

BIO326 Genome sequencing; tools and analysis

Marie Saitou

Goal of today's class

- Learn how to analyze genome sequence data on

We will learn

- How to do the “cleaning” of the NGS genome data
- How to map raw sequence to the reference genome
- How to identify and analyze the genetic variants



Stop me whenever something is unclear.
Comments and questions are encouraged.

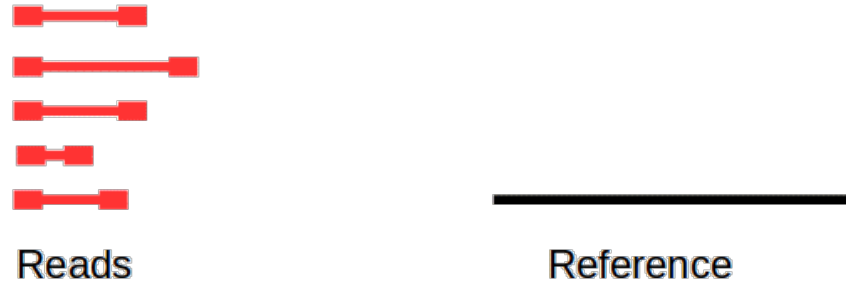
Today's schedule:

- **[Lecture]** Review the genome sequencing pipeline
- **[Lecture]** Go over how to make the genome data ready for analysis
- **[Group work]** Do the analysis by yourself
- **[Short break]**
- **[Group work]** Do the analysis by yourself

Review: Coverage vs Read Depth

Review: Coverage vs Read Depth

(A) coverage in terms of redundancy

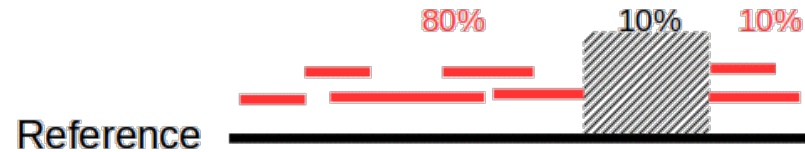


$$C = \frac{\text{\# sequenced bases}^1}{\text{\# bases of reference}}$$

(= # bases of all mapped reads)

The average depth of sequencing coverage: LN/G ,
where L : read length, N : the number of reads, G : the genome length.

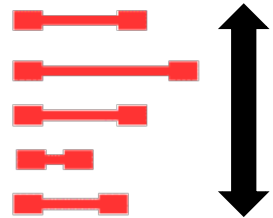
(B) percentage of coverage



$$C = \frac{\text{\# area covered by reads}}{\text{\# reference area}}$$

sequencing depth

(C)



Reads for mapping

Sequencing depth = total read number

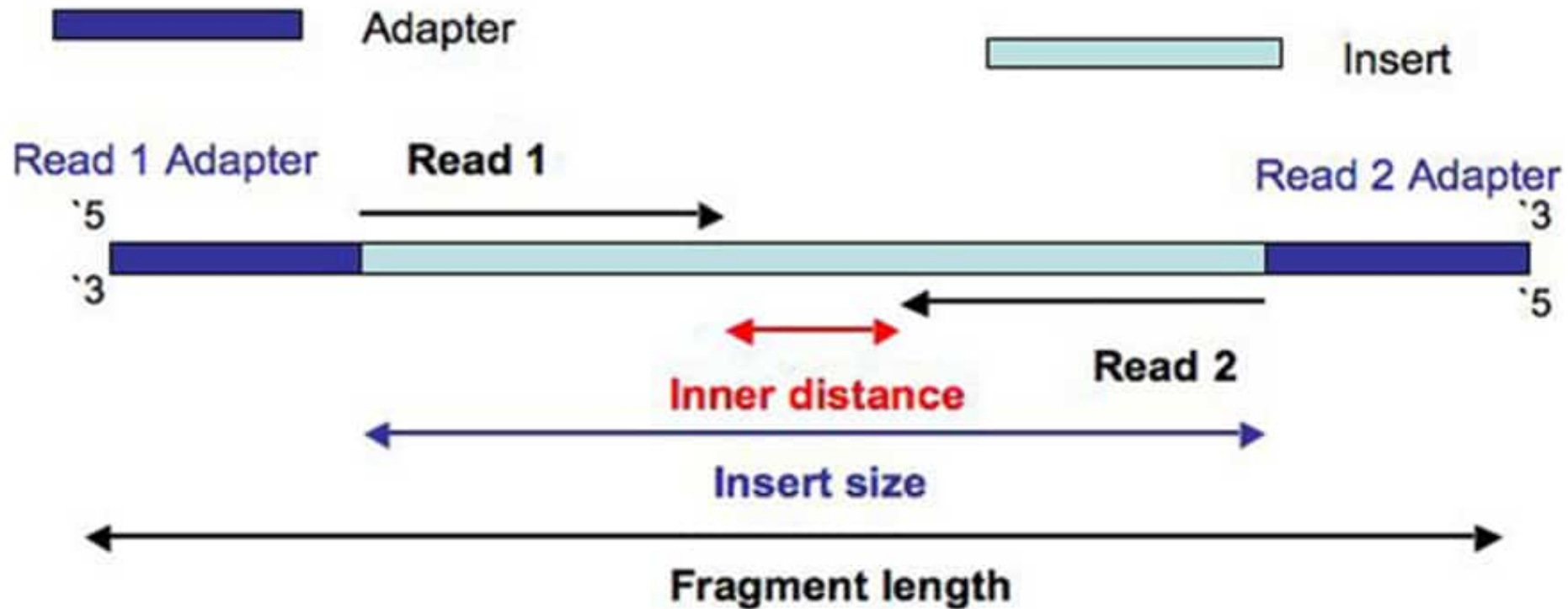
<https://www.ecseq.com/support/ngs/how-to-calculate-the-coverage-for-a-sequencing-experiment>

Brief Review of the previous lesson

- What is adapter?
- What is paired-end?

<https://www.youtube.com/watch?v=fCd6B5HRaZ8>

- **Adapters** include platform-specific sequences for fragment recognition by the sequencing
- **Paired-end** sequencing allows users to sequence both ends of a fragment and generate high-quality sequence data

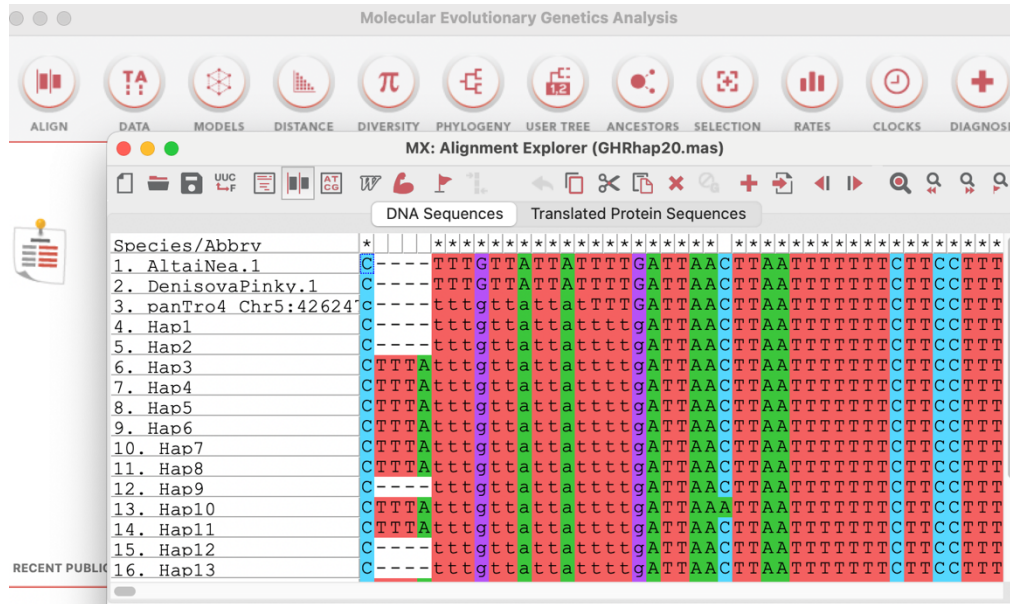


<https://thesequencingcenter.com/knowledge-base/what-are-paired-end-reads/>

Intro to Bioinformatics

- a field of biology that develops/uses computational tools to understand biological data -

GUI: Graphical User Interface



- Easier to learn/Intuitive

CUI: Character User Interface

```
[omergokc@vortex2:/projects/academic/omergokc/ogshared]$ tar -xvf omer_RNAseq.tar -C
20201020_20-lee-007/
20201020_20-lee-007/Ghr-0026_GT20-15754_GAACCGCG-TAAGGTCA_S77_L002_R2_001.fastq.gz
tar: 20201020_20-lee-007/Ghr-0026_GT20-15754_GAACCGCG-TAAGGTCA_S77_L002_R2_001.fastq.
e left on device
20201020_20-lee-007/Ghr-0036_GT20-15752_TCATCCTT-AGCTCGCT_S92_L002_R2_001.fastq.gz
tar: 20201020_20-lee-007/Ghr-0036_GT20-15752_TCATCCTT-AGCTCGCT_S92_L002_R2_001.fastq.
left on device
20201020_20-lee-007/Ghr-0015_GT20-15725_ATATGGAT-TAATACAG_S90_L002_R1_001.fastq.gz
tar: 20201020_20-lee-007/Ghr-0015_GT20-15725_ATATGGAT-TAATACAG_S90_L002_R1_001.fastq.
left on device
20201020_20-lee-007/Ghr-0077_GT20-15739_TTGCCTAG-TAAGTGGT_S75_L002_R1_001.fastq.gz
tar: 20201020_20-lee-007/Ghr-0077_GT20-15739_TTGCCTAG-TAAGTGGT_S75_L002_R1_001.fastq.
left on device
20201020_20-lee-007/Ghr-0148_GT20-15762_CGTCTGCG-ATTGTGAA_S71_L002_R1_001.fastq.gz
tar: 20201020_20-lee-007/Ghr-0148_GT20-15762_CGTCTGCG-ATTGTGAA_S71_L002_R1_001.fastq.
left on device
```

- Easier to custom/run bulk jobs
- Orion -> Later in this course

Notes to start bioinformatics

- When you get errors:
 1. Ask colleagues
 2. Ask google and use forum
- Gain multitasking skills
- Make backup of your files/scripts
- Laziness is the father of invention
- There are many online courses



Common errors

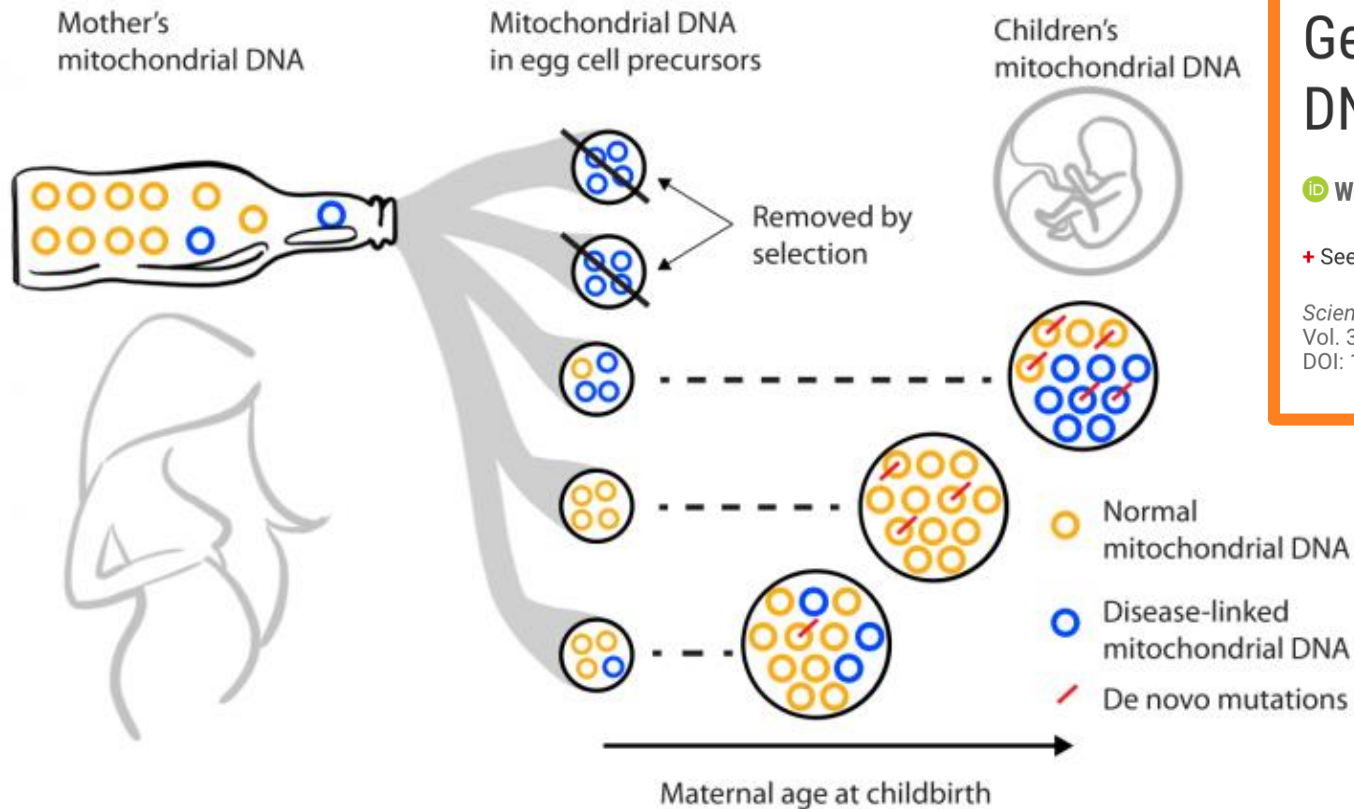
- Typo
- Input file is missing / specified wrongly

command -i input.file -o output.file

- Input file format (tab/space. Mac/Windows etc.)
- Software version issue
- Unknown error
 1. Somehow deal with it
 2. Give up and search for another software

Today's question:

How does frequency of mitochondrial variants change from mother to child?



RESEARCH ARTICLE

Germline selection shapes human mitochondrial DNA diversity

[Wei Wei](#)^{1,2}, [Salih Tuna](#)^{3,4}, [Michael J. Keogh](#)¹, [Katherine R. Smith](#)^{5,*}, [Timothy J. Aitman](#)^{6,7}, [Phil L. Be...](#)

[+ See all authors and affiliations](#)

Science 24 May 2019:
Vol. 364, Issue 6442, eaau6520
DOI: 10.1126/science.aau6520

Seven to ten of the mother's thousands of copies of mitochondrial DNA get passed on to each child

Arslan Zaidi and Kateryna Makova, Penn State

Today's question:

How does frequency of mitochondrial variants change from mother to child?

- Start: “FastQ format (sequence data with a quality score)”

```
@M01368:8:000000000-A3GHV:1:1101:6911:8255/1
ATCTGGTTCCTACTTCAGGGCCATAAAACCTAAATAGCCCACACGTTCCCC
+
BCCCCFFFFFFFGGGGGGGGGGHHHHGHGHHHHHHHHHHGGGGGGHHHHGH
M01368:8:000000000-A3GHV:1:1101:14518:9998/1
GTTATTATTATGTCCTACAAGCATTAAATTAACACACTTTAGTAAGTA
+
AAAAAFFFFFFFGGGGGGGGGHGGHHHHGHGHHHHHHHGCCHHHHHHHHHH
M01368:8:000000000-A3GHV:1:1101:18422:19051/1
GTATCCGACATCTGGTTCCTACTTCAGGGTCATAAAACCTAAATAGCCCAC
.
```

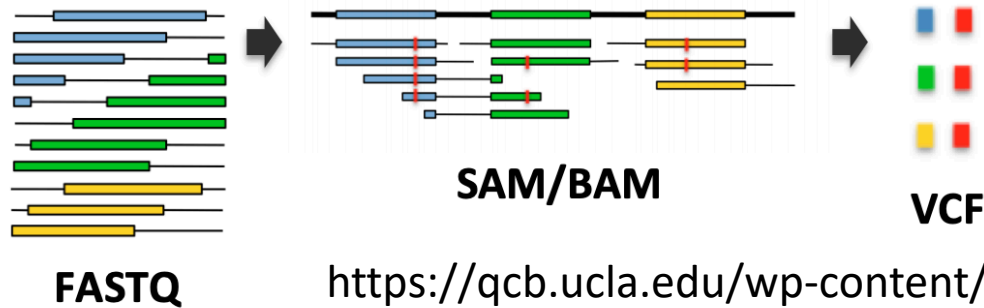
- Goal: Calculate “Allele frequency of variants in Mother and Child”

Position: Where is/are the variants?

Child : How much is the observed frequency?

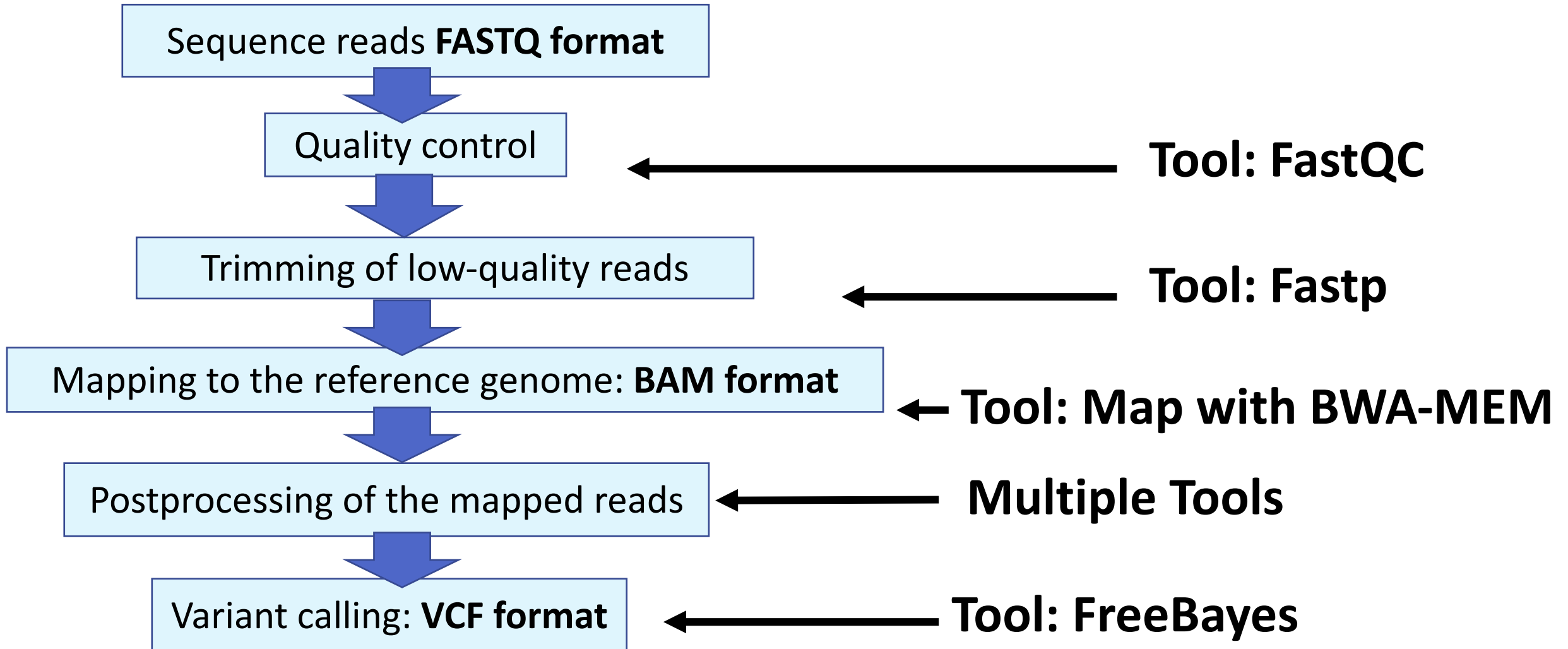
Mother : How much is the observed variant frequency?

Reads-to-variants workflows



https://qcb.ucla.edu/wp-content/uploads/sites/14/2016/03/GATKwr12-1-GATK_primer.pdf

Variant calling workflow



There are numerous tools for the genome analysis

- Which one is the best tool?

- Suitable to your data format
- Works in your environment
- Friendly manual
- Fast
- Follow-up or forum
- Frequently updated
(co-evolution with lab technology)
- Well-cited

Tools for variant identification.

Tools	Input files
Germline caller tools	
Galaxy platform	BAM/SAM
SanGeniX platform*	BAM/SAM
VarScan2	pileup/mpileup
SNVer	BAM/SAM
CRISP	BAM/SAM
GATK(Unified Genotyper)	BAM/SAM
SAMtools	BAM/SAM, FASTA
Somatic callers tools	
Galaxy platform	BAM/SAM
SanGeniX platform*	BAM/SAM
VarScan2	pileup/mpileup
GATK (Somatic Indel Detector)	BAM/SAM
SAM tools	BAM/SAM, FASTA
CNV identification tools	
ExomeCNV	BAM/SAM, pileup + bed + FASTA
CNVnator	BAM/SAM, FASTA
CONTRA	BAM/SAM, FASTA
RDXplorer	BAM/SAM, FASTA
SV identification tools	
GASVPro (GASVPro-HQ)	BAM/SAM
CLEVER	BAM/SAM, FASTA
BreakDancer	BAM/SAM, config file
Breakpointer	BAM/SAM
SV-M	BAM/SAM, FASTA



- **Accessibility**

- Users without programming experience can easily upload/retrieve data, run complex tools and workflows, and visualize data

- **Reproducibility**

- Galaxy captures information so that any user can understand and repeat a complete computational analysis

- **Transparency**

- Users can share or publish their analyses (histories, workflows, visualizations)
- Pages: online Methods for your paper

<https://usegalaxy.no/> (but they are so nice!)

<https://usegalaxy.org/>

<https://training.galaxyproject.org/>

Menu Bar

Browser navigation: back, forward, refresh, home, address bar (usegalaxy.no), search, star, zoom, extensions, user profile, menu.

Galaxy Norway navigation: NeLS, Galaxy Norway, Analyze Data, Workflow, Visualize, Shared Data, Help, User, grid icon, Using 6%

Tools sidebar with search bar (search tools) and categories:

- Get Data
- Send Data
- Collection Operations
- Lift-Over
- Text Manipulation
- Convert Formats
- Filter and Sort
- Join, Subtract and Group
- Fetch Alignments/Sequences
- Operate on Genomic Intervals
- Statistics
- Graph/Display Data
- Phenotype Association
- Interactive Tools
- Mapping
- SAM/BAM

Tools to analyze data

elixir NORWAY logo and "Welcome to usegalaxy.no" header.

Web-based platform for data science research that provides users with a unified, easy-to-use graphical interface to a host of different analysis tools. These tools can be run interactively, one by one, or combined into multi-step workflows that can be executed as a single analysis.

If this is your first time using Galaxy, you might want to read the [Quick Start Guide](#). Additional documentation and tutorials on using Galaxy can be found [here](#).

This Galaxy server has limitations on disc usage, and you have currently used **12.4 GB** of your total quota of **200.0 GB**. To free up disc space, please move your files to the NeLS Storage after you are finished with them. If you require a larger disc quota, contact the [Help Desk](#).

Main window comes here

Tweets by @elixirnorway

ELIXIR Norway @elixirnorway
Ready to contribute on the data side with sequencing in Norway is scaled up with core facilities #ELIXIRvsCOVID19 #COVID19
















ELIXIR Norway @elixirnorway
Few spots left on the @swcarpentry course by ELIXIR Norway and @DigitaltLiv for PhD researchers korbinib.github.io/2021-02-18

History Your workflow

search datasets

- mtDNA
41 shown, 24 deleted, 6 hidden
292.59 MB
- 9: FastQC on data 2: Web page
- 8: FastQC on data 4: Raw Data
- 7: FastQC on data 4: Web page
- 6: FastQC on data 1: Raw Data
- 5: FastQC on data 1: Web page
- 4: mother2.fq
- 3: mother1.fq

Job Status

Colour	Status
 <u>5: Filter on data 1</u>   	Queued
 <u>4: Filter on data 1</u>   	Running
<u>3: Filter on data 1</u>   	OK
 <u>2: Filter on data 1</u>   	Error

Be patient and wait

Let's go to next step

Hmm... Let's examine it

If you get an error, don't panic...

✖ 65: BamLeftAlign on data 61 (alignments)

error
An error occurred with this dataset.

```
/data/part0/tmp/jobs/015/15271/tool_scr  
/data/part0/tmp/jobs/015/15271/tool_scr
```

⌵ ⓘ ?

display at UCSC main test
display with IGV local
display in IGB View

The screenshot shows a red error dialog box. At the top, it has a red 'X' icon and the text '65: BamLeftAlign on data 61 (alignments)'. Below this, it says 'error' and 'An error occurred with this dataset.' followed by two lines of file paths. At the bottom of the dialog, there is a menu with three icons: a downward arrow, an information 'i' icon, and a question mark. A blue box highlights this menu, and a blue callout box points to it with the text 'Click to expand the menu'. Below the dialog, there are three lines of text: 'display at UCSC main test', 'display with IGV local', and 'display in IGB View'. A blue arrow points from the text 'Let's examine it' to the information icon in the menu.

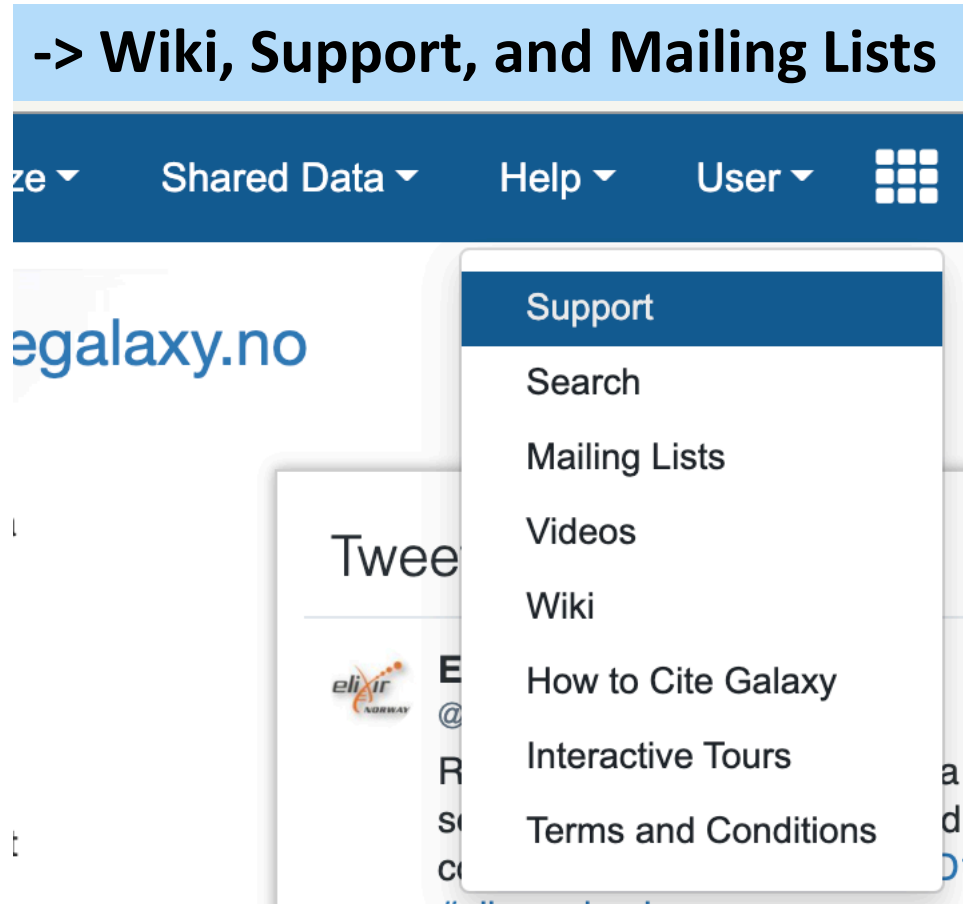
Click to expand the menu

Let's examine it

Common issues:

- Wrong command
- Input data is not suitable
- Software version issue

If can not figure it out, ask colleagues and experts



contact@bioinfo.no

Create new history

The screenshot shows a software interface with a dark blue header bar containing 'Help', 'User', and a hamburger menu icon. To the right of the header is a light blue box with the text 'New history: click +' and a green progress indicator showing 'Using 6%'. Below the header is a 'History' panel with a search bar labeled 'search datasets' and a '+ Create new history' button. The 'Create new history' button is highlighted with an orange border. Below the search bar is an entry for 'Unnamed history (empty)'. A callout box with a black background and white text says 'Click here to rename it as you like' with an arrow pointing to the 'Create new history' button. Below the 'Unnamed history' entry is a blue information box that reads: 'This history is empty. You can load your own data or get data from an external source'. In the background, a tweet from @elixirnorway is visible, mentioning 'Ready to contribute on the data side...' and 'sequencing in Norway is scaled up with...'. The tweet text is partially obscured by the interface elements.

Let's get started...

- Get the data!

Mitochondrial DNA of the child and the mother

https://zenodo.org/record/1251112/files/raw_child-ds-1.fq

https://zenodo.org/record/1251112/files/raw_child-ds-2.fq

https://zenodo.org/record/1251112/files/raw_mother-ds-1.fq

https://zenodo.org/record/1251112/files/raw_mother-ds-2.fq

- Q: Why there are two data per individual?

Import data set

1. Click here to import data set

Download from web or upload from disk

Regular Composite Collection Rule-based

You added 1 file(s) to the queue. Add more files or click 'Start' to proceed.

Name	Size	Type	Genome	Settings	Status
New File	-	Auto-de...	----- Additional ...	0%	

Download data from the web by entering URLs (one per line) or directly paste content.

3. Paste the four URLs here all at once -> start

2. Click here

Type (set all): Auto-detect Additional ...

Choose local files Paste/Fetch data Start Select Pause Reset Close

Rename the data set as you like

The screenshot displays a data management interface. At the top, there are buttons for 'Edit attributes', 'Auto-detect', and 'Save'. The 'Save' button is highlighted with an orange box. Below this, the 'Name' field contains 'raw_child-1.fq' and is also highlighted with an orange box. The 'Info' section shows 'uploaded fastqsanger file'. The 'Annotation' section is empty. Below the annotation, there is a note: 'Add an annotation or notes to a dataset; annotations are available when a history is viewed.' The 'Database/Build' field is partially visible. On the right side, a list of datasets is shown, with the first item highlighted by an orange box. The first item is '1: https://zenodo.org/record/1251112/files/raw_child-ds-1.fq' and has an 'Edit attributes' button next to it.

Edit attributes Auto-detect Save 535.27 MB

Name
raw_child-1.fq

Info
uploaded fastqsanger file

Annotation

Add an annotation or notes to a dataset; annotations are available when a history is viewed.

Database/Build

7: FastQC on data 2: Web page

5: FastQC on data 1: Web page

4: https://zenodo.org/record/1251112/files/raw_mother-ds-2.fq

3: https://zenodo.org/record/1251112/files/raw_mother-ds-1.fq

2: https://zenodo.org/record/1251112/files/raw_child-ds-2.fq

1: https://zenodo.org/record/1251112/files/raw_child-ds-1.fq Edit attributes





Let's examine the FastQ file






This dataset is large and only the first megabyte is shown below.

[Show all](#) | [Save](#)






```
@M01368:8:000000000-A3GHV:1:1101:6911:8255/1
ATCTGGTTCCTACTTCAGGGCCATAAAACCTAAATAGCCACACGTTCCCCTTAAATAAGACATCACGATGGATCACAGGTCTATCACCTATT
+
BCCCCFFFFFFFGGGGGGGGGGGHHHHGHGHHHHHHHHHHGGGGGGHHHHGHGHHHHHHHHHHGHHHHHHGGHGGHHHHGHGHHHHFHHGHHHHHHHHHC
@M01368:8:000000000-A3GHV:1:1101:14518:9998/1
GTTATTATTATGTCCTACAAGCATTAAATTAACACACTTTAGTAAGTATGTTGCGCTGTAATATTGAACGTAGGTGCGATAAATAATAGGA/
+
AAAAAFFFFFFFGGGGGGGGGGGHGGHHHHGHGHHHHHHHGCCHHHHHHHHHHHHHHHHHGGGGGGHHHHHHHHHHGHGHHGFHFE5BGEEHFGGGHHHHHHH
@M01368:8:000000000-A3GHV:1:1101:18422:19051/1
GTATCCGACATCTGGTTCCTACTTCAGGGTCATAAAACCTAAATAGCCACACGTTCCCCTTAAATAAGACATCACGATGGA
+
CCCCCFDDDDDFGGGGGGGGGGHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHH
@M01368:8:000000000-A3GHV:1:1101:25545:21098/1
ATTAATTAACACACTTTAGTAAGTATGTTGCGCTGTAATATTGAACGTAGGTGCGATAAATAATAGGATAAGGCAGGAATCAAAGACAGATAC
+
33AA?DFD5BDFGGGFEBDGEHGEHGEHCEGGHHCHGHHFFHHGFGAGE53FF2FAFFGDE5FFFE5GFBFGAEE1GHHHGHHEHE3FGHF
@M01368:8:000000000-A3GHV:1:1101:5446:12248/1
AATTAACACACTTTAGTAAGTATGTTGCGCTGTAATATTGAACGTAGGTGCGATAAATAATAGGATGAGGCAGGAATCAAAGACAGATACTGCC
+
CCCCDFFFFCFGGGGGGGGGFGHHHHHHGGGGHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHC
@M01368:8:000000000-A3GHV:1:1101:5001:2172/1
```




History    

search datasets  

Genome
4 shown
85.25 MB   

Click the eye icon to view data

4: <https://zenodo.org/rec>   
3: https://zenodo.org/rec/1251112/files/raw_mother-ds-1.fq    **View data**

2: https://zenodo.org/rec/1251112/files/raw_child-ds-2.fq   
26

Label

@M01368:8:00000000-A3GHV:1:1101:6911:8255/1

ATCTGGTTCCTACTTCAGGGCCATAAAACCTAAATAGCCCACACGTTCCCCTTAAATA

+

BCCCCFFFFFFGGGGGGGGGGGGHHHHGHHGHHHHHHHHGGGGGGHHHHGHHHHHHHH

Sequence

Quality score as ASCII symbol

Quality scores

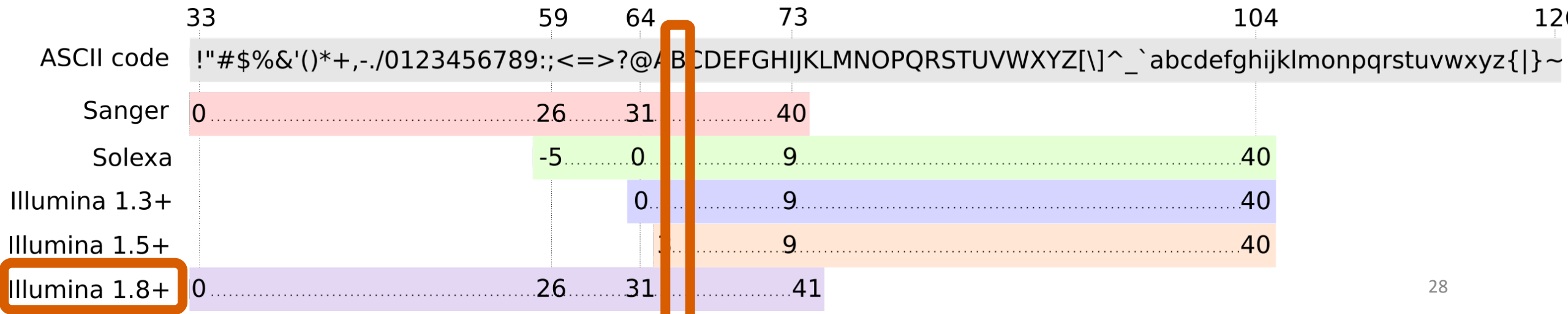
@M01368:8:000000000-A3GHV:1:1101:6911:8255/1

ATCTGGTTCCTACTTCAGGGCCATAAAACCTAAATAGCCCACACGTTCCCCTTAAATA

+

BCCCCFFFFFFFGGGGGGGGGGGHHHHGHHGHHHHHHHHHGGGGGGHHHHGHHHHHHHHI

Phred Quality Score	Probability of incorrect base call	Base call accuracy
10	1 in 10	90%
20	1 in 100	99%
30	1 in 1000	99.9%
40	1 in 10000	99.99%
50	1 in 100000	99.999%



Run FastQC for the quality check

1. search "FastQC"

Tools

FastQC

Show Sections

FastQC Read Quality reports

fastp - fast all-in-one preprocessing for FASTQ files

2. Multiple datasets

Short read data from your current history

Multiple datasets

4: https://zenodo.org/record/1251112/files/raw_mother-ds-2.fq

3: https://zenodo.org/record/1251112/files/raw_mother-ds-1.fq

2: https://zenodo.org/record/1251112/files/raw_child-ds-2.fq

1: https://zenodo.org/record/1251112/files/raw_child-ds-1.fq

3. Select the four data

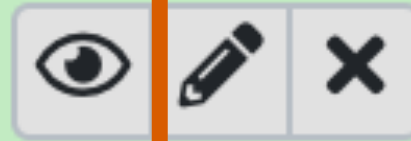
This is a batch mode input field. Separate jobs will be triggered for each dataset selection.

→ Execute

Run FastQC for the quality check

Click the “eye” icon to view data

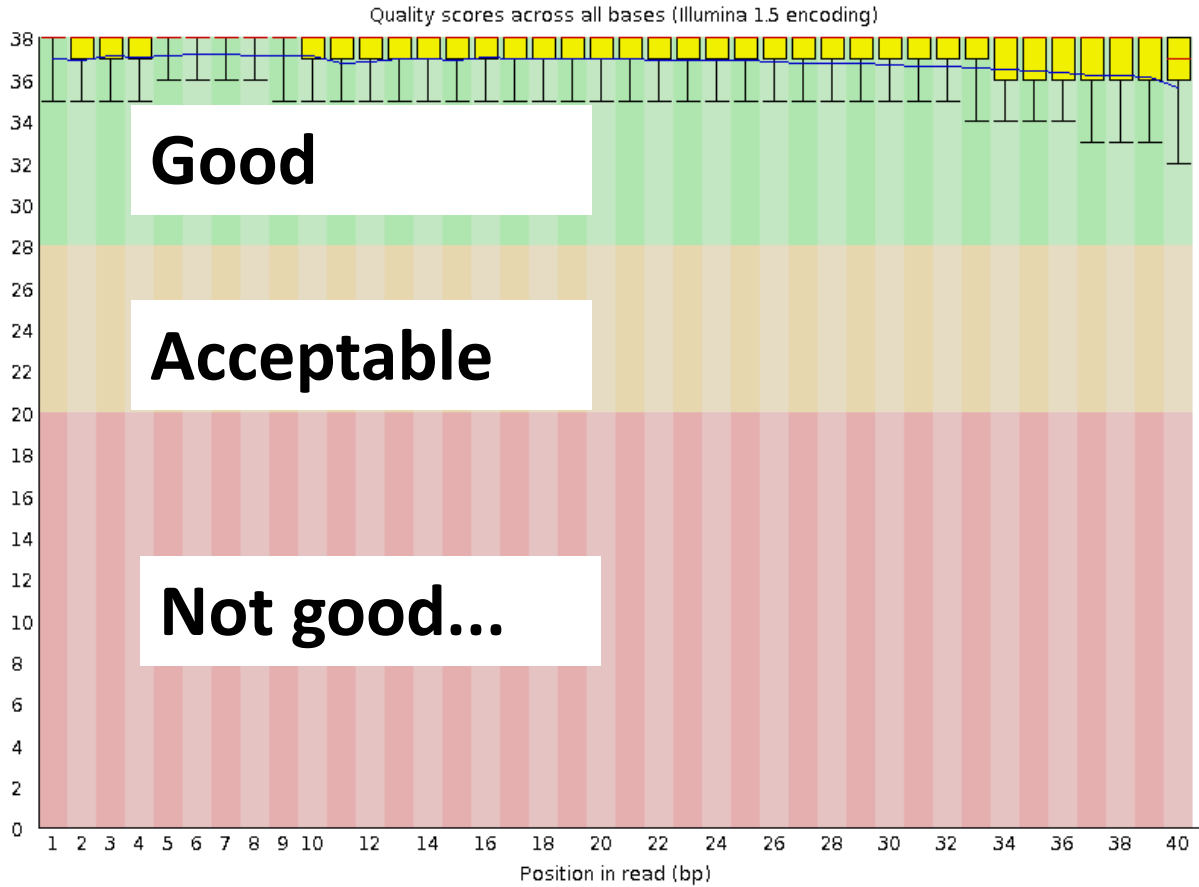
9: FastQC on data 2: We
bp page



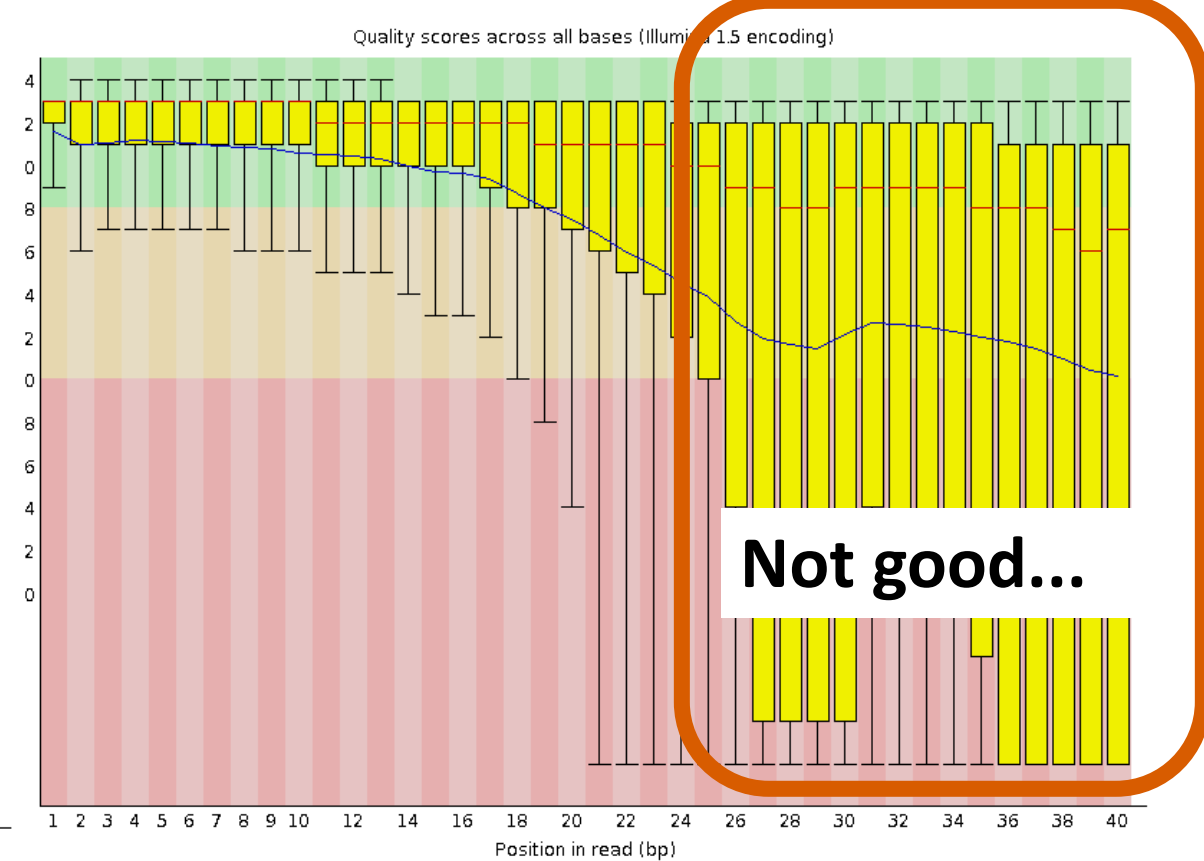
Good Illumina data

Bad Illumina data

Quality



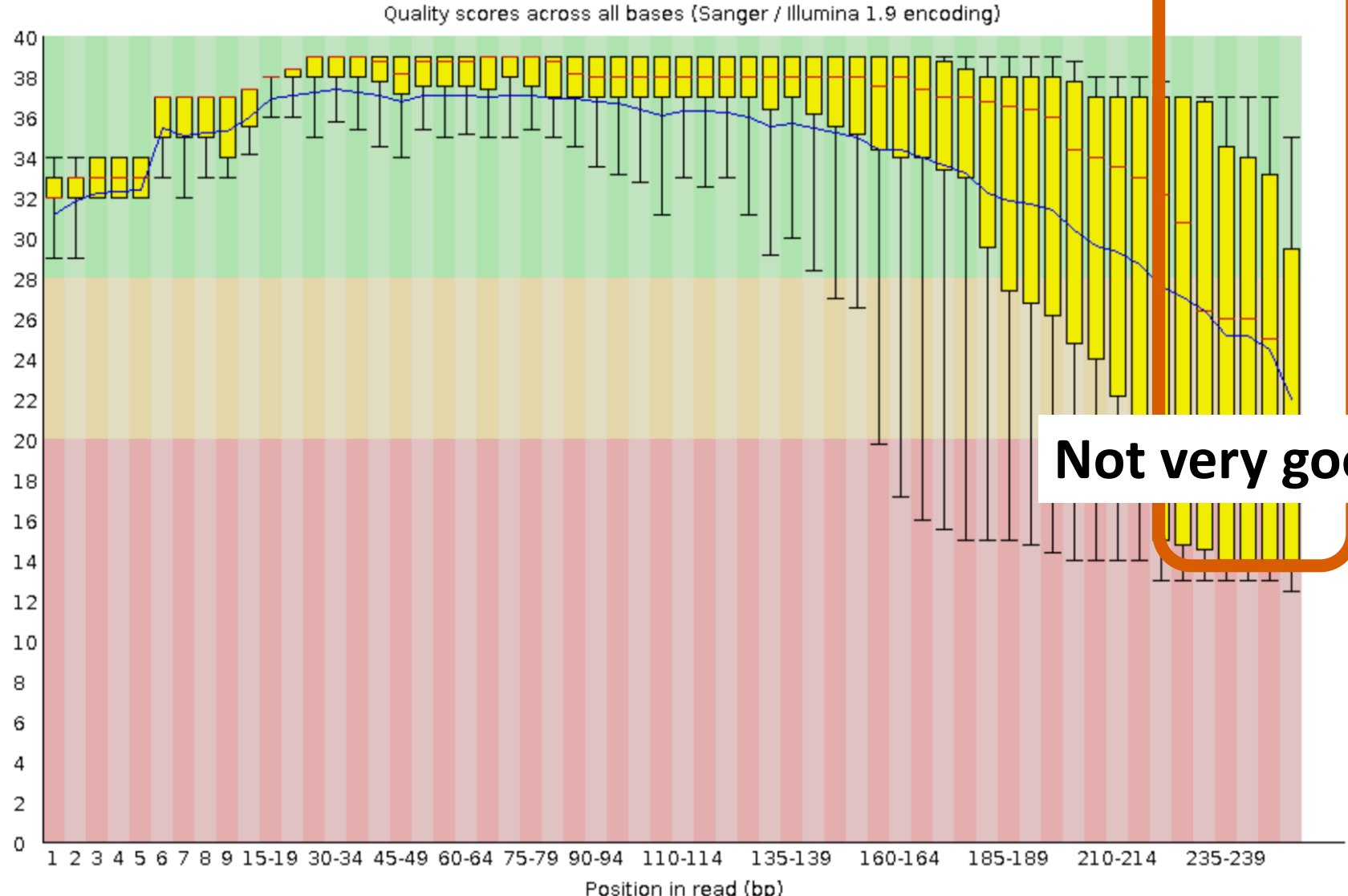
Nucleotide position in read



Nucleotide position in read

FastQC Data Interpretation

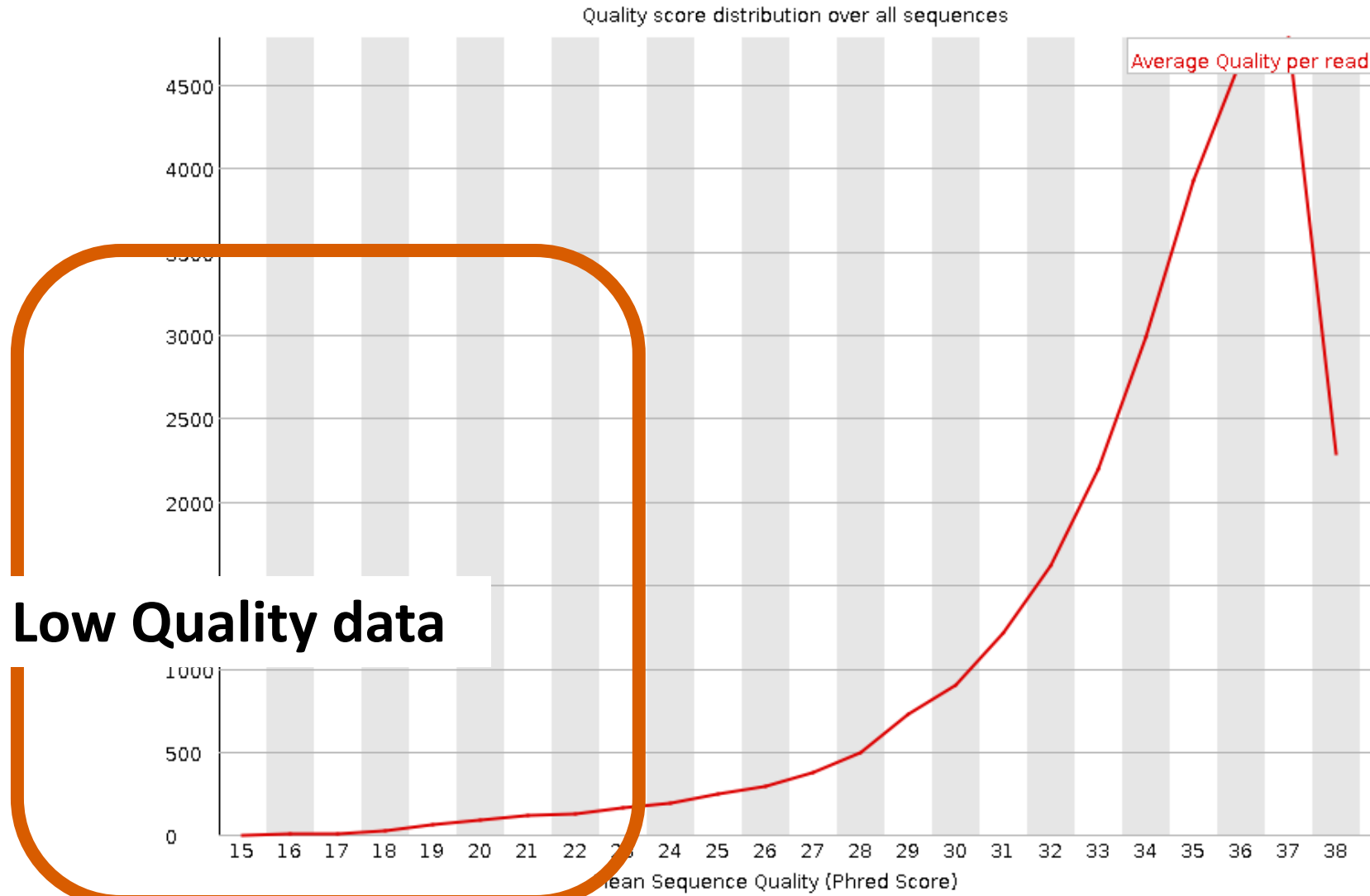
Per base sequence quality



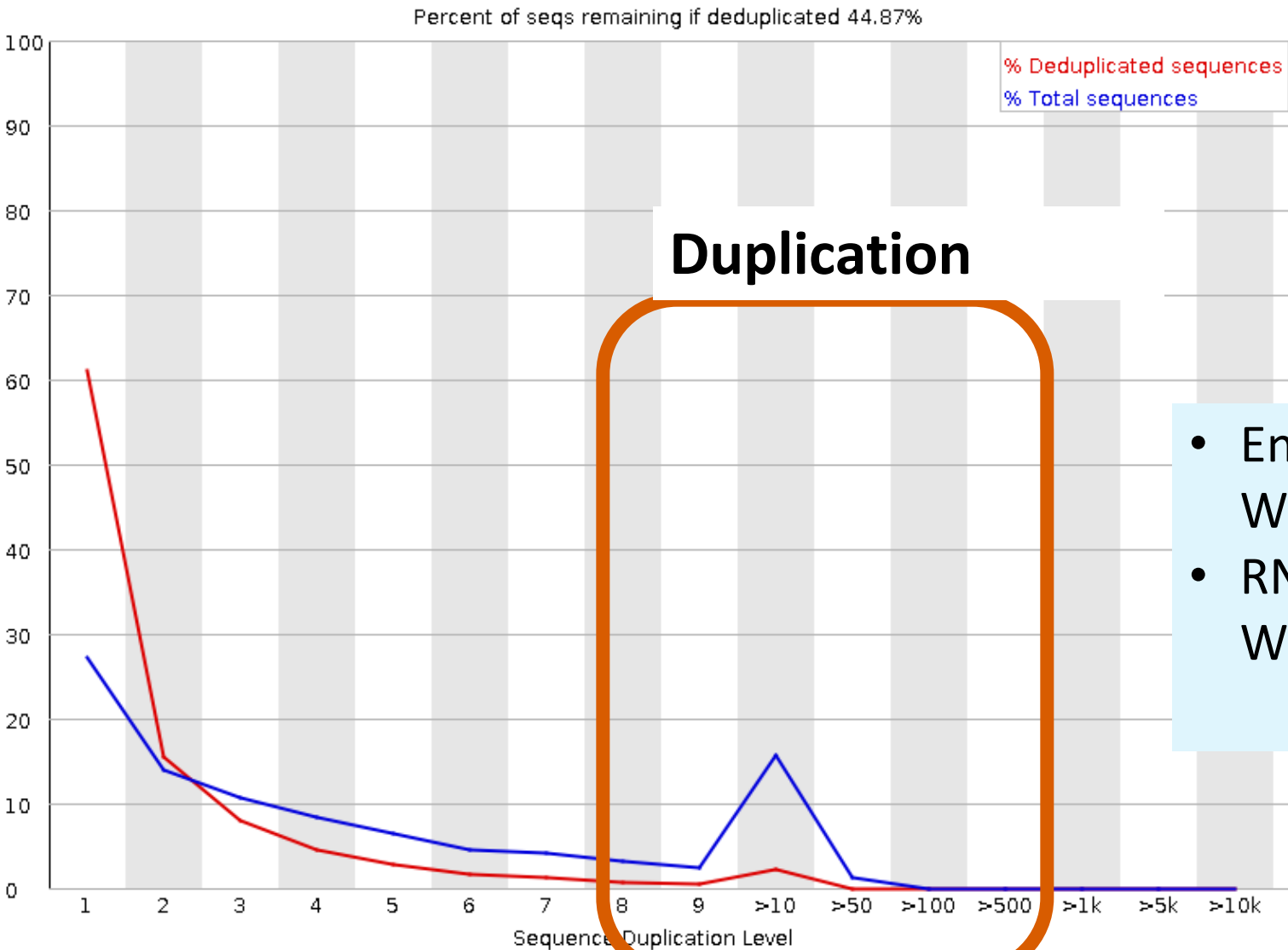
Not very good...

FastQC Data Interpretation

Per sequence quality scores



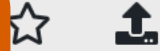
Sequence Duplication Levels



- Enrichment bias (PCR over amplification). Will be removed in a future step
- RNA-sequencing We will handle that data next time

Trimming low quality data with fastp

Tools



fastp



Show Sections

fastp - fast all-in-one preprocessing for FASTQ files

fastpca - dimensionality reduction of MD simulations

Pedigree check for mendelian errors

WORKFLOWS

All workflows

fastp - fast all-in-one preprocessing for FASTQ files (Galaxy Version 0.20.1+galaxy0)

☆ Favorite

▼ Options

Single-end or paired reads

Paired

Input 1



1: child1.fq



Input FASTQ file #1 (-i)

Input 2

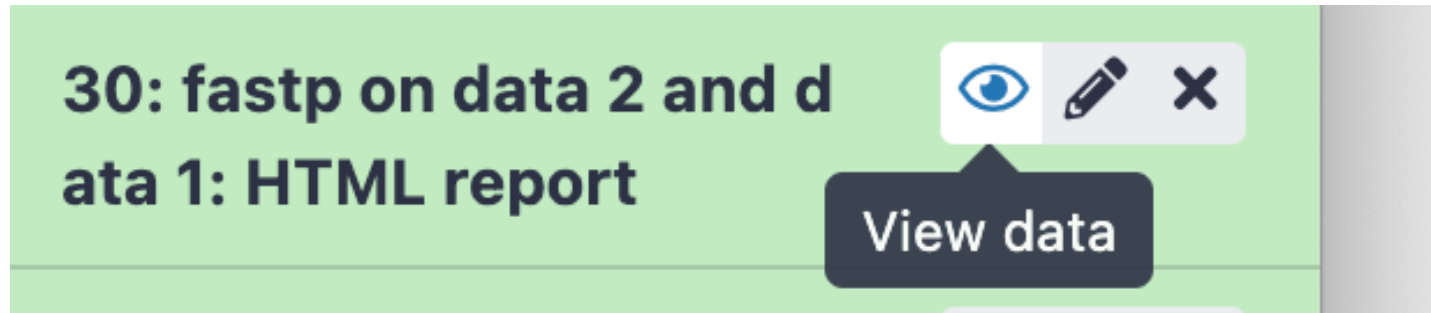


2: child2.fq



Input FASTQ file #2 (-l)

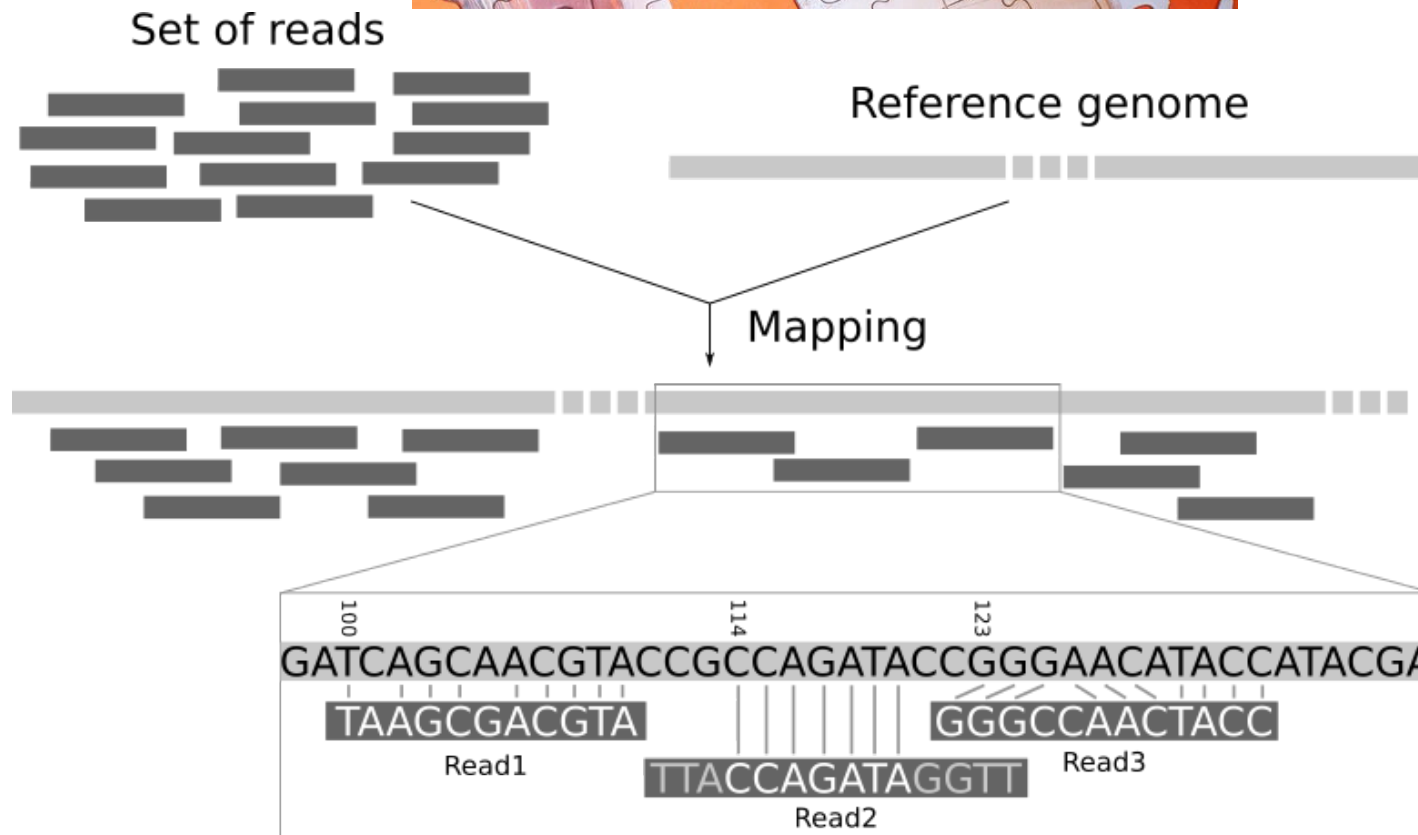
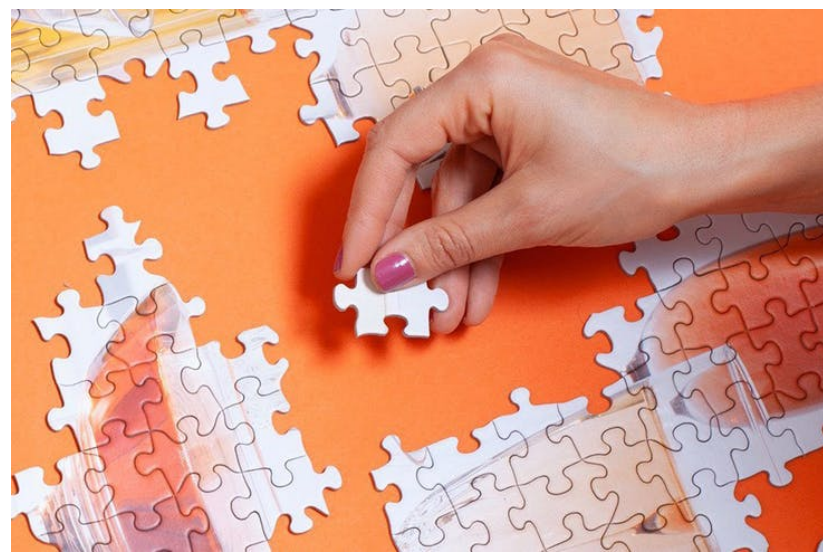
Examine the filtering result of fastp



Filtering result

reads passed filters:	54.352000 K (98.442368%)
reads with low quality:	726 (1.314932%)
reads with too many N:	134 (0.242701%)

Mapping!



Map the cleaned reads to the reference genome with BWA-MEM

Tools ☆ ↑

Map with BWA-MEM ✕

Show Sections

Map with BWA - map short reads (< 100 bp) against reference genome

Map with BWA-MEM - map medium and long reads (> 100 bp) against reference genome

Map with **BWA-MEM** ☆ Favorite 🔄 Versions ▾ Options

- map medium and long reads (> 100 bp) against reference genome (Galaxy Version 0.7.17.1)

When you select a reference genome from your history or use built-in index?

Use a built-in genome index ▾

Built-ins were indexed using default options. See `Indexes` section of help below

Using reference genome

Human (Homo sapiens) (b38): hg38 ▾

Select genome from the list

Single or Paired-end reads

Paired ▾

Select between paired and single end data

Select first set of reads

57: fastp-child1 ▾

Specify dataset with forward reads

Select second set of reads

58: fastp-child2 ▾

Specify dataset with reverse reads

assign the value

Platform/technology used to produce the reads (PL)

ILLUMINA ▾

Merging two BAM data with MergeSamFiles

The image shows a screenshot of the Galaxy web interface. On the left, a 'Tools' sidebar is visible with 'MergeSamFiles' highlighted by an orange box. Below the sidebar, a 'Show Sections' button is present. The main content area displays the 'MergeSamFiles' tool details, including its description: 'merges multiple SAM/BAM datasets into one (Galaxy Version 2.18.2.1)'. Below this, a section titled 'Select SAM/BAM dataset or dataset collection' shows a list of datasets. Two datasets are highlighted with an orange box: '64: mapped.mother' and '63: mapped.child'. Other visible datasets include '52: Filter on data 50: Filtered BAM', '50: BamLeftAlign on data 49 (alignments)', and '49: MarkDuplicates on data 47: MarkDupl'.

After this process, sample names are now incorporated and you don't have to keep renaming data sets.

Removing the PCR duplicates

Tools

MarkDuplicates

Show Sections

MarkDuplicatesWithMateCigar
examine aligned records in BAM datasets to locate duplicate molecules

MarkDuplicates examine aligned records in BAM datasets to locate duplicate molecules

AddOrReplaceReadGroups add or replaces read group information

FastqToSam convert Fastq data into unaligned BAM

Map with BWA - map short reads (< 100 bp) against reference genome

Map with BWA-MEM - map medium and long reads (> 100 bp) against reference genome

QualiMap BamQC

WORKFLOWS

All workflows

MarkDuplicates examine aligned records in BAM datasets to locate duplicate molecules (Galaxy Version 2.18.2.2)

Favorite Versions Options

Select SAM/BAM dataset or dataset collection

65: MergeSamFiles on data 64 and data 63: M... with

If empty, upload or

Comment

+ Insert Comment

You can provide multiple comments

If true do not write appropriate flags

Yes No

REMOVE_DUPLICATES

Assume the input file is already sorted

Yes No

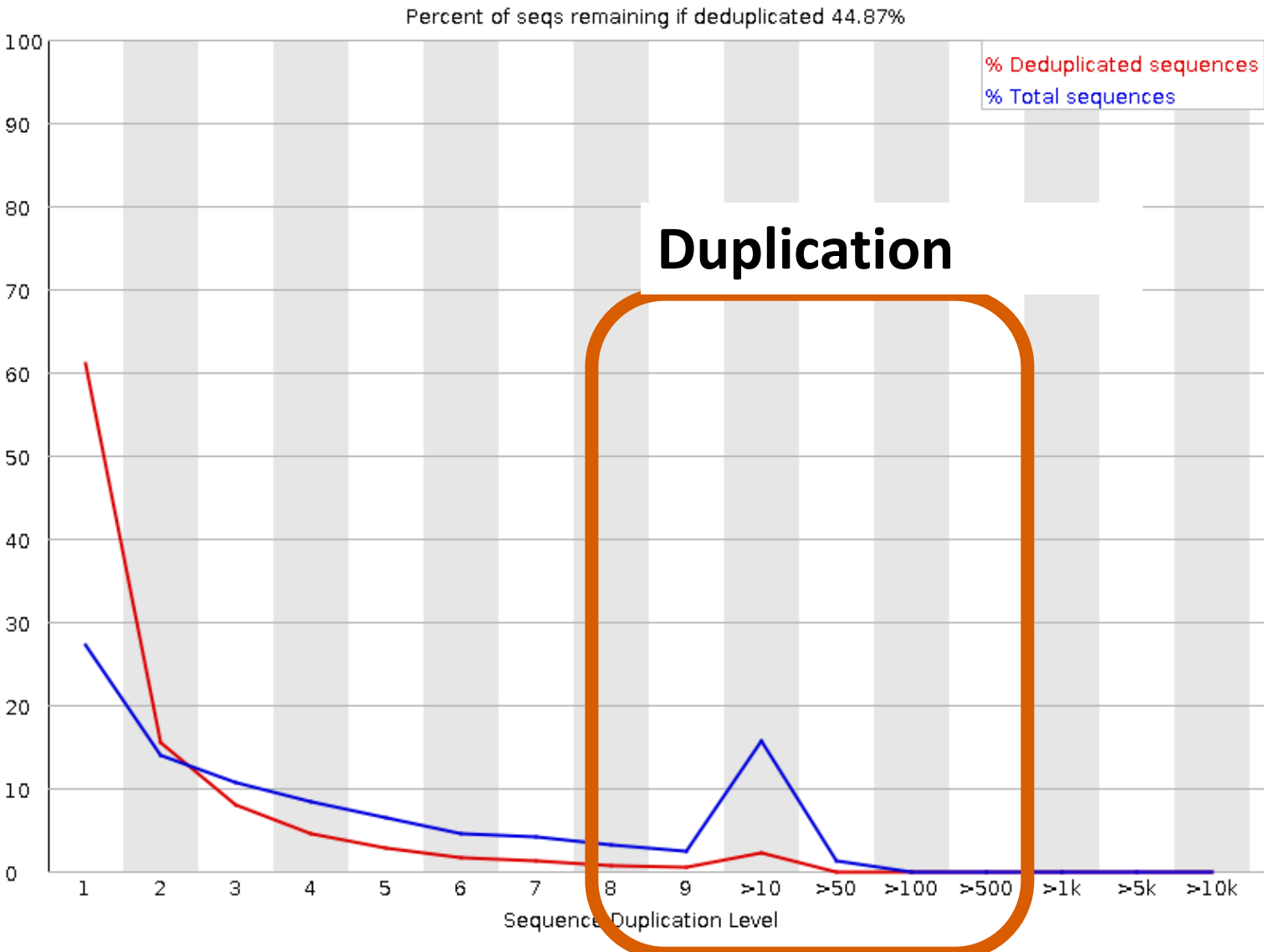
ASSUME_SORTED; default=True

The scoring strategy for choosing the non-duplicate among candidates

SUM_OF_BASE_QUALITIES

- “The scoring strategy for choosing the non-duplicate among candidates”: **SUM_OF_BASE_QUALITIES**
- “The maximum offset between two duplicate clusters in order to consider them optical duplicates”: **100**
- “Select validation stringency”: **Lenient**

Review: Sequence Duplication Levels



Left-aligning indels (insertions/deletions)

	Alignment	Variant Call
Reference sequence Your sequence (Alternative sequence)	GGGCACACACAGGG GGGCAC--ACAGGG	Ref: CAC Alt: C
Reference sequence Your sequence	GGGCACACACAGGG GGGCA--CACAGGG	Ref: ACA Alt: A
Reference sequence Your sequence	GGGCACACACAGGG GGG--CACACAGGG	Ref: GCA Alt: G

Same

Different Calls!

The deletion is left-aligned

Left-aligning indels (insertions/deletions)

BamLeftAlign

Show Sections

BamLeftAlign indels in BAM datasets

WORKFLOWS

All workflows

BamLeftAlign indels in BAM datasets (Galaxy Version 1.3.1) ☆ Favorite 🔄 Versions ▾ Options

Choose the source for the reference genome

Locally cached ▾

Select alignment file in BAM format

📄 📄 📁 **67: MarkDuplicates on data 65: MarkDuplica...** 📁

Using reference genome

Human (Homo sapiens): hg38 ▾
(--fasta-reference)

Maximum number of iterations

5

Iterate the left-realignment no more than this many times (--max-iterations)



Email notification


Yes No


Send an email notification when the job completes.

✓ Execute



Filtering BAM file

Tools  

Filter BAM datasets on a variety of 

 Show Sections


Filter BAM datasets on a variety of attributes **updated**

Filter BAM datasets on a variety of attributes (Galaxy Version 2.4.1)  Favorite  Versions

BAM dataset(s) to filter

   68: BamLeftAlign on data 67 (alignments)

Set the following four filters

4: Filter 

Select BAM property to filter on

reference

Filter on the reference name for the read

chrM

You can use ! (not) in your expression

+ Insert Filter

+ Insert Condition

Would you like to set rules?

Yes No

Allows complex logical constructs. See Example 4 below.

Email notification

Yes No

Send an email notification when the job completes.

✓ Execute

What is does

1. “Select BAM property to filter on”: mapQuality
“Filter on read mapping quality (phred scale)”: ≥ 20
2. “Select BAM property to filter on”: isPaired
“Selected mapped reads”: Yes
3. “Select BAM property to filter on”: isProperPair
“Select reads with mapped mate”: Yes
4. “Select BAM property to filter on”: reference
“Select reads with mapped mate”: chrM

Calling variants with FreeBayes

Tools ☆ ↑

FreeBayes ✕

Show Sections

FreeBayes bayesian genetic variant detector

BamLeftAlign indels in BAM datasets

SnEff build: database from Genbank or GFF record

Map with BWA - map short reads (< 100 bp) against reference genome

Map with BWA-MEM - map medium and long reads (> 100 bp) against reference genome

Call specific mutations in reads: Looks for reads with mutation at known positions and calculates frequencies and stats.

DCS mutations to SSCS stats: Extracts all tags from the single stranded consensus sequence (SSCS) bam file that carry a mutation at the same position a mutation is called in the duplex consensus sequence (DCS) and calculates their frequencies

DCS mutations to tags/reads: Extracts all tags that carry a mutation in the duplex consensus sequence (DCS)

WORKFLOWS

All workflows

FreeBayes bayesian genetic variant detector (Galaxy Version 1.3.1) ☆ Favorite Versions Options

Choose the source for the reference genome

Locally cached

Run in batch mode?

Run individually
 Merge output VCFs

Selecting individual mode will generate one VCF dataset for each input BAM dataset. Selecting the merge option will produce one VCF dataset for all input BAM datasets

BAM dataset

70: Filter on data 68: Filtered BAM

Using reference genome

Human (Homo sapiens): hg38

Limit variant calling to a set of regions?

Limit to region

Sets --targets or --region options

Region Chromosome

chrM

(--region)

Region Start

1

Region End

16569

Calling variants with FreeBayes

Choose parameter selection level

5. Full list of options

Population model options

Set population model options

Sets `--theta`, `--ploidy`, `--pooled-discrete`, and `--pooled-continuous` options

The expected mutation rate or pairwise nucleotide diversity among the population under analysis

0.001

This serves as the single parameter to the Ewens Sampling Formula prior model (`--theta`)

Set ploidy for the analysis

1

(`--ploidy`)

Assume that samples result from pooled sequencing

Yes No

Model pooled samples using discrete genotypes across pools. When using this flag, set `--ploidy` to the number of alleles in each sample or use the `--cnv-map` to define per-sample ploidy (`--pooled-discrete`)

Output all alleles which pass input filters, regardless of genotyping outcome or model

Yes No

Allelic scope options

Set allelic scope options

Sets `-l`, `i`, `-X`, `-u`, `-n`, `--haplotype-length`, `--min-repeat-size`, `--min-repeat-entropy`, and `--no-partial-observations` options

Ignore SNP alleles

Yes No

(`--no-snps`)

Ignore indels alleles

Yes No

(`--no-indels`)

Ignore multi-nucleotide polymorphisms, MNPs

Yes No

(`--no-mnps`)

Ignore complex events (composites of other classes)

Yes No

(`--no-complex`)

Calling variants with FreeBayes

Input filters

Set input filters

Sets -4, -m, -q, -R, -Y, -Q, -U, -z, -\$, -e, -0, -F, -C, -3, -G, and -! options

Include duplicate-marked alignments in the analysis

Yes No

(--use-duplicate-reads)

Exclude alignments from analysis if they have a mapping quality less than

20

(--min-mapping-quality)

Exclude alleles from analysis if their supporting base quality less than

30

How many variants did you get?

**71: FreeBayes on data 70
(variants)**



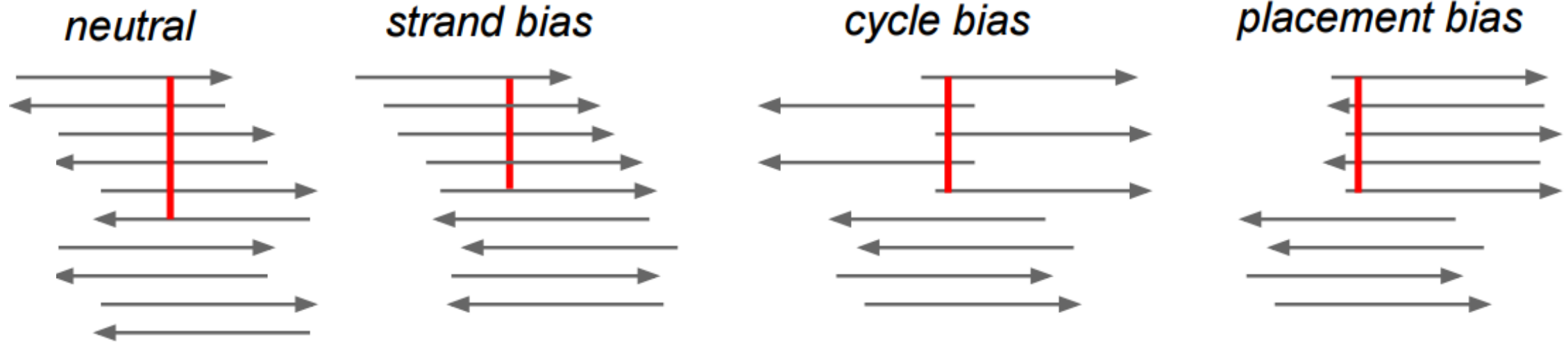
31 lines, 516 comments

format: **vcf**, database: **hg38**

We're almost there...

Filtering variants:

“false-positive” variants due to read-alignment bias example



Filtering variants

The screenshot shows the Galaxy VCFfilter tool interface. On the left, the 'Tools' sidebar has 'VCFfilter' highlighted with an orange box. Below it is a 'Show Sections' button. The main panel shows the tool description: 'VCFfilter: filter VCF data in a variety of attributes (Galaxy Version 1.0.0_rc3+galaxy3)'. Below this, the 'VCF dataset to filter' dropdown is set to '53: FreeBayes on data 52 (variants)', also highlighted with an orange box. Under 'more filters', the 'Select the filter type' is 'Info filter (-f)' and the 'Specify filtering value' is 'SRP > 20', both highlighted with orange boxes. A light green callout box on the right contains the text: 'Set the following filters' followed by five filter rules: 'Specify filtering value': SRP > 20, 'Specify filtering value': SAP > 20, 'Specify filtering value': EPP > 20, 'Specify filtering value': QUAL > 20, and 'Specify filtering value': DP > 20.

Filtering FreeBayes VCF for strand bias (SPR and SAP), placement bias (EPP), variant quality (QUAL), and depth of coverage (DP).

How many variants survived?

72: VCFfilter: on data 71



2 lines, 517 comments

format: **vcf**, database: **hg38**

Reformat the VCF file

Tools ☆ ↑

VCFtoTab-delimited ✕

Show Sections

VCFtoTab-delimited: Convert VCF data into TAB-delimited format

Tabular-to-FASTA converts tabular file to FASTA format

Kraken assign taxonomic labels to sequencing reads

Compute an expression on every row

Kernel Canonical Correlation Analysis

Compare two Datasets to find common or distinct rows

Nonpareil to estimate average coverage and generate Nonpareil curves

Megablast compare short reads against htgs, nt, and wgs databases

Reverse columns in a tabular file

VCFtoTab-delimited: Convert VCF data into TAB-delimited format (Galaxy Version 1.0.0_rc3+galaxy0) ☆ Favorite Versions Options

Select VCF dataset to convert

72: VCFfilter: on data 71

Report data per sample

Yes No

-g option

Fill empty fields with

Nothing

-n option

Email notification

Yes No

Send an email notification when the job completes.

Execute

Reformat the VCF file

Tools



Cut



Show Sections

Condense consecutive characters

Cut columns from a table (cut)

seqtk_cutN cut sequence at long N

Clearcut Generate a tree using relaxed neighbor joining

Generate all possible combination of STR length profile of the consecutive allele from given error profile

Cutadapt Remove adapter sequences from Fastq/Fasta

cutseq Removes a specified section from a sequence

Differential Cleavage

Cut columns from a table (Galaxy Version 1.0.2)

Favorite

Options

Cut columns

c2,c4,c5,c52,c54,c55

Delimited by

Tab

From



73: VCFtoTab-delimited: on data 72



Email notification

Yes

No

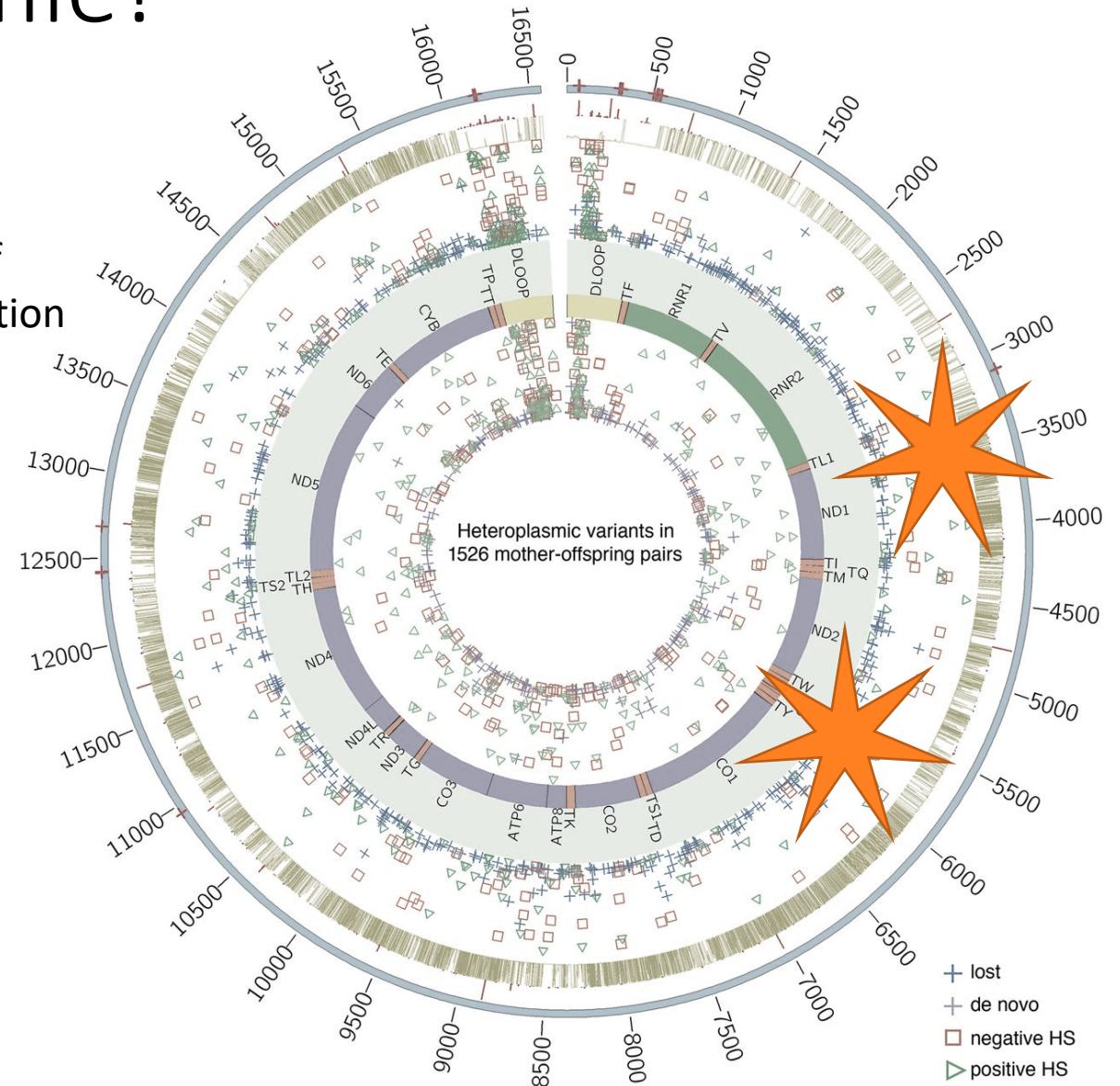
Send an email notification when the job completes.

Execute

Take a look on the VCF file!

Science 24 May 2019:
DOI: 10.1126/science.aau6520

Position	Ref	Alt	Sample	AO Number of alternative observations	DP Depth of this position
1	2	3	4	5	6
3243	A	G	fastp-child	666	995
3243	A	G	fastp-mother	651	1949
5539	A	G	fastp-child	77	296
5539	A	G	fastp-mother	302	489

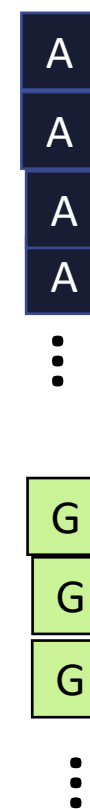


Take a look on the VCF file...

Position	Ref	Alt	Sample	AO Number of alternative observations	DP Depth of this position
1	2	3	4	5	6
3243	A	G	fastp-child	666	995
3243	A	G	fastp-mother	651	1949
5539	A	G	fastp-child	77	296
5539	A	G	fastp-mother	302	489

Position 3243
Reference: A

Child



Total Read Counts
995

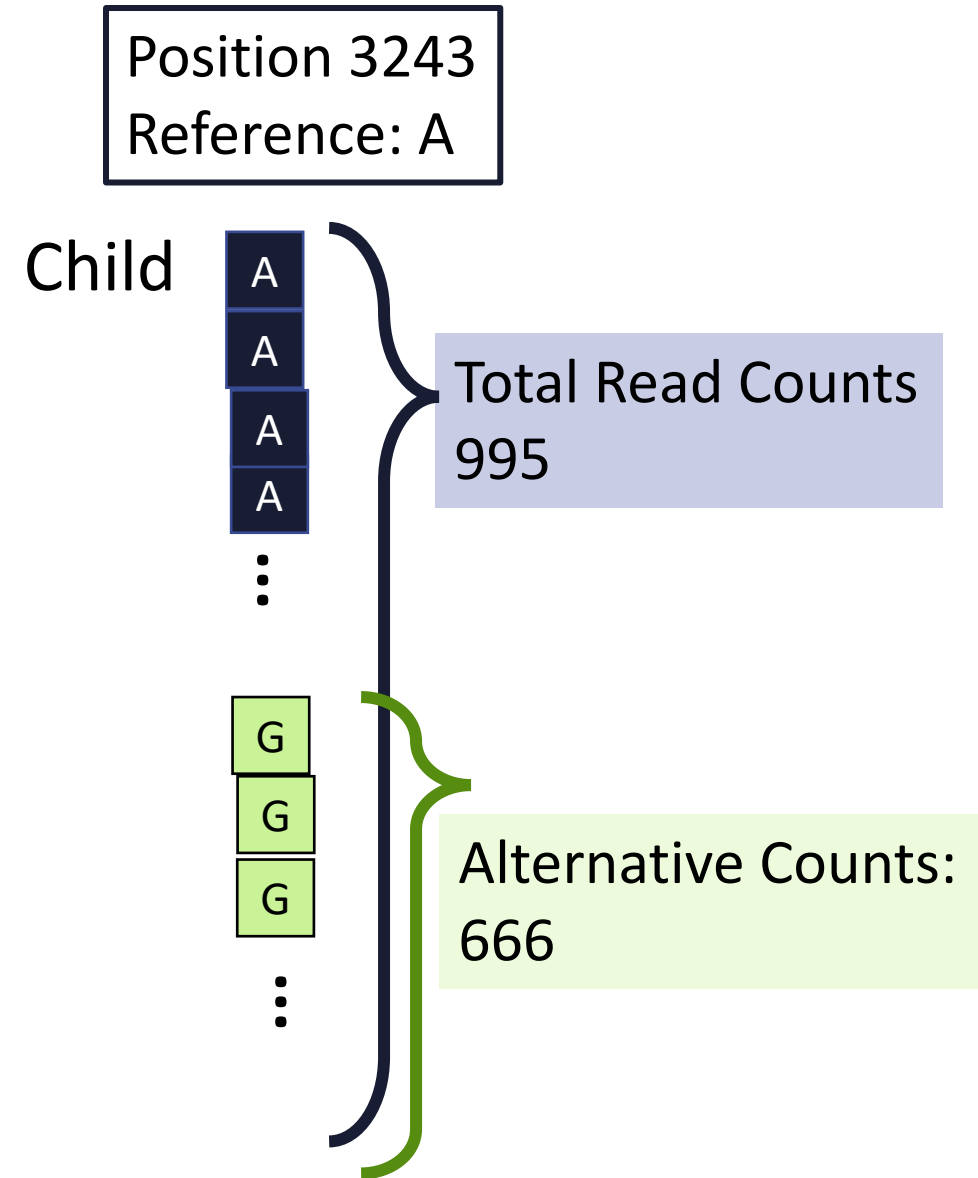
Alternative Counts:
666

Take a look on the VCF file...

Reference Allele: A, Read counts = 995 - 666 = 329
Alternative Allele: G, Read counts = 666



Reference Allele Frequency: $A = 329 / 995 = 0.33$
Alternative Allele Frequency : $G = 666 / 995 = 0.67$



Allele frequency of variants in Mother and Child

Position 3243

Child : $G = 0.67$

Mother : $G = ???$

Position 5538

Child : $G = ???$

Mother : $G = ???$

Allele frequency of variants in Mother and Child

Position 3243

Child : $G = 0.67$

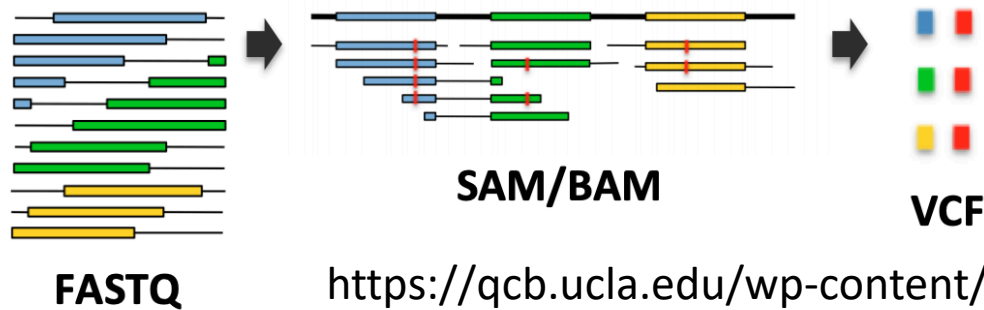
Mother : $G = 0.33$

Position 5538

Child : $G = 0.26$

Mother : $G = 0.61$

Summary: Reads-to-variants workflows



https://qcb.ucla.edu/wp-content/uploads/sites/14/2016/03/GATKwr12-1-GATK_primer.pdf

Hands-on materials are edited from:

- **[Quality Control]**
(<https://training.galaxyproject.org/training-material/topics/sequence-analysis/tutorials/quality-control/tutorial.html>)
- **[Mapping]**
(<https://training.galaxyproject.org/training-material/topics/sequence-analysis/tutorials/mapping/tutorial.html>)
- **[Variant Analysis]**
(<https://training.galaxyproject.org/training-material/topics/variant-analysis/tutorials/non-dip/tutorial.html>)

Preview of the next time: RNA-sequencing