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.

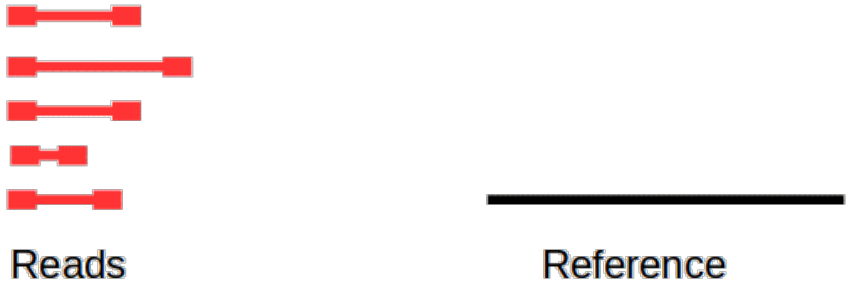
Today's schedule:

- [**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

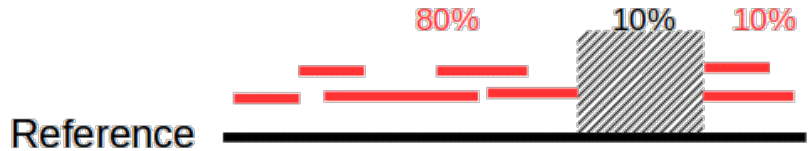
(A) **coverage in terms of redundancy**



$$C = \frac{\#\ \text{sequenced bases}^1\ (=\ \#\ \text{bases of all mapped reads})}{\#\ \text{bases of reference}}$$

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

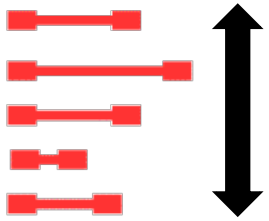
(B) **percentage of coverage**



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

sequencing depth

(C)



Sequencing depth= total read number

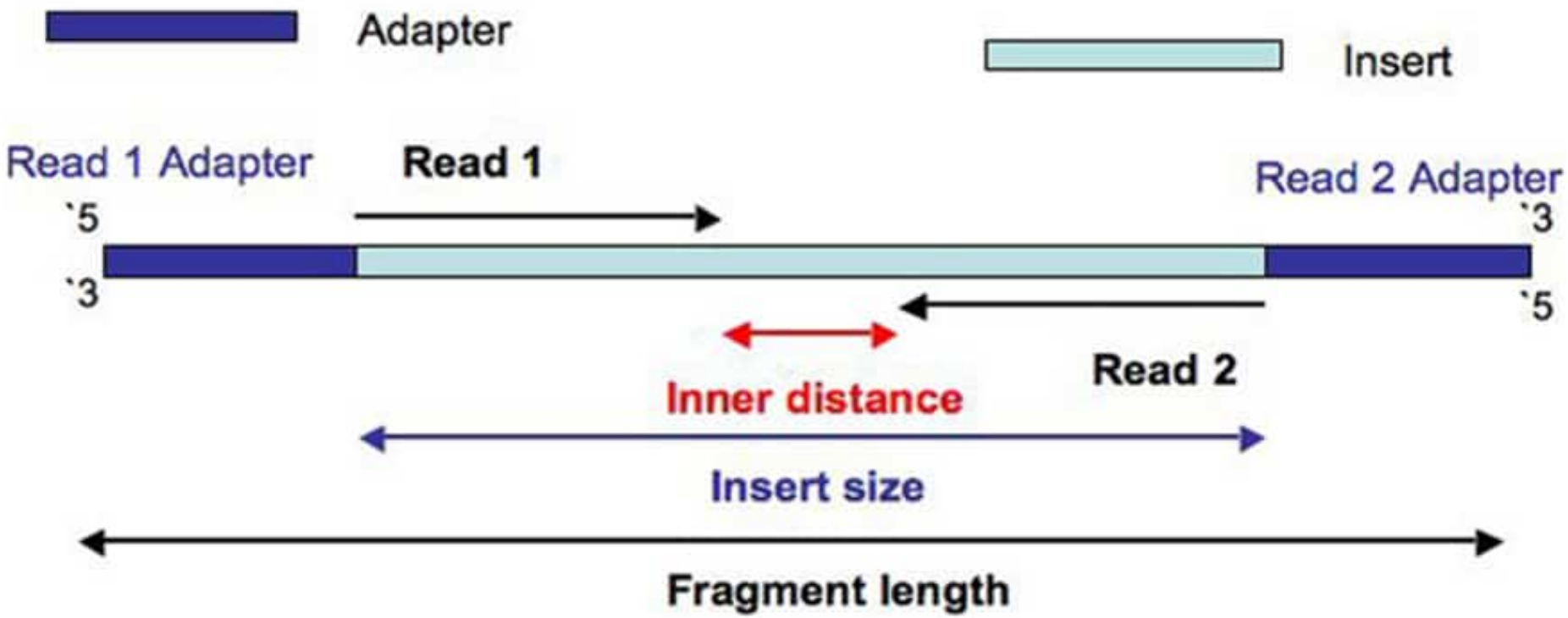
https://www.ecseq.com/support/ngs/how-to-calculate-the-coverage-for-a-sequencing-experiment

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

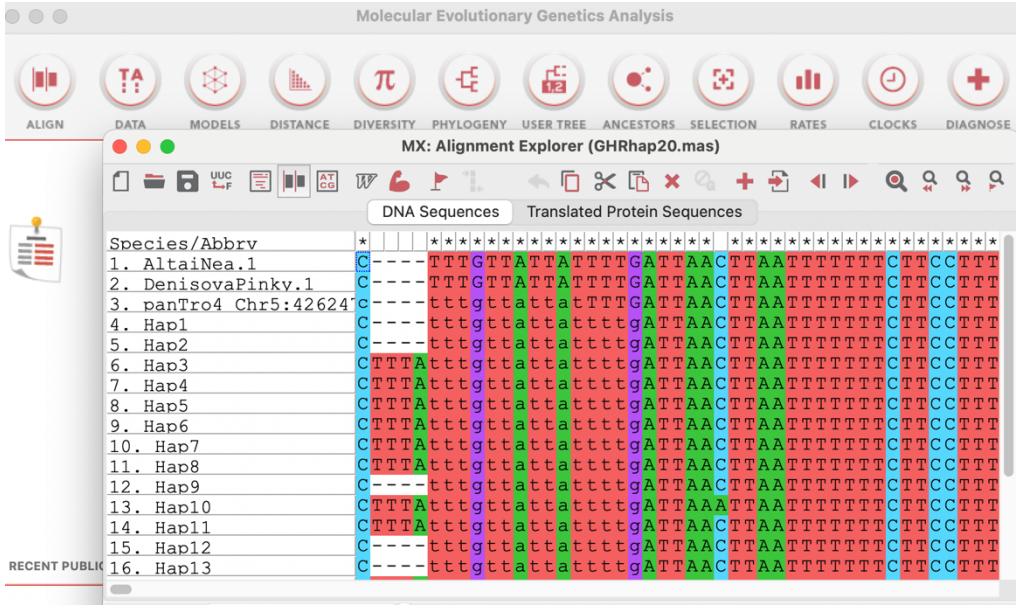


https://thesequencingcenter.com/knowledge-base/what-are-paired-end-reads/

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/stripts
- Laziness is the father of invention
- There are many online courses

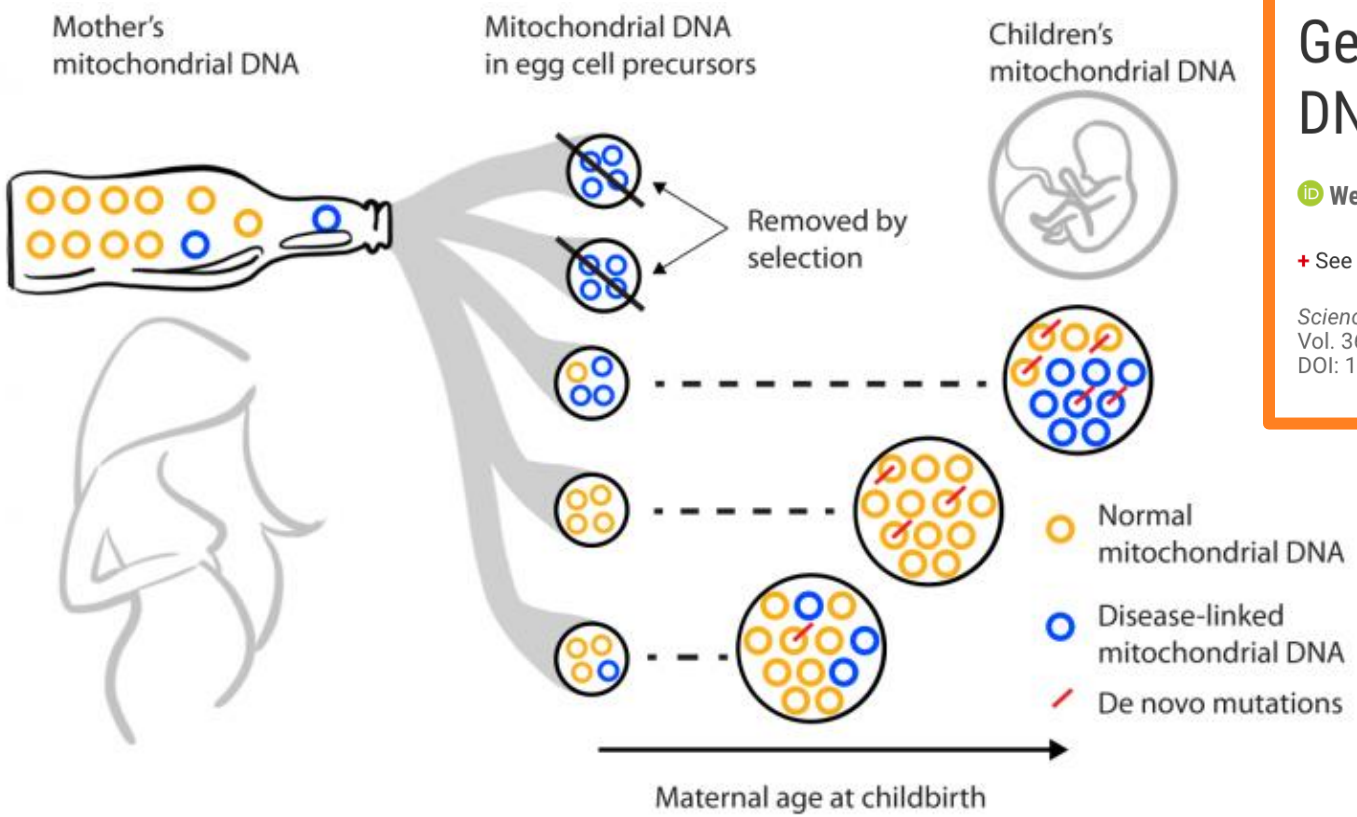
Common errors

- 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
+
BCCCCFFFFFFFGGGGGGGGGGGHHHHGHGHHHHHHHHHHGGGGGGHHHHGH
@M01368:8:000000000-A3GHV:1:1101:14518:9998/1
GTTATTATTATGTCCTACAAGCATTAATTAATTAACACACTTTAGTAAGTA
+
AAAAAFFFFFFFGGGGGGGGGGHGGHHHHGHHHHHHHHGCGHHHHHHHHHHH
@M01368:8:000000000-A3GHV:1:1101:18422:19051/1
GTATCCGACATCTGGTTCCTACTTCAGGGTCATAAAACCTAAATAGCCCAC
.
```

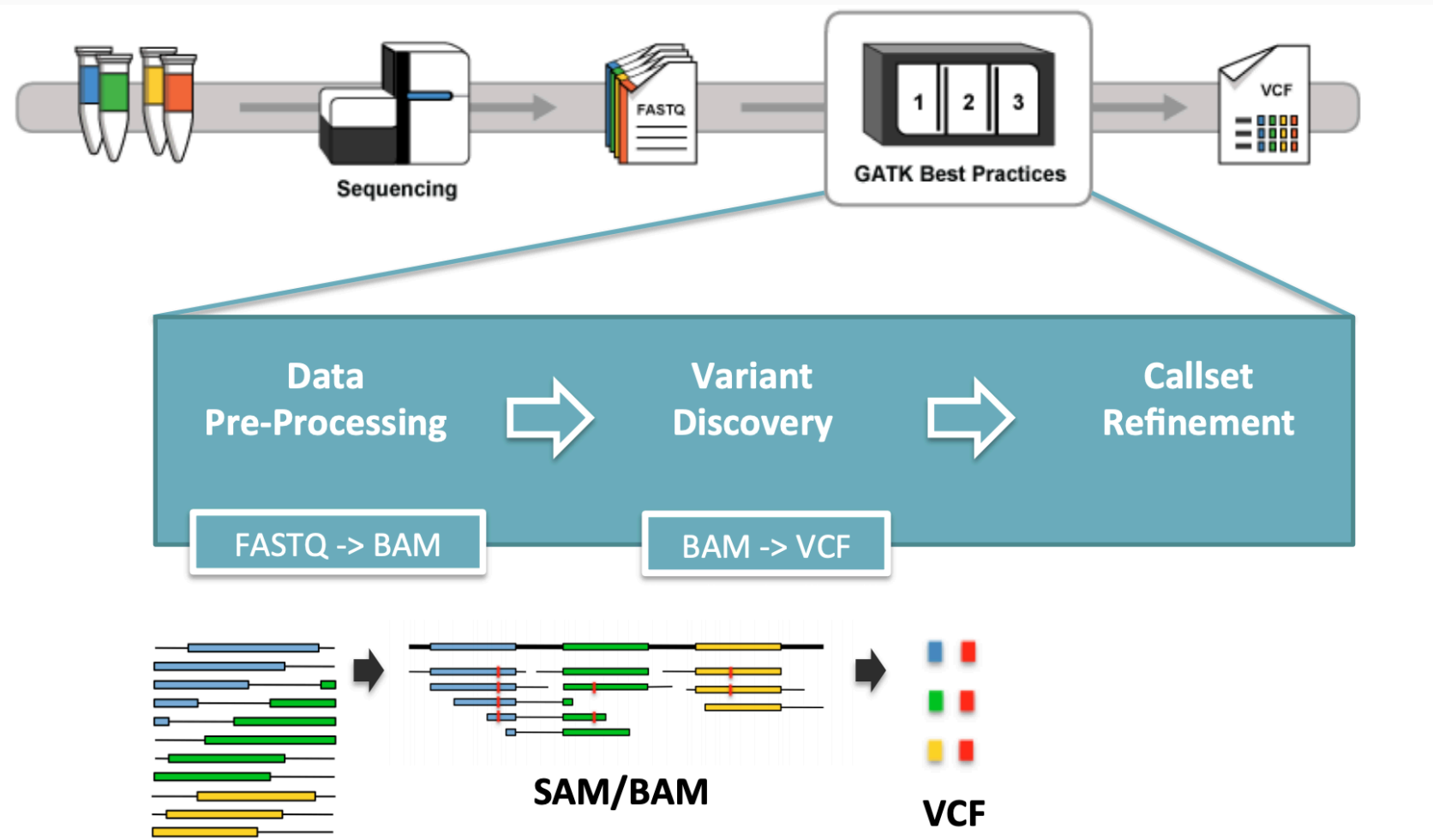
- Goal: Calculate "Allele frequency of variants in Mother and Child"

  Posision: Where is/are the variants?

  Child : How much is the observed frequency?

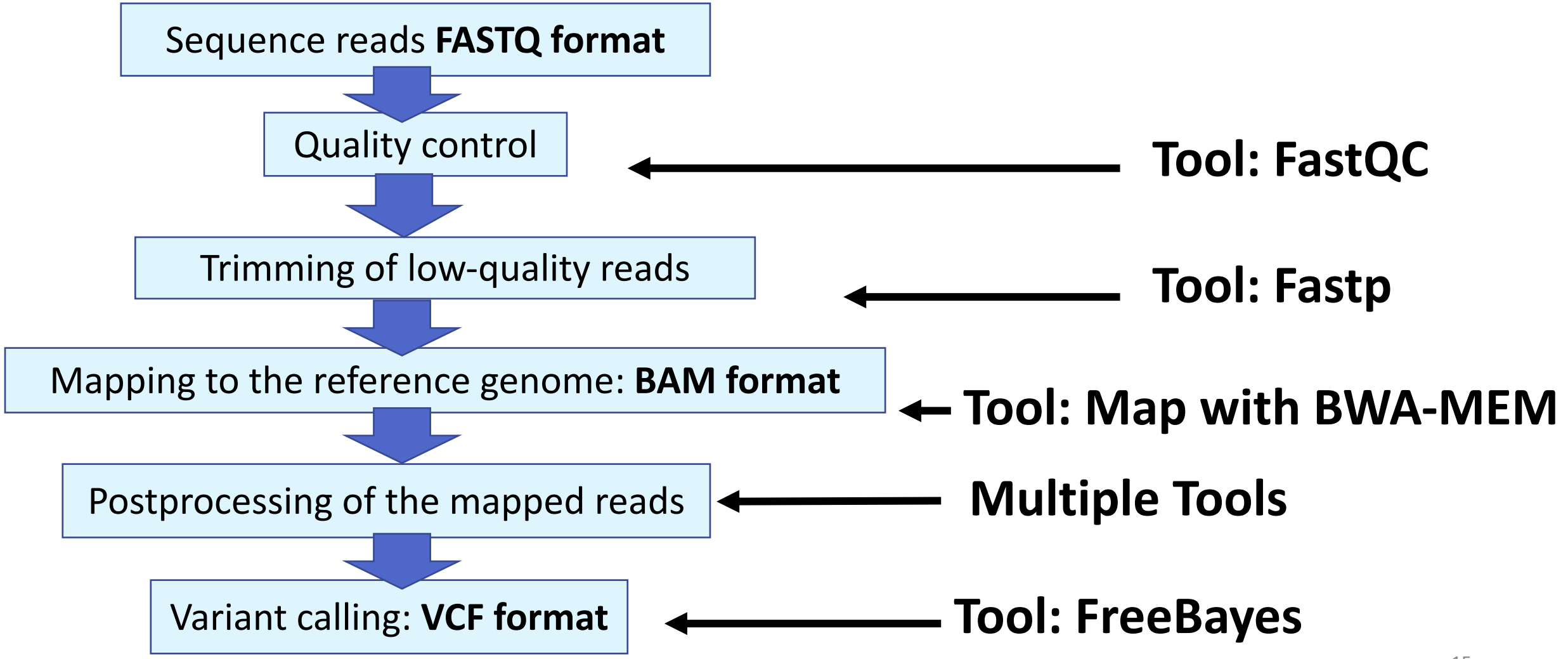
  Mother : How much is the observed variant frequency?

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

Wadapurkar and Vyas,
Informatics in Medicine unlocked, 2018

Galaxy

- **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

usegalaxy.n

NeLS | Galaxy Norway | Analyze Data | Workflow | Visualize | Shared Data | Help | User | Using 6%

Tools to alalyze data

Tools

search tools

- 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

Main window comes here

Welcome to usegalaxy.no

eb-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

If this is
might w

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.

Tweets by @elixirnorway

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

ssen/statu

ELIXIR Norway
@elixirnorway
Few spots left on the @swcarpentry
ELIXIR Norway and @DigitalLiv for
researchers korbinib.github.io/2021-

Embed

Your workflow

History

search datasets

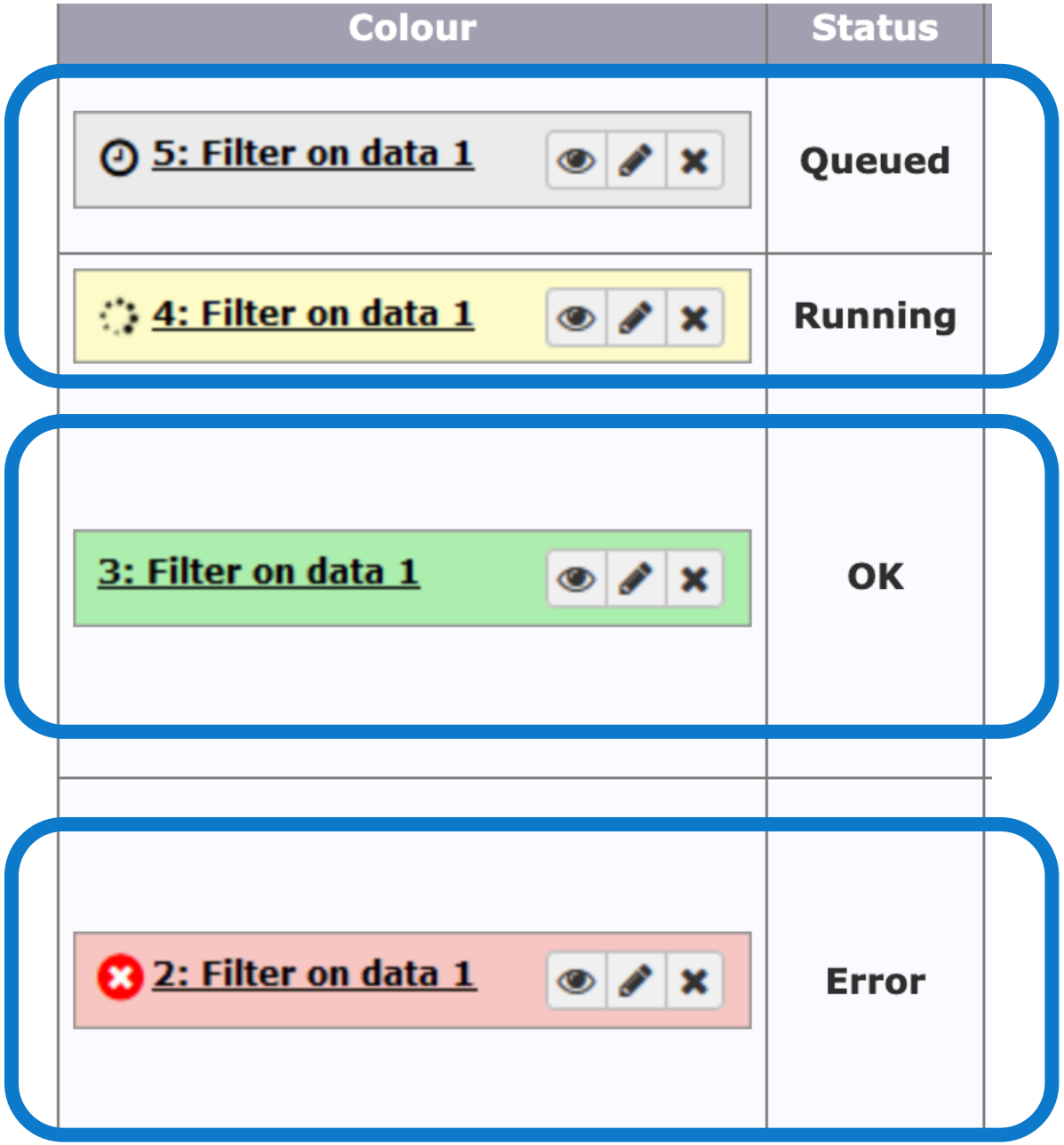
mtDNA

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



Be patient and wait

Let's go to next step

Hmm… Let's examine it

If you get an error, don’t panic…

Click to expand the menu

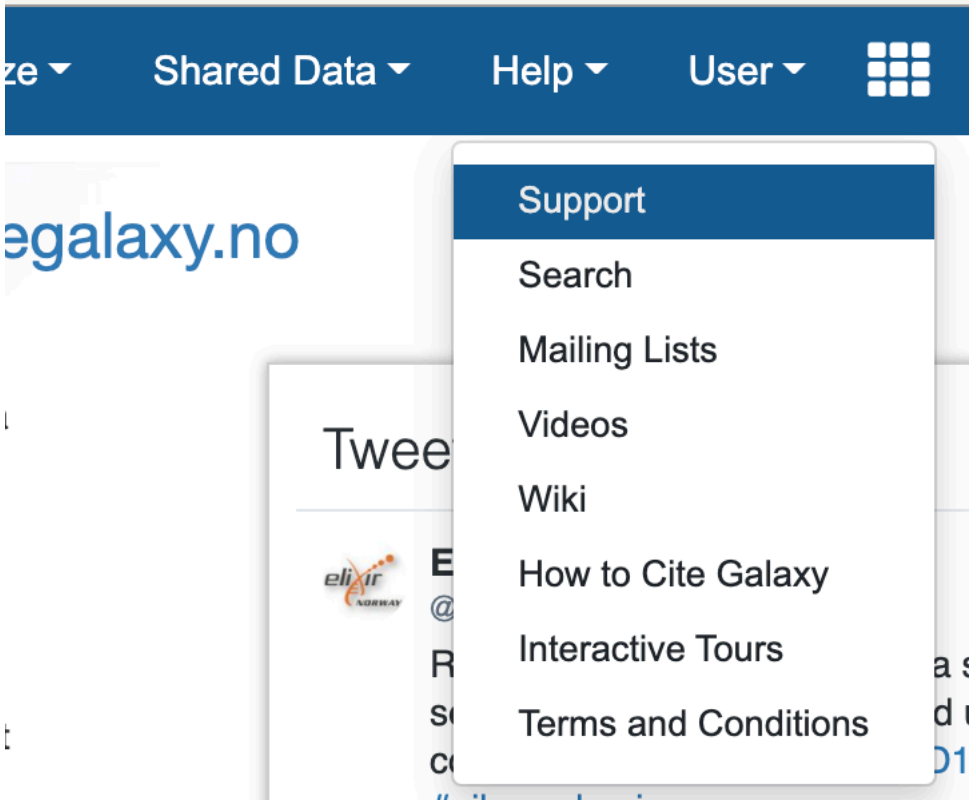
Let’s examine it

Common issues:

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

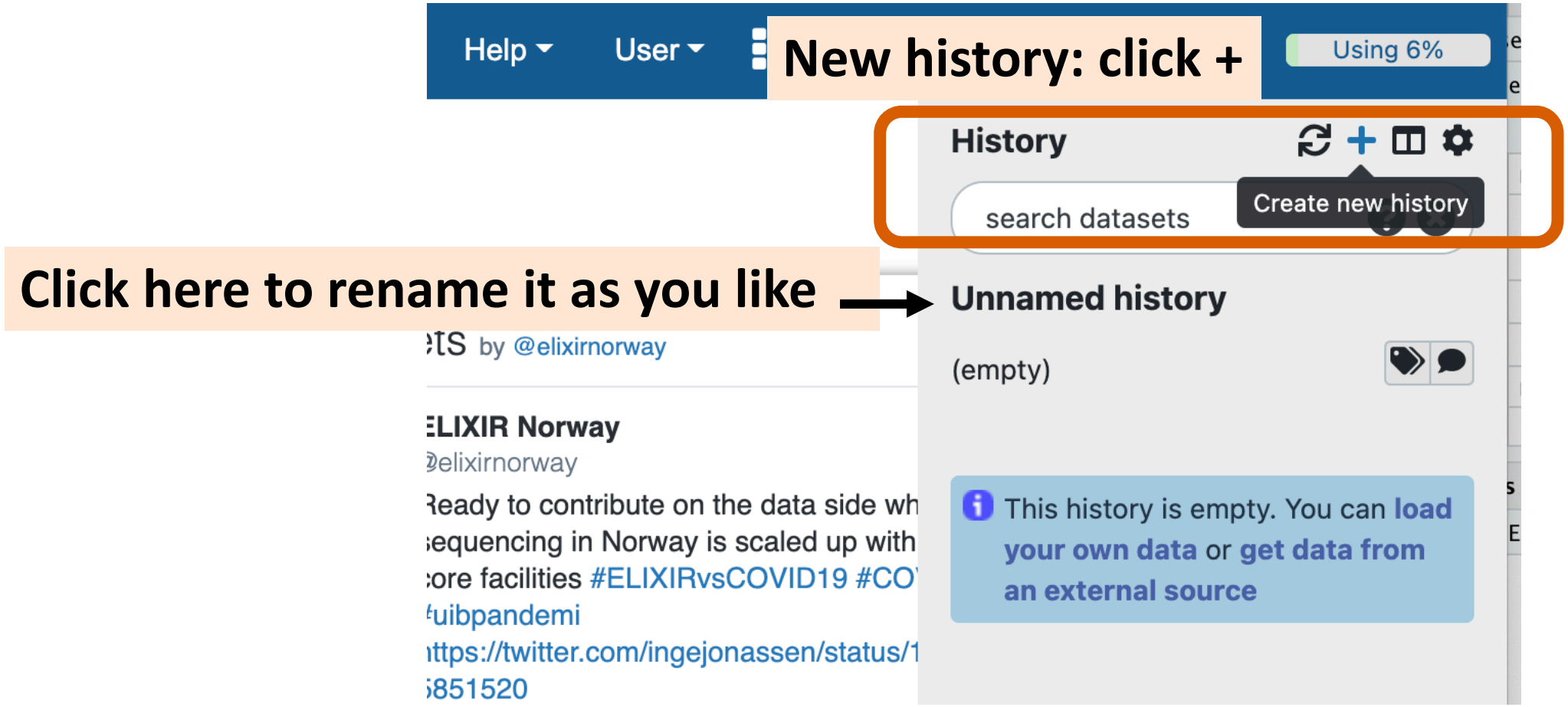
If can not figure it out, ask colleagues and experts

-> Wiki, Support, and Mailing Lists



contact@bioinfo.no

Create new history



Let’s get started…

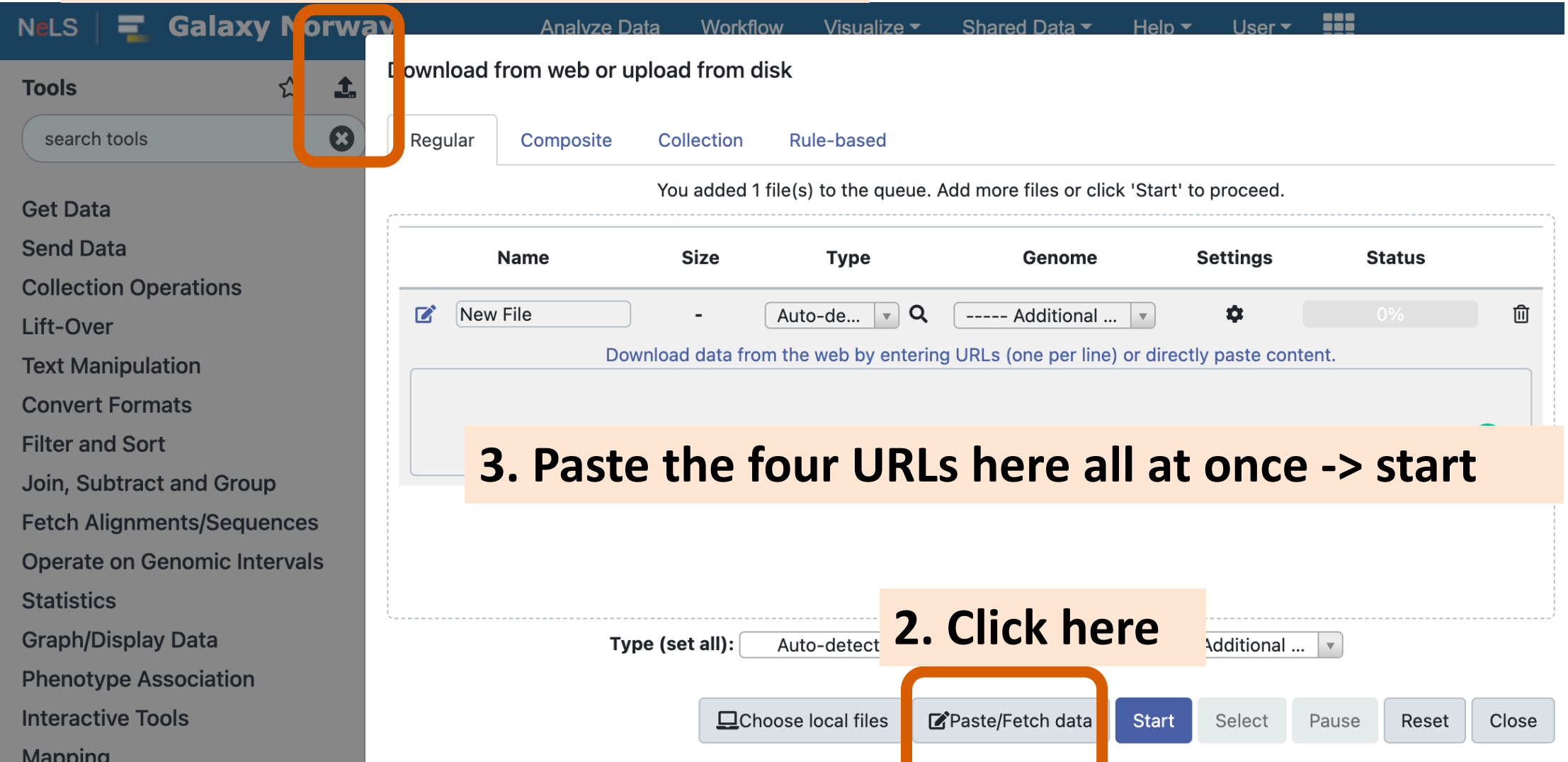
- Get the data!
  Mitchondrial DNA of the child and the mother

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

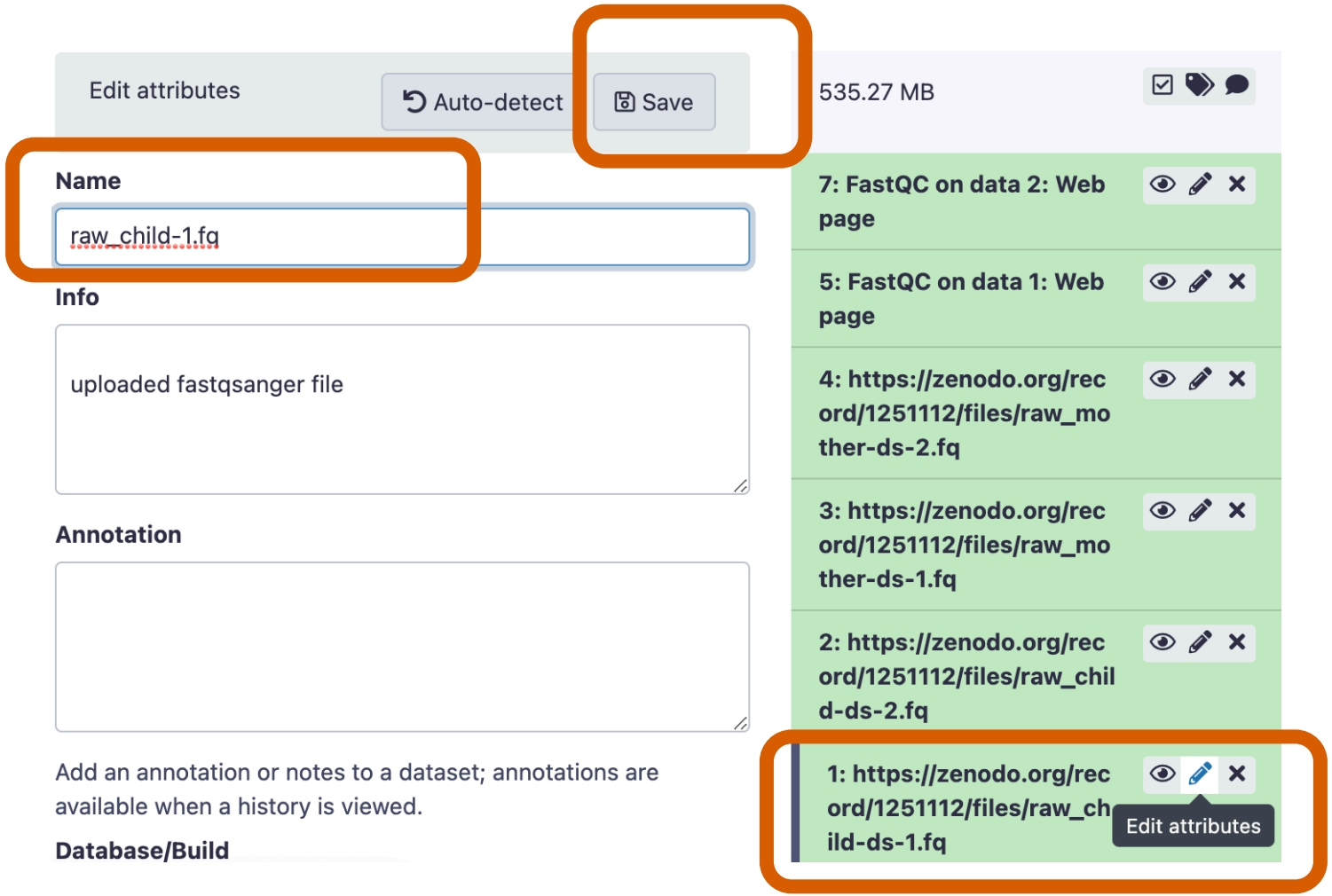
- Q: Why there are two data per individual?

Import data set

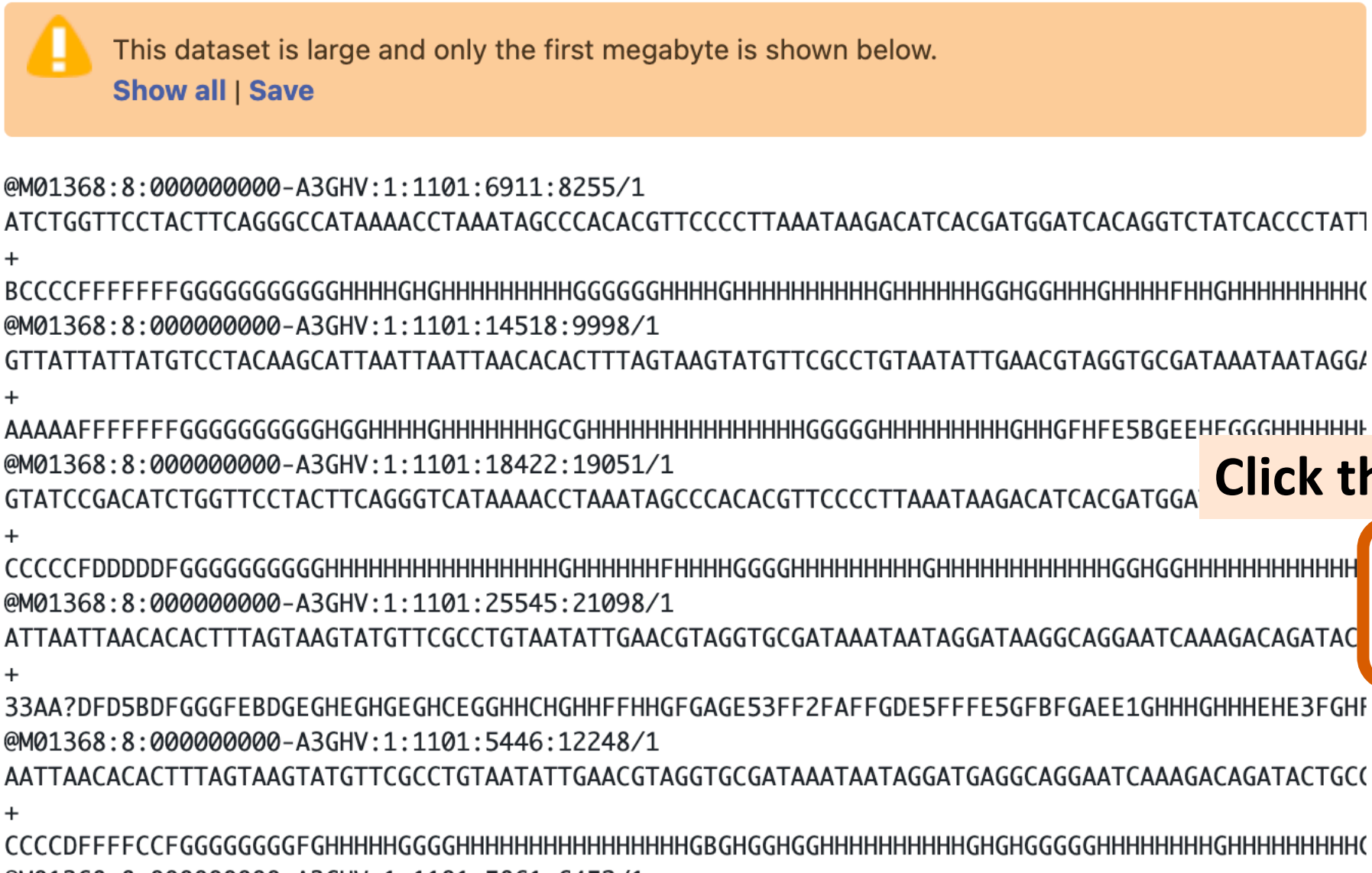
1. Click here to import data set

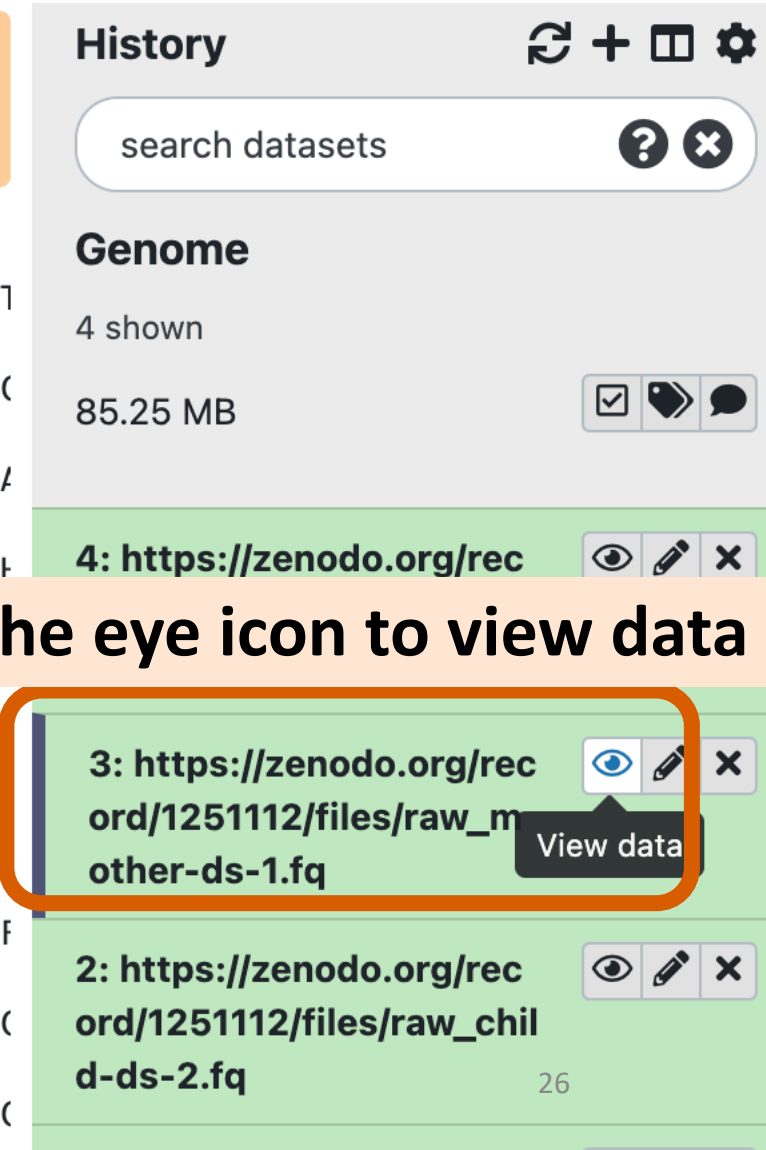


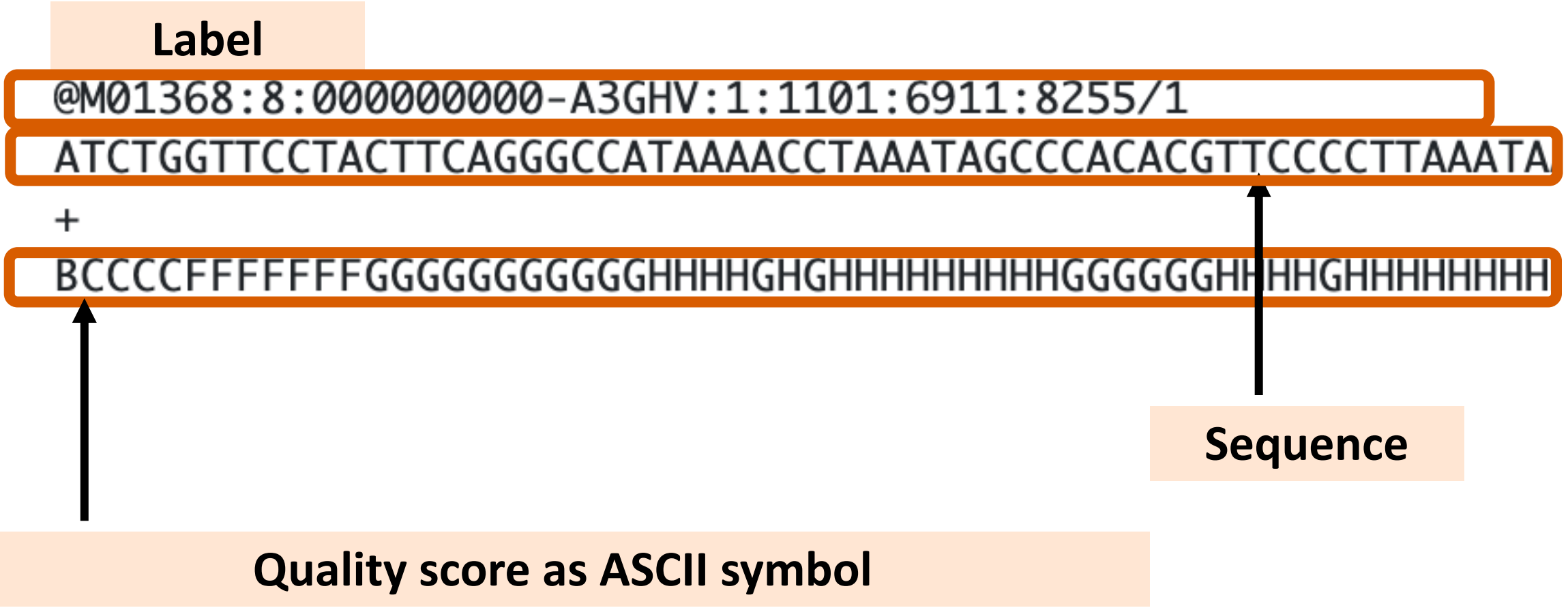
Rename the data set as you like



Let’s examine the FastQ file



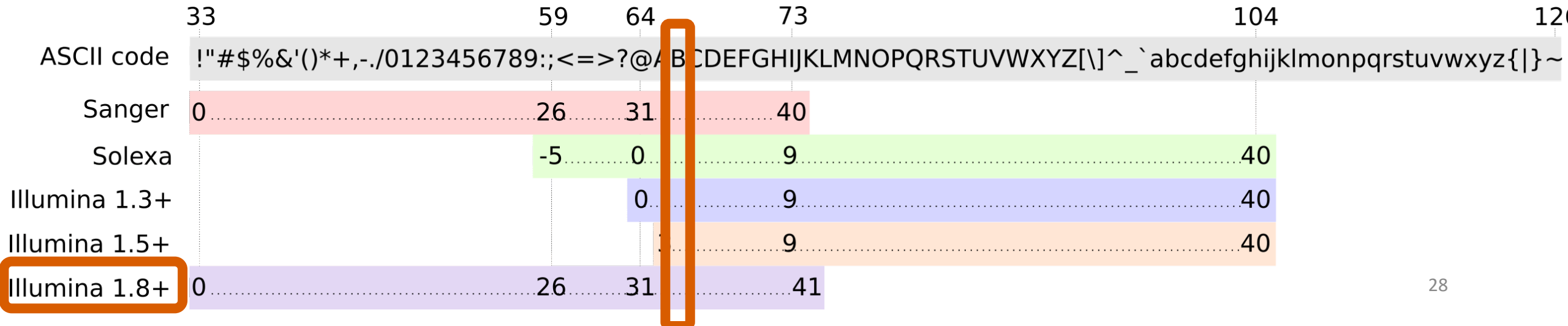



Label
@M01368:8:000000000-A3GHV:1:1101:6911:8255/1
ATCTGGTTCCTACTTCAGGGCCATAAAACCTAAATAGCCCACACGTTCCCCTTAAATA
+
BCCCCFFFFFFFGGGGGGGGGGGHHHHGHGHHHHHHHHHHGGGGGGHHHHGHHHHHHHH
Sequence
Quality score as ASCII symbol

Quality scores

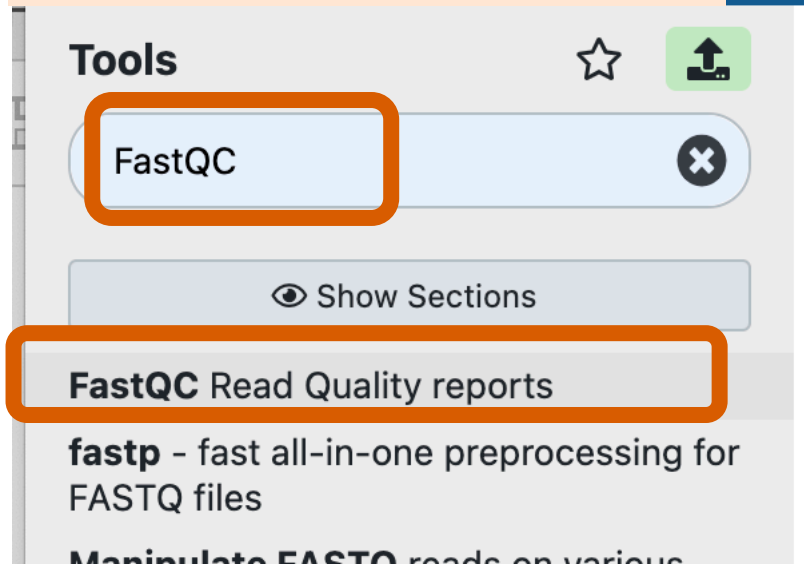
```
@M01368:8:000000000-A3GHV:1:1101:6911:8255/1
ATCTGGTTCCTACTTCAGGGCCATAAAACCTAAATAGCCCACACGTTCCCCTTAAATA
+
BCCCCFFFFFFFGGGGGGGGGGGHHHHGHGHHHHHHHHHHGGGGGGHHHHGHHHHHHHHI
```

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%

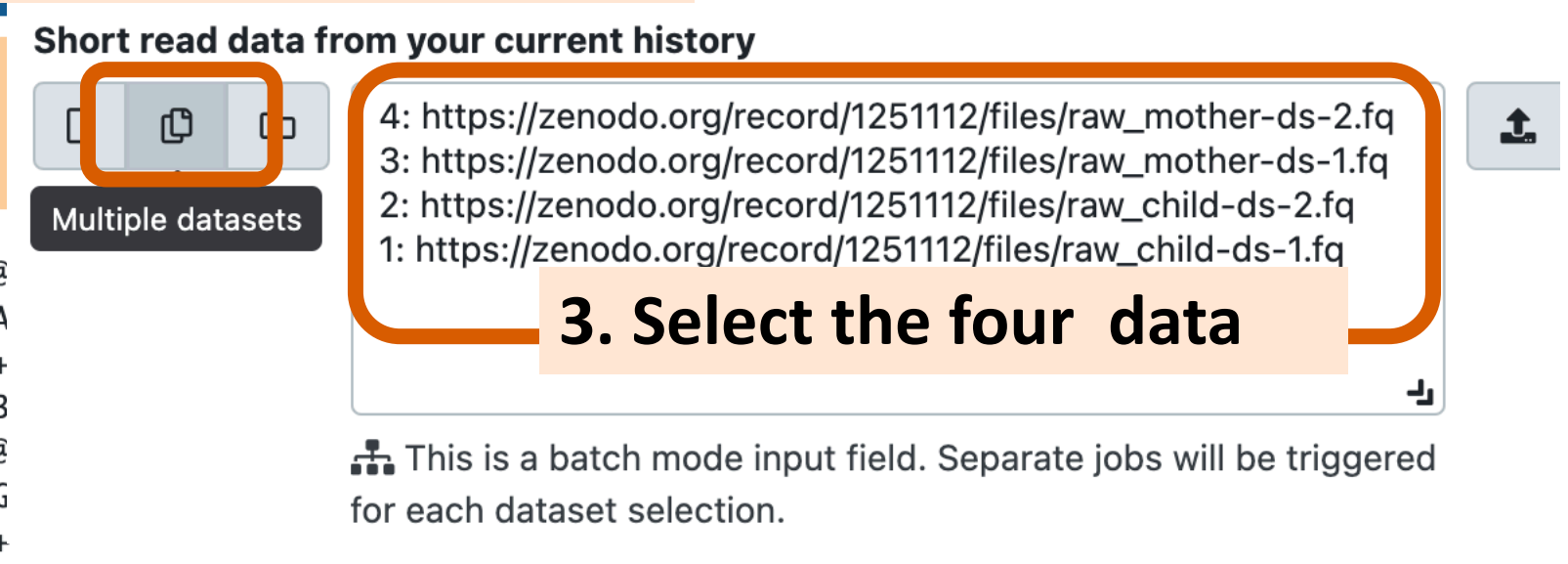


Run FastQC for the quality check

1. search “FastQC”



2. Multiple datasets

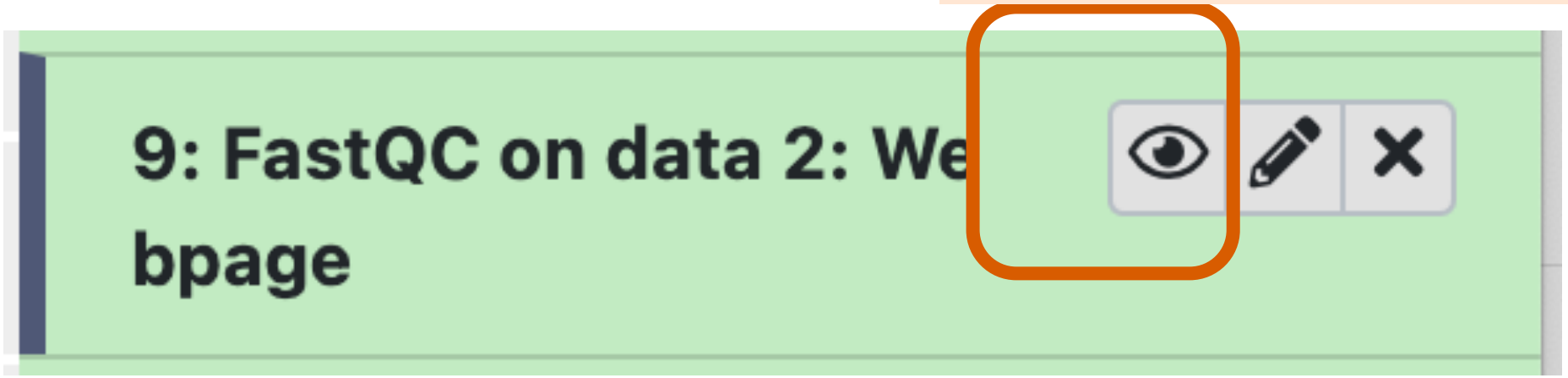


3. Select the four data

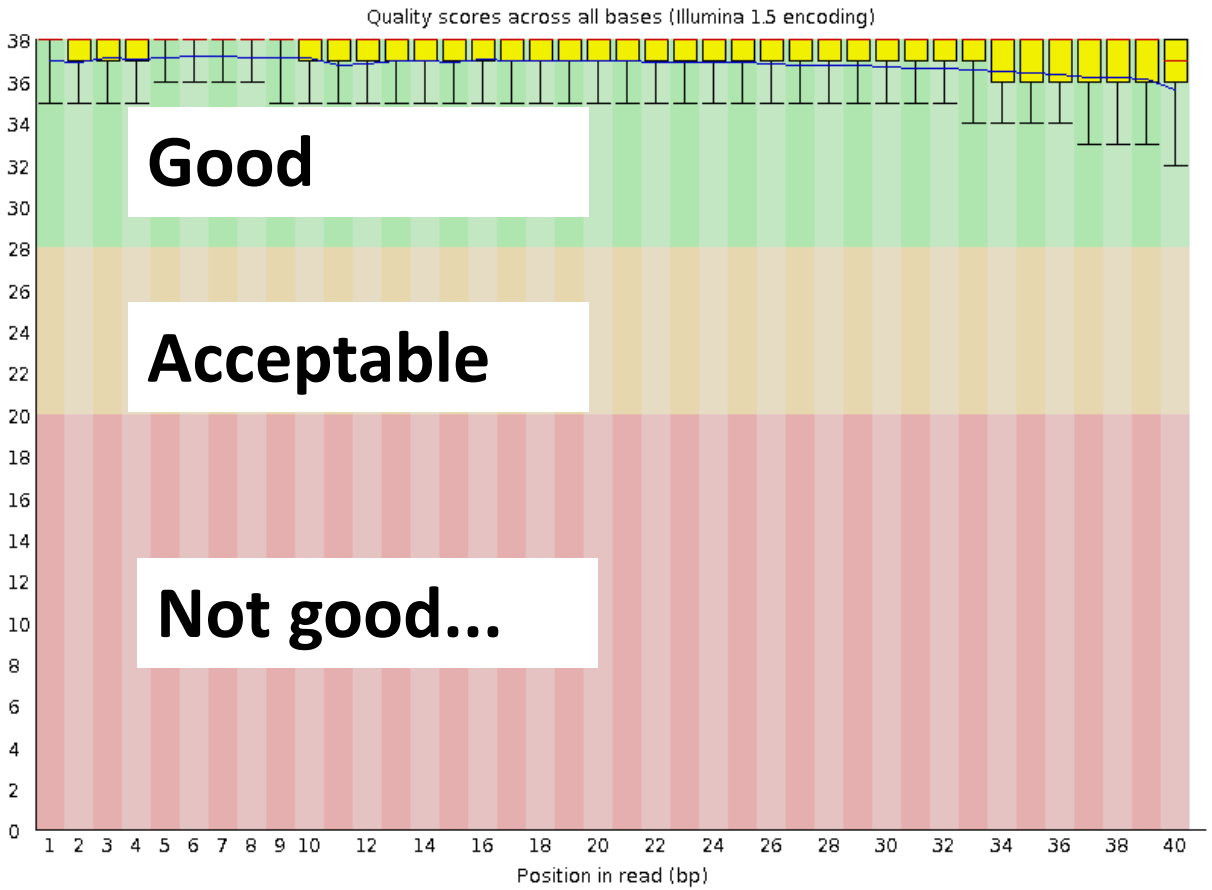
→ **Execute**

Run FastQC for the quality check

Click the “eye” icon to view data

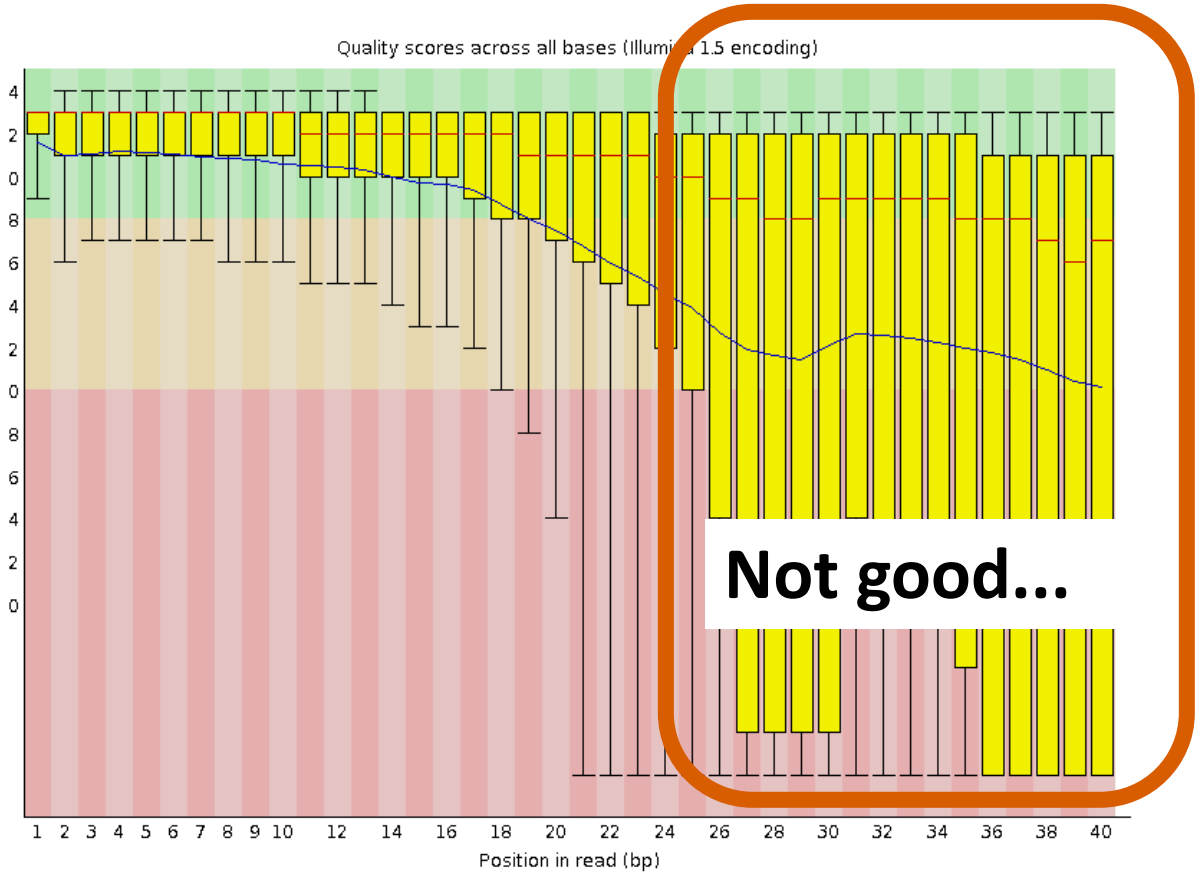


Good Illumina data



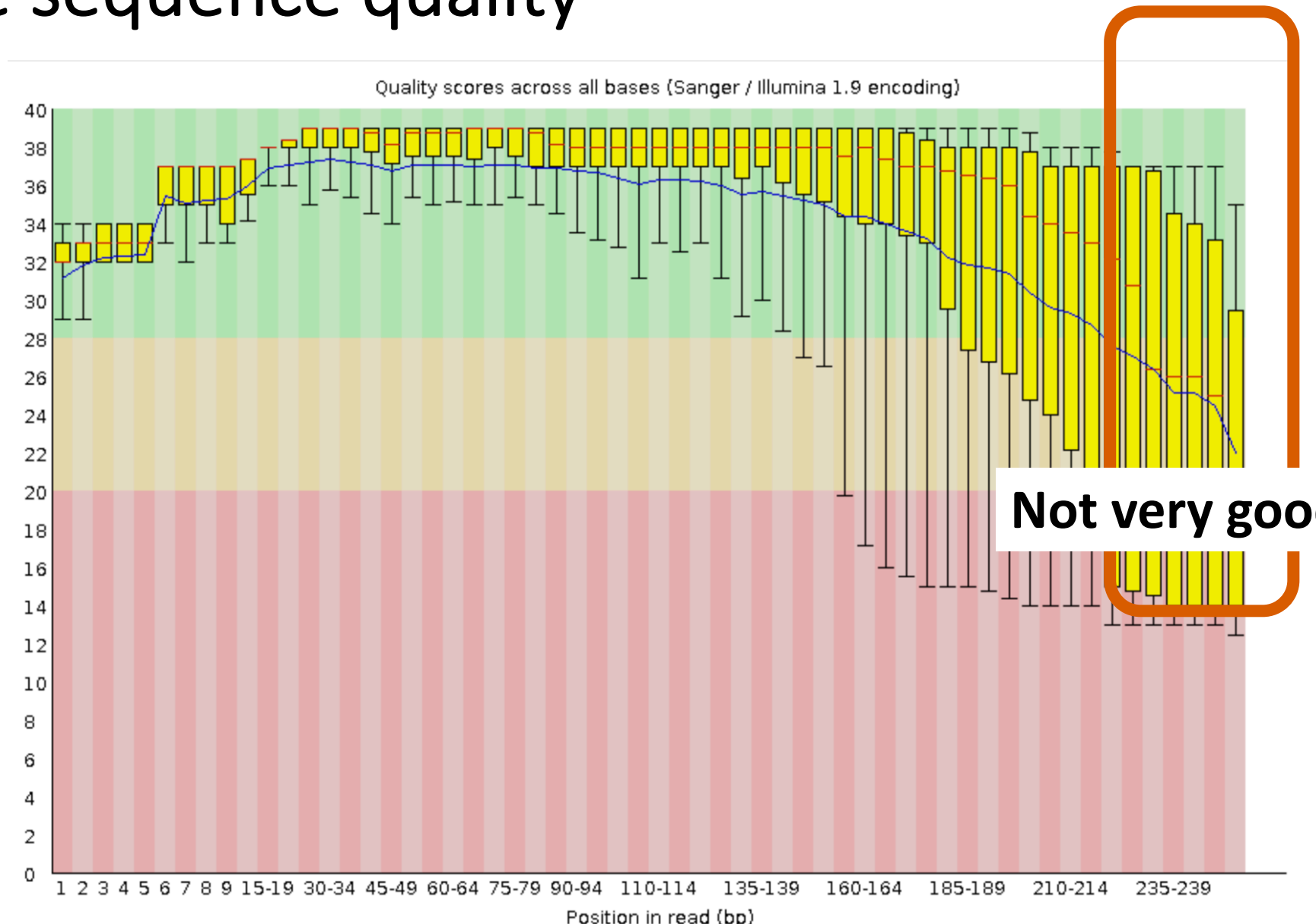
Nucleotide position in read

Bad Illumina data



Nucleotide position in read

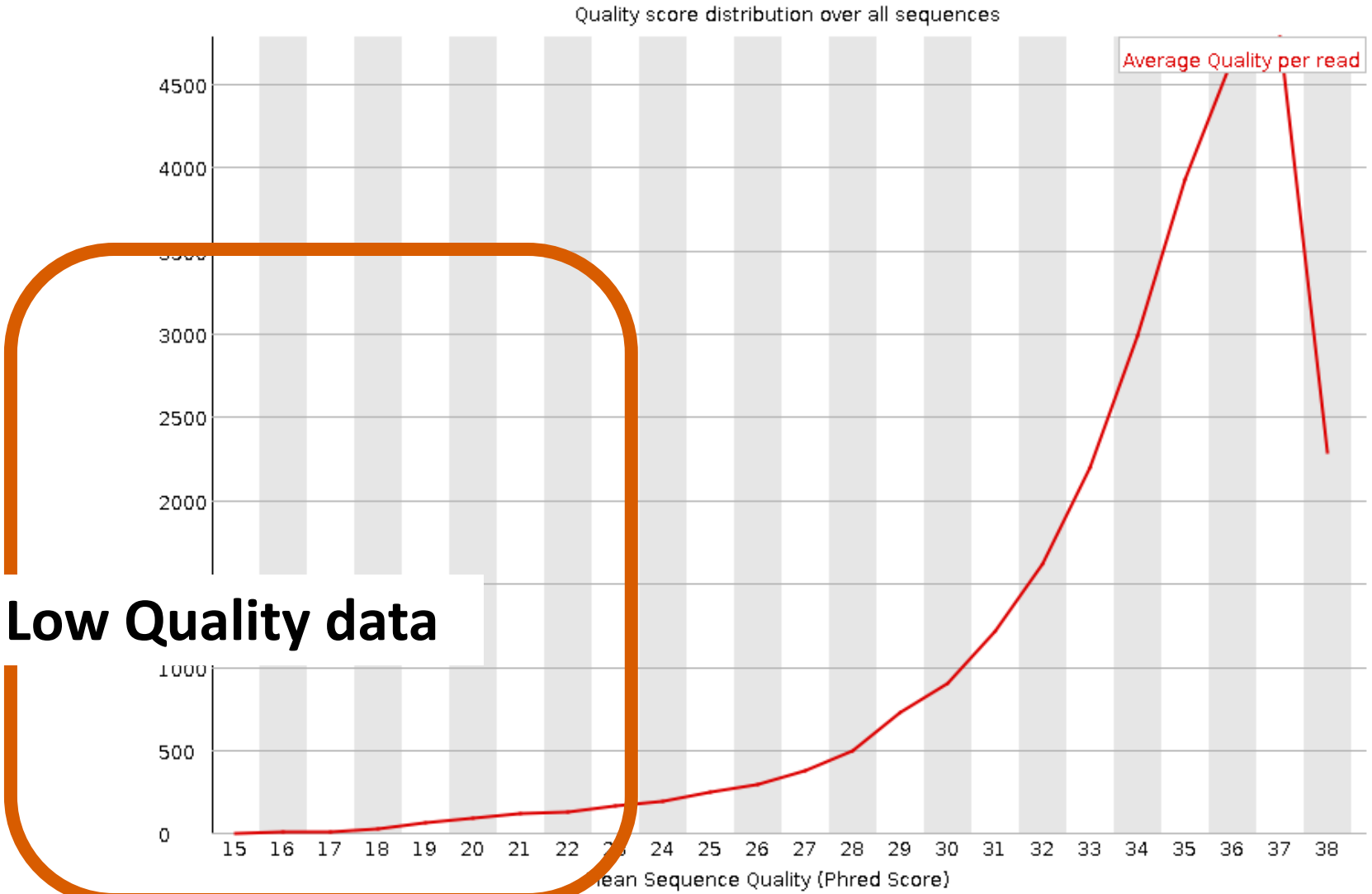
FastQC Data Interpretation
Per base sequence quality



Not very good...

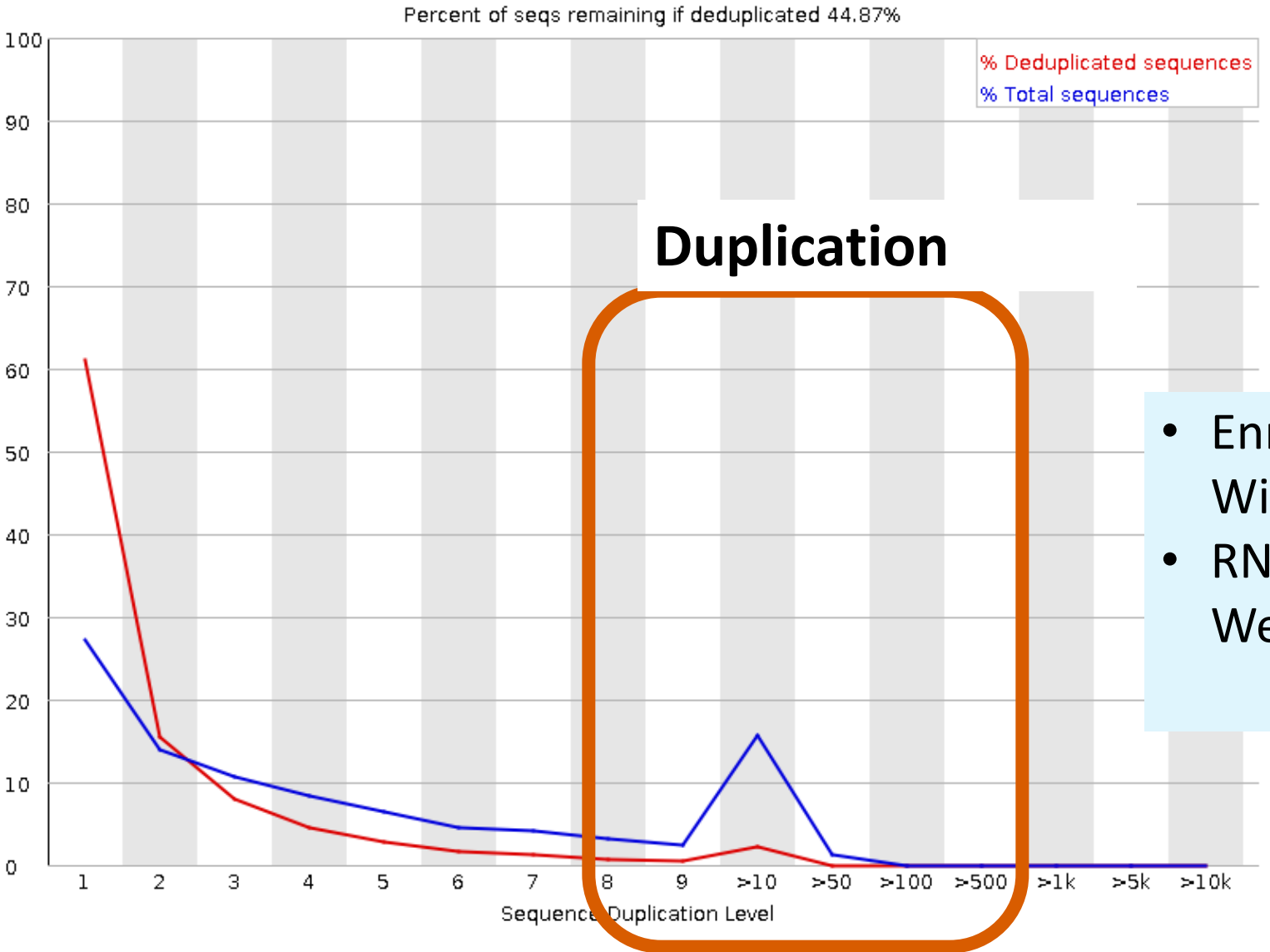
FastQC Data Interpretation

Per sequence quality scores



Low Quality data

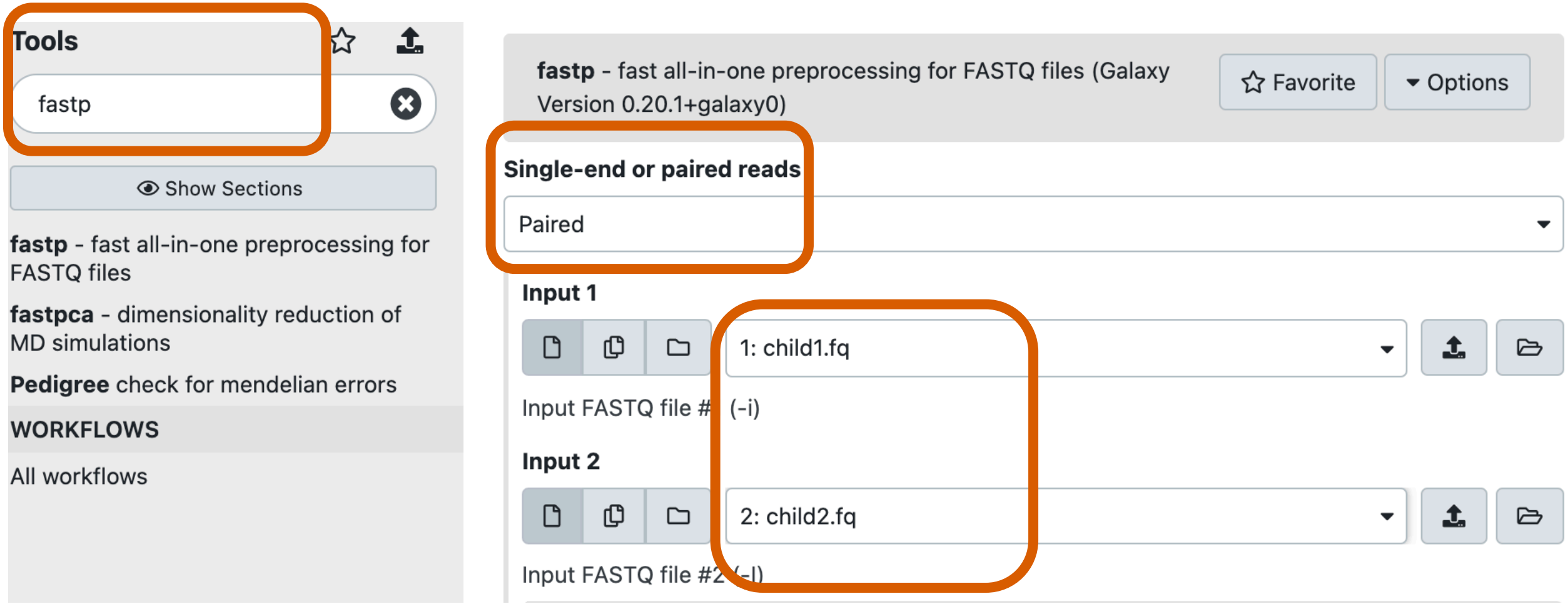
Sequence Duplication Levels



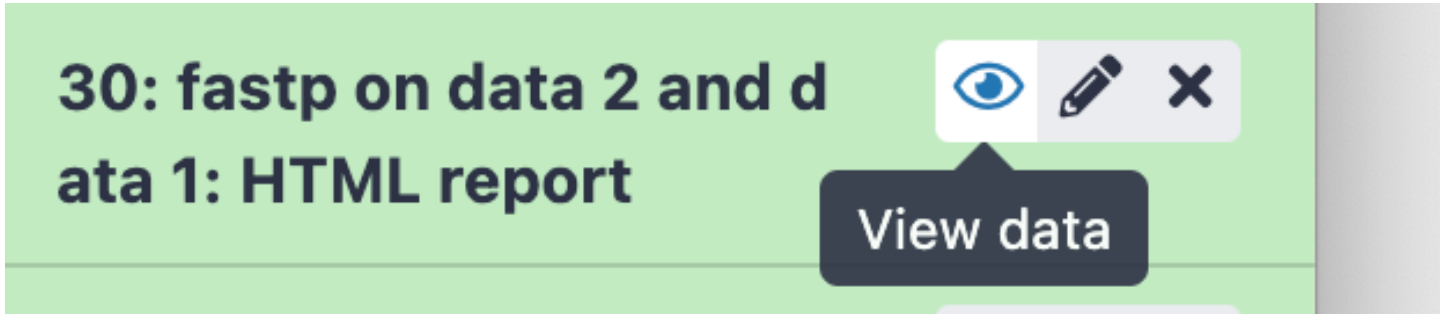
Duplication

- 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



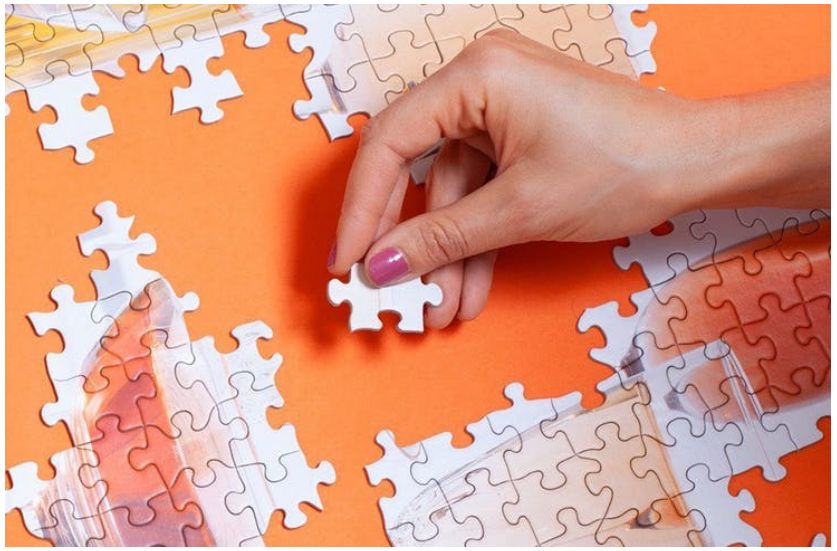
Examine the filtering result of fastp

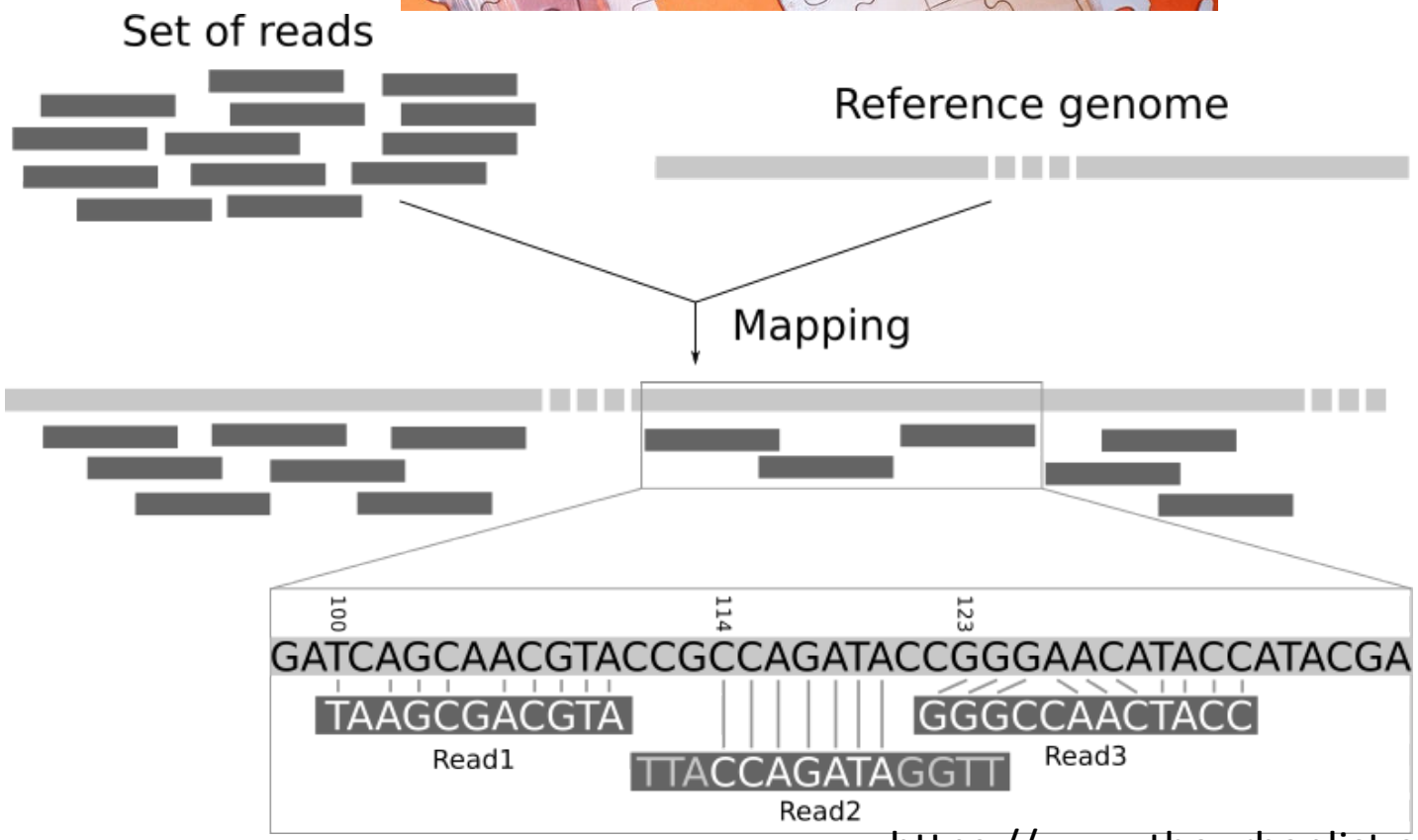


Filtering result

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

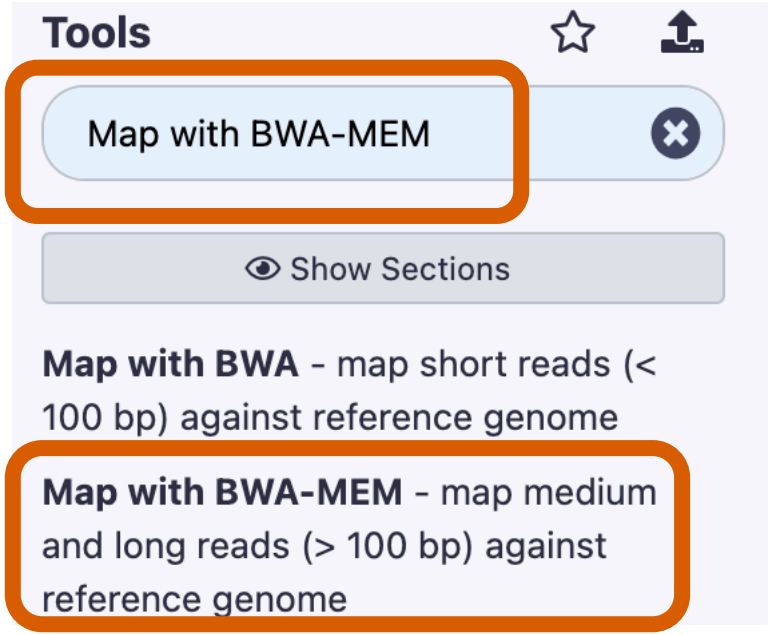
Mapping!

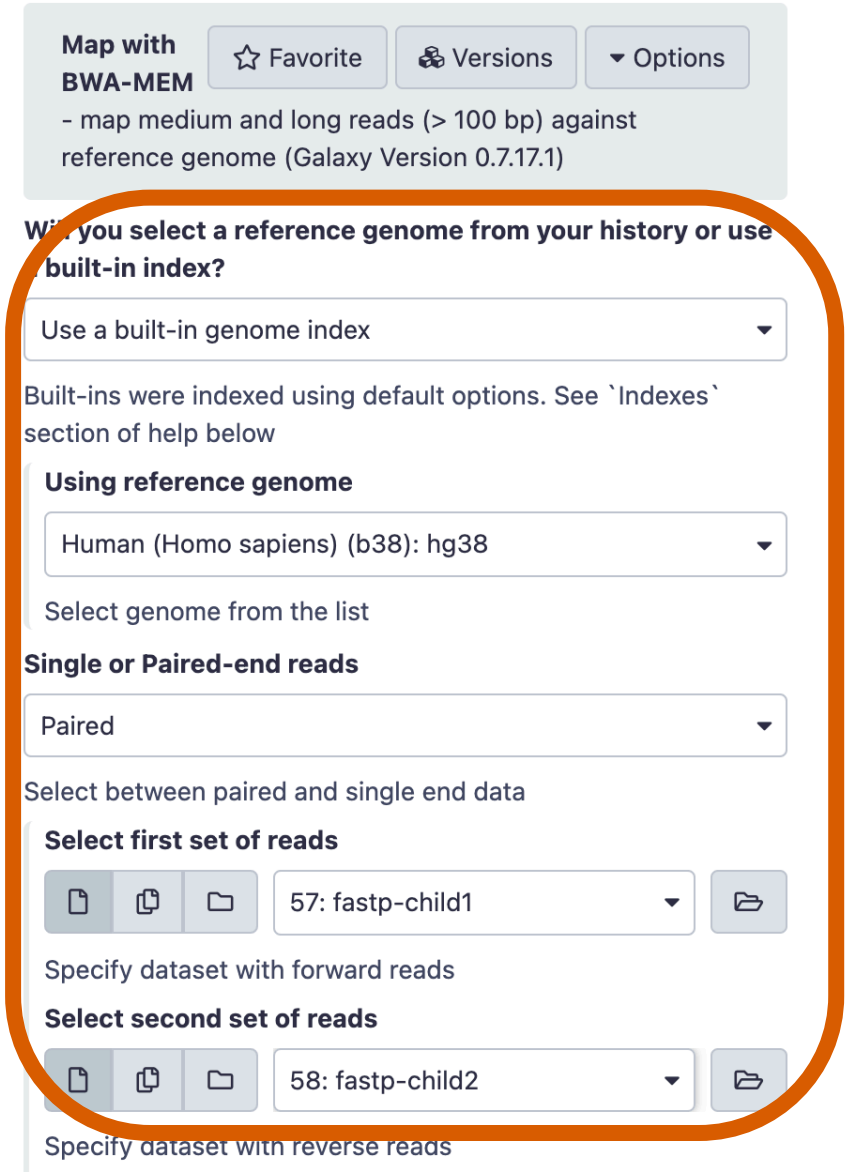




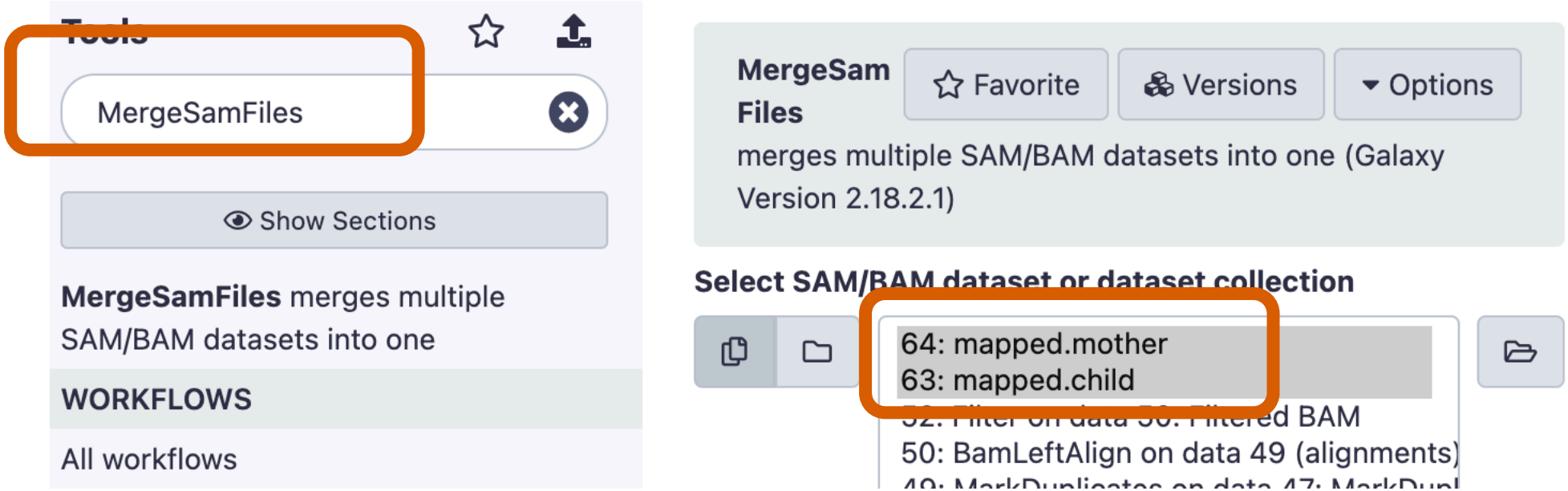
https://www.theurbanlist.com/a-list/best-adult-jigsaw-puzzle

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



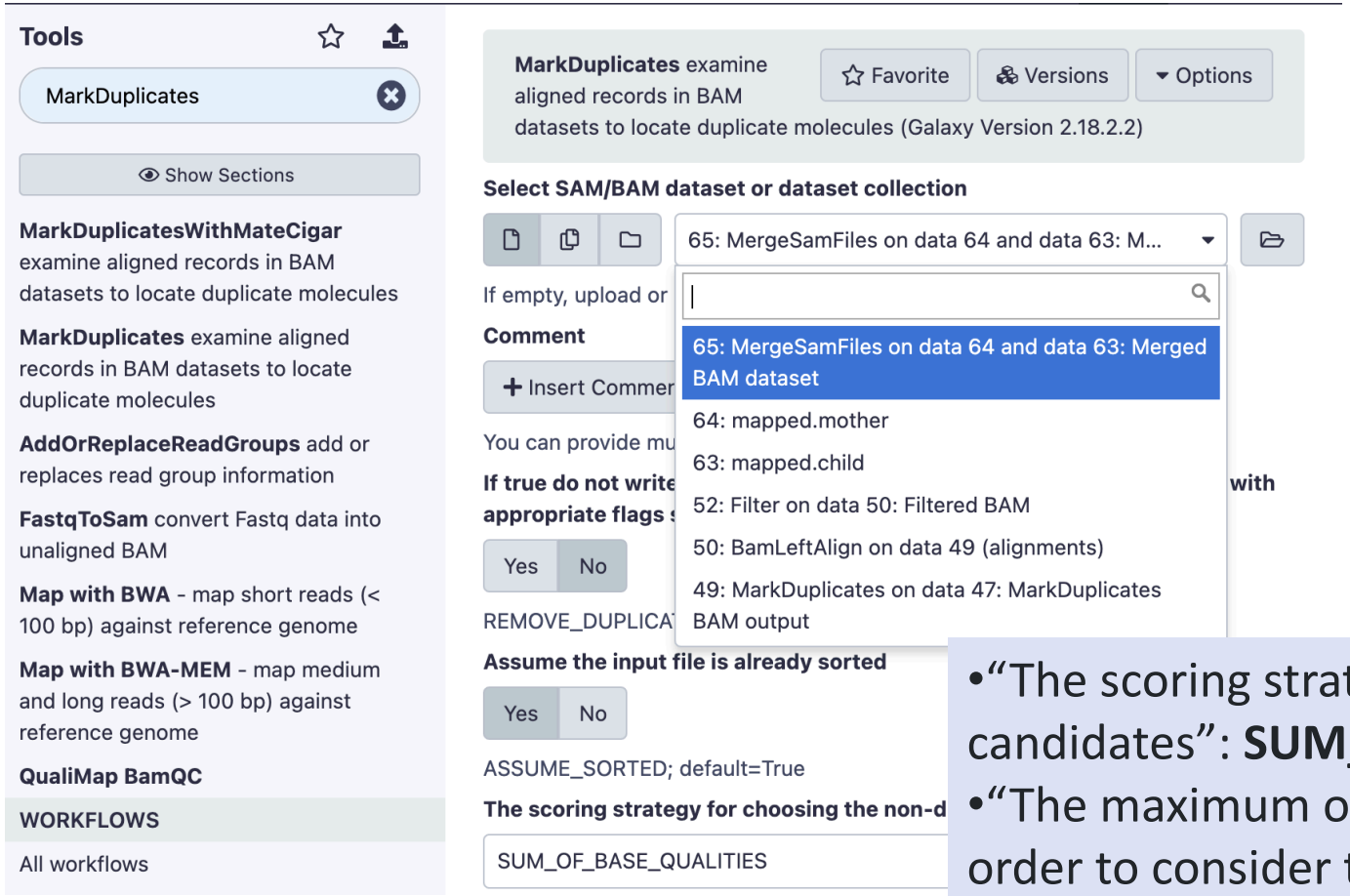


Merging two BAM data with MergeSamFiles



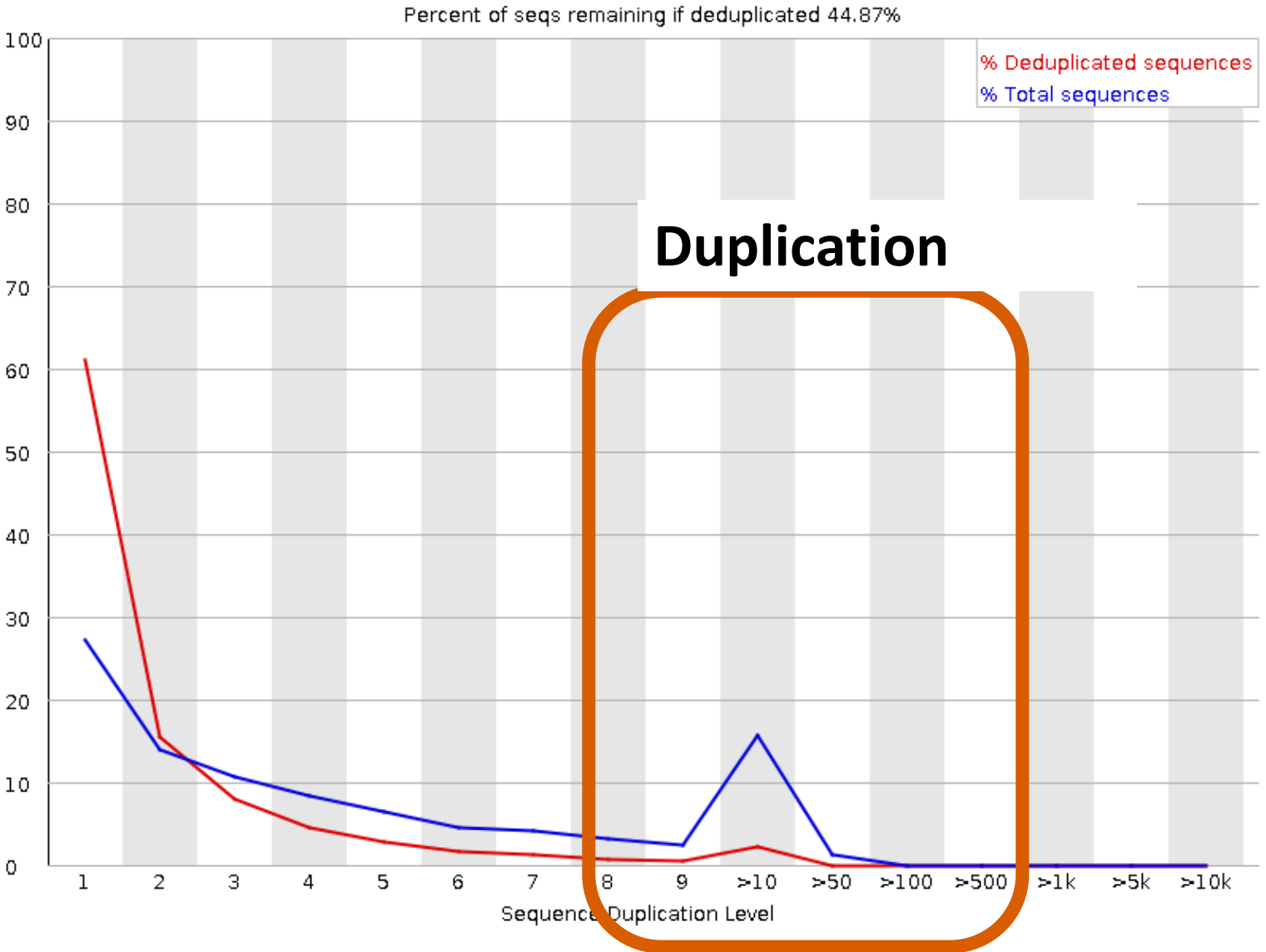
After this process, sample names are now incoorporated and you don’t have to keep renaming data sets.

Removing the PCR duplicates

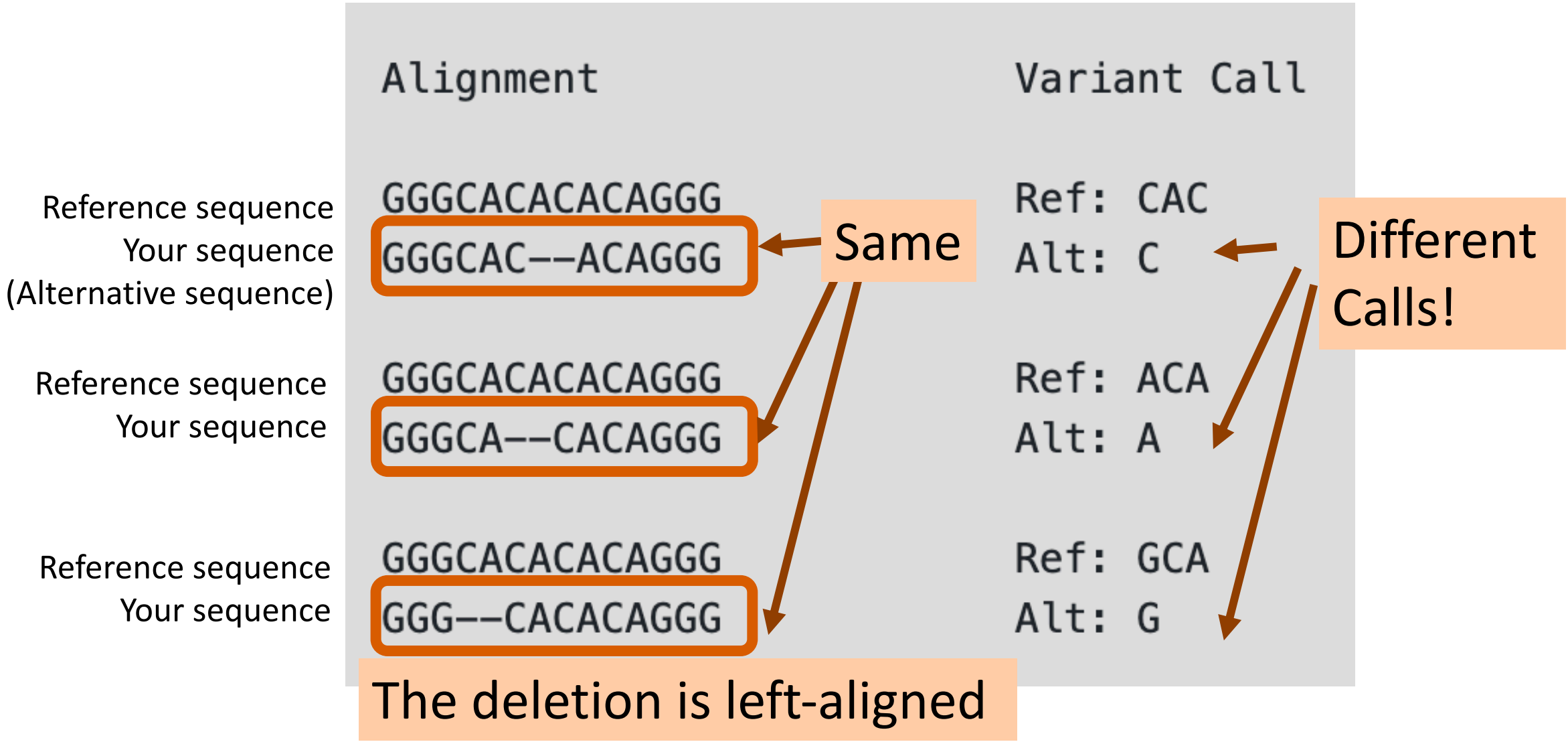


- "The scoring strategy for choosing the non-duplicate among candidates": **SUM_OF_BASE_QUALITIES**
- "The maximum offset between two duplicate clusters in order to consider them optical duplicates": **100**
- "Select validation stringency": **Lenient**

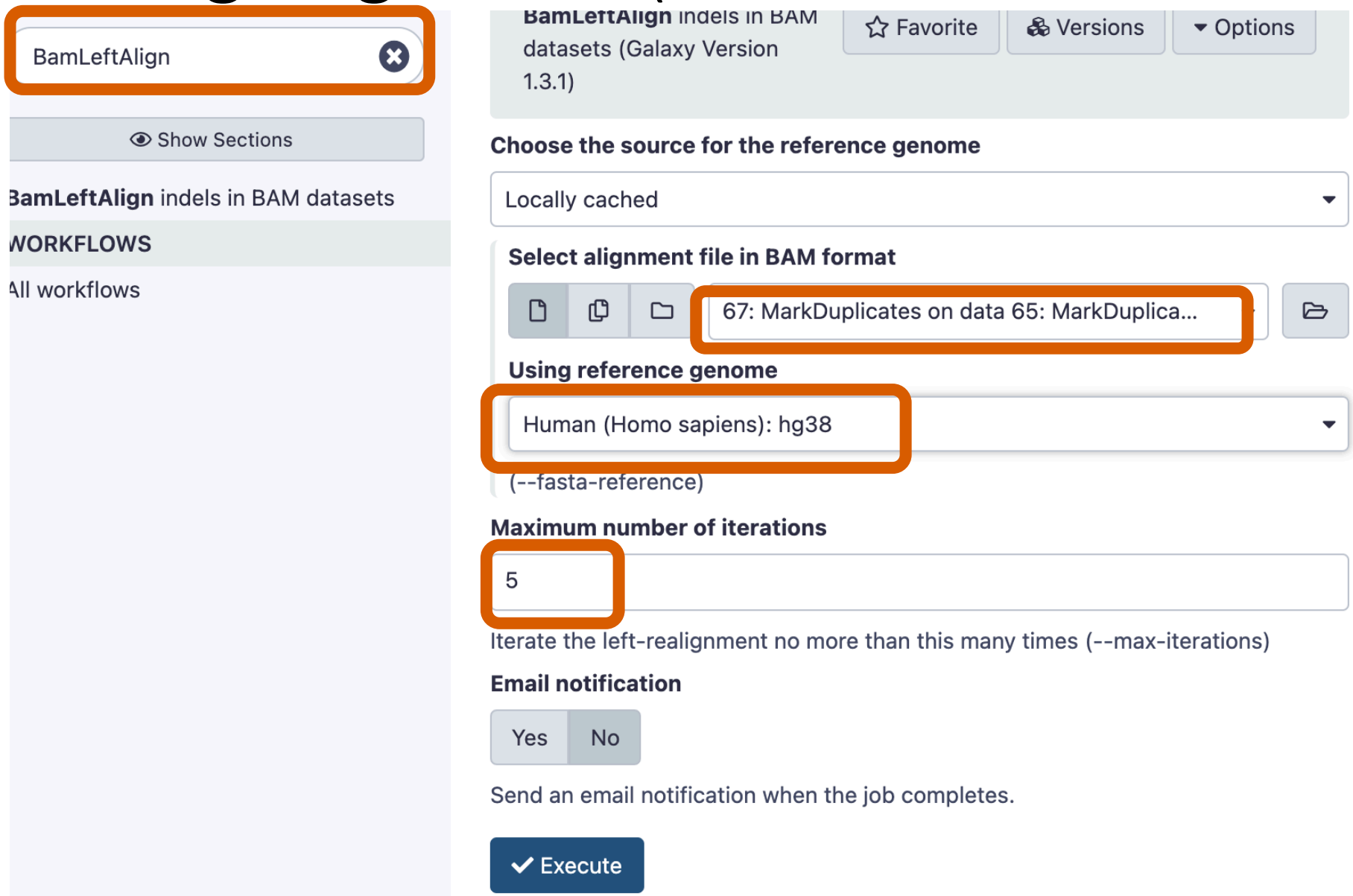
Review: Sequence Duplication Levels



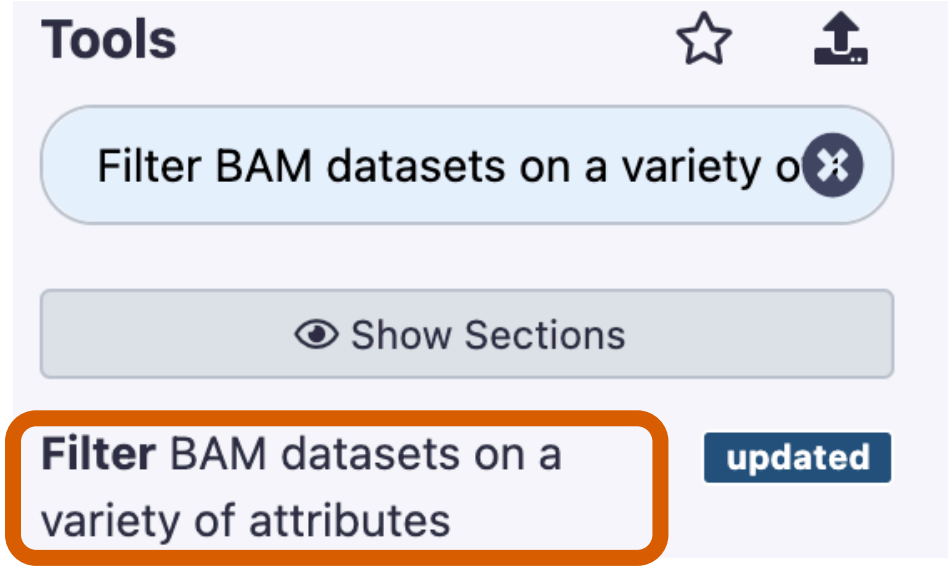
Left-aligning indels (insertions/deletions)

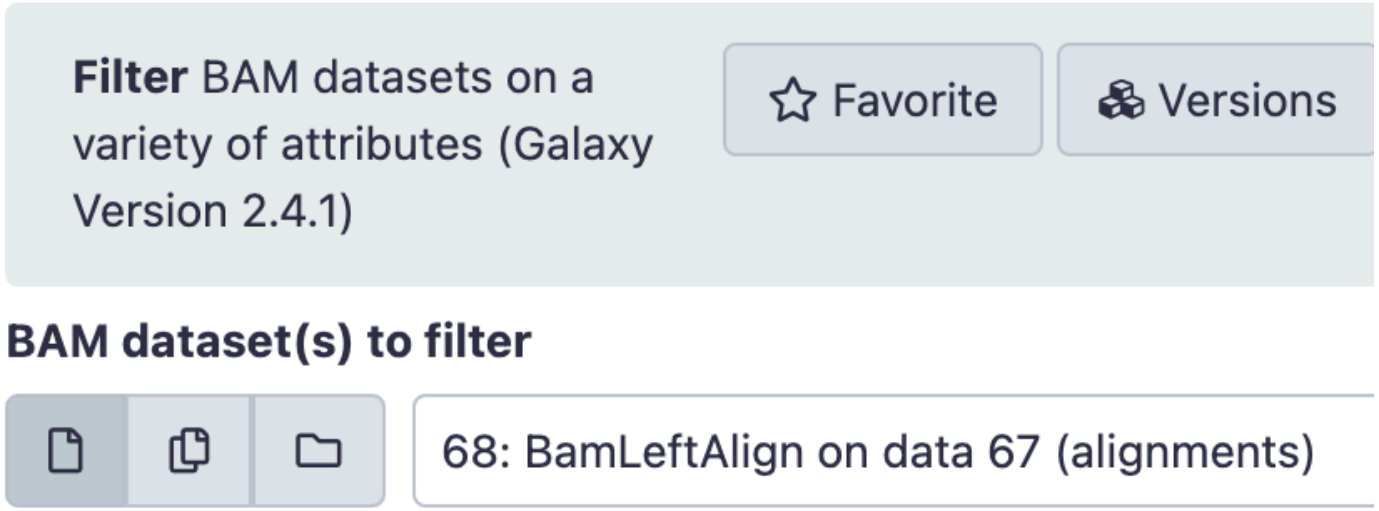


Left-aligning indels (insertions/deletions)

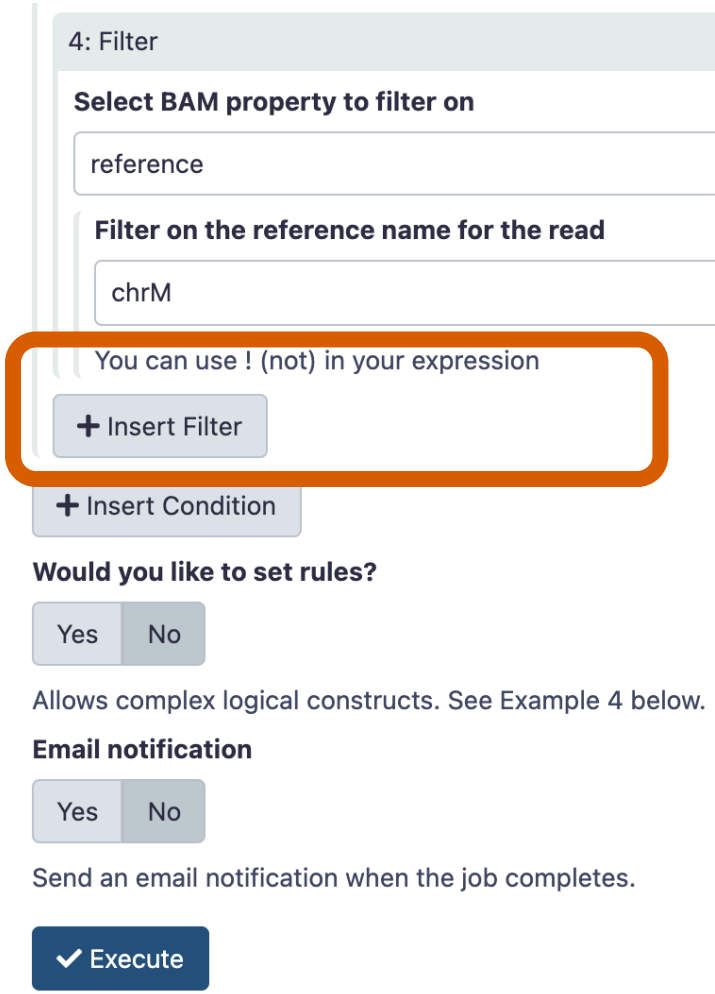


Filtering BAM file



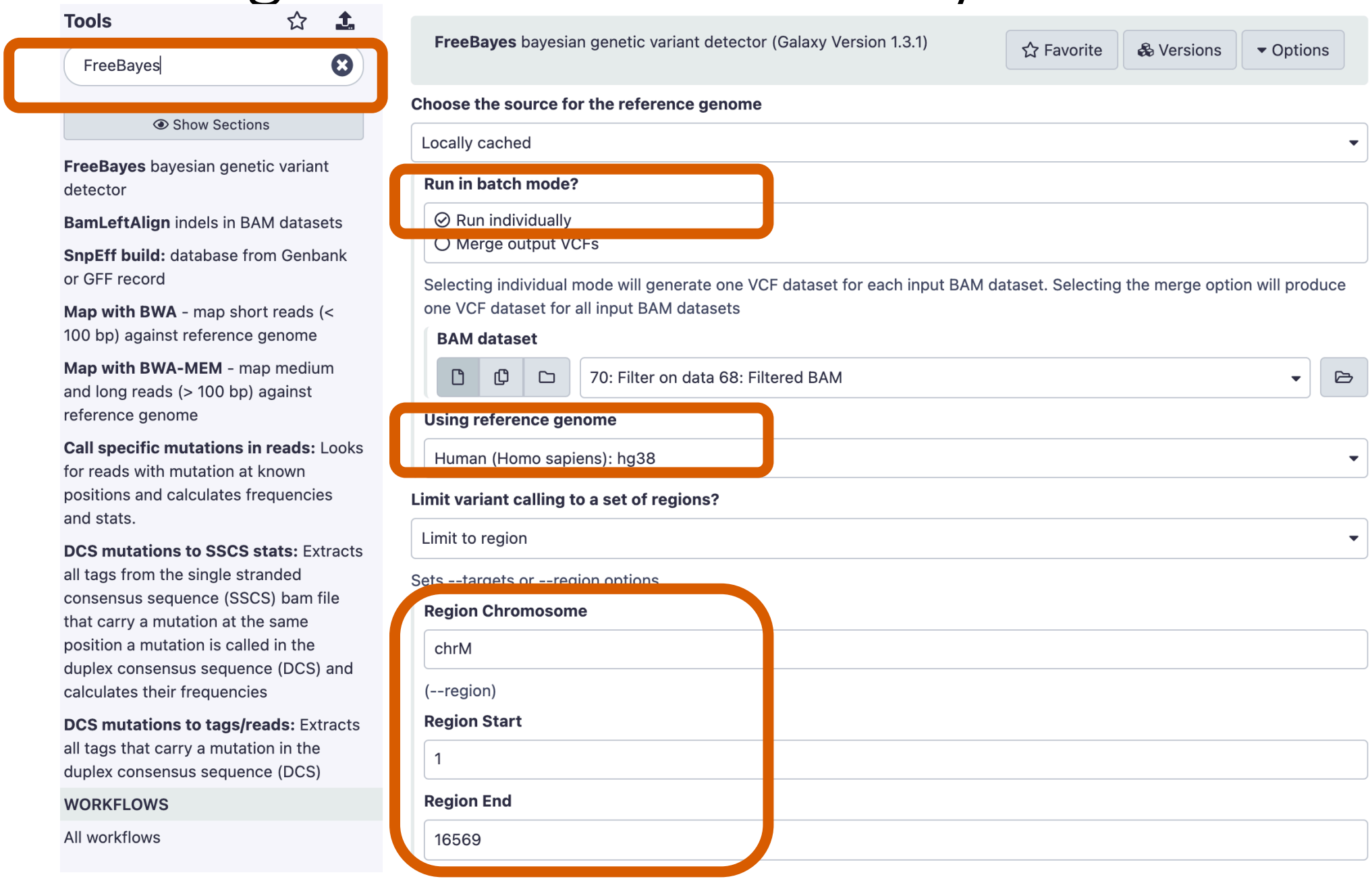


Set the followeing four filters



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



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, regardles of genotyping outcome or model

Yes | No

Allelic scope options

Set alleic scope options

Sets -I, 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

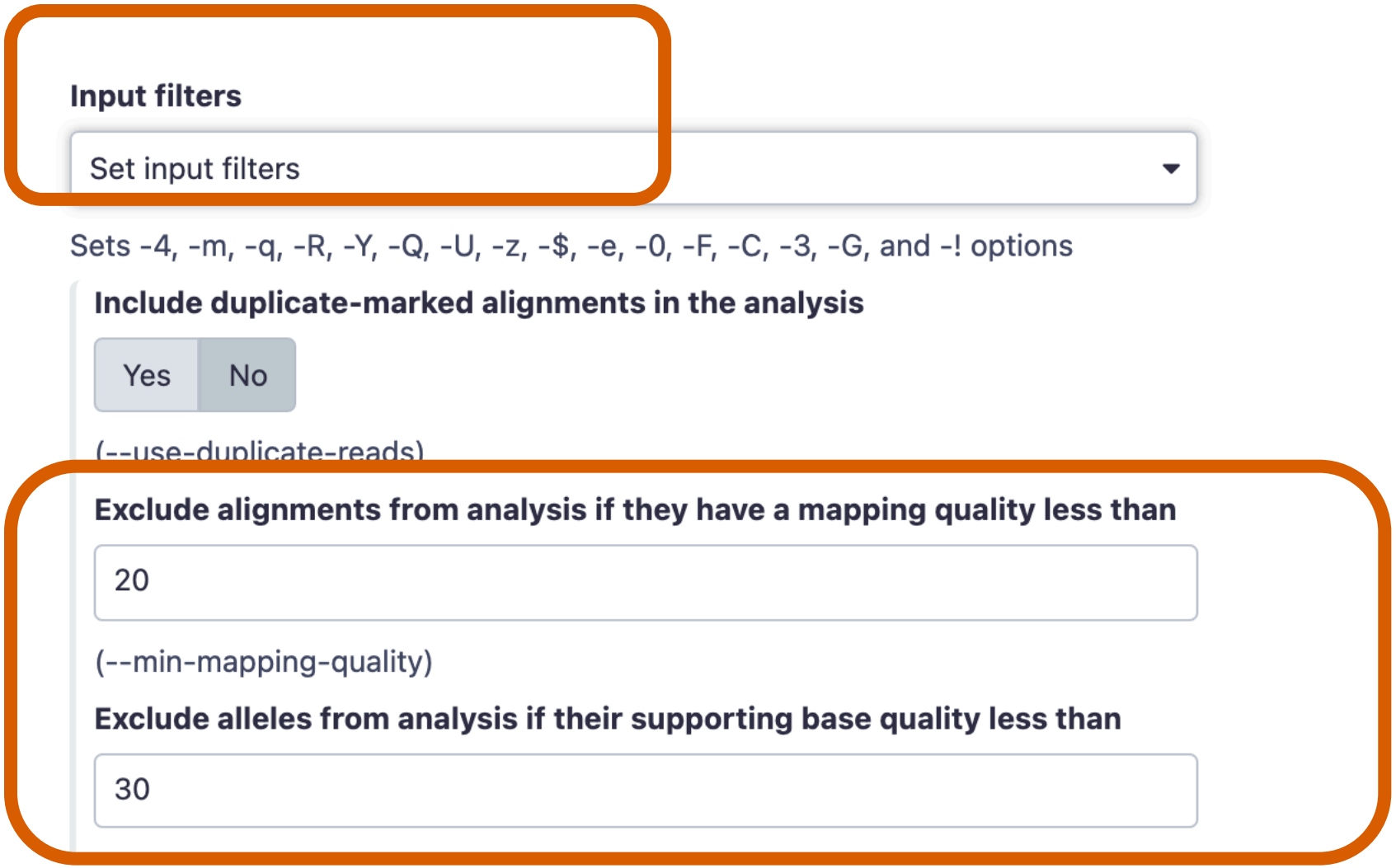
(--no-mnps)

Ignore complex events (composites of other classes)

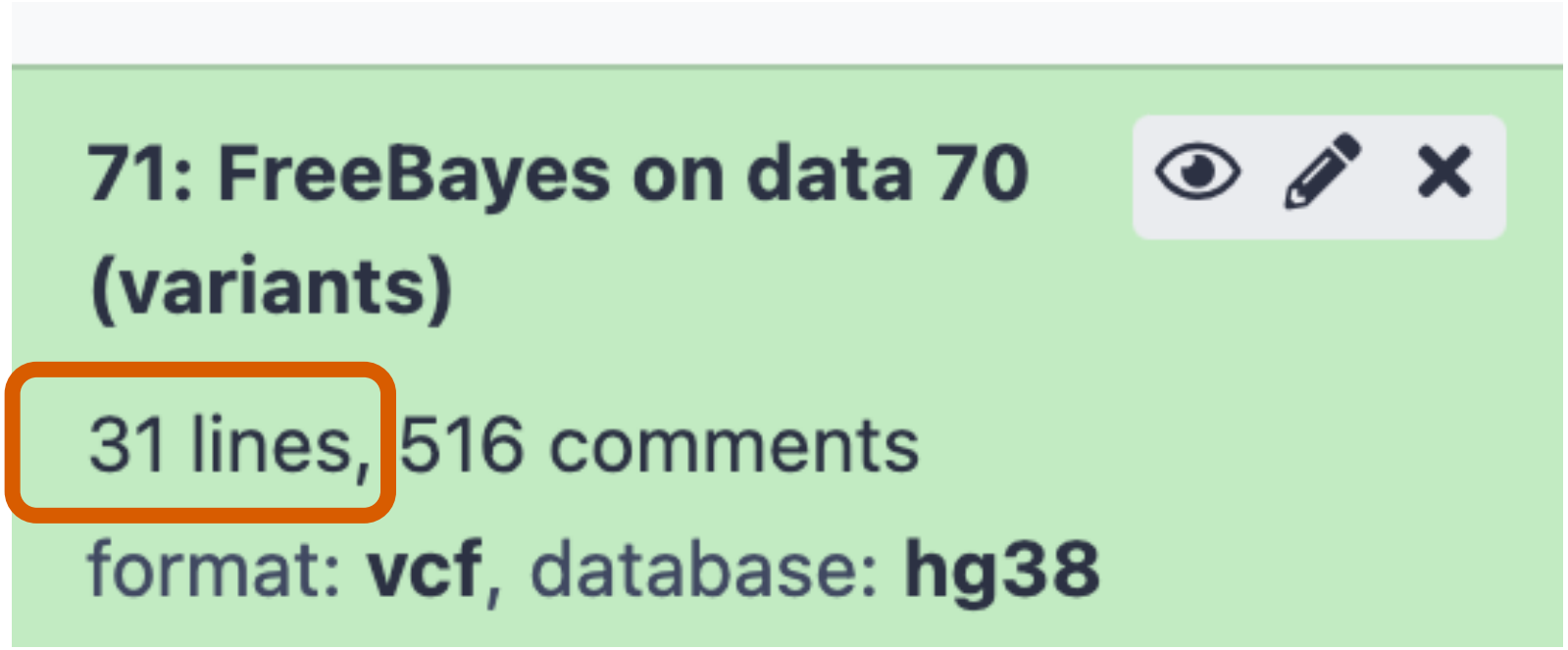
Yes | No

(--no-complex)

Calling variants with FreeBayes

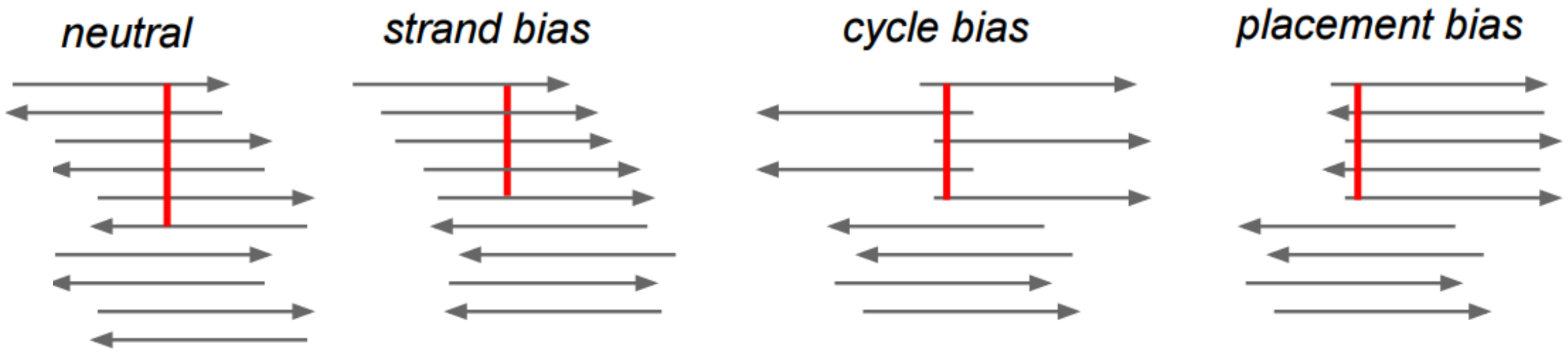


How many variants did you get?

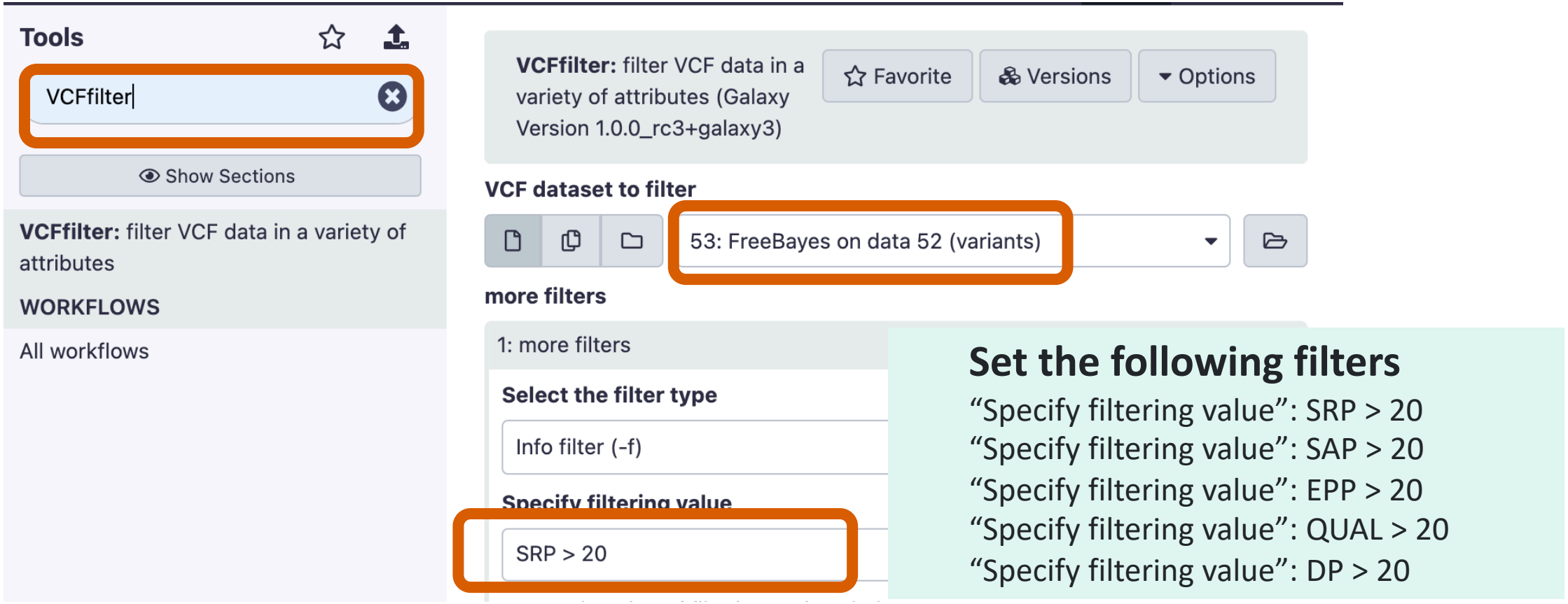


We’re almost there…

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



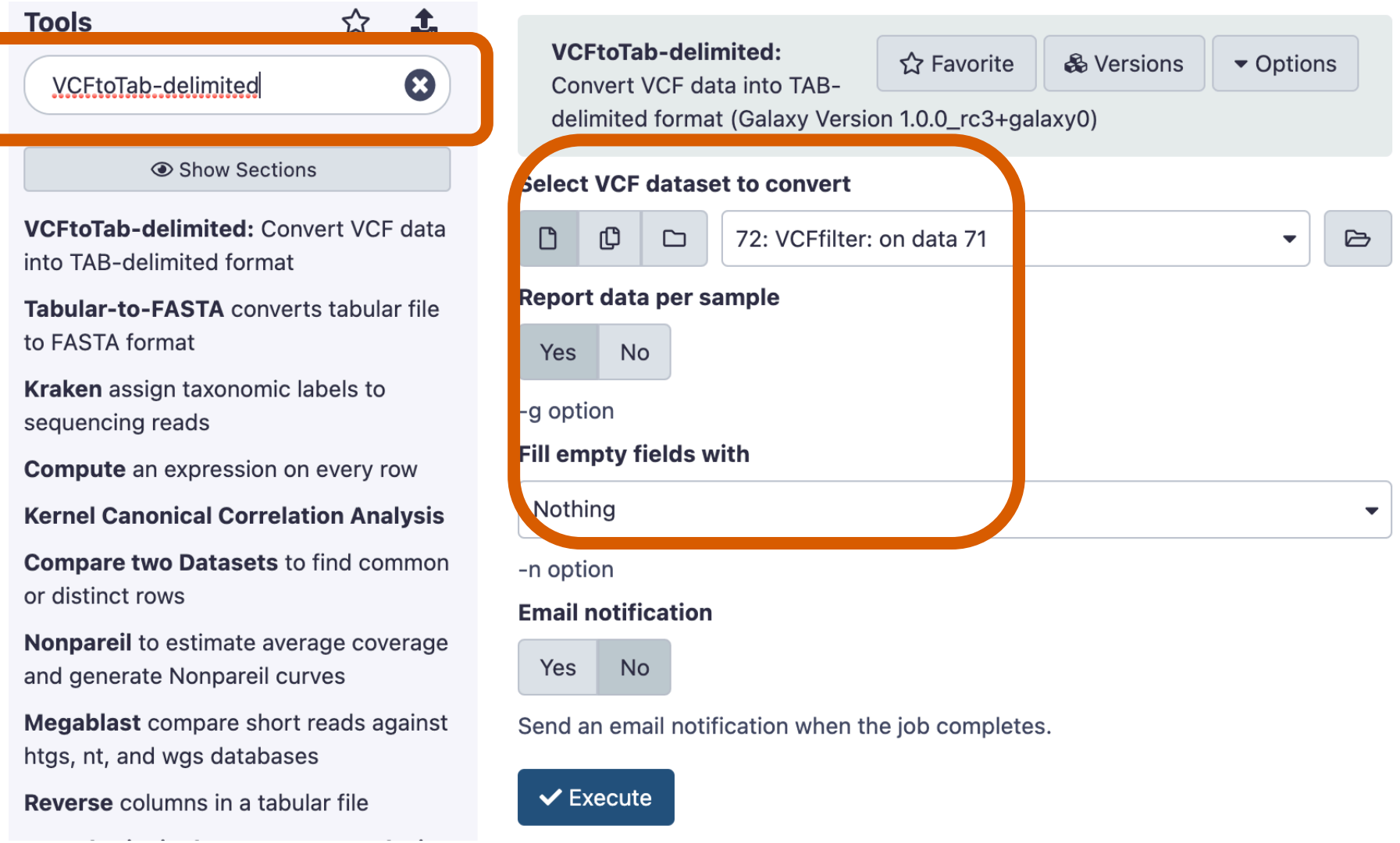
Filtering variants



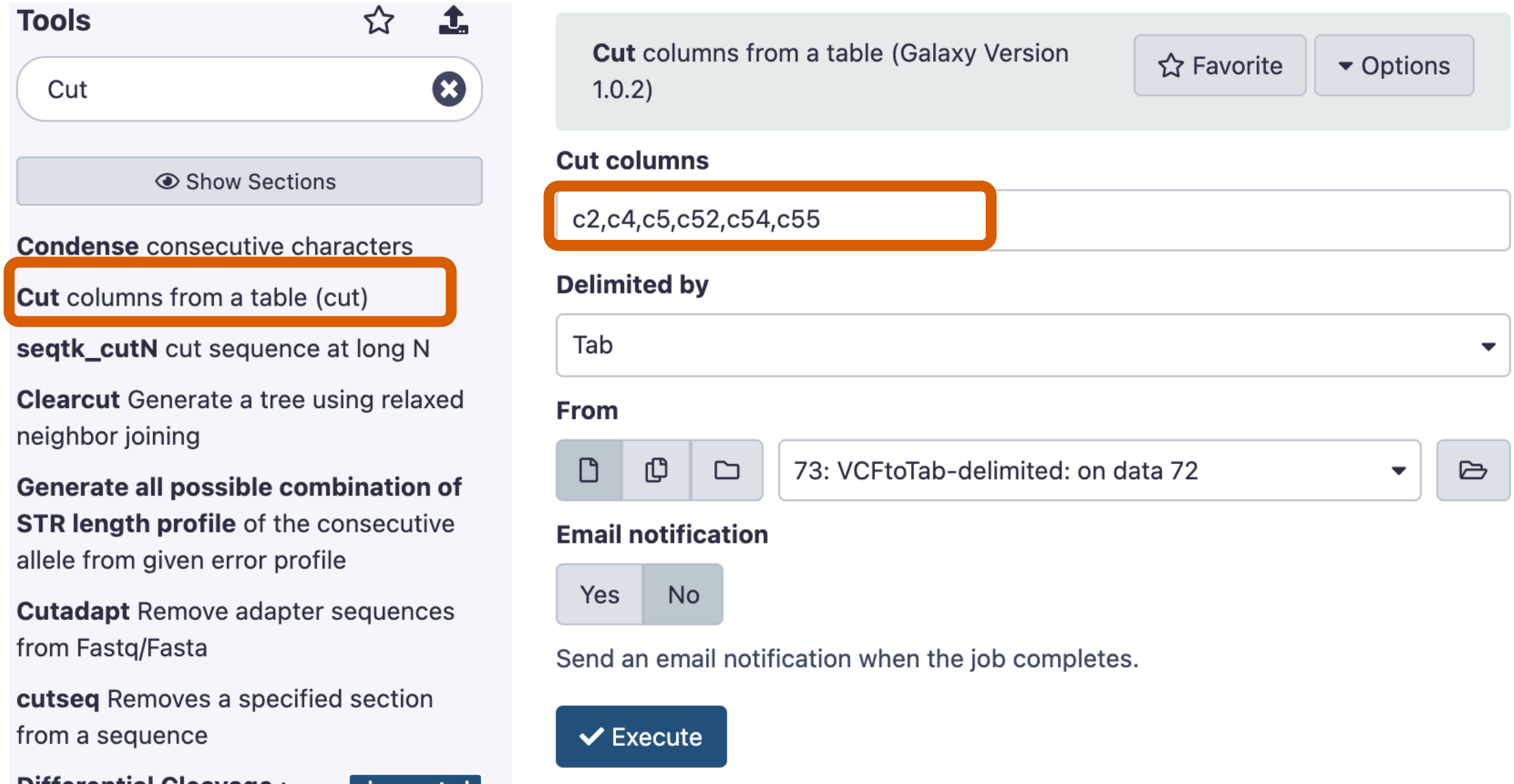
Filtering FreeBayes VCF for strand bias (SPR and SAP), placement bias (EPP), variant quality (QUAL), and depth of coverage (DP).

How many variants survived?

Reformat the VCF file



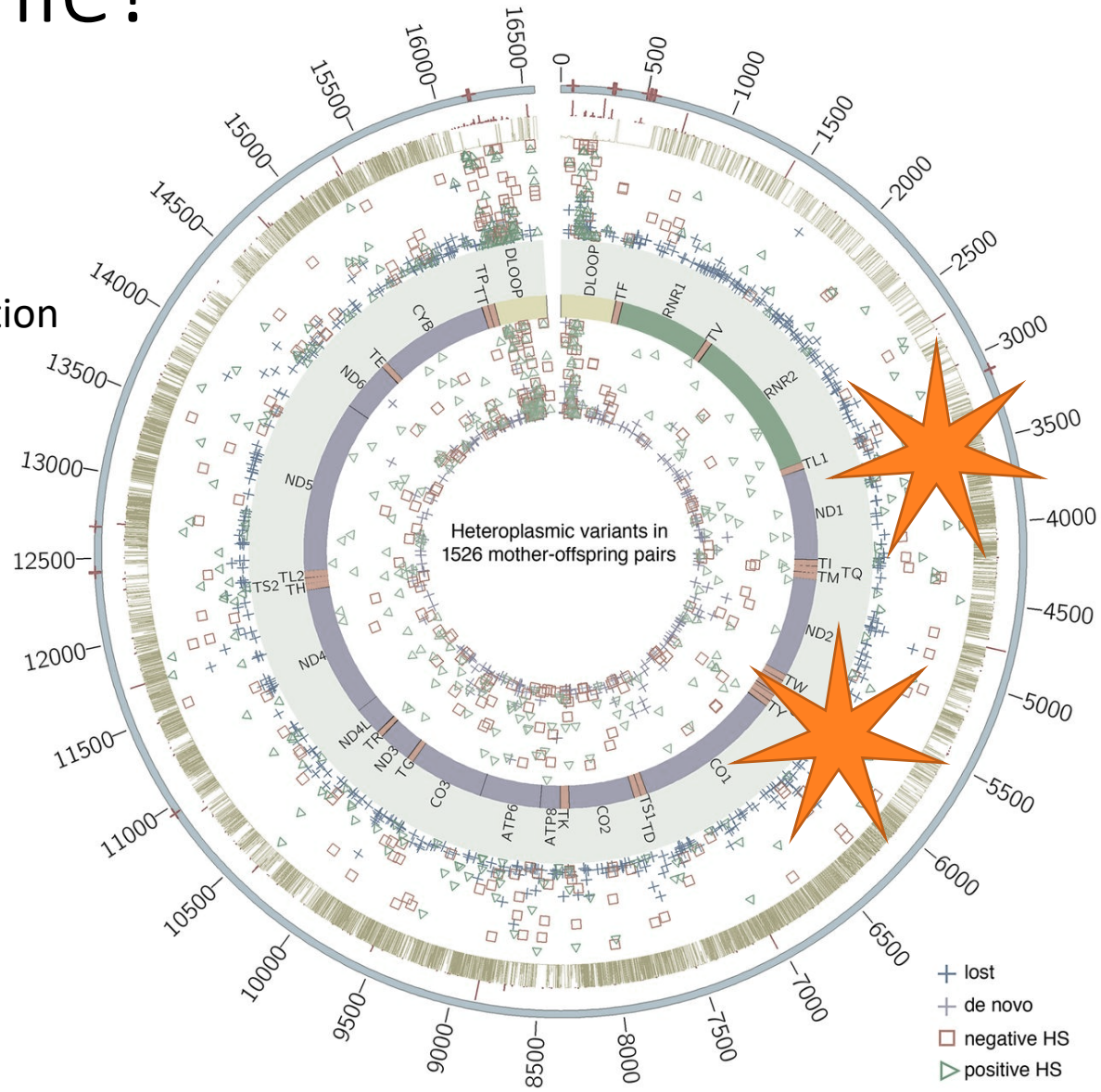
Reformat the VCF file



Take a look on the VCF file!

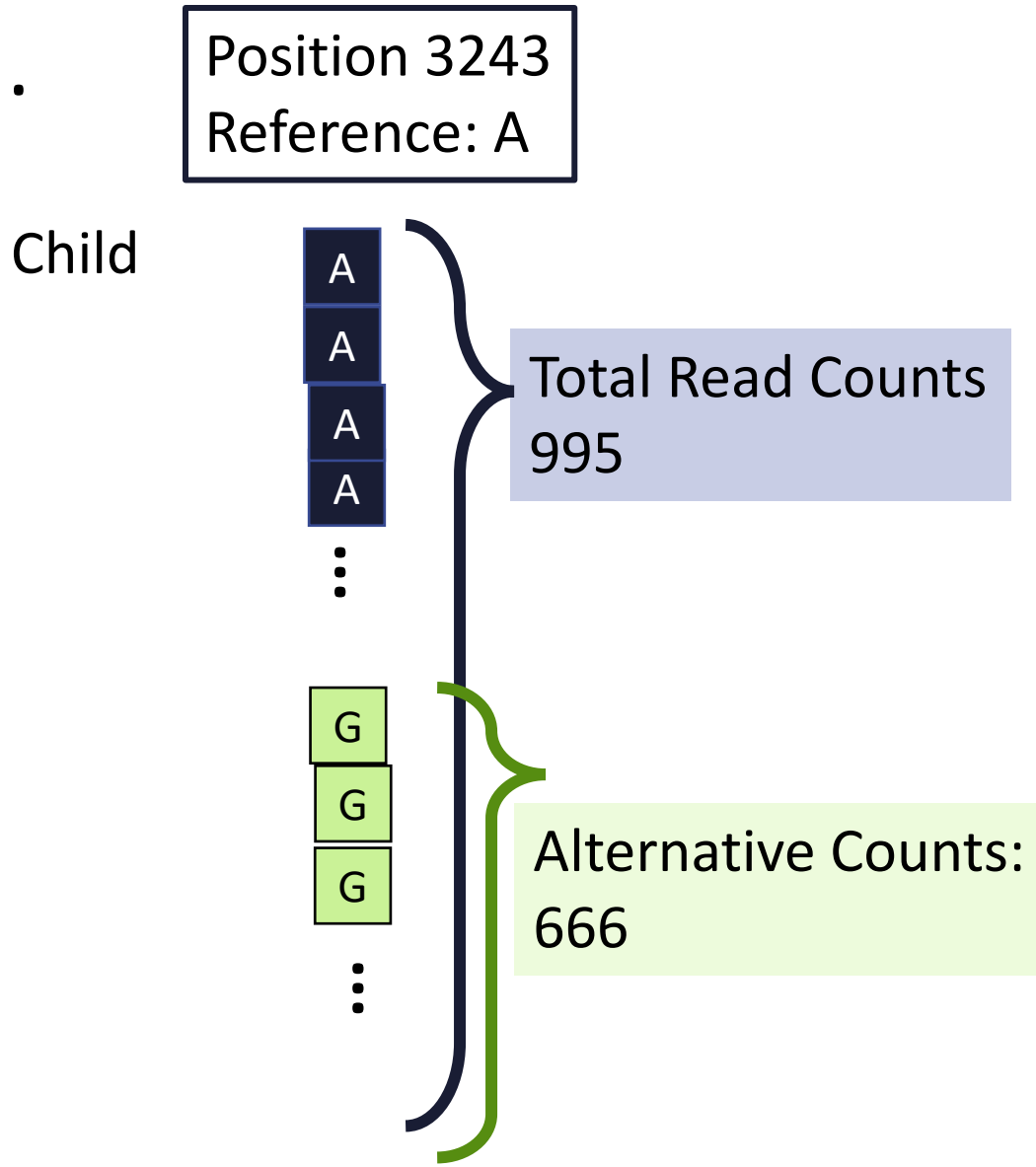
Science 24 May 2019:
DOI: 10.1126/science.aau6520

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

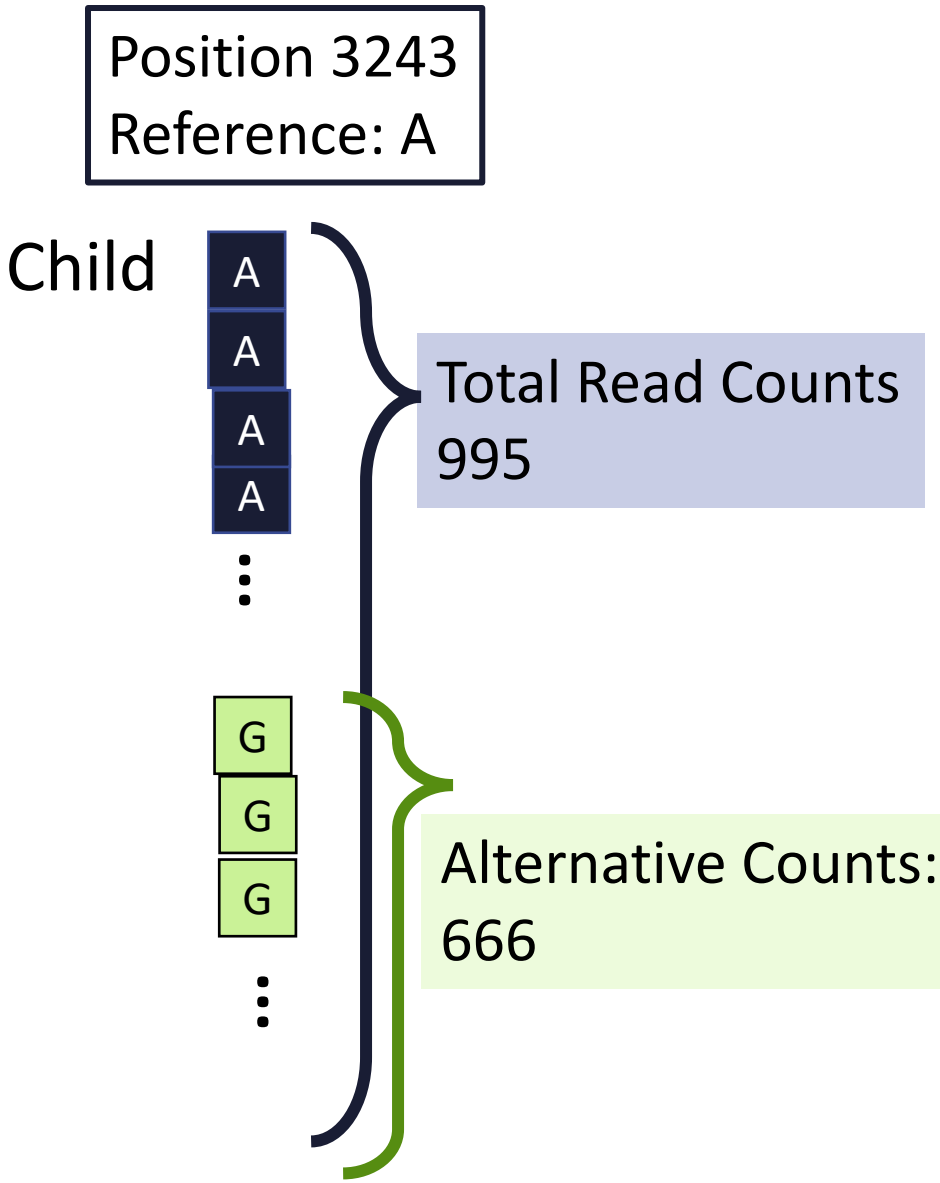


Take a look on the VCF file…

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



Take a look on the VCF file...



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

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

Allele frequency of variants in Mother and Child

Posision 3243
Child : G = 0.67
Mother : G = ???

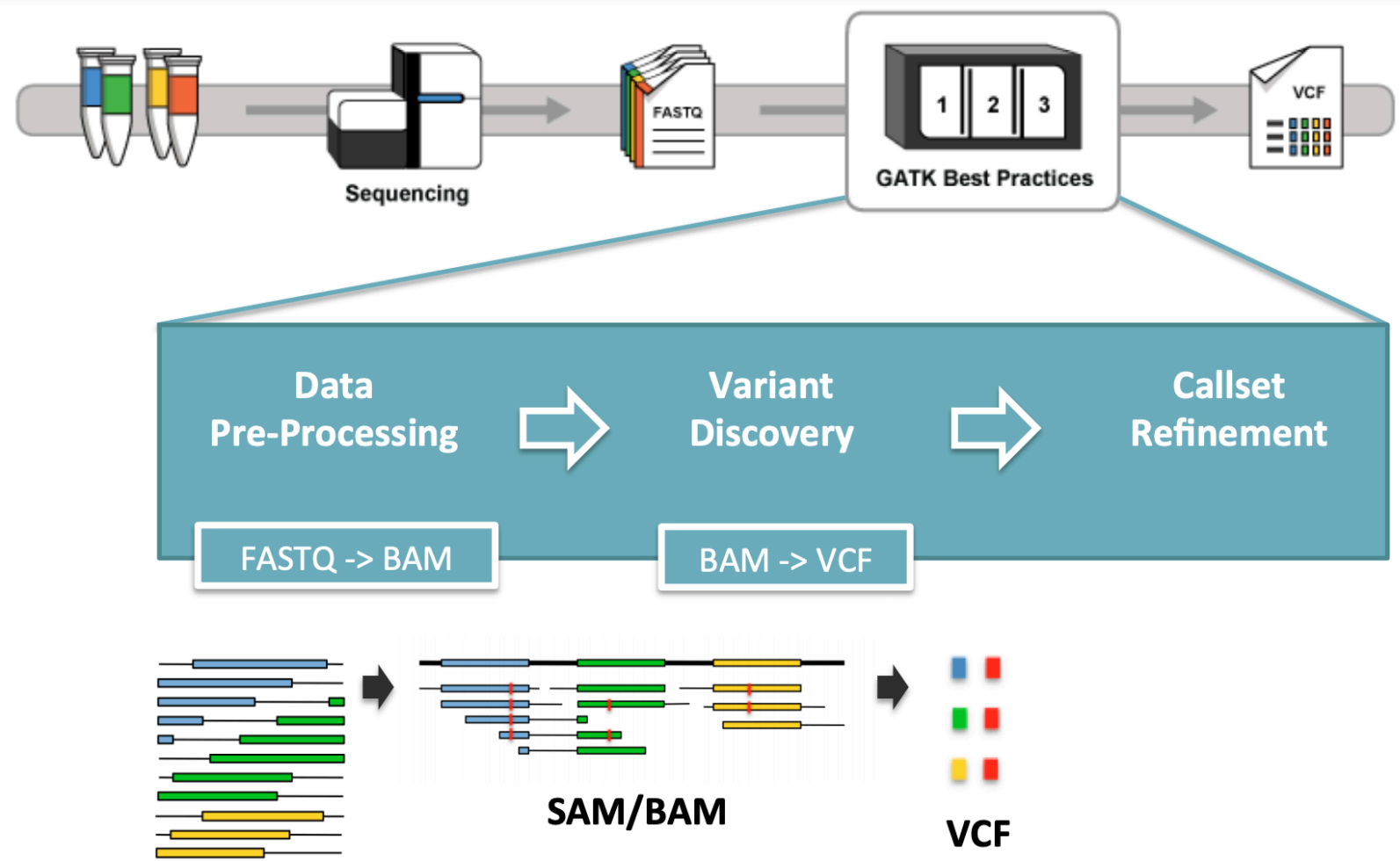
Posision 5538
Child : G = ???
Mother : G = ???

Allele frequency of variants in Mother and Child

Posision 3243
Child : G = 0.67
Mother : G = 0.33

Posision 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