medicine Service UK. [450](#)
genotype, definition [110](#)

genotype-phenotype correlations [124](#)

allelic heterogeneity and [125–6](#), [126F](#)

anticipation and [129](#), [129F](#)

difficulties in Mendelian disorders [231–2](#)

difficulties in mitochondrial disorders [232](#)

environmental/epigenetic factors and [233–234](#)

imprinting and nonpenetrance [127](#), [127F](#)

locus heterogeneity and [125](#), [125F](#), [126F](#)

modifier genes and [232–3](#), [234F](#)

projects, [88B](#)

threshold effects [232](#), [232F](#)

variable expressivity of Mendelian phenotypes [128](#), [128F](#)

genotyping point mutations/SNVs

methods for [436–7](#), [437F](#), [438F](#), [439F](#)

multiplex genotyping [438–40](#), [440F](#)

germ line gene therapy [488](#)

germ cell layers [332B](#)

germ cells/germ line/germline [13](#)

number of cell divisions to make sperm and egg cells [189](#),
[190F](#)

see also [mosaicism, _germline](#); [mutation rates, _germline](#);
[primordial germ cells](#)

Gerstmann-Straussler-Scheinker syndrome [229B](#)

Gleevec (imatinib) [408](#), [409T](#)

Giemsa staining, [35](#), [204B](#)

glioblastoma (multiforme) [403](#), [406](#), [407F](#), [411](#)

2 major signaling pathways [406](#), [407F](#)

α -globin [51](#)

in thalassemias, [51](#), [233](#)

α -globin genes, copy number variation [233](#)

β -globin [51](#)

β -globin gene/gene clusters [22T](#), [47F](#), [48](#)

see also [sickle-cell disease](#); [thalassemia](#)

globin superfamily [50–1](#)

glucose-6-phosphate dehydrogenase deficiency, heterozygote advantage [97](#)

glutamate/glutamic acid

chemical class [185T](#)

structure [25F](#)

glutamine

chemical class [185T](#)

structure [25F](#)

glutathione S-transferase (GST) superfamily [318](#)

glutaric aciduria type 1 [465](#)

glycine

chemical class [185T](#)

role in protein folding [185](#)

structure [25F](#)

gnomAD (Genome Aggregation Database Consortium) [88B](#), [443F](#), [447F](#)

gonadal dysfunction [225](#)

gonadal mosaic/mosaicism *see* [mosaicism](#), [germline](#)

gout [232F](#)

GM2 ganglioside [309](#), [467](#)

graft-versus-host disease (GVHD) [340](#)

Graves disease [288](#)

great apes [94](#)

Gregor Mendel [251](#)

growth factor receptor-RAS signal transduction pathway [191](#)

growth factors

transforming growth factor β (TGF β) [394](#)

vascular endothelial growth factor (VEGF) [328T](#), [369T](#)

see also [fibroblast growth factor](#)

GT(U)-AG splice sequence signal [24](#)

see also [alternative splicing](#)

guanine

8-oxo [-7,8-dihydro-] [81F](#)

structure of [4F](#)

guide RNAs [345](#)

guide sequence [350](#)

gut microbiome [291–2](#)

benefits of [262](#)

Guthrie card [419T](#)

GWAS, *see* [genomewide association](#)

gynogenetic embryos [171](#)

H

Hantigen [290](#)

H2A.2 histone variant [154T](#), [155](#)
H2A.X histone variant [154T](#), [155](#)
H2A.Z histone variant [154T](#), [155](#)
H3.3 histone variant [154T](#), [155](#)
H19 gene [173F](#)
Hailey-Hailey disease [246B](#)
hand-foot-genital syndrome [194T](#)
haplogroups, see [mitochondrial DNA \(mtDNA\)](#)
haploid cells [11](#)
haploinsufficiency

dominant disorders and [221](#), [221B](#) [241](#), [304](#)
dosage-sensitivity and [220](#), [221B](#)
tumorigenesis due to [386](#)

haplotype blocks [268](#), [269–70B](#), [274](#), [298–9](#)
Haplotype Reference Consortium [272](#)
haplotype [242](#)

deriving HLA haplotypes in families [106B](#)
disease-associated [242](#), [243F](#)
use in linkage analysis [247F](#)

HapMap project [268](#), [269–70B](#), [272–3](#)
Hardy-Weinberg law [130–1](#), [131–2B](#), [136](#)
HBA1 and *HBA2* α -globin genes [233](#)
HBB β -globin gene [22T](#), [47F](#), [95T](#), [125](#), [233](#), [234F](#), [488](#)
HBBP1 pseudogene [47F](#)
 β -HCG (human chorionic gonadotropin) levels [464](#)
HDL-C (high-density lipoprotein cholesterol) [278](#), [291F](#)
heart attack [322](#)
heat-shock proteins [226](#)
heatmaps [289](#)
helicase [5](#), [6F](#), [142](#)

Helicobacter pylori infection

strongest risk factor for gastric cancer [407](#)

helper T lymphocytes [100](#), [102–3](#), [265B](#), [282B](#), [352T](#)

hematopoietic stem cells [306](#), [322B](#)

bone marrow as a source of [333](#), [335](#), [340](#), [341F](#), [342F](#)

as gene therapy targets [335](#), [340](#), [341F](#)

origin of some tissue immune system cells [341F](#)

source of all blood cells [341F](#)

hemizyosity/hemizygous [116](#), [123](#), [207T](#), [225](#)

definition of hemizygous [78](#), [110](#)

see also [autozygosity](#); [heterozygosity](#); [homozygosity](#).

hemochromatosis [234F](#), [308T](#), [439](#), [482](#)

hemoglobin [95T](#)

aggregation in sickle-cell disease [227](#)

HbF (hemoglobin F) [233](#)

HbS (hemoglobin S) [115](#), [227](#), [227F](#)

tetrameric structure [23](#)

see also [globins](#); [globin genes](#); [sickle-cell disease](#); [thalassemia](#)

hemolytic disease of newborn [462](#)

hemophilia A

common inversion in Factor VIII gene [203](#), [204F](#)

intrachromatid recombination [204F](#)

hemophilia B, successful gene therapy [344](#)

hepatitis B virus [376](#)

HER-2 (ERBB2) oncogene [486T](#)

HER2 and HER3 receptors [412](#)
HERVs (human endogenous retroviruses) [52](#), [52T](#)
hereditary neuropathy with liability to pressure palsies (HNPP) [203T](#)
hereditary nonpolyposis cancer, *see* [Lynch syndrome](#)
hereditary transthyretin amyloidosis [348](#), [349T](#)
heritability

definition [256](#)
of epigenetic marks [151](#)
estimating via family/twin studies [256–7](#)
missing heritability *see under* [genomewide association studies \(GWAS\)](#)
revealing variable genetic contributions for different disorders [258](#)
variability in changing environment [258–9](#)

heterochromatin

centromeric, poorly conserved DNA [43F](#), [51](#)
heritable epigenetic settings for centromeric and telomeric [150](#), [151T](#)
constitutive, % of human genome [43F](#)
constitutive, locations on human chromosomes [36F](#)
dysregulation causing disease [168–70](#), [170–1B](#)
DNA sequencing of, difficulties with [37](#)
facultative [37B](#)
heterochromatin spreading [169](#), [169F](#)
pericentromeric, extensively methylated [155](#), [156B](#)
telomeric, strongly conserved DNA [51](#)

heterochromatin protein 1 [155F](#)

heterodisomy [171](#), [172F](#)

mixed heterodisomy/isodisomy [427–8B](#)

heteroduplex [65F](#), [66](#), [67F](#), [68F](#)

heteroplasmy *see* [mitochondrial DNA \(mtDNA\)](#)

heterozygosity/heterozygote

compound heterozygotes [113](#), [114F](#), [221](#), [228](#), [317F](#)

definition of heterozygote/heterozygous, [77](#), [110](#)

loss of, *see* [tumors, loss of heterozygosity](#)

manifesting heterozygote [117–8](#), [136](#)

see also [autozygosity](#); [hemizyosity](#); [heterozygosity](#)

heterozygote advantage [136–7](#).

distinguishable from founder effect [137](#)

HEXA gene (hexosaminidase A) [467](#)

hexosaminidase A [309](#), [467](#)

HFE gene

modifier gene for β -thalassemia [234](#)

variants causing hemochromatosis [439](#)

HGMD *see* [Human Gene Mutation Database](#)

high-density lipoprotein cholesterol

and LIPC variants [291F](#)

Hirschsprung disease [224](#), [240](#)

histidine

chemical class [185T](#)

structure [25F](#)

histocompatibility testing [104–5](#)

family studies [106B](#)

histones

modifying enzymes [153](#)

substitution [153](#)

histone acetyl transferases (HATs) [153](#), [167T](#)

histone acetylation and deacetylation [152](#), [152F](#)

histone deacetylases (HDAC) [153](#), [155](#)

histone demethylase [153](#), [167T](#)

histone H3 variant [151T](#)

histone lysine methyltransferases (KMTs) [153](#)

histone modifications [150](#), [152](#), [153F](#)

aging and [294](#)

characteristic of different chromatin states [154T](#)

classes of modified amino acids [153F](#)

effects on chromatin structure [154](#), [154F](#), [155F](#)

epigenetic mechanism [150T](#)

in nucleosomes [153F](#)

histone N-terminal tails [153F](#)

histone proteins and DNA compaction [155F](#)

histone substitutions [150](#), [153](#), [154T](#)

epigenetic mechanism [150T](#)

in open and condensed chromatin [155F](#)

histone variants [154](#), [154T](#)

hitchhiker alleles, in selective sweeps [96B](#)

HIV/AIDS [352T](#), [413](#)

HLA (human MHC)

allele nomenclature [105B](#)

class I and class II HLA regions [105B](#)

class I gene family [47T](#)

class I genes [354](#)

silenced by deleting *B2M* gene [354](#), [354F](#)

class II genes

silenced by deleting *CIITA* gene [354](#), [354F](#)

classical class I proteins [99](#), [99F](#)

classical class II proteins [99F](#), [100](#)

disease associations [106](#), [265–6B](#)

strongest risk factors in autoimmune disease [287–8](#)

extreme heterozygosity, due to natural selection [104](#)

gene organization [105B](#)

haplotypes [105–6B](#)

histocompatibility testing/HLA typing [104–5](#), [105B](#)

donor-patient matching in transplantation [105](#), [340](#),
[353–4](#)

serological vs. DNA typing [264](#)

low recombination within HLA complex [267](#)

medical importance of [104–5](#)

polymorphism statistics [104T](#)

protection against infectious disease [95](#)

structural similarity of proteins to Igs and TCRs [99F](#)

transplantation and graft-versus-host disease [104](#)

see also [MHC](#); [MHC-peptide binding](#); [MHC restriction](#); [β2-microglobulin](#)

HLA alleles (in disease associations)

HLA-B27 [266T](#)

HLA-Cw6 [266T](#)
HLA-DQ2, DQ6, DQ8 [266T](#)
HLA-DR4 [266T](#)

HLA selection, in preimplantation diagnosis [484](#)
HMG CoA reductase (hydroxymethylglutaryl CoA) [318](#), [322](#)
hMutS α protein [392](#), [393F](#), [395B](#)
hMutS β protein [392](#), [393F](#)
HNF4A [471](#)
hnRNP proteins, splicing regulators [145](#)
homoduplex [65F](#), [67F](#)
homogeneously staining regions (in cancer) [337](#)
HomoloGene computer program [39T](#), [41](#)
homologous chromosomes (homologs)

pairing of paternal and maternal autosomal homologs, [15F](#) [16](#)
X-Y pairing at pseudoautosomal regions [119](#), [119F](#), [120](#), [120F](#)

homologous genes (homologs) [39T](#)
homologous recombination (HR)-mediated DNA repair [83](#), [84F](#), [85B](#),
[391](#), [391F](#)
homologous sequences, on X and Y chromosomes [119](#), [119F](#)
homologs, *see* [homologous chromosomes](#); [homologous genes](#)
homoplasmy, *see under* [mitochondrial disorders](#)
homoplasmy, *see* [mitochondrial DNA \(mtDNA\)](#)
homozygosity

definition of homozygote/homozygous [78](#), [110](#)
see also [autozygosity](#); [hemizygosity](#); [heterozygosity](#).

Hongerwinter [296](#)
HOTAIR (HOX antisense intergenic RNA) [159T](#)
HOTTIP RNA [159T](#)
Housekeeping genes [144](#)
HOXC homeobox gene cluster [159T](#)

HP1 heterochromatin protein [155F](#)
HPRT (hypoxanthine guanine phosphoribosyltransferase) [232F](#)
HPRT gene, variable disease phenotypes according to residual function
[232F](#)
Hsp60, Hsp70 molecular chaperones [226](#)
HTT (huntingtin) gene [47F](#), [195F](#)
human chromosomes

 acrocentric, with ribosomal DNA [208](#), [209F](#)
 ancestral chromosome segments [267–8](#), [268F](#)
 average length and DNA content [8](#)
 chromosome banding methods [204B](#)
 consequences of additional [124](#)
 constitutive heterochromatin, locations of [36F](#)
 G-banding pattern, and gene density in [36F](#)
 gene-rich vs gene-poor [46](#)
 ideogram showing chromosome banding [36F](#), [205B](#)
 karyotyping, standard [204](#), [204–5B](#)
 spectral karyotyping of, in tumor cells [390](#), [390F](#)
 nomenclature of chromosomes/chromosome banding [204–5B](#)
 nomenclature of chromosome abnormalities [206T](#)
 proximal and distal locations [205B](#)
 spectral karyotyping used for tumor cells [390](#), [390F](#)
 see also [chromosomes](#); [sex chromosomes](#); [chromosome abnormalities](#)

human endogenous retroviruses (HERVs) [52](#), [52T](#), [52F](#)
Human Fertilisation and Embryology Authority (HFEA, UK) [484](#), [487](#)
Human Gene Mutation Database (HGMD) [192T](#), [447T](#)
Human Genetics Commission UK [471](#)
human genetic maps [242](#)
Human Gene Nomenclature Committee (HGNC) [39T](#), [40–1](#), [40F](#)
Human Gene Nomenclature Database [39T](#), [40–1](#), [40F](#)

human genome

% coding DNA [43](#), [43F](#)

% constitutive heterochromatin [43F](#)

% functionally significant [44](#)

% highly conserved (functionally constrained) [43F](#)

analysis and interpretation

electronic resources for interrogating [39–42](#)

evolutionary conservation [43–4](#), [43F](#)

gene density [36F](#), [45–6](#)

mitochondrial genome, *see* [mitochondrial DNA \(mtDNA\)](#)

multigene families [47](#), [47T](#)

mutation rates [189](#)

noncoding DNA [21](#)

coding and noncoding RNA transcripts per protein-coding gene

protein-coding genes, total number (GENCODE) [46T](#)

pseudogenes, number in genome (GENCODE) [46T](#)

reference sequence [77](#)

repetitive DNA [21](#), [43F](#)

 extent of [46](#)

 retrotransposon repeats [52T](#), [156](#)

 RNA genes, total number (GENCODE) [46T](#)

 size of [21](#)

 transcription pervasive [43](#)

Human Genome Project (HGP) [22](#), [36–7](#), [87](#)

Human Genome Variation Society (HGVS) [444](#), [445B](#)

human herpesvirus-8 [376](#)

human population, idealized for genetic studies [130](#)

human population genomics [88](#)

 projects [88B](#)

humanized antibodies [334](#)
huntingtin (protein), nuclear aggregates in Huntington disease [230B](#)
Huntington disease [128](#), [195](#), [197T](#), [474](#)

- as amyloid disease [230B](#)
- CAG repeat expansion [47F](#), [194T](#), [195F](#)
- consent versus duty of care to family members [474B](#)
- founder effect, [134T](#)
- genetic testing [421](#), [428](#), [453](#), [456](#)
- variable age at onset [127](#), [127F](#), [128](#)

hybridization *see* [nucleic acid hybridization](#)

hybridomas [326](#), [327F](#)

hydatidiform mole [163F](#)

hydrogen bonding

- disruption of (denaturation) [65F](#), [66](#)
- in α -helices and β -sheets [31F](#)
- in A-T and G-C base pairs [4](#), [4F](#)
- in single-stranded RNAs [27](#), [29F](#)

hydrolytic damage to DNA [81](#)

hypermutation [400](#)

hyperphenylalaninemia [235B](#)

hyperplastic growth/cell proliferation [362](#)

hypogonadism [225](#)

hypomorphic mutations [180](#), [221B](#)

hypoxanthine phosphoribosyl transferase) [231](#), [232F](#)

hypoxia response [95T](#)

I

IAPP islet amyloid peptide [230B](#)

ICF syndrome [167T](#)

ICF1 (immunodeficiency with centromeric instability and facial anomalies) [396](#)

ICR (imprinting control regions)

ICR1 and ICR2 [172](#), [173F](#)

identity testing [419T](#)

IDH1 and *IDH2* genes (isocitrate dehydrogenase) [395](#), [403](#)

IGF2 gene [161](#), [173F](#)

IGH gene/locus [100](#)

IGH-MYC fusion gene [379F](#)

IGK, *IGL* immunoglobulin genes/loci [100](#)

IL-17, IL-23 (interleukins) [282B](#)

IL2RG gene, and severe combined immunodeficiency [341](#)

Illumina DNA sequencing [75T](#), [435B](#)

Illumina genomic methylation scans [295](#)

Illumina Infinium HumanMethylome450 BeadChip [295](#)

Illumina TruSight Pan-Cancer Panel [430](#)

Illumina TruSight One gene panel (clinical exome) [442](#), [442B](#)

imatinib (Gleevec) [408](#), [409T](#), [412](#)

immune cloaking [354](#), [354F](#)

immune privilege/immune privileged sites [288](#), [353](#)

immune system/immune responses

adaptive immune system [99–106](#), [265–6B](#)

in cancer therapy [409](#), [409T](#), [410](#)

genetic variation in [99](#), [99F](#), [100](#)

innate immune system [262](#), [277T](#), [287–8](#)

see also [autoimmune diseases/autoimmunity](#).

immune system cells and CNS

immune system cell types in brain [288](#)

importance of immune system cells in neurodevelopment [288](#)

immune tolerance [265B](#)
immunodeficiency, [86B](#), [167T](#), [397](#)

HIV/AIDS [413](#)

SCID *see* [severe combined immunodeficiency](#).

immunohistochemistry analyses [393](#), [394B](#)

immunogenicity problems

in cell therapy [336T](#), [337](#), [343](#)

in gene therapy [353–4](#), [354F](#)

immunoglobulin(s) (Ig) [99](#), [160T](#)

B-cell receptors vs. soluble antibodies [99](#)

cell-specific production in mature B cells [100–102](#)

constant regions, structure and function [99](#), [99F](#), [101](#)

extreme genetic variation of [99](#), [99F](#)

functions of [99](#)

isotype switching to make different antibody classes [102](#)

protein chain combinatorial diversity [102](#)

structural similarities to HLA proteins, T-cell receptors [99F](#)

immunoglobulin (Ig) genes

allelic and light chain exclusion [102](#)

C (constant region) gene segments [101](#), [101F](#)

class-switching and [102](#)

D (diversity) gene segments [101](#), [101F](#)

junctional diversity in recombination events [102](#)

often involved in oncogene translocations [378](#), [378T](#)

programmed DNA rearrangements in B cells [100–1](#), [101F](#), [102](#)

somatic recombinations in mature B cells [101](#), [101F](#), [102](#)

V (variable) gene segments [101](#), [101F](#)

VDJ exon created by somatic recombination [101F](#)
immunosuppressive drugs [104](#), [359T](#)
immunosurveillance, to kill cancer cells [366](#)
immunotherapy, to treat cancer [409](#), [409T](#)

CAR-T cell therapy [410–11](#), [410F](#)
immune checkpoint therapy [410](#)

imprinting (genomic)

assisted reproduction and [175](#)
cis-regulation by noncoding RNA [161](#), [173F](#)
establishing imprints by differential DNA methylation [163](#)
extent and significance of [160](#)
imprinted genes in mice [161](#)
maternal-paternal conflict theory [161](#)
and nonpenetrance [129](#), [129F](#)
parent-of-origin effects [129](#), [129F](#)
reversal of imprints between generations [161](#), [162F](#)
uniparental disomy and [161](#), [163F](#)

imprinting control region [174F](#)

imprinting marks

erasure of parental [158F](#)
established in early embryo [158F](#)

in vitro fertilization (IVF; assisted reproduction) [420](#), [460](#)

and preimplantation diagnosis [460](#), [460F](#), [461](#)

inborn errors of metabolism

different types of pathogenesis [305ff](#)
newborn screening for [464](#)

phenotype classes [306](#)
very different treatment options [305–9](#)
inbreeding [132](#)
incidental findings, genetic testing [475](#)
inclusions/inclusion bodies (protein aggregation) [227–8](#), [228F](#), [233](#),
[234F](#)
incontinentia pigmenti [118F](#), [119](#)
indel(s) [89](#), [89F](#)

and copy number variants [90](#)
modern definition [90](#)

induced pluripotent stem cells (iPSCs) [151](#), [332–3B](#), [353](#)

dedifferentiation [332–3B](#)
producing iPSCs [332–3B](#)
transdifferentiation [332–3B](#)

infertility [225](#)
inflammation

Alzheimer disease genes with role in [287](#)
cause of in Alzheimer disease [286](#)
cause of in Crohn's disease [262](#)
due to increased immune response [261](#)
due to infection with *H. pylori*, causing gastric cancer [397](#)
induction of tumor-promoting (cancer hallmark) [369T](#), [397T](#)
present in type 2 diabetes, as well as in type 1 [288](#)

inflammatory bowel disease (IBD)

chronic inflammation by activation of Th17 cells [282B](#)
identifying a recessive gene by exome sequencing [443F](#), [444](#)
genes / pathways regulating mucosal immunity [283B](#)

genetic susceptibility factors [282B](#)
importance of IL-23 pathways [282B](#)
pathogenesis [282–3B](#)
see also [Crohn's disease](#); [ulcerative colitis](#)

inheritance patterns

5 types of Mendelian inheritance [112–22](#)
difficult to define in small pedigrees [122](#)
Fisher's infinitesimal model [253](#)

initiation codon (start codon) [26](#), [28](#), [485](#)

innate immune system [262](#)

importance in brain, eye disorders [288](#)

inner cell masses [332B](#), [339B](#)

insulators, form of boundary element [143](#)

at imprinting control region ICR1 [173F](#)

insulin

intra- and inter-molecular disulfide bridges [32F](#)

human recombinant [326](#). [326T](#)

intellectual disability [174](#), [194T](#)

novel genes identified by exome sequencing [248](#), [251F](#)

phenylketonuria [235B](#)

Syndromic Intellectual Disability gene panel [175](#)

see also [mental retardation](#)

interchromosomal recombination (for ribosomal DNA) [50](#)

interleukin-23 [282B](#)

International Cancer Genome Consortium (ICGC) [398](#)

International HapMap Consortium [268](#)

International HapMap Project [269B](#), [270](#)

interphase, definition [11](#)

interphase FISH [428–9](#), [429F](#)

intrabodies (intracellular antibodies) [327F](#), [327–8](#), [328T](#)

different from small molecule drugs in how they block protein-protein interactions [328](#)

nanobodies [328](#)

scFv intrabodies [328](#)

intrachromatid recombination [199](#), [202](#), [203F](#), [204F](#), [206](#)

intracytoplasmic sperm injection (ICSI) [121](#), [175](#), [358F](#)

intravasulation [363F](#)

intron retention [187](#), [187F](#)

intronic splice silencer (ISS) [347B](#)

introns

absent in mtDNA genes [44](#)

absent from some nuclear genes [24](#)

genes located within [174](#)

phase 0, phase 1 and phase 2 introns [26B-27B](#)

splitting coding sequences [26–7B](#)

inversion(s) [87](#), [92F](#), [198–9](#), [202](#), [203F](#), [207](#)

causing heterochromatin spreading [169](#), [169F](#)

causing position effect [151](#), [161](#), [169F](#)

common in *F8* (factor VIII) gene [203](#), [204F](#)

detection of [428](#)

in disease gene identification [248](#)

paracentric versus pericentric [208F](#)

in tumor cells [469F](#)

inverted repeats [202–3](#), [204F](#), [349F](#)
ionizing radiation [79](#), [82](#)
ipilimumab [409T](#), [410](#)
iron-response elements (IREs) [147](#), [148F](#)

IRE-binding proteins [148F](#)

ISCN (International Standing Committee on Human Cytogenetic Nomenclature) [205B](#), [206](#)
isochromosomes [209](#), [209F](#)
isocitrate dehydrogenase [403](#)
isodisomy [171](#)
isoforms [93](#), [93T](#), [98](#), [146](#), [146F](#), [147](#)
Isoleucine

chemical class [185T](#)
structure [25F](#)

isoniazid [317](#), [317F](#), [320T](#)
isotype switching (class-switching), B cells [102](#)
Isovaleric acidemia [465](#)
ivacaftor [324](#), [487](#)
IVF treatment *see* [in vitro fertilization](#)

J

J (joining region) gene segment (in Ig genes) [101](#), [101F](#), [102](#)
Joint Committee on Genomics in Medicine (UK) [474–5](#)
junctional diversity (Ig, T-cell receptor genes) [102](#)
'junk DNA' [43](#)

K

Kabuki syndrome type 1 [251T](#)
KAL gene, pseudoautosomal region [120](#)

Kallmann syndrome [120](#)
karyotyping [204](#), [390](#), [390F](#), [420F](#), [423](#), [425](#), [428](#), [430](#)

see also [chromosome banding](#)

kataegis [389](#), [400](#), [400F](#)
KCNQ1 gene [15](#), [172](#), [173F](#)

variant associated with long QT syndrome

KCNQ1OT1 antisense RNA [159T](#), [172](#), [173F](#)

KDM5C gene [167T](#)

Kearns-Sayre syndrome [215](#)

Kennedy's disease (spinal bulbar muscular atrophy) [194T](#), [224](#)

keratins, in epidermolysis bullosa [222](#)

kinetochores

keratinocytes [314](#)

KLF4 gene [396](#)

Klinefelter syndrome [225](#)

knockout mice [338B](#)

Kozak consensus sequence [26](#)

KRAS gene [395](#)

L

labeling of nucleic acids and oligonucleotides

biotinylation [69B](#)

fluorescence labeling [69B](#)

principles of [68B](#)

lactase persistence in adults, [95](#)

lactose tolerance, selection for [97–8](#)

Langer mesomelic dysplasia [120](#)

late-onset single-gene disorders, variable age at onset [127](#), [127F](#)

latent splice site, *see* [cryptic splice site](#)

LCT (lactase) gene [98](#)

LDLR gene (low-density lipoprotein receptor) [322](#), [453](#), [482](#)

leader sequence (signal peptide) [31](#)

Leber congenital amaurosis, type 2, gene therapy for [345](#)

Leber hereditary optic neuropathy (LHON) [216](#)

homoplasmy in [216](#)

Leigh syndrome [216](#), [216–7B](#)

mitochondrial replacement therapy for, [356F](#)

lentiviruses, in gene therapy [336T](#)

lentivirus vectors, self-inactivating [342](#)

Leri-Weill dyschondriostosis [120](#)

Lesch-Nyhan syndrome, [232F](#)

leucine

chemical class [185T](#)

structure [25F](#)

leucine zipper [144](#)

DNA binding domains [144F](#)

monomer [144F](#)

leukemias

6-mercaptopurine treatment [318](#)

acute myeloid leukemia (AML) [378T](#), [379F](#), [409T](#)

cancer stem cell evidence [374–5B](#)

chronic myelogenous leukemia (CML) [375B](#), [378T](#), [408](#), [411](#),
[428](#), [429F](#)

CLL (chronic lymphocytic leukemia) [388](#), [400](#), [405F](#), [409T](#), [468T](#)

CML (chronic myelogenous leukemia) [375B](#), [378T](#), [408](#), [411](#), [428](#), [429F](#)

resulting from retroviral gene therapy [342](#)

Lewy bodies [232F](#)

Li-Fraumeni syndrome [385T](#), [387B](#), [453](#)

liability threshold, to explain dichotomous traits [253–4B](#)

light chain exclusion [102](#)

limb girdle muscular dystrophy [188F](#)

LIN28 gene [396](#)

LINES (long interspersed nuclear elements)

LINE-1 family [52](#), [52T](#), [53](#), [53F](#)

linkage analysis [240–7](#)

with affected sib-pairs in complex diseases

autozygosity mapping in recessive diseases [247](#), [284](#)

defining minimum candidate region [247F](#)

identifying recombinants and non-recombinants [245](#), [245F](#)

informative and uninformative meioses [245](#), [245F](#), [246B](#), [247F](#)

likelihood ratios and lod scores [246B](#)

limited success for complex genetic disease [262–3](#)

nonparametric [260–1](#)

obtaining statistical evidence [245](#), [246B](#), [247](#)

parametric [259–60](#)

principles of genetic linkage [242](#), [243F](#), [244](#)

standard genomewide [244–7](#)

linkage disequilibrium [263](#)

as an explanation for allelic association [266–7](#)

explained by shared ancestral chromosome segments [267](#)

mapping genes with [261](#)

liposomes [335](#)

liquid biopsies [469](#)

liver

 cirrhosis due to inclusion bodies [228](#), [228F](#)

 drug metabolism in [312](#)

 gene therapy target via hepatic portal vein [343](#)

liver transplantation, treating some inborn errors of metabolism [306](#)

LMNA gene (lamin A)

 extreme phenotype heterogeneity [126B](#)

locus, definition [77](#), [110](#)

locus heterogeneity [125](#), [125F](#), [126F](#), [247](#)

Locus Reference Genomic (LRG) database [444B](#)

lod scores [245](#), [246B](#)

long noncoding RNA (lncRNAs)

 chromatin-modifying [159](#)

cis-acting regulators [159](#), [160F](#)

 different classes [159T](#)

 in cancer [389](#)

trans-acting regulators [159](#), [160F](#)

long QT syndrome [481](#)

long terminal repeats (LTR) [52T](#)

loss-of-function (LOF) mutations/variants [93](#), [218–20](#), [221B](#), [231](#), [241](#)

 average number inherited by a person [189](#)

 definition [218](#)

 inherited in familial cancers [381](#)

 and gain-of-function in one gene [223–4](#)

 making gene knockouts by [338B](#)

loss of heterozygosity [380](#), [382](#), [401](#)
low-copy-number repeats [199](#), [203F](#), [204F](#)
LPA gene encoding lipoprotein Lp(a) [47F](#), [277T](#)
LRG (Locus Reference Genomic) database [444B](#)
LRRK2 gene [284T](#), [285](#)
luciferase color reaction [439F](#)
luciferin [439F](#)
lung cancer

mutational signatures [400](#)

lupus *see* [SLE](#)

Lynch syndrome [85B](#), [385T](#), [393](#), [454](#), [470](#)

gene panel [442B](#)

lysine

acetylation [153](#), [153F](#)
chemical class [185T](#)
methylation [153](#), [153F](#)
structure [25F](#)

M

M (mitosis) phase, of cell cycle [11–12](#), [12F](#), [13](#)

‘mad cow disease’ (vCJD) [228](#)

macrosatellite repeat D424 [169–70](#), [170F](#)

Mainstreaming Cancer Genetics Programme

Mainstream Genomic Medicine Service [450](#), [452F](#)

Major Histocompatibility Complex *see* [MHC](#)

malaria [95](#), [95T](#), [233](#)

heterozygote advantage [97](#), [136](#)

MALDI-TOF mass spectrometry

Agena MassArray [440](#)

multiplex genotyping using [439–40](#)

male breast cancer [455T](#)

male-specific region, of Y chromosome [119](#), [119F](#), [120](#)

mammalian genomes

maternal insufficiency [161](#)

paternal insufficiency [161](#)

manifesting heterozygotes [117](#), [136](#)

Maple syrup urine disease [450](#)

MAPT (microtubule-associated protein tau) gene [230B](#)

Marfan syndrome [222](#), [240](#), [324](#)

Mary Lyon [164](#)

mass spectrometry, *see* [MALDI-TOF mass spectrometry](#).

massively parallel DNA sequencing (Next Generation Sequencing)
[74–5](#), [75T](#), [433](#), [435B](#)

and cancer classification

different to dideoxy sequencing [74–5](#)

identifying rare variants using [88B](#), [281](#)

Illumina workflow [435B](#)

population-based *see* [human population genomics](#)

sequencing-by-synthesis [435B](#)

and structural variation [433](#)

see also [whole genome sequencing](#); [whole exome sequencing](#)

Mastermind Professional database [447T](#)

maternal age and Down syndrome [212](#)

maternal circulation, fetal DNA in [461](#)

mating, nonrandom [132–3](#)

matrilineal inheritance [121](#), [121F](#)
MBNL1 (muscleblind protein) [198](#)
MDM2 regulator [383F](#), [384](#), [387B](#), [388F](#)
MDM4 regulator [387B](#), [388F](#)
MECP2 gene, and Rett syndrome [167](#), [168B](#)
MECP2 protein, function of [167](#)
MECP2 gene [167T](#), [168B](#)
medium-chain acyl-CoA dehydrogenase deficiency (MCAD) [465](#)
medullary thyroid carcinoma [224](#)
megakaryocytes, polyploid [11](#)
meiosis [13–14](#), [15F](#)

- asymmetric cell divisions in females [14](#)
- average number of cross overs in male and female [16](#)
- bivalents [15F](#), [16–17](#)
- distinguished from mitosis [13–14](#)
- epigenetic effects in plants transmitted through [150](#)
- independent assortment [17](#), [17F](#)
- informative and uninformative [245](#), [245F](#), [246B](#), [247F](#)
- meiosis I and II [14](#), [15F](#), [16–17](#)
- nondisjunction [210–11](#), [211F](#), [212](#)
- oogonia and oocytes [14](#)
- pairing of paternal and maternal homologs, [15F](#) [16](#)
- polar body [14](#)
- spermatogonia [14](#)
- spermatocytes [14](#)
- synapsis [14](#)
- X-Y pairing [119F](#), [120F](#)
- zygote [14](#)

meiotic crossovers, mapping in humans [243F](#)
meiotic recombination frequencies

differences in individuals [244F](#)

sex differences [244](#), [244F](#)

melanoma(s)

BRAF oncogene mutations frequent [401](#)

C → T transitions (UV light) [400](#)

familial [385T](#)

high mutation prevalence [398–9](#)

TERT promoter mutations frequent [403](#)

MELAS (mitochondrial encephalopathy, lactic acidosis, stroke-like episodes) [479](#)

Mendelian vs. monogenic characters/traits [110](#)

Mendelian disorders

abnormal epigenetic regulation [165ff.](#)

genotype-phenotype correlations [231ff.](#)

Mendelian inheritance, five patterns of [112](#)

Mendelian subsets of complex genetic disease [260](#), [285](#)

Alzheimer disease [284](#), [284T](#), [285](#)

Parkinson disease [284](#), [284T](#), [285](#)

mental retardation, *see* [intellectual disability](#)

messenger RNA *see* [mRNA](#)

metabolic block(s) [234–5B](#), [309F](#)

metabolic factors, inducing primary epimutations [166](#), [166F](#)

metabolism

changes in cancer cells [369T](#), [403](#), [404F](#)

drug metabolism, influenced by genetic variation [313ff.](#)

see also [inborn errors of metabolism](#)

metaphase chromosomes [9](#), [14F](#), [15F](#)

human ideogram [36F](#)
metaphase FISH [377F](#)
metaphase plate [15F](#)
metastasis [362](#)

angiogenesis not always necessary for [369](#)
definition of [363](#)
intravasation and extravasation [363F](#)
metastatic spread, single-cell analyses [405F](#)
re-differentiation of metastases [368](#)
seeding secondary tumors [363F](#)

methionine

chemical class [185T](#)
initiator amino acid during translation [24](#), [26](#), [26B](#), [27](#)
in N-terminal cleavage [31](#)
S-adenosyl- (SAM)
structure [25F](#)

methylation

of DNA, see [DNA methylation](#)
of histones, see [histone modifications](#)
of proteins, [30T](#)

methylation-sensitive MLPA [449](#), [450F](#)

methylome, screening of [295](#)

MGMT gene [396](#)

MHC (Major Histocompatibility Complex) [99–100](#), [102–4](#)

classical MHC genes [103](#)
class I and class II proteins: different functions [103](#)
see also [HLA complex \(human MHC\)](#)

MHC-peptide binding [265B](#)

from endogenous proteins: class I MHC [103](#)

from exogenous proteins: class II MHC [103](#)

MHC polymorphism [104](#)

MHC restriction [265B](#), [410–11](#)

microarray-based hybridization

chromosome SNP micorarrays [426](#)

and copy number analysis [425](#)

a feature in a microarray [70](#), [71F](#)

in GWA studies, *see* [GWA studies](#)

genotyping variants with “SNP-chips” [471](#)

methylation scans [295](#)

overview of [69–70](#), [71F](#)

see also [oligonucleotide microarrays](#)

microbial pathogens, natural selection and [79](#)

microbiome [291–2](#)

microbiota (gut flora) [262](#), [282B](#), [290T](#), [291](#)

microcephaly [85–6B](#), [167T](#), [174](#)

microglia [288](#)

origin of [341F](#)

β 2-microglobulin [99F](#), [105B](#); *see also* [HLA](#)

microRNAs *see* miRNAs [35](#)

microsatellites/microsatellite DNA [90](#)

microsatellite instability [392–3](#), [394B](#), [399F](#), [421T](#), [454B](#)

microsatellite markers/polymorphisms [91F](#), [96B](#), [242](#), [425F](#)

microtubules [14](#)

attached to kinetochore [10](#), [10F](#)

migration

founder effects [133](#)

out-of-Africa [95](#)

Miller syndrome [251T](#)

minimal residual disease [471](#)

minisatellite DNA [90](#)

minor nucleotides, in RNA [29F](#)

minority bases [32](#)

MIR15A and *MIR16-1* genes [388](#)

MIR96, *MIR184* and *MIR204* genes [191T](#)

miRNAs (microRNAs) [35](#)

in cancer [398-9](#)

in gene regulation [149F](#)

multigenic regulation [148](#), [149F](#)

negative regulation by competing endogenous RNAs [148-9](#),
[149F](#)

production in cells [148](#), [149F](#)

roles in cancer [388](#)

seed sequence [148](#)

trans-acting regulators at RNA level, [140](#), [141F](#), [148](#), [148F](#)

miRNA sponges (competing endogenous RNAs) [149F](#)

misattributed maternity [483B](#)

mismatch repair system [80](#)

basic mechanism of [392](#), [393F](#)

consequences when defective [392](#), [394](#), [394B](#), [421T](#)

defective in Lynch syndrome [391](#), [394-5B](#)

hMutS α [392](#), [393F](#)

hMutS β [392](#), [393F](#)

hMutL α [392](#), [393F](#)

missense mutations [183](#), [183T](#), [185](#)

average number inherited by a person [189](#)

conservative substitutions [184](#)

common in oncogenes and narrowly distributed [379](#), [380F](#)

evaluating pathogenicity of [443F](#), [444](#), [445](#)

dominant negative effects [222](#), [223F](#)

harmful in one human genome [189](#)

nonconservative substitutions [184](#), [444](#)

p53 mutants [386](#), [386–8B](#)

Pittsburgh variant [219–220](#)

selfish spermatogonial selection and [130](#), [190](#), [190T](#)

mitochondrial DNA (mtDNA)

7S DNA [45F](#)

and common disease [293B](#)

circular nature, [45F](#)

clonal expansion [213](#), [370](#)

common diseases and [293B](#)

copy number variation [11](#), [121](#)

CR/D control region displacement loop [45F](#)

deletion hotspots [216T](#)

frequent large de novo deletions in mtDNA [215](#)

evolution explained by endosymbiont theory [42](#), [212–3](#)

gene-rich genome [44](#), [45F](#)

genetic code, different for mtDNA [184F](#)

haplogroup evolution, [292–3B](#)

heteroplasmy/heteroplasmic [13F](#), [78](#), [121](#), [213–4](#), [214F](#), [215](#),
[232](#), [487](#)

thresholds for disease [216](#)

variable causing clinical variability [121–2](#)

homoplasmy [293B](#)
human mitochondrial genome and gene map [44](#), [45F](#)
lack of introns [44](#)
L and H strands [45F](#)
maternal transmission of [121](#)
mutation rate, elevated [121](#)
mitochondrial pseudogenes present in nuclear DNA, *see*
 HUMT sequences
multigenic transcripts [44](#), [45F](#)
rapid evolution of variants [122](#)
repeats in mtDNA predisposing to large deletions [182T](#)
replication and segregation of [213–4](#)
sequencing of human [35](#)
size of human [35](#)
stochastic segregation into daughter cells [13](#), [13F](#)
unequal replication [13](#)
mitochondrial disorders [479](#), [487](#)

 arising from point mutations [216](#)
 arising from deletions [215](#)
 clinical variability [213](#)
 common biochemical phenotype [215](#)
 due to pathogenic variants in mitochondrial DNA [213–7](#)
 due to pathogenic variants in nuclear DNA [213](#)
 deletion disorders due to large mtDNA deletions [215](#)
 heteroplasmy causing clinical variability [121](#), [214](#), [214F](#), [216](#),
 [217B](#), [232](#)
 homoplasmy in some disorders [216](#), [232](#)
 incomplete penetrance [121F](#)
 matrilineal inheritance [121](#), [121F](#)
 mitochondrial DNA variants in common disease [293B](#)
 prevention by mitochondrial donation therapy [485](#), [487](#)

mitochondrial genetic bottleneck [122](#), [214](#), [214F](#), [487](#)
mitochondrial replacement therapy

to treat severe mitochondrial disorders [355](#), [356F](#)
a form of germline gene therapy [355](#), [356F](#)

mitochondrial segregation, stochastic nature of [13](#)
MITOMAP database [45](#), [192T](#), [216](#), [292B](#)

mitomycin C, inducing interstand crosslinks in DNA [86B](#), [421T](#)
mitosis/mitotic division

stages of [13](#), [14F](#)
total number of divisions in human lifetime [124B](#)

mixoploidy [206T](#), [212](#)

MLPA (multiplex ligation-dependent probe amplification) [394B](#), [423T](#),
[430–1](#), [431F](#), [432](#), [432F](#)

MS (methylation-sensitive)-MLPA [448](#)

modifier genes / loci [232–3](#), [253](#)

the example of β -thalassemia [232–3](#), [234F](#)

molecular pathology

protein structure abnormalities [225ff.](#)
genotype-phenotype correlations [231ff.](#)

monoallelic expression, natural

according to parent of origin [160](#), [160T](#)
independent of parent origin [160](#), [160T](#)

monoclonal antibodies (mAbs)

different types, by genetic engineering [326–7](#), [327F](#)
licensed, examples of [327](#), [328T](#)
targeted cancer therapies using [409T](#), [409–10](#)

monogenic disorders

abnormal epigenetic regulation [165ff.](#)
genotype-phenotype correlations [231ff.](#)
variable expression in [121–2](#), [128](#), [128F](#)
see also [Mendelian disorders](#); [mitochondrial disorders](#)

monooxygenases

in phase I drug metabolism [313](#), [313F](#)

monosomy/monosomies [163](#), [206T](#), [210T](#), [211](#), [224–5](#)

monosomy rescue [171](#)

viable in the case of the X chromosome [224–5](#)

monosomy, lethal except for [45X](#) [163](#)

monozygotic twins [17](#)

epigenetic changes in [295–6](#)

mosaics, why all of us are [124B](#)

mosaicism [100](#), [116](#), [123](#), [124B](#)

chromosome abnormalities and [212](#)

copy number variation mosaicism in neurons [100](#)

diploid/triploid [212](#)

germline [123](#), [124B](#)

in female mammals due to X-inactivation [164](#)

myxoploidy, aneuploidy and [206T](#)

post-zygotic variation and [100](#)

X-chromosome inactivation and [116](#)

mouse models of disease [337](#), [339–40](#)

construction of transgenic models [338B](#)

construction of gene knockouts [339B](#)

glioblastoma [411](#)

Hirschsprung disease [240](#)

mdx muscular dystrophy [337](#)

Parkinson disease [353](#)

MRN (MRE11-RAD50-NIBRIN) complex, [391F](#)

mRNA (messenger RNA = coding RNA)

poly(A) tails at 3' end [28F](#), [29](#)

post-translational capping at 5' end [28](#)

translation process [26–9](#), [28F](#)

translational reading frames [26](#)

translation start and stop sites [26](#), [28](#)

see also [5'](#), [3' untranslated regions](#)

mRNA surveillance [186–7B](#)

MSI-positive (or MIN-positive) colorectal cancer [393–4](#)

mtDNA, *see* [mitochondrial DNA](#)

MT-ND1, *-ND4*, *-ND5*, *-ND6* genes [217B](#)

MT-RNR1 gene [191T](#)

MTOR gene (mammalian target of rapamycin) [401](#)

mTOR signaling [324](#), [362](#)

mTOR protein [324](#)

mTORC1 growth signaling [324](#)

multigene families [47](#), [47T](#), [48](#), [50](#)

multifactorial diseases (complex diseases) [252](#)

multiple endocrine neoplasia types 2A or 2B [224](#)

multiple sclerosis

% concordance in MZ and DZ twins [257T](#)

HLA association [266B](#)
multiplex testing [469](#)
muscleblind regulatory proteins [198](#)
muscle fibre cells, polyploidy [11](#)
muscular dystrophies

congenital muscular [126T](#)
Emery-Dreifuss muscular, types 2 and 3 [126T](#)
mdx mouse model [337](#)
see also [Becker](#); [Duchenne](#); [facioscapulohumeral](#)

Mutalyzer computer program [444B](#)
mutation load, *see* [pathogenic mutation load](#)
mutation rates, germline [189](#)

effect of parental age/sex

mutation(s)

advantages of [79](#)
databases [92T](#), [192](#), [192T](#)
de novo [123](#), [189](#), [444–5](#), [446T](#)
number inherited from each parent [189](#)
driver mutations, *see under* [cancer evolution](#)
dual meaning [79](#)
due to DNA replication errors [80](#)
dynamic mutations, causing disease [194–8](#)
C → T substitution very frequent in vertebrates [84](#), [85F](#)
hotspots [180](#)
human mutation rates [189](#)
hypomorphic [221B](#)
hypermutation, *see* [somatic hypermutation](#)
interpretation [443ff.](#)
missense *see* [missense mutations](#)

new mutations [123](#)
nonsense *see* [nonsense mutations](#)
nonsynonymous classes [183T](#)
number in different cancers [398–9](#); [399F](#)
origins of [79ff.](#)
paternal transmission bias [190](#), [190T](#)
post-zygotic [123](#), [124F](#)
purifying selection (against harmful mutations) [44](#), [132](#)
selfish mutations (spermatogonial selection) [130](#), [190](#), [190T](#)
splicing, pathogenic [187](#), [187F](#), [188F](#)
stop-gain [183T](#)
stop-loss [183T](#)
synonymous (silent) [183](#)
testing, *see under* [genetic testing](#)
see also [pathogenic mutations](#); [heteroplasmy](#).

MUTYH gene [392](#)

MYC gene [378T](#), [379F](#), [383F](#)

MYCN (onco)gene [428](#)

amplification in neuroblastoma cells [377](#), [377F](#)

myelodysplastic syndrome (MDS) [389](#), [400](#)

myeloproliferative disease [468](#)

myotonic dystrophy [198](#)

pathogenesis [198](#)

type 1, [194T](#), [195–6](#), [196–7B](#), [197T](#)

type 2 [194T](#), [195](#), [197T](#)

N

N-terminal ends (polypeptides) [27](#)

function of [27](#)

N-terminal tails (histones) [153F](#)
NANOG gene [396](#)
narcolepsy, HLA association [266B](#)
NAT1, NAT2, N-acetyl transfers, [317](#), [317T](#)
National Genome Test Directory (UK) [469](#)
natural killer (NK) cells [354](#)

in immunosurveillance against cancer [366](#)

natural selection

after gene/exon duplication [50](#)
cancer versus whole organism [365–6](#)
causing gene amplification [97](#), [98F](#)
causing gene upregulation [97–8](#)
causing high genetic variation in drug metabolism genes [312](#),
[317](#)
invading pathogens and [79](#)
see also [purifying selection](#); [positive selection](#); [balancing selection](#); [overdominant selection](#)

NCBI (US National Center for Biotechnology Information) [40](#)

negative selection, *see* [purifying selection](#)

neonatal diabetes, transient [173T](#)

neoplasms, *see* [tumor types](#)

NER, *see* [nucleotide excision repair](#)

neural tube defect [255](#)

neurodegenerative disorders

amyotrophic lateral sclerosis [230B](#)
due to unstable expansion of short tandem repeats [194T](#)
FXTAS (fragile X tremor-ataxia syndrome) [198](#)
potential of intrabodies [328](#)
predictive testing (Huntington disease) [456](#)

prion and prion-like diseases [229–30B](#)
Tay-Sachs disease [463](#), [467](#)
vulnerability of neurons [198](#)
see also [Alzheimer](#); [Huntington and Parkinson disease](#)

neurofibrillary tangles [286](#), [287B](#)
neurons, mosaic CNV patterns [100](#)
neurofibromatosis types 1 and 2 [385T](#)
neutrophils [219](#), [228](#), [373T](#)
newborn genome sequencing [484–5](#)
newborn screening [463–6](#)
Next-Generation Sequencing (NGS) *see* [massively_parallel_DNA_sequencing](#)

NF1 (neurofibromatosis type 1) gene, [47T](#), [385T](#), [387B](#), [407F](#)

and dispersed pseudogenes [48B](#)

NIPT, *see* [noninvasive_prenatal_testing](#)
nitisinone [307F](#)
nivulomab [409T](#), [410](#)
NOD2 gene [261–2](#), [262F](#)
NOD2 protein, in innate immune system [262](#)
nomenclature

DNA, RNA and protein sequence variants [444](#), [444–5B](#)
histone modifications [154T](#)
HLA alleles [105B](#)
human chromosomes/chromosome banding [204–5B](#)
human chromosome abnormalities [206T](#)
human gene symbols [40](#)
human pedigree symbols [112F](#)

non-integrating viral vectors [337](#), [343](#)
nonallelic homologous recombination (NAHR) [199](#), [202](#), [203F](#)
noncoding DNA

highly repetitive (satellite DNA) [51](#), [90](#)
minisatellites [90](#)
noncoding RNA (ncRNA)
defective in single gene disorders [191](#), [191T](#)
genes specifying, *see* [RNA genes](#)
versatility of [33](#), [34F](#), [34–5](#)
see also [long noncoding RNA](#), [miRNA](#); [piRNA](#); [ribosomal RNA](#); [transfer RNA](#)

nonconservative substitutions [184–5](#), [444](#)
nondisjunction (NDJ) [211–2](#)
nonhistone proteins [154](#)
nonhomologous end joining (NHEJ) [83](#), [85B](#) [351](#)
noninvasive cancer testing (“liquid biopsies”) [412–3](#)
noninvasive fetal aneuploidy screening
noninvasive prenatal testing (NIPT) [461–2](#)
nonparametric linkage analysis [260–1](#)
nonpenetrance, single-gene disorders, [127–8](#), [132](#), [135](#), [455](#)

due to imprinting [127F](#)

nonrandom mating [131–2](#)
nonsense-mediated decay (NMD) [186–7B](#)
nonsense mutations [181F](#), [183T](#), [185](#)
nonsynonymous substitutions/mutations [183](#), [185](#), [399](#)

classes of [183T](#)

Noonan syndrome [190T](#)
norovirus [290](#)
Norwegian population, HLA disease associations in [266B](#)
NOTCH1 gene

oncogene in lymphomas and leukemias [379](#)
tumor suppressor in squamous cell carcinomas [379](#)

NRG1 (neuregulin) gene [261](#)

NRXN1 gene [278](#)

nuchal translucency [464](#)

nucleic acid hybridization

annealing (hybridization) and denaturation [62](#), [65](#), [65F](#)

hybridization stringency [67](#), [68F](#)

principles of [65–70](#)

two fundamental types of assay [68–9](#), [69F](#)

nucleic acid hybridization, assays

chromosome in situ FISH [70T](#), [377F](#)

Southern blot [171B](#), [423T](#), [430](#)

tissue in situ [70T](#)

see also [allele-specific oligonucleotide \(ASO\) hybridization](#);
[microarray-based hybridization](#)

nucleic acid labeling, *see* [labeling of nucleic acids and oligonucleotides](#)

nucleic acids

5' and 3' ends [3B](#)

see also [DNA](#); [RNAs](#)

nucleolar RNA polymerase (RNA polymerase I)

nucleosomes [152F](#)

in chromosome organization [9](#), [9F](#)

histone modification and variants in [152–3](#), [153F](#) [154](#), [154T](#)

looped domain [9F](#)

N-terminal histone tails [152](#), [153F](#)

structure of [152](#), [153F](#)
nucleosome repositioning [150](#)

epigenetic mechanism [150T](#)
nucleosomes [9](#), [9F](#)
nucleosomal filament [9](#), [9F](#)

nucleotide excision repair [83–4](#), [85T](#)
nucleotides [2](#)

dideoxy analogs (ddNTPs) [72](#), [72F](#), [440](#), [440F](#)
minor nucleotides in RNA [29F](#)

nucleotide substitutions

conservative and non-conservative [184](#)
see also [synonymous substitutions/mutations](#)
see also [nonsynonymous substitutions/mutations](#)

null alleles [218](#)
nulliploid cells [11](#)
nullisomy [172F](#), [210T](#), [211F](#)
NUMT (**n**uclear-**m**itochondrial sequences) [212](#)
nutrition [296](#), [470](#)

O

obesity [125](#), [174](#), [291](#), [296](#), [326T](#)
OCT4 gene [396](#)
oculopharyngeal muscular dystrophy [194T](#)
odds ratios [266B](#)

a worked example in case-control studies [264T](#)

Okazaki fragments [6](#), [6F](#), [7](#), [7T](#)

olfactory neurons [160T](#)

olfactory receptor genes/proteins [98](#), [160T](#)

high-frequency of deleterious variants [98F](#)

importance of gene duplication [98](#)

largest human gene family [98](#)

oligonucleotide ligation assay [437](#), [437F](#)

oligonucleotides, allele-specific, *see* [allele-specific oligonucleotides \(ASO\)](#)

oligonucleotide therapeutics [344–5](#), [344F](#), [346–7B](#), [349T](#)

olaparib [409T](#)

OMIM (Online Mendelian Inheritance in Man) database [111B](#)

oncogenes

detecting amplification of [377F](#)

dominant acting [375](#)

gain-of-function mutations [378–9](#), [380F](#)

nature of [375](#)

origin from proto-oncogenes [375](#)

translocation-induced activation of [377–8](#), [378T](#), [379F](#)

viral and cellular oncogenes [376](#)

see also [proto-oncogenes](#)

oncogene activation

mechanisms of [367–369](#)

oncotype DX [468](#)

oocyte development

oogonia [14](#), [164F](#)

open reading frames (ORF) [26](#), [28F](#)

oculopharyngeal muscular dystrophy

optical genome mapping [420T](#), [433](#), [434F](#)

ornithine transcarbamylase deficiency, treatment [309F](#)
orthologs [37](#), [43](#)
osteogenesis imperfecta (brittle bone disease) [222](#), [223F](#)
osteogenesis imperfecta type VI [251T](#)
ovarian cancer [454B](#)
overdominant selection, promoting MHC polymorphism [104](#)
oxidative damage [81](#), [86B](#), [292B](#), [367B](#)
oxidative phosphorylation system (OXPHOS) [44](#)
2-oxoglutarate cofactor [403](#), [404F](#)
8-oxoguanine [81F](#)

P

P14 and p16 isoforms from

CDKN2A gene [146F](#), [147](#)

p53 tumor suppressor protein

activating apoptosis [384](#), [385F](#)
evolutionary origins of encoding gene [43](#)
guardian of the genome [384](#), [386B](#)
human mouse comparison [38F](#)
as nonclassical tumor suppressor [386–8B](#)
missense mutants [388B](#)
regulator of cell growth [383](#), [383F](#)

PAH (phenylalanine hydroxylase) gene [235B](#)

PALB2 protein [391F](#), [454](#)

Pan-Cancer Analyses of Whole Genomes (PCAWG) Consortium [398](#),
[403](#)

palindrome nature, of many restriction nuclease target sequences [61B](#)

PAPP-A (pregnancy-associated plasma protein A) [464](#)

paralogs [346B](#)

parametric linkage analyses [259](#)

see also [nonparametric linkage analyses](#)

Parkinson disease

as amyloid disease [230B](#)

cytoplasmic aggregation of α -synuclein [230B](#)

% concordance in MZ and DZ twins [257T](#)

genes involved in Mendelian subsets [284](#), [284T](#)

LRRK2 variants in common and Mendelian subsets [284T](#), [285](#)

PARP (poly[ADP-ribose] polymerase) [452](#)

parthenogenesis [161](#)

Patau syndrome [211](#)

paternal-age-effect disorders [190](#), [190T](#)

paternal-maternal conflict theory

paternity, misattributed [474](#), [483](#)

pathogenesis, at protein structure level [226–8](#), [229–30B](#)

pathogenic load, total per person

average number of damaging DNA variants [189](#)

average number of loss-of-function mutations [189](#)

average number of missense mutations [189](#)

pathogenic mutations

affecting multiple genes simultaneously [218](#)

causing gain of function [219–20](#), [219F](#)

causing loss of function [218–9](#), [220–1](#), [221B](#), [222–4](#)

curating and databases of [192](#), [192T](#)

different classes altering *amount* of product [181–2](#), [181F](#)

dominant negative effects [222](#), [223F](#)

dynamic mutations [194–8](#)

effect due to interaction with alleles [218](#)

effect due to interaction with modifier genes [218](#)

effect due to interaction with epigenetic/environmental factors [218](#)
evaluating candidate pathogenic mutations [443](#), [443F](#), [444–6](#), [446F](#)
nonconservative substitutions [184–5](#), [444](#)
nonsynonymous, classes of [183](#), [183T](#)
pathogenic load across a human genome [189](#)
in RNA genes [191](#), [191T](#)
synonymous substitutions occasionally pathogenic [187](#), [188F](#)
triggered by repetitive DNA [182](#), [182T](#)
two fundamental classes of [180–81](#), [181F](#), [182](#)
see also [frameshifting mutations](#); [missense mutations](#);
[nonsense mutations](#); [splicing mutations](#)

pathogens, natural selection to counter [79](#)
PAX6 gene, regulation of [143F](#)
PCR (polymerase chain reaction) [58T](#) [62–3](#), [64F](#)

- allele-specific PCR [45](#)
- basics [62–3](#)
- methylation-specific PCR [448](#), [449F](#)
- reaction mechanism [63F](#)
- phases of reaction [64F](#)
- primers for [62](#), [64F](#)
- quantitative PCR [63](#)
- real-time PCR [63](#)

using TaqMan genotyping [437](#), [438F](#)

- reverse transcriptase (RT-PCR) [63](#)
- triplet repeat-primed PCR (TP-PCR) [195](#), [196–7B](#)
- see also* [digital PCR](#); [droplet digital PCR](#)

PD1 receptor [410](#)

pediatric tumors *see* [childhood cancers](#)

pedigree

definition [111](#)

types of inheritance [112–122](#)

early-onset Alzheimer disease [259F](#)

familial cancer (Li-Fraumeni syndrome) [387B](#)

matrilineal inheritance [121](#), [121F](#)

recording of [111–2](#)

symbols used [112F](#)

PEG [poly(ethylene glycol)] [325](#), [349F](#)

PEGylation [325](#)

penetrance

in complex diseases [255](#)

in single-gene disorders [126–7](#)

see also [nonpenetrance](#)

pentose phosphate pathway (PPP) [366](#)

peptide bonds [24](#), [25F](#), [28F](#), [29F](#), [31B](#)

formation [25F](#)

formation [24](#), [24F](#)

peptidyltransferases [29F](#)

personal genome sequencing [87](#), [88B](#), [92](#)

personal genomics testing

personalized medicine [471](#)

pharmacodynamics [311F](#), [312](#), [318](#)

pharmacogenetics [310](#)

pharmacogenomics [312](#)

pharmacokinetics [311F](#), [312](#), [318](#)

Phase I drug metabolism [312–3](#), [313F](#), [315–6](#)

Phase II drug metabolism [312–3](#), [313F](#), [317](#), [317F](#), [317T](#), [318](#)

phenocopies [256](#)

phenotype(s) [78](#)

broad and narrow usages [109](#)

causes of variation [110](#)

classification of, and phenocopies [256](#)

correlations with genotype, see [genotype-phenotype correlations](#)

different due to gain and loss of function in one gene [223–4](#)

disease phenotype concept [109](#)

dominant and recessive [110–11](#)

effects of environmental factors [233–4](#)

effects of modifier genes [232–3](#), [234F](#)

general effects of genetic variation [93–4](#)

variable due to anticipation [129](#), [129F](#), [194](#)

variable due to nonpenetrance [127](#), [127F](#)

variable due to variable heteroplasmy [121–2](#)

phenylalanine

chemical class. [185T](#)

structure [25F](#)

phenylketonuria [226](#), [234](#), [235–6B](#), [258](#), [306](#), [308T](#), [464](#), [465T](#)

environmental factors and [234–5B](#)

embryofetopathy [234–5B](#)

as multifactorial condition [234–5B](#)

variable heritability [258–9](#)

Philadelphia chromosome (CML-associated) [374–5B](#), [378](#), [408](#), [409T](#),
[411](#)

phlebotomy [308](#), [308T](#)

phosphodiester bonds

5'–5' bonds connect neighboring nucleotides in a strand [3B](#)

5'–5' bonds in cap of mRNA [28](#)

state and protein function

phosphatidyl inositol-3-kinase [147](#)

phosphorothioate bonds [344F](#)

phosphorylation, of histone tails [153F](#)

phosphorylation of proteins [30T](#)

PIK3CA gene, [380F](#), [402](#), [402F](#)

Pittsburgh variant, of α 1-antitrypsin (α 1-AT) [219–20](#)

Piwi protein-interacting RNAs (piRNAs) [34F](#), [35](#)

placenta [158](#), [160T](#), [161](#), [235B](#), [457F](#), [461](#)

plasma, circulating DNA in [461](#), [469](#)

plasmids, as cloning vectors [59](#)

pLoF computer program [447T](#)

ploidy

definition [10](#)

variability [11](#)

diploid cells [10](#)

pluripotent stem cells [331–3B](#), [354F](#)

induced pluripotent stem cells (iPS) [332–3B](#), [353](#)

PMID (PubMed Identifier), *see* Glossary

PMP22 gene [207](#)

polar body [14](#)

poly (A) polymerase [28](#)

poly (A) tail, of mRNAs [28F](#), [29](#)

polyalanine expansion, pathogenic [193](#), [194T](#)

Prader-Willi syndrome (PWS) [173T](#), [174](#), [174F](#)

polycomb group proteins [155F](#), [165](#)

polycomb repressive complex-1 (PRC1) [159](#)

polycomb repressive complex-2 (PRC2) [159](#), [160F](#), [396](#)

polygenic disorders [252](#)

 differences from monogenic disease [253](#)

polygenic risk scores [279–80](#), [280F](#), [470](#)

polygenic theory, liability threshold [253–4B](#)

polyglutamine repeats, expansion of [193–4](#), [194T](#), [195](#)

polymerase chain reaction *see* [PCR](#)

polymorphism(s)

 compared to variants [87](#)

 copy number polymorphisms (CNP) [277](#), [277T](#)

 different meanings [87](#)

 microsatellite polymorphism [91F](#)

 protein polymorphism and sequence variation [93](#), [93T](#)

 SNPs (single nucleotide polymorphisms) [89](#), [242](#)

polypeptides

 N- and C-terminal ends [27](#)

 N-terminal methionine, sometimes cleaved [28F](#)

 post-translational modifications

 chemical modifications [29–30](#), [30T](#)

 cleavage [29](#)

 structure of [25F](#)

 synthesis of [24](#), [25F](#), [28F](#)

PolyPhen-2 program [443F](#), [447T](#)

polyploidy

 origin of natural [11](#)

population bottlenecks [133](#), [134F](#)

population genomics, *see* [human population genomics](#)

population stratification [263](#)

populations, human

- allele frequency changes [132–3](#)

- Askenazi Jewish [113](#)

- Finnish [113](#)

- human population, variable meaning [130](#)

position effects [151](#), [151T](#), [168](#)

positional cloning [240–1](#)

positive selection [94](#)

- adaptation to new environments [94](#), [94T](#), [95](#)

- and human evolution [94](#)

- in response to microbial pathogens [94–5](#)

- and selective sweeps [96](#). [96–7B](#)

- see also* [heterozygote advantage](#)

positively harmful metabolites [306](#)

post-zygotic genetic variation

- extensive copy number variation in olfactory neurons [98](#), [98F](#)

- not identical to somatic variation [78](#), [78T](#)

- see also* [immunoglobulin genes](#), [T cell receptor genes](#)

Potocki-Lupski syndrome [203T](#)

Prader-Willi syndrome (PWS) [173T](#), [174](#), [203](#), [207T](#) [426](#), [427B](#), [448](#)

Precision Medicine Initiative (China) [484](#)

precision oncology [413](#)

preconception couple carrier screening [467](#)

predictive genetic testing [453](#), [455](#), [457](#)

predictive value, genetic testing [422T](#), [471](#), [481](#)

pregnancies

- and confidentiality [474](#)
- genetic screening [463](#)
- termination [458B](#), [467–8](#), [474B](#), [483](#)
- preimplantation embryo [116](#)
- preimplantation genetic testing/diagnosis [419T](#), [420](#), [457](#), [460–1](#), [460F](#),
[484](#), [487–8](#)
- premature termination codons (PTCs)
 - consequences of [192T](#)
 - mutations producing [181F](#), [185](#)
 - nonsense-mediated decay [186](#), [186–7B](#)
 - translational readthrough [183T](#), [325](#)
- premutation, [198](#)
- prenatal genetic testing [457](#), [483](#)
 - ethical considerations [474–5](#), [475F](#)
 - invasiveness [461–2](#), [462T](#)
 - prospective parents [456–7](#), [457F](#), [460–1](#), [460F](#)
- prenatal HLA selection [484](#)
- presenilin 1 and [2](#)
 - PSEN1* and *PSEN2* genes [284T](#), [286](#)
- presymptomatic testing [453](#)
- prevention of disease
 - in familial hypercholesterolemia [308T](#)
 - mitochondrial DNA disorders [356](#), [357F](#)
 - newborn screening and [464](#), [465T](#), [465–6](#)
 - in inborn errors of metabolism [306](#), [308](#)
 - in phenylketonuria [308T](#)
 - prenatal in 21-hydroxylase deficiency [307F](#)

strategies for [303F](#)

see also [carrier screening](#)

primary biliary cirrhosis [256](#)

primary structure, proteins [25F](#)

primordial germ cells (PGC) [14](#), [158](#), [158F](#) [189–90](#), [214](#), [214F](#)

prion proteins

abnormal aggregation of mutant proteins [228](#), [229B](#)

disease due to mutant proteins [229B](#)

PrP^C and PrP^{Sc} prion proteins [229B](#)

see also [amyloid proteins](#)

prionoid neurodegenerative diseases [229–30B](#)

private variants [89](#)

PRNP prion protein gene [229B](#)

proband [112](#)

probes, hybridization assay [66](#), [67F](#)

in MLPA [449](#), [450F](#)

procaspases 8 and 9 [385F](#)

prodrugs [314](#)

progeria [85–6B](#), [126T](#)

proline

chemical class [185T](#)

in protein folding [185](#)

structural role [185](#)

unusual amino acid structure [25F](#)

prometaphase [13](#), [14F](#)

prometaphase chromosome preparations [204](#), [204B](#), [248](#)

promoter, internal in RNA pol III-transcribed genes [50](#)

promoter sequences [141–2](#), [142F](#)

core elements, [142](#), [142F](#), [144](#)

downstream promoter element [142F](#)

histone modifications of [154T](#)

internal promoters for some genes [49B](#)

TATA box [142F](#)

promyelocytic leukemia [378T](#), [429](#), [468T](#)

pronuclear microinjection [338B](#)

pronucleus [338B](#)

prophase (meiosis I) [15F](#)

mitosis [13](#), [14F](#)

propositus [112](#)

prospective cohort studies *see under* [gene-environment interactions](#)

prostate cancer [468T](#)

proteasomes (103)

protective factors

importance in reducing disease risk [289–90](#), [289T](#), [290F](#)

protein aggregation, causing disease [226–8](#), [229–30B](#)

protein sequence

factors causing variation in [103T](#)

protein structure

different levels of [30B](#)

secondary structure of proteins [30–1B](#)

protein-coding genes

containing/overlapping RNA genes
gene organization [22](#), [22T](#)
number in human genome [46T](#)

protein folding

diseases caused by misfolding [226–8](#), [229–30B](#)
environmental influences [185](#), [232–3](#)
misfolded proteins as templates *see under* [prion](#)
regulation of [226](#)
roles of glycine, cysteine and proline in [185](#)

protein isoforms, origin of [93](#), [93T](#)

protein polymorphisms/variants

functional genetic variation and [93ff.](#)
gene duplication and [98](#)
HLA proteins [102–4](#), [104T](#)
MHC polymorphism [104](#)

protein-protein interactions

blocking of by intrabodies [308](#)
transcription modulation

protein structure, four classes of [30–31B](#)

proteins

chemical modification [29–30](#), [30T](#)
as drug targets [325](#)
factors causing sequence variation in [93T](#)
see also [polypeptides](#)

proto-oncogenes

as normal genes [376](#)

three major ways of being activated [376](#), [376F](#), [377](#), [377F](#),
[378T](#), [378–9](#), [379F](#)

see also [oncogenes](#)

protospacer motif (PAM) [351](#), [351F](#)

PROVEAN computer program [224](#), [447T](#)

proximal locations, on chromosomes [205B](#)

PSEN1 and *PSEN2* genes see [presenilin 1 and 2](#)

pseudoautosomal inheritance [16](#), [119–20](#), [120F](#)

pseudoautosomal regions, [16](#), [119–20](#), [225](#)

major (PAR1) [119](#), [119F](#)

minor (PAR2) [119](#), [119F](#)

obligatory crossover in PAR1 in male meiosis, [16](#), [119](#), [120F](#)

sites of X-Y crossover [119](#), [120F](#)

pseudogenes

dispersed across the genome (example of *NF1* family) [48B](#)

mitochondrial pseudogenes in nuclear genome, see [NUMT
sequences](#)

number of in human genome (GENCODE) [46T](#)

retropseudogenes [49B](#)

tumor suppressor function for *PTENP* “pseudogene”

unprocessed, arising by gene duplication [48](#), [48F](#), [48B](#)

pseudouridine [29F](#)

pseudohermaphroditism [224](#)

psoriasis

HLA association [266B](#)

PSM2 endonuclease [393T](#)

PTEN tumor suppressor gene [49B](#), [148–9](#), [149F](#)

binding site for miRNAs [149F](#)
PTENP1 functional “pseudogene”

as a regulator of *PTEN*, [148–9](#), [149B](#), [149F](#)

ptosis [215](#)

PTPN11 gene [190T](#)

PTPN22 protein [289](#)

R620W variant [289](#)

purifying selection [132](#)

functional pseudogenes and [148](#)

operating on harmful mutations [135](#)

and proportion of human genome under functional constraint
[44](#)

purines, and structures of [4](#), [4F](#)

pyloric stenosis [260B](#)

pyrimidines, and structures of [4](#), [4F](#)

pyrophosphate [437](#), [439F](#)

pyrosequencing [437](#), [439F](#)

Q

Q/R editing [147](#)

QT interval [319–20B](#), [481B](#)

quantitative fluorescence PCR (QF-PCR) [423T](#), [424](#), [425F](#)

detecting sex chromosome aneuploidies [426F](#)

quantitative PCR, *see* [PCR](#)

quantitative trait loci (QTL) [253B](#)

quaternary structure, of proteins [30B](#)

R

RAD50 gene [269B](#)

RAD51 in DNA repair [391F](#)

rapamycin (sirolimus) [324](#), [324F](#)

see also [mTOR](#)

rarity, key parameter in possible pathogenicity of sequence variants
[443](#), [443F](#)

RAS family of oncogenes [378–9](#), [385T](#)

RAS signal transduction pathways [406](#), [407F](#)

RB1 retinoblastoma protein [368](#), [383](#), [383F](#), [384](#), [3](#)

RB1 tumor suppressor gene, regulator of cell growth [380F](#), [381–2](#),
[385T](#), [387B](#)

reactive oxygen species (ROS) [233](#)

apoptosis pathways and [384](#)

chemistry of [81](#), [83](#)

mitochondria and [121](#), [189](#), [292B](#)

released by metastatic cells [363F](#)

real-time PCR, *see* [PCR](#)

receiver-operating characteristic (ROC) curves [278–9B](#)

area under the curve (AUC) [278–9B](#)

recessive phenotypes

definition of [111](#)

reciprocal translocations [207–8](#), [209F](#)

recombinant DNA [58](#), [58F](#), [59F](#), [61B](#)

advantage of “sticky ends” in ligating vector to DNA [61B](#)

recombinant proteins, therapeutic use [326](#), [326T](#), [486](#)
see also recombination [15](#)

recombination

errors in [80](#)

frequencies [244](#), [244F](#)

intrachromatid [202](#), [203F](#), [204F](#), [206](#)

see also [crossover](#); [nonallelic homologous recombination](#);
[unequal crossover](#); [unequal sister chromatid exchange](#)

redundancy, in the genetic code [27](#)

RefSeq database [39T](#) [444](#)

RefSeqGene database [39](#), [39T](#)

regenerative medicine [329](#), [353](#)

related people, meaning of [267](#)

relative risk and lifetime risk of disease

contrasting values for monogenic and multifactorial conditions
[254](#), [255T](#)

relatives, degree of genetic relationship [314–5B](#)

see also [family members](#)

renal cancer [396](#), [401](#)

repetitive sequences, human

Alu repeats [52](#), [52T](#)

Alu repeat structure [52F](#)

interspersed [182F](#)

LINES (long interspersed nuclear elements) [52](#), [52T](#)

LINE-1 (L1) repeat, structure [52F](#)

LINE-1 (L1) repeats and exon shuffling [53](#), [53F](#)

low copy number variation [91](#), [92F](#)

noncoding [51–3](#)

overview of how they predispose to disease [182](#), [182F](#)

satellite DNAs [51](#), [90](#)
SINES (short interspersed nuclear elements) [52](#), [52T](#)
SVA repeats [52](#), [52T](#), [52F](#)
and tandem duplication [182F](#)
transposon-derived, human classes [52T](#)
see also [multi-gene families](#)

replication fork [82](#)

replication origins [10](#)

replication slippage, *see* [DNA replication](#)

replicons [59](#)

reprogramming, epigenetic

artificially induced in pluripotent cells [332B](#)

in cancer cells [395](#), [397T](#), [403](#), [404F](#)

in the early embryo [151T](#)

in the germ line [151T](#)

restriction endonucleases

in DNA cloning [61](#), [61B](#)

natural role [60–1B](#)

sequence specificity [61B](#)

type II [61B](#)

restriction fragment length polymorphism(s) (RFLP) [89](#), [89F](#)

restriction sites [60–1B](#)

RET gene [224](#)

retinal disorders, gene panel [442B](#)

retinitis pigmentosa [125](#), [207T](#)

retinoblastomas [385T](#), [386](#)

familial vs. sporadic [381–2](#)

retrogenes, *see* [DUX4 retrogene](#)

retropseudogenes [47T](#), [49B](#)
retrotransposons/retrotransposition [51–2](#), [52T](#), [53](#), [53F](#), [182T](#)
retrotransposon elements, suppression by DNA methylation [155](#)
retrovirus-like LTR elements [52](#), [52T](#), [396](#)
retroviruses, gene delivery using [336T](#)
retroviruses/retrovirus vectors

 gammaretroviruses [336T](#), [336–7](#)
 human endogenous retroviruses (HERVs) [52](#), [52T](#)
 lentiviruses [336T](#), [337](#)
 oncoretroviruses [376](#)

Rett syndrome [167](#), [167T](#), [168B](#)
REVEL program [443F](#), [447T](#)
reverse transcriptases [51](#), [63](#)

 and cDNA libraries [62](#)
 in genome evolution [7T](#), [53F](#)
 source of retropseudogenes and retrogenes [49B](#)
 telomerase endonuclease reverse transcriptase (TERT) [7](#), [368B](#)

rheumatoid arthritis

 as amyloid disease [230B](#)
 HLA association [266B](#)
 protective variants for [289T](#)

ribose and deoxyribose, structures of [3](#)
ribosomal RNAs

 28S, 5.8S and 18S rRNA and RNA polymerase I
 5S rRNA and RNA polymerase III

ribosomal DNA regions, human [208](#)

satellite stalks on acrocentric chromosomes, [209F](#)
ribosomes [24](#), [28F](#)

and 5' untranslated region [28](#), [28F](#)
mitochondrial ribosomes
peptidyltransferase in 28S rRNA

in translation [28F](#)
ribosomal RNA genes

interchromosomal recombination [50](#)

ribozymes

RNase MRP, [34F](#)
RNase P [34F](#)
28S rRNA (peptidyltransferase) [33](#)

ring chromosome [206T](#), [207](#), [208F](#)

risk assessment [457](#), [458B](#)

risk ratios *see* [disease risk](#)

RNA(s)

antisense RNAs [35](#)
coding vs noncoding [8](#)
circular RNAs [34F](#), [149F](#)
noncoding RNA, versatility of [33](#), [34F](#)
primary transcript [23](#), [23F](#)
secondary structure [33](#), [33F](#), [147](#)

RNA classes

long noncoding RNA, *see* [long noncoding RNA](#)
mRNA, *see* [messenger RNA \(mRNA\)](#)
miRNA, *see* [miRNA \(microRNA\)](#)

piRNAs (Piwi protein-interacting RNAs) [34F](#), [35](#)
ribosomal RNAs, *see* [ribosomal RNA \(rRNA\)](#)
short interfering RNAs (siRNA), *see* [siRNA](#)
scaRNA (small Cajal-body RNA) [33](#), [34F](#)
snRNA (small nuclear RNA) *see* [snRNA](#)
snoRNA (small nucleolar RNA) *see* [snoRNA](#)
tRNA *see* [transfer RNA \(tRNA\)](#)

RNA editing [93T](#), [400](#)

A to (Q/R) editing [147](#)

C to U editing [147](#)

transamination of certain nucleotides [147](#)

U to C editing [147](#)

RNA enzymes *see* [ribozymes](#)

RNA fusion panels [423T](#)

RNA genes, general

disease loci in single gene disorders [191](#), [191T](#)

difficult to identify [45](#)

number in human genome (GENCODE) [46T](#)

gene families, example of U6 snRNA family [47T](#)

mutated in single-gene disorders [191](#), [191T](#)

polymerases transcribing [142](#)

RNA-induced silencing complex (RISC) [347](#), [348F](#)

RNA interference (RNAi)

function [347](#), [348F](#)

therapeutic gene silencing [347–8](#), [349F](#), [349T](#)

triggered naturally by certain viruses and transposon transcripts
[347](#)

RNA polymerase(s) eukaryotic nuclear

RNA polymerase I [142](#)
RNA polymerase II [142](#)
RNA polymerase III [49B](#), [50](#), [142](#)

RNA and oligonucleotide therapeutics, an overview [344](#), [344F](#)
RNA splicing [24](#), [145–7](#)

evolutionary value [24](#)
back-splicing to make circular RNAs [149F](#)
branch site [145](#), [145F](#)

bound by U2 snRNA [145F](#)

nonsense-mediated decay and [186–7B](#)
regulation of [145](#), [145F](#), [146–7](#), [146F](#)
splice acceptor site [24](#), [145](#), [145W](#)
splice donor site, [124](#), [45](#), [145F](#)

bound by U1 snRNA [145F](#)

splicing modulation therapy [345](#), [346–7B](#)
see also alternative splicing, [24](#)

RNase Hi ribonuclease [345](#)
RNA therapeutics [344F](#)
RNA world hypothesis [33](#)
Robertsonian translocation [206](#), [208](#), [209F](#)
Ronald Fisher [252](#)
RPE65 genes [344](#)
retinal pigment epithelial cells [344](#)
RT-PCR (reverse transcriptase-polymerase chain reaction)
Rubinstein-Taybi syndrome [167T](#)

S

S-adenosylmethionine (SAM) [296](#)
Sanger/dideoxy sequencing [71–2](#), [74B](#), [75](#), [88](#), [433](#), [435B](#), [436](#)
satellite DNA, families [51](#)
scFv (single-chain variable fragment) antibodies [410](#)
Schinzel-Giedion syndrome [251T](#)
schizophrenia

- % concordance in MZ and DZ twins [257T](#)
- adoption studies [258](#)
- and immune system susceptibility factors [288–9](#)
- effect of reduced complement C4 on synapse number [288–9](#)
- genome-wide linkage analysis [261](#)
- GWA studies [288](#)
- protective variants for [289T](#)

Seckel syndrome [85B](#)
secondary findings, genetic testing *see* [incidental findings](#)
secondary structure, of proteins [30–1B](#)
 β -secretases (BACE1) [286](#)
 γ -secretases [286](#)
segmental aneuploidies [207T](#), [224–5](#)
segmental duplications [199](#)
segregation of DNA molecules

- mtDNA [13](#), [13F](#)
- nuclear DNA [14F](#)

segregation ratio [122](#)
selection pressure and cancer [65–6](#)
selective sweeps [96](#), [96B](#)
selenocysteine [184F](#)
selfish mutation(s) [130](#), [190](#), [190T](#)
selfish spermatogonial selection [190](#)
senescence, of cells [367B](#)

sensitivity of a genetic test, *see* [genetic test parameters](#)
sequence conservation, *see* [evolutionary conservation](#)
sequencing-by-synthesis [74](#), [435B](#)
serine

chemical class [185T](#)
phosphorylation [153](#)
structure [25F](#)

SETD2 gene [401](#)

severe combined immunodeficiency (SCID) [85B](#), [308T](#), [341](#), [342F](#)
sex chromosomes

aneuploidies [210T](#)
male-specific region on Y [119](#), [119F](#)
pseudoautosomal (PAR) regions [116](#), [119–20](#), [119F](#), [120F](#), [225](#)
recombination, obligatory crossover in major PAR [120F](#)
X-specific region [119](#), [119F](#)
X-Y pairing confined to pseudoautosomal regions [16](#), [119](#)
see also [X-chromosome inactivation](#)

sex differences

cell divisions needed for gametogenesis [189](#), [190F](#)
recombination frequency differences [244](#), [244F](#)
see also [sex chromosomes](#)

sex-determining region [120](#)

SF3B1 gene [400](#)

short hairpin RNA (shRNA) [347](#), [349F](#)

short interfering RNAs (siRNA) [347–8](#), [348F](#)

in artificial gene suppression/silencing [345](#), [347–8](#), [349F](#), [349T](#)
delivery to cells [347](#), [349F](#)

endogenous siRNAs [33F](#)
in RNA interference [347](#), [348F](#)
short tandem repeats, unstable expansions of [194–8](#)
SHOX homeobox gene [129](#)
sib (sibling) and sibship [112](#)
sickle-cell disease [97](#), [115–6](#), [130](#), [421](#), [436](#)

due to disruptive protein fibers [227](#), [227F](#)
genetic testing and [463–5](#)
heterozygote advantage [137](#)
mutationally homogeneous [438](#)
pregnancy screening [463](#)
sickle-cell trait [115–6](#)

SIFT program [443F](#), [447T](#)
signal peptides (leader sequences), role in protein export, [31–2](#)
signal recognition particle (containing 7SL RNA) [52](#)
silencer, cis-acting regulatory element [143](#)
silent substitution *see* [synonymous substitution](#)
Silver-Russell syndrome [172](#), [173T](#)
SINES (short interspersed nuclear elements) [52](#), [52T](#)
single-cell genomics

in cancer, [404](#), [405F](#)

single-cell transcriptomics

in cancer, [404](#), [405F](#)
for classifying tumor cells [406–7](#)
in identifying rare tumor cells [406–7](#)

single-gene disorders *see* [monogenic disorders](#)

abnormal epigenetic regulation [165](#)

single nucleotide polymorphism(s) (SNP) [89](#), [470](#)
single nucleotide variant(s) (SNV) [87](#), [89F](#), [92](#), [437](#), [440](#)
single-nucleotide variation

nonrandom features of [89T](#)

single-strand DNA breaks [81](#), [83](#), [85B](#), [409T](#)

SIRT6 tumor suppressor [403](#)

sister chromatids [12](#), [12F](#), [13](#)

in HR-mediated DNA repair [83](#), [84F](#)

sister chromatid exchange

high frequency of in Bloom syndrome cells [86B](#)

Sjogren syndrome

protective variants for [289T](#)

skin pigmentation [95T](#), [97B](#)

SLC24A5 gene, positive selection for advantageous variant [96](#), [96–7B](#)

SLC24A5 protein, function [96](#)

SLE (lupus), protective variants for [289T](#)

small bowel cancer [454](#)

small Cajal-body RNA (scaRNA)

small molecule drugs [303F](#), [310–11](#)

assays and trials needed [311F](#)

developing from gene-protective factors [281](#)

effects of genetic variation on metabolism and performance
[311ff.](#)

method of action [310](#)

major stages of drug development [310–11F](#)

and targeted cancer therapies [408–9](#), [409T](#)

therapies to counter mutant gene product [305](#)
translating genetic advances [322–5](#)
see also [adverse drug reactions](#); [pharmacogenetics](#);
[pharmacokinetics](#)

small nuclear RNA (snRNA) *see* [snRNA](#)
small nucleolar RNA (snoRNA) *see* [snoRNA](#)
SMARCA4 gene [395](#)
SMCHD1 gene [170](#)
SMCHD1 protein [170F](#)
Smith-Magenis syndrome [203T](#)
SMN1 gene [346B](#)
SMN2 gene [346–7B](#)
smoking

risk factor in complex disease [290T](#)

SNHG14 (*SNRPN*) [159T](#)
SNORD116 gene [174](#), [174F](#)
snoRNA (small nucleolar RNA)
snoRNA genes [173T](#), [174](#)
SNP microarray hybridization [270](#), [270F](#), [271](#), [272F](#), [423T](#), [425–6](#),
[427–8B](#)
SNPs (single nucleotide polymorphisms) [270](#)
SNP chips (microarrays) [268](#), [270](#), [270F](#), [271](#), [273](#), [277](#), [471](#)
snRNA (small nuclear RNA) [23](#), [33](#), [34F](#), [145F](#), [191T](#)

U6 snRNA gene family [47T](#)

SNURF-SNRPN gene [174F](#)
SOD1 (superoxide dismutase 1)

cytoplasmic aggregates in amyotrophic lateral sclerosis [230B](#)

sodium bisulfite [396](#), [448](#), [449F](#)

software *see* [computer programs](#); [databases](#)
solid supports, hybridization assays
somatic cells, distinguished from germline cells [13](#)
somatic genetic variation *see* [post-zygotic genetic variation](#)
somatic mutations

COSMIC database [398](#), [398T](#)
not identical to post-zygotic mutations [78](#), [78T](#)
major role in cancers [364](#)

somatic hypermutation

due to excess cytidine deaminase in activated B cells [102](#)
due to excess cytidine deaminase in cancer cells [102](#)

somatic recombinations, importance of in B and T cells [100–2](#)
Sotos syndrome [203T](#)

gene identification [248T](#)

Southern blot-hybridization [171B](#)

SOX2 gene [333B](#), [396](#)

SOX2-related disorders [479](#)

SOX10 gene [240](#)

Spastic paraplegia type-30 [251T](#)

specificity of a genetic test, *see* [genetic test parameters](#)

spectral karyotyping (SKY) [390](#)

sperm cells

each genetically unique [77](#)

haploid [11](#)

number of cell divisions to make [190F](#)

spermatocytes, primary [17](#)

spermatogonia [14](#)

spinal bulbar muscular atrophy (Kennedy disease) [194T](#), [224](#)

spinal muscular atrophy [462T](#)

spinal muscle atrophy, exon skipping therapy [346–7B](#), [349T](#)

spindle checkpoints, defects [390](#)

spinocerebellar ataxia

 type 7 (SCA7) [194T](#)

 type 10 (SCA10) [194T](#)

splice enhancer sequences [145](#), [145F](#)

 bound by SR proteins [145](#)

splice junctions

splice acceptor site [24](#), [144](#), [145F](#)

 alternative acceptor sites [146F](#)

splice branch site [145](#)

 bound by U2 snRNA [145F](#)

 cryptic, see [cryptic splice sites](#)

splice donor site [144](#), [145F](#)

 alternative sites [146F](#)

 bound by U1 snRNA [145F](#)

splice suppressor sequences [145](#), [145F](#)

 bound by hnRNP proteins [145](#)

SpliceAI program [447T](#)

SpliceDisease Database [192T](#)

spliceosomes [145F](#)

splicing see [RNA splicing](#)

splicing modulation therapy [344F](#), [345](#), [345–6B](#)
splicing mutations, pathogenic [187](#), [187F](#), [188F](#)
sporadic cases, due to new mutation [123](#)
SR proteins, splicing regulators [145](#)
statins, and HMG CoA reductase inhibition [318](#), [322](#)
stem cells [11](#)

- an overview [331–313B](#)
- asymmetric versus symmetric cell division [331B](#)
- cancer stem cells, *see* [cancer stem cells](#)
- embryonic stem cells (ES cells) [331–2B](#), [338–9B](#)
 - as gene therapy targets [331](#)
- protection of stem cell genome [374B](#)
- somatic stem cells [332B](#)
- spermatogonial stem cells [189](#), [190F](#), [219](#)
- transit amplifying cells [331B](#), [374B](#)
 - see also* [cancer stem cells](#); [embryonic stem cells](#); [hematopoietic stem cells](#); [induced pluripotent stem cells](#)

stem cell therapies

- banks of iPSC lines [354](#)
- cell sources for [353](#)
- minimizing immune responses [354](#), [354F](#)
- obstacles [353](#)
- problems with pluripotent stem cells [353–4](#), [354F](#)

stem-loop structures, in RNA

steroid 21-hydroxylase deficiency [199](#), [201–2B](#)

- prenatal treatment of virilization [307F](#)
- salt-wasting phenotype [231](#)
- simple virilizing phenotype [231](#)
- steroid supplementation [306](#), [307F](#)

steroid 21-hydroxylase gene, *see* [CYP21A2 gene](#)
sticky ends, helpful in making recombinant DNA [61B](#)
stop codons

drugs that suppress [325](#)
four in the human mitochondrial genetic code [184F](#)
in universal genetic code [184F](#)
see also [premature termination codons](#)

stop-gain mutation [183T](#)
stop-loss mutation [183T](#)
streptavidin, *see* [biotin-streptavidin](#)
stringency, hybridization assays [67](#), [68F](#)
stromal cells/stroma, as support for cancer cells [363](#), [363F](#), [364T](#)

different cell types in [372](#), [373F](#), [373T](#)

STRPs (short tandem repeat polymorphisms) [90](#)
structural variation [88](#)

balanced and unbalanced [91](#), [92F](#)
and low copy number variation [91-2](#), [9](#)

subependymal giant cell astrocytomas (SEGAs) [324](#)
substitutions, *see* [nucleotide substitutions](#)
suicide gene [330F](#), [355](#)
sugar-phosphate backbone asymmetry
supplementation therapy [303F](#), [304](#)
susceptibility allele, meaning of [267](#)
susceptibility factors *see* [disease susceptibility](#)
SVA repeats [52](#), [52T](#)

structure of [52F](#)

symmetric cell division versus asymmetric cell division [331B](#)

synapsis [16](#)

Syndromic Intellectual Disability gene panel [175B](#)

synonymous (silent) substitution [183](#)

occasionally pathogenic [187](#), [188F](#)

synpolydactyly type II [194T](#)

synthetic lethality [409T](#)

α -synuclein, cytoplasmic aggregates in Parkinson disease [230B](#)

systemic lupus erythematosus [277T](#), [287](#), [290F](#)

T

T-cell leukemia, oncogene activation in [378T](#)

T-cell receptor genes (TCRs)

often involved in oncogenic translocations [378](#), [378T](#)

programmed rearrangements in T cells [100–102](#)

T-cell receptors [99](#), [103](#), [160T](#)

cell-specific production of in T cells [100–2](#)

extreme genetic variation of [99F](#), [100–1](#), [101F](#), [102](#)

functional roles [99](#)

T-cells

cell-specific production of T- cell receptors [100](#)

genetically engineered as “living drugs” [410–11](#), [410F](#)

inhibited by ligand activation of PD1 and CTL4A [410](#)

lack of in ADA deficiency [341](#)

somatic recombination [100–2](#)

see also [cytotoxic T lymphocytes \(CTLs\)](#); [helper T cells](#)

T loop [10](#), [10F](#)

TALEN (TALE nuclease) [352T](#)

tamoxifen [409T](#)

tandem duplication of exons [47F](#), [50](#)

tandem duplication of genes [47F](#)

tandem repeats

containing genes [46](#), [47F](#)

evolutionary advantage [50](#)

in exons [47F](#)

segmental duplication [46](#), [199](#), [346B](#)

see also [macrosatellite DNA](#); [microsatellite DNA](#); [satellite DNA](#); [short tandem repeats](#)

TaqMan genotyping [437](#), [438F](#), [439](#)

targeted DNA sequencing

for mutation scanning [441–2B](#)

and exome capture [250F](#)

targeted RNA sequencing [441B](#)

and fusion gene transcripts in cancer [429–30](#)

TATA box [142F](#)

tau mRNA, alternative splicing [147](#)

tau protein, cytoplasmic aggregates in frontotemporal lobar degeneration [230B](#)

Tay-Sachs disease [463](#), [467](#)

treatment difficulty [307](#)

TBLASTN program [41](#)

TCGA (The Cancer Gene Atlas) [398](#), [398T](#)

TCP10L gene [43](#)

telomerase

in cancer [403](#)
cancer cell immortality and [367B](#)
function [367B](#), [368B](#)
solving end-replication problem [367B](#), [368B](#)
TERT (telomerase reverse transcriptase) [368B](#)
TERC (telomerase RNA complex) [368B](#) [191T](#)

telomeres

copy number [367B](#)
evolutionary conservation of telomeric DNA [10](#), [51](#)
function [10](#)
G-rich and C-rich strands [10](#), [10F](#)
reduced at cell division [367B](#)
structure of [10](#), [10F](#)
T-loops [10](#), [10F](#)
TTAGGG repeats [10](#), [10F](#), [51](#)

telomeric heterochromatin [51](#)
teratogen [290T](#)
teratomas [163F](#), [354F](#), [355](#)
TERC RNA, [34F](#)
TERRA telomerase RNA [34F](#)
terminally differentiated cells [11](#), [331](#), [332B](#)
termination codons *see* [stop codons](#)
TERT (telomerase reverse transcriptase) [7T](#)
TERT gene, in cancer [403](#)
tertiary structure, proteins [30](#)
testicular feminization syndrome (androgen-insensitivity syndrome)
[224](#)
TET2 demethylase [403](#), [403F](#)
tetrahydrobiopterin (BH₄) [235B](#)
tetraploidy [210](#), [210F](#)

TGF β (transforming growth factor β), inhibitor of cell proliferation
[394](#)

TGFBR2 protein [393–4](#)

Th17 helper T cells [282B](#)

thalassemia(s) [95T](#), [97](#), [130](#), [136](#), [450](#), [463](#)

 alpha-thalassemia X-linked mental retardation [167](#)

β -thalassemia [125](#), [233](#), [234F](#), [252](#), [352](#), [466–7](#)

 modifier genes in [232–3](#), [233F](#), [234F](#)

The Cancer Genome Atlas (TCGA) [398](#), [398T](#)

therapeutic antibodies

 chimeric (V/C) [327](#), [327F](#)

 genetically engineered [326–7](#), [327F](#)

 humanized and fully human [327](#), [327F](#)

see also [intrabodies \(intracellular antibodies\)](#).

therapeutic “recombinant” proteins [325–6](#), [326T](#)

 PEGylation of [325](#)

therapeutic splice modulation [345](#), [346–7B](#)

therapeutic windows [313](#), [313F](#)

thiopurine methyltransferase [313F](#), [318](#), [320B](#)

thrifty phenotype (thrifty gene) hypothesis [296](#)

threonine

 chemical class [185T](#)

 structure [25F](#)

thymidine glycol [81F](#)

thymine

 structure [4F](#)

thyroid hormone [308T](#)
tight junctions [343](#)
tissue typing, *see* [HLA](#), [histocompatibility testing](#)
tobacco (carcinogen) [234](#), [399–400](#), [407](#)
Toll-like receptors [262](#)
TopMed (Trans-Omics for Precision Medicine) Program [272](#), [277](#)
torsades de pointes [320B](#)
total pathogenic load, in an average human genome [189](#)
toxic protein (aggregates) [199](#), [229B](#)
toxic RNA [344](#)
toxins, delivered to kill cancer cells [327](#)
TP53 gene [395](#)

apoptosis and [365F](#)
database [397](#)
missense mutations [387B](#)
role in cancer [453](#), [468](#)

TRA, *TRB*, *TRD*, *TRG* T-cell receptor genes [100](#)
tracRNA [350](#), [350F](#)
trait(s)109

continuous vs dichotomous [253B](#)

trans-acting

definition [140](#)
gene regulation [140–1](#), [141F](#)
long noncoding RNAs acting on DNA [159T](#)
miRNAs binding to cis elements in UTRs of mRNAs [141F](#)
regulatory proteins binding to cis elements in DNA [140](#), [141F](#)
regulatory proteins binding to cis elements in UTRs of mRNAs
[147](#), [148F](#)

transcription [7](#), [7F](#)

primary transcript [7F](#)

transcription-coupled repair [83](#)

transcription factors [143–4](#)

combinatorial action to increase specificity [144](#)

DNA-binding motifs [144](#), [144F](#)

ubiquitous versus tissue-specific [142](#)

transcription initiation complexes [142](#)

transcription unit, DNA [23](#), [23F](#)

transdifferentiation [353](#)

transduction [334](#)

transfection [334](#)

transfer RNA (tRNA) [26](#), [27](#)

acceptor arm [29F](#)

as adaptors [27](#)

amino acid covalently linked to 3' end [27](#), [29F](#)

anticodon in tRNA [27](#), [29F](#)

binding of specific amino acids [26](#)

clover leaf structure [17](#), [29F](#)

minor nucleotides in [29F](#)

genes for mitochondrial [44](#), [44F](#)

transferases in phase II drug metabolism [313](#)

transferrin, translational regulation [24](#), [26](#) [147](#), [148F](#)

transformation

of bacterial cells for DNA cloning [58](#), [58F](#), [59F](#)

by oncoretroviruses [376](#)

transgenes [330](#), [338B](#)

definition [330](#)
expression possibly eliciting immune response [334](#)
transgenic animals

- as disease models [337](#)
- and producing recombinant proteins [326](#), [326T](#)
- pronuclear microinjection [338B](#)

transit amplifying cells, produced by stem cells [331B](#)
transitions [89](#)
translation [24](#), [26](#), [26B](#), [27](#)

- RAN (repeat-associated non AUG) mechanism [195–6](#), [196F](#)

translational reading frame [26B](#), and deletion of coding exons [27B](#)

see also [frameshifting mutations](#)

translational readthrough [183T](#), [325](#)

- and Ataluren (PTC124) [325](#)

translesion synthesis [84](#)
translocations

- balanced [87](#), [94](#)
- versus unbalanced [91](#), [92F](#)
- derivative chromosomes [208](#)
- including oncogene activation [377–8](#), [378T](#)
- serial in chromoplexy [391](#)
- see also* [reciprocal and Robertsonian translocations](#)

transplantation [104–5](#)

allogenic vs autologous [340](#)
bone marrow transplantation [304](#), [306](#), [308T](#), [329](#), [340](#)
graft-versus-host disease [104](#)
histocompatibility testing [104–5](#), [105B](#)
immune rejection [353](#)
transposon repeats [43F](#)

evolutionary value [53](#), [53F](#)

transposons [35](#), [51](#)

see also [retrotransposons](#)

transversion [89](#)
trastuzumab (Herceptin) [412](#)
treatment of genetic disease

altering genetic susceptibility [303F](#), [305](#)
an overview [303–5](#)
different strategies [303F](#)
for disorders producing positively harmful effect [303F](#), [304](#)

tricarboxylic acid (Krebs) cycle [395](#), [405F](#)
trinucleotide (triplet) repeat expansion [193ff.](#)

see also [polyalanine expansion](#)
see also [polyglutamine repeats](#)

trio testing [440](#), [474](#), [475F](#), [479](#), [483B](#)
triplet repeat-primed PCR, *see* [PCR](#)
triploidy [216T](#)
trisomies [206T](#)
trisomy [11](#)
trisomy 13 and 18, [211](#), [225](#)
trisomy 21 [116](#), [211](#), [225](#); *see also* [Down syndrome](#)

trophectoderm [460](#)
trophoblasts [332B](#)
tropism of viruses [336](#), [343](#)
truncated protein [183T](#)
tryptophan

chemical class [185T](#)
structure [25F](#)

TSC1, *TSC2* genes, [128F](#)
TTR transthyretin gene [348](#)
tuberous sclerosis [324](#), [324F](#)

treatment [324](#)
variable expressivity [128](#), [128F](#)

tumor biopsies

checking mutations governing tumor response to targeted drug
[436](#)
invasive versus liquid biopsies [412–3B](#)
screening for residual disease [436](#)

tumor cells

circulating, single-cell analyses of [405F](#)
spectral karyotyping of [390](#), [390F](#)

tumor necrosis factor receptor superfamily [384](#)

tumor recurrence, basis of [411–2](#)

tumor subclones [372](#), [373F](#), [412](#), [468](#)

tumor suppressor proteins
non-classical [386](#)
haploinsufficiency [386](#)

epigenetic silencing [386](#)
gain-of-function mutations in some [386](#)
p53, a non-classical tumor suppressor [386–8B](#)
tumor suppressor genes [396](#)

caretaker genes [380](#), [385T](#), [391](#)
gatekeeper genes [380](#), [385T](#)
landscaper genes [380](#)
loss of heterozygosity (LoH) [382](#), [382F](#)
mapping by LOH [382](#)
miRNA genes as [388](#)
normal function [380](#)
recessively-acting [380](#)
two-hit paradigm [381–2](#), [381F](#), [386](#)

tumor types

benign and malignant [362](#), [363F](#)
classification of recently improved [405–6](#)
development of malignant [362](#)
hereditary and sporadic [386](#)
major categories by tissue of origin [364T](#)
multiple levels of cell heterogeneity [372](#), [373F](#), [373T](#)
nonsynonymous mutations, numbers in different tumor types
[399](#), [399F](#)
MSI-positive (MIN-positive) [393](#)
solid vs. “liquid” [362](#)

tumors

intertumor heterogeneity [401–2](#)
intratumor heterogeneity [401](#), [405F](#)
variation in mutation number in different types [399](#), [399F](#)

Turner syndrome (45,X) [116](#), [163](#), [211](#), [224–5](#)
twins

dizygotic (DZ) [257](#)

monozygotic (MZ) [257](#)

twins studies

and complex disease

concordance between DZ twins [257T](#)

concordance between MZ twins [257T](#)

diamniotic twins [295–6](#)

for estimating heritability [257–8](#)

monozygotic twins [78](#)

two-hit paradigm [386](#)

tyrosine

chemical class [185T](#)

structure [25F](#)

tyrosinemia, type 1, treatment for [307F](#), [308](#), [308T](#)

U

U1 snRNA, U2 snRNA [145F](#)

U6 snRNA, gene family [47T](#)

UBE3A gene [161](#), [173T](#), [174](#)

and Angelman syndrome [175B](#)

ubiquitin [121](#), [153](#), [383–4](#)

ubiquitin-protein ligase [174](#)

UDP glucuronyltransferase superfamily [318](#)

UGT1A1 enzyme [234F](#), [318](#)

UK Biobank Project [88B](#) [293–4](#), [294T](#), [463](#), [471](#)

UK10K Project [272](#)

ulcerative colitis

% concordance in MZ and DZ twins [257T](#)

ultrasound scanning [457F](#), [464](#)

ultraviolet/UV light/radiation [73B](#), [79](#), [81F](#), [82](#), [86B](#), [95–6](#), [95T](#), [397T](#),
[399–400](#), [407](#)

as a mutagen [81F](#)

and vitamin D₃ [95](#)

unequal crossover (UEC) [182T](#), [200](#), [200F](#)

unequal sister chromatid exchange (UESCE) [182T](#), [200](#), [200F](#)

uniparental diploidy [163F](#)

androgenetic embryo [163F](#), [171](#)

ovarian teratoma [163F](#)

uniparental disomy (UPD) [171](#), [172F](#)

arising by trisomy rescue [172F](#)

arising by monosomy rescue [172F](#)

untranslated regions (UTRs)

5' and 3' untranslated regions [28](#)

cis-acting regulatory elements in [141F](#)

uracil, structure [4F](#)

uracil DNA glycosylase [84](#), [85F](#)

urea cycle [309F](#)

urea cycle disorders [308](#)

urinary tract cancer [454B](#)

ustekinumab [282B](#)

V

V (variable) gene segments, in immunoglobulin genes, [101](#), [101F](#), [102](#)
valine

chemical class [185T](#)

structure [25F](#)

variable clinical expression

in mtDNA disorders

in mendelian disorders [128](#), [128F](#), [129](#), [129F](#)

variance, of a phenotype [256](#)

variants *see* [DNA variants](#); [histone variants](#)

variant of uncertain significance (VUS) [436](#), [464F](#), [447–8](#), [472](#), [475F](#)

vascular endothelial growth factor/VEGF [328T](#), [363F](#), [369T](#), [373F](#),
[409T](#)

vCJD (variant Creutzfeldt-Jakob disease) [229B](#)

VDJ coding unit [101–2](#), [379F](#)

vector DNA *see* [cloning vectors](#)

venetoclax [409T](#)

verumafenib [409T](#)

VHL gene (von Hippel-Lindau) [396](#), [401](#)

viral oncogenes [376](#)

viral vectors (gene therapy)

AAV (adeno-associated virus) [336T](#), [337](#)

adenoviruses [336T](#), [343](#)

gammaretroviruses [336T](#), [337](#)

integrating and non-integrating [336T](#)

lentiviruses [336T](#), [337](#)

see also [tropism](#)

virtual gene panels [442–3](#), [472](#), [480](#)

see also [gene panels](#)

viruses

integrating vs non-integrating [337](#)

genetic material in [2](#)

RNAi and [347](#)

vitamin D₃ [95](#)

vitamin K epoxide reductase complex subunit 1 (VKORC1) [321](#)

VNTR (variable number of tandem repeats) polymorphism [90](#), [92F](#)

W

Waardenburg syndrome type I [113](#)

Waardenburg syndrome (Waardenburg-Hirschsprung disease) [240](#)

WAGR syndrome (Wilm's tumor, aniridia, genito-urinary abnormalities, and developmental delay) [207T](#)

Warburg effect [367](#)

warfarin

metabolism by CYP2C9 [321](#), [321F](#)

example of drug where multiple loci are important [321](#), [321F](#)

Wellcome Trust Case Control Consortium (WTCCC) [273](#)

Werner syndrome [85–6B](#)

whole-exome sequencing [88B](#)

whole-genome duplication [42](#)

whole-genome screening, prenatal

whole-genome sequencing (WGS) [88B](#), [433](#), [435B](#)

challenge of sequence interpretation [436](#), [472](#)

compared to whole exome sequencing [442](#)

need to filter data from [442–443](#), [443F](#), [472](#)
in newborn screening [466](#)
prospects in routine healthcare [451](#), [451B](#)
Wiley database of clinical gene therapy trials [340](#)
Williams-Beuren syndrome [203T](#)
Wilms tumor [385T](#)
Wnt signaling pathway

aberration of, driving adenoma formation [370F](#)

WT1 Wilms tumor gene [207](#), [385T](#)
WT1 Wilms tumor mRNA, U → C editing in [147](#)
WT1 Wilms tumor protein

isoforms of [93T](#), [146](#), [146F](#)

X

Xchromosome

regions showing homology to Y chromosome [119](#), [119F](#)
size of [119F](#)
X-specific region [119](#), [119F](#)
see also [pseudoautosomal regions](#)

X-(chromosome) inactivation [116](#), [117F](#), [150](#), [151T](#), [160T](#), [163–4](#),
[164F](#), [165](#)

Barr bodies [116](#), [117F](#)
genes escaping [165](#)
initiation of [165](#)
mosaicism due to [116](#), [117F](#)
nonrandom [117–8](#)
persistence [165](#)
skewing of, by X-autosome translocation [117–8](#)

X-inactivation center (XIC) [165](#)
X-chromosome counting mechanism [165](#)
X-linked dominant inheritance [118](#), [118F](#), [119](#)
X-linked recessive inheritance [116–7](#), [117F](#), [123](#)
X-linked severe combined immunodeficiency [341–2](#), [342F](#)
X-Y crossover

limited to pseudoautosomal regions (PAR) [119](#), [120F](#)
obligate after X-Y pairing in PAR1 [16](#), [119](#), [120F](#)

X-Y gene pairs [120](#)
xenobiotics [262](#), [312](#), [318](#)
xeroderma pigmentosum [85–6B](#)
XIAP gene, and recessive inflammatory bowel disease [443F](#)
XIST (X-inactivation-specific transcript) gene [165](#)
Xist mouse gene, as tumor suppressor [389](#)
XIST RNA [151T](#), [159T](#), [165](#)
XYY males [112](#), [120–1](#)

Y

Y chromosome

evolution of [120](#)
few genes [120](#)
interstitial deletions [22](#)
male-specific region [119](#), [119F](#)
see also [pseudoautosomal region\(s\)](#).

Y-linked inheritance [120–1](#)
YAP1 gene [395](#)
yeast artificial chromosome (YAC)

Z

zinc finger domains [144](#)

DNA binding motif [144F](#)

zinc finger nucleases [352T](#)

zona pellucida [460](#)

zygote [13-14](#)

genetically unique [17](#)