

BIO326 Genome sequencing; tools and analysis

Marie Saitou

Goal of today's class

- Learn how to analyze genome sequence data

We will learn

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



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

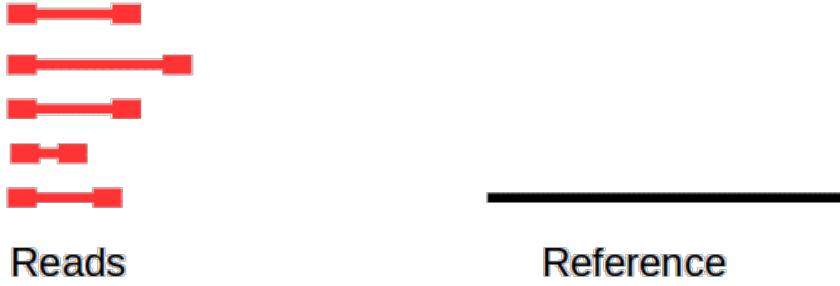
Today's schedule:

- **[Lecture]** Review the genome sequencing pipeline
- **[Lecture]** Go over how to analyze the genome data today
- **[Group work]** Do the analysis by yourself
- **[Short break]**
- **[Group work]** Do the analysis by yourself
- **[Lecture]** Summarize today's lesson

Review: Coverage vs Read Depth

Review: Coverage vs Read Depth

(A) coverage in terms of redundancy

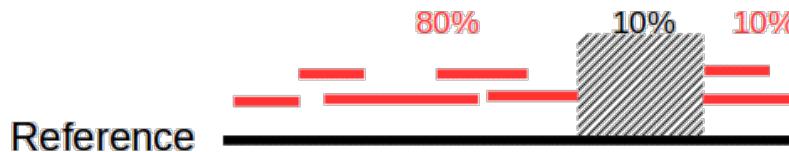


$$C = \frac{\text{\# sequenced bases}^1}{\text{\# bases of reference}}$$

(= \# bases of all mapped reads)

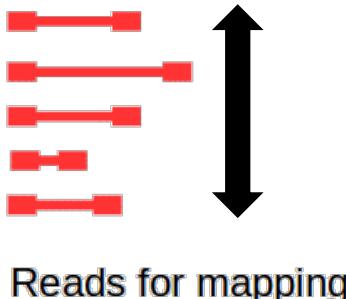
The average depth of sequencing coverage: **LN/G** ,
where L : read length, N : the number of reads, G : the genome length.

(B) percentage of coverage



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

(C) sequencing depth



Sequencing depth = total read number

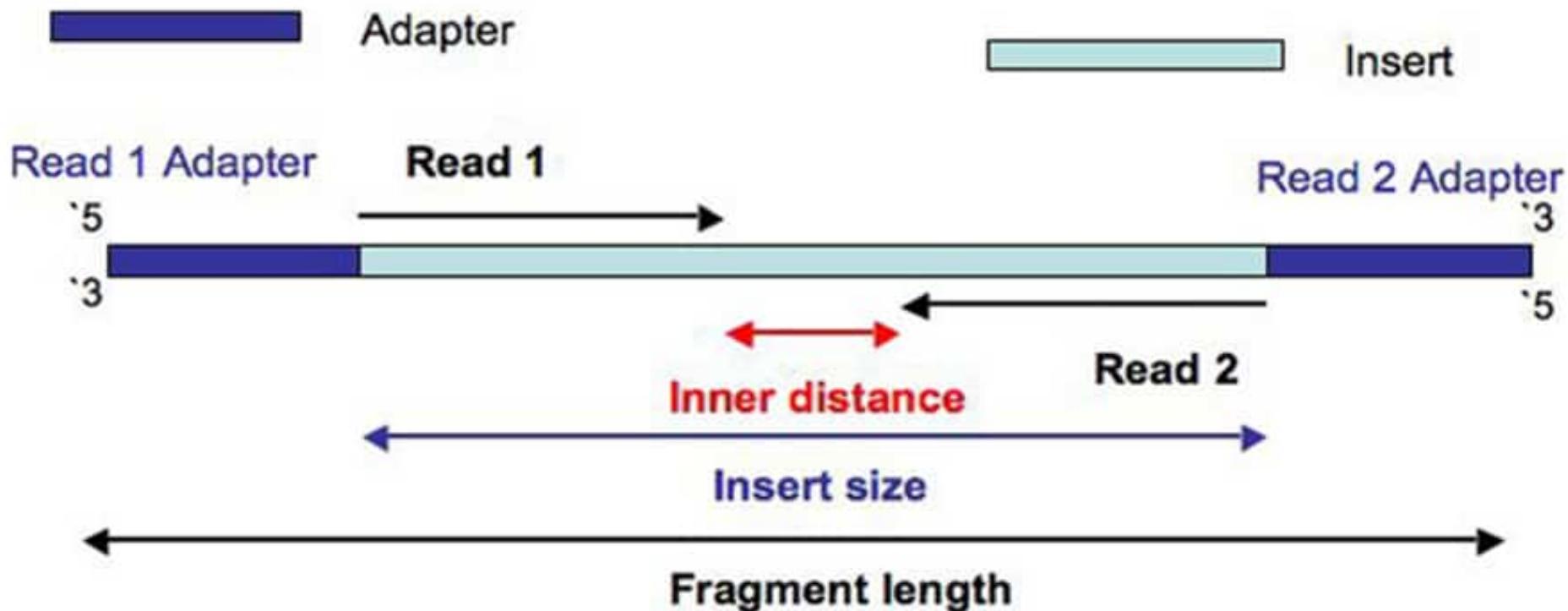
<https://www.ecseq.com/support/ngs/how-to-calculate-the-coverage-for-a-sequencing-experiment>

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

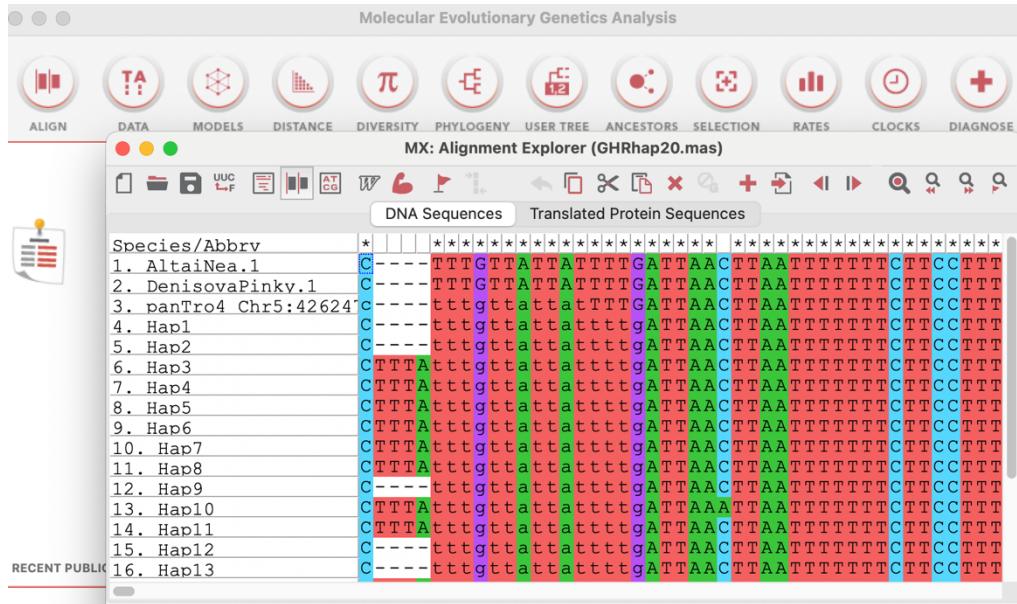


<https://thesequencingcenter.com/knowledge-base/what-are-paired-end-reads/>

Intro to Bioinformatics

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

GUI: Graphical User Interface



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.gz 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.gz 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.gz left on device
20201020_20-lee-007/Ghr-0077_GT20-15739_TTGCTCTAG-TAAGTGGT_S75_L002_R1_001.fastq.gz
tar: 20201020_20-lee-007/Ghr-0077_GT20-15739_TTGCTCTAG-TAAGTGGT_S75_L002_R1_001.fastq.gz 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.gz left on device
```

- Intuitive
- Easier to learn

- 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

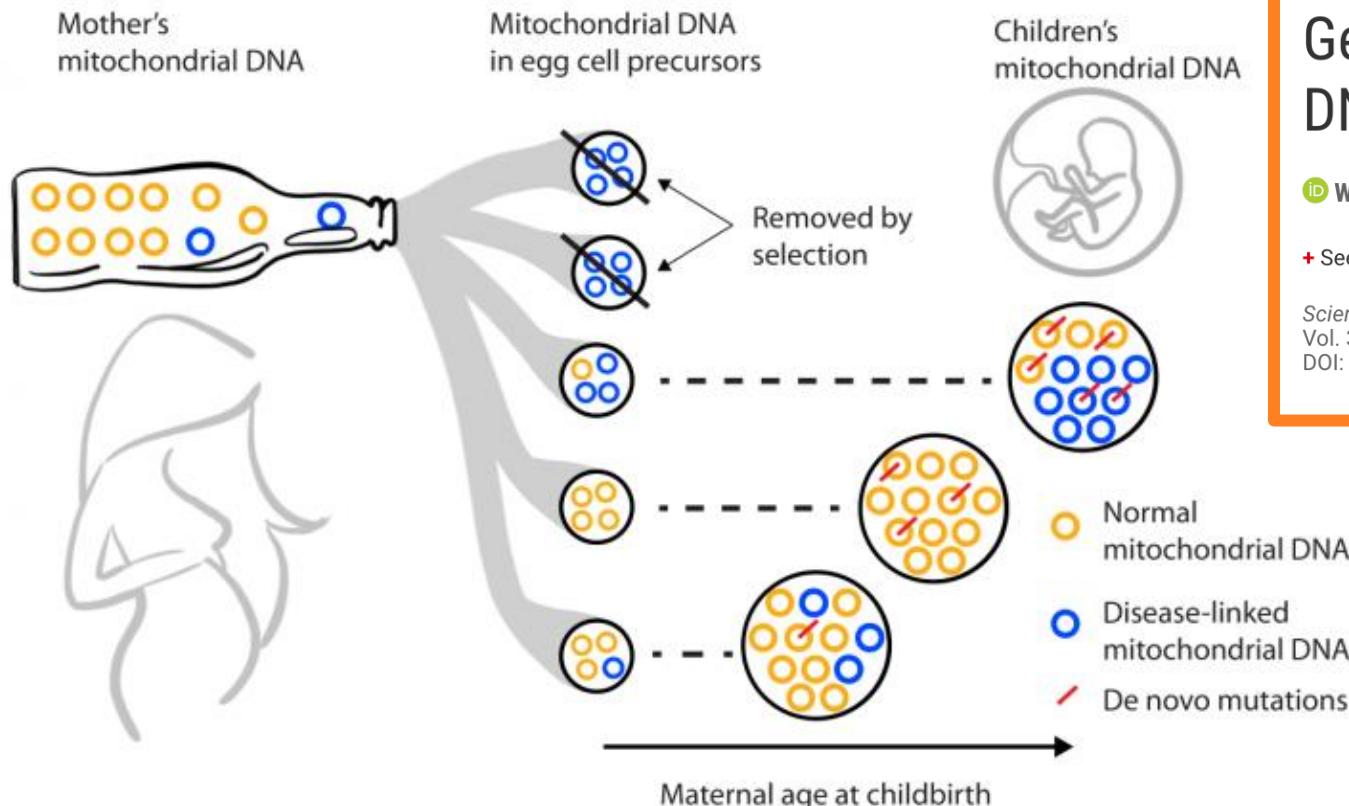


Common errors

- Typo
- Input file is missing / specified wrongly
- 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. Keogh¹, 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

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
ATCTGGTTCCTACTTCAGGGCATAAAACCTAAATAGCCCACACGTTCCCC
+
BCCCCFFFFFFGGGGGGGGGGHHHHGHGHHHHHHHGGGGGGHHHHGH
@M01368:8:000000000-A3GHV:1:1101:14518:9998/1
GTTATTATTATGTCTACAAGCATTAATTAACACACTTAGTAAGTA
+
AAAAAFFFFFFGGGGGGGGGGHGGHHHHGHGHHHHHHGCGHHHHHHHHHH
@M01368:8:000000000-A3GHV:1:1101:18422:19051/1
GTATCCGACATCTGGTTCCTACTTCAGGGTCATAAAACCTAAATAGCCCAC
.
```

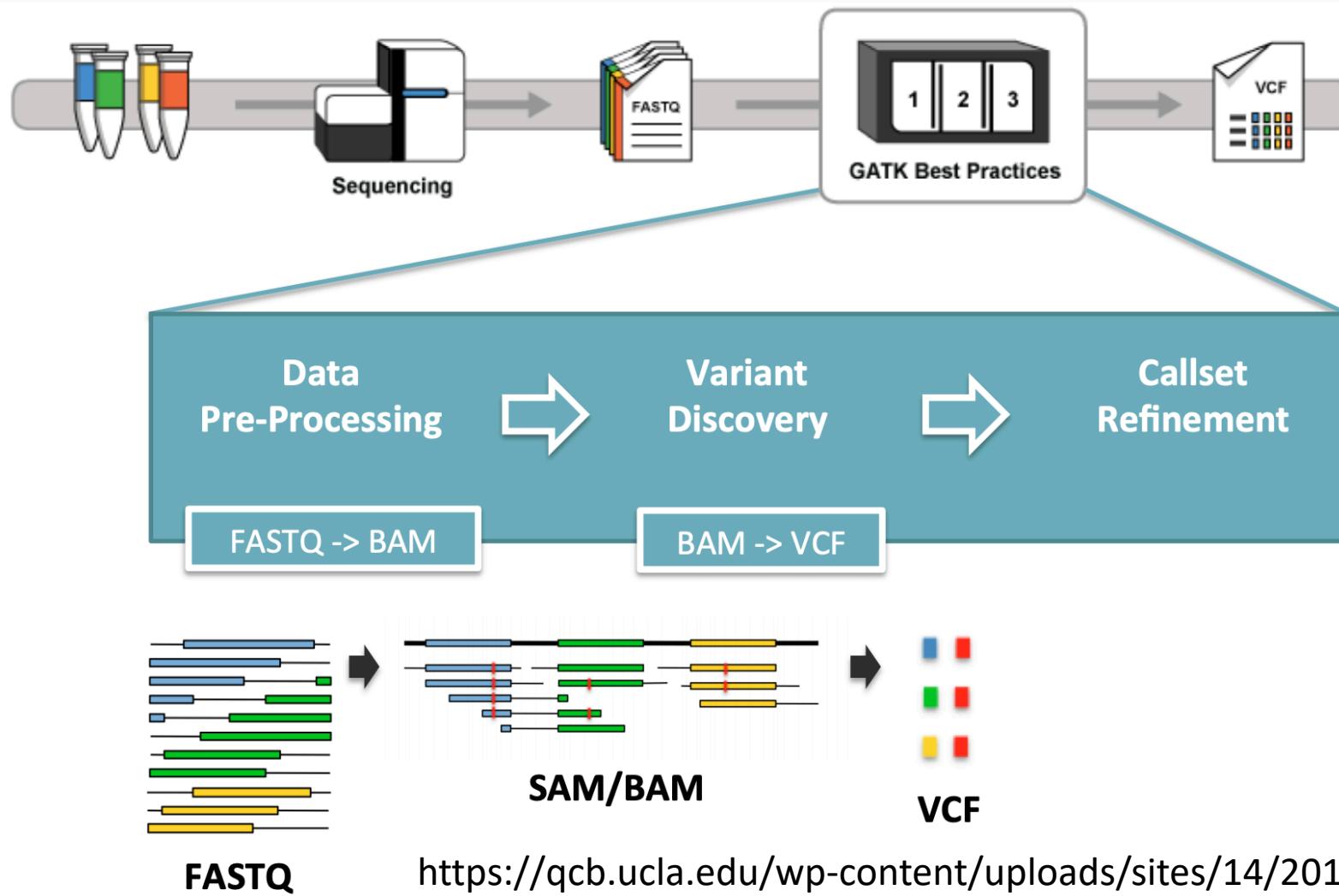
- Goal: Calculate “Allele frequency of variants in Mother and Child”

Position: Where is/are the variants?

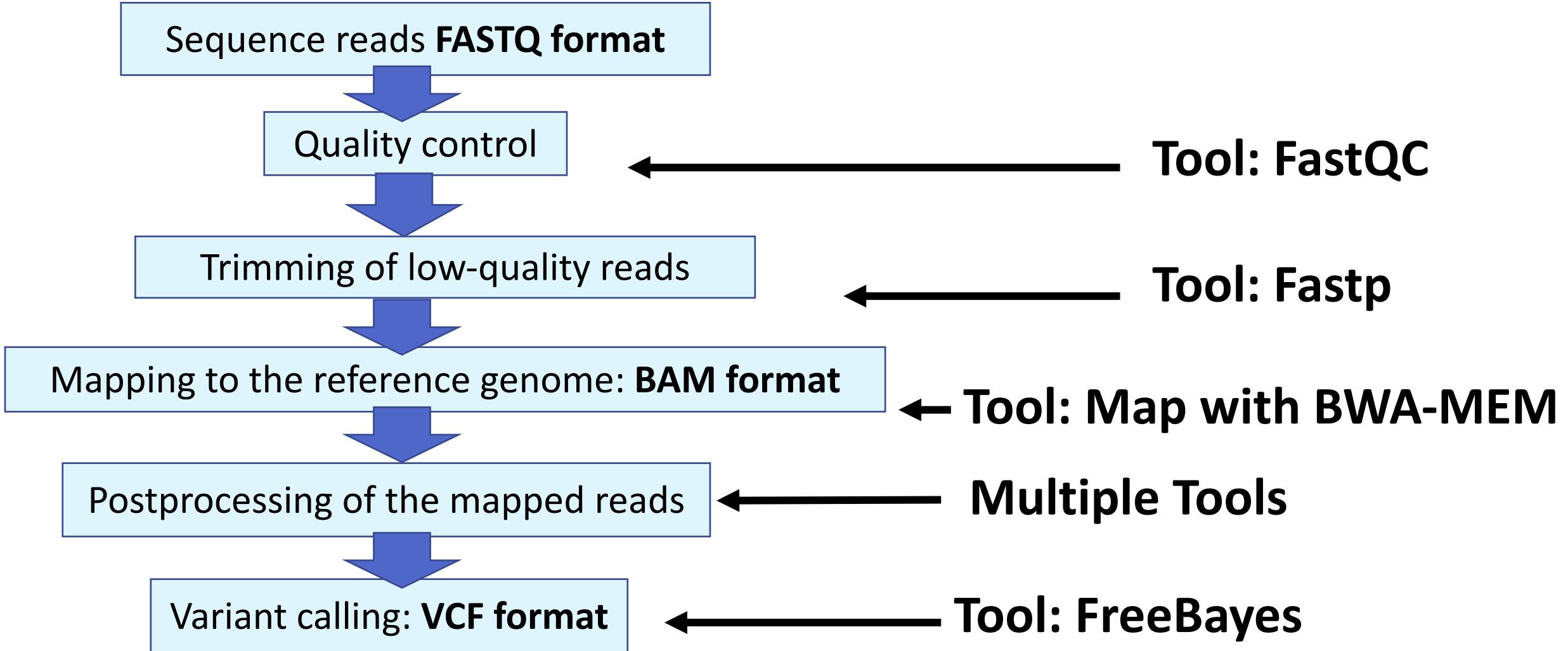
Child : How much is the observed frequency?

Mother : How much is the observed variant frequency?

Reads-to-variants workflows



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 BAM/SAM, FASTA



Accessibility/Reproducibility/Transparency

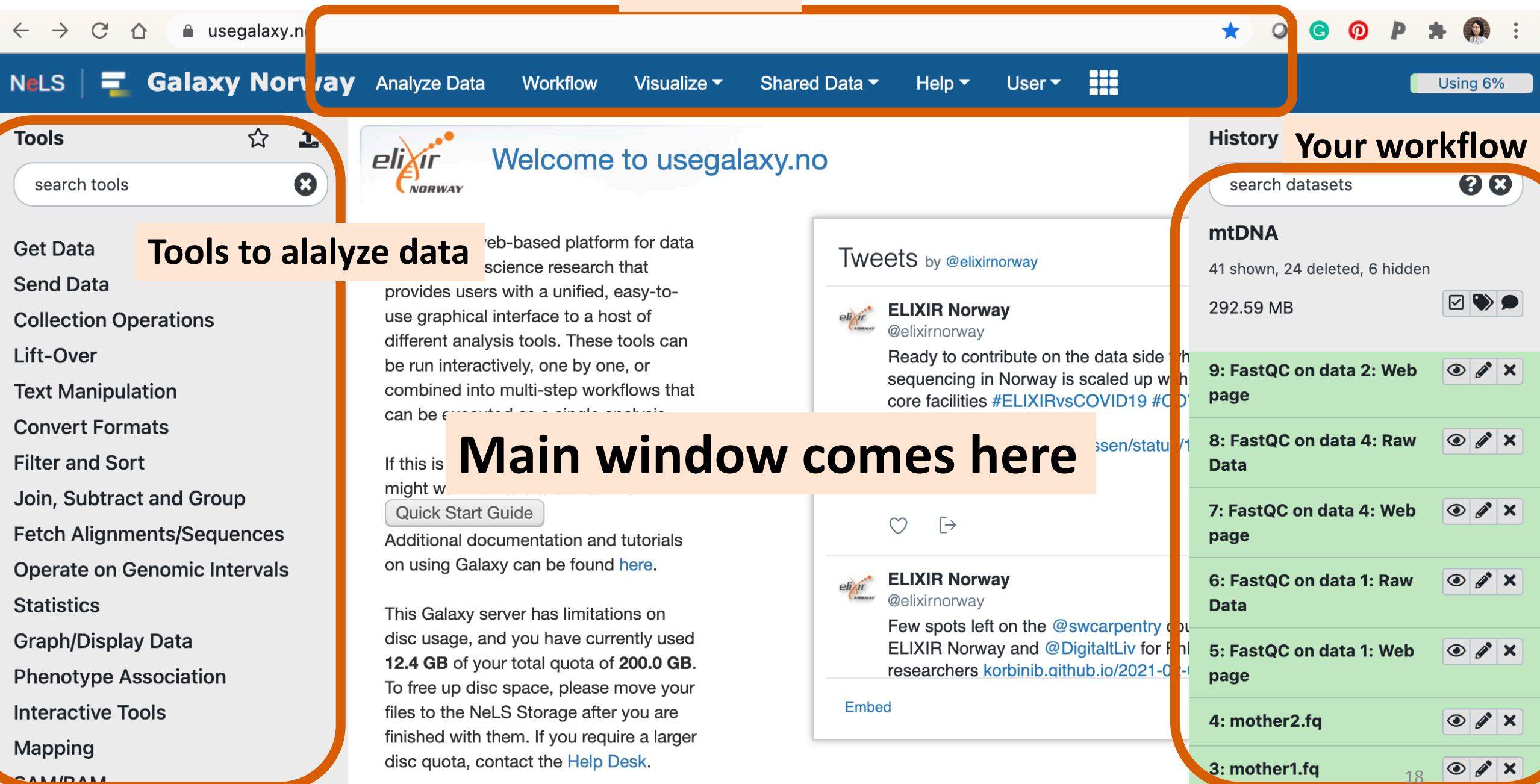
- <https://usegalaxy.no/>
- <https://usegalaxy.eu/>
- <https://usegalaxy.org/>

We have increased the machine
from 20 cores/200 GB RAM to 84 cores/456 GB RAM.

- Tue, 2 Feb, 19:52

Select one platform you like, and if one gets stuck and another runs smoothly, let's move to the smooth one.

Menu Bar



Job Status

Colour	Status
 5: Filter on data 1	Queued
 4: Filter on data 1	Running
 3: Filter on data 1	OK
 2: Filter on data 1	Error

Be patient and wait

Let's go to next step

Hmm... Let's examine it

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

The screenshot shows a bioinformatics tool interface with a red header bar. The header contains a red circle with a white 'X' icon, followed by the text "65: BamLeftAlign on d" and "ata 61 (alignments)". To the right of the text are three icons: a magnifying glass, a pencil, and a close button. Below the header, the word "error" is displayed. A blue callout box with the text "Click to expand the menu" is positioned over the error message. The main body of the interface displays the error message: "An error occurred with this dataset." followed by two lines of log output: "/data/part0/tmp/jobs/015/15271/tool_scr" and "/data/part0/tmp/jobs/015/15271/tool_scr". Below this, there is a row of three icons: a gear, an info symbol, and a question mark. This row is highlighted with a blue rectangular border and has a blue arrow pointing upwards from the bottom left towards it. To the right of this row are two smaller icons: a speech bubble and a gear. At the bottom, there are three links: "display at UCSC main test", "display with IGV local", and "display in IGB View". A blue callout box labeled "Let's examine it" is located at the bottom left.

- ## 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 user interface for managing datasets. At the top, there are 'Help' and 'User' dropdown menus. A prominent orange header bar displays the text 'New history: click +' and 'Using 6%'. Below this, a navigation bar includes 'History' (highlighted with an orange box), a refresh icon, a plus sign for creating new datasets, a search icon, and a gear icon for settings. A 'Create new history' button is also visible. To the left, a sidebar contains a message: 'Click here to rename it as you like' with an arrow pointing to the 'Unnamed history' section. The main area shows a dataset card for 'ELIXIR Norway' with a thumbnail, the author '@elixirnorway', and a detailed description. A blue callout box at the bottom right of the main area says: 'This history is empty. You can **load your own data** or **get data from an external source**'.

Help ▾ User ▾ New history: click + Using 6%

History refresh + search gear

search datasets Create new history

Click here to rename it as you like → Unnamed history

(empty)

ELIXIR Norway
@elixirnorway

Ready to contribute on the data side when sequencing in Norway is scaled up with core facilities #ELIXIRvsCOVID19 #COVID19pandemi
<https://twitter.com/ingejonassen/status/13851520>

This history is empty. You can **load your own data** or **get data from an external source**

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

The screenshot shows the Galaxy Norwave web interface. On the left, a sidebar lists various tools: 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, and Mapping. The 'Get Data' tool is currently selected. At the top, there is a navigation bar with links for NeLS, Galaxy Norwave, Analyze Data, Workflow, Visualize, Shared Data, Help, User, and a grid icon. Below the navigation bar, a message says 'Download from web or upload from disk'. A table shows one file added to the queue: Name (New File), Size (-), Type (Auto-de...), Genome (Additional ...), Settings (gear icon), and Status (0%, trash bin icon). A note below the table says 'Download data from the web by entering URLs (one per line) or directly paste content.' At the bottom, there are buttons for 'Choose local files', 'Paste/Fetch data' (which is highlighted with an orange box), 'Start', 'Select', 'Pause', 'Reset', and 'Close'. A large orange box highlights the 'Paste/Fetch data' button with the instruction '2. Click here'. Another large orange box highlights the 'Get Data' tool in the sidebar with the instruction '3. Paste the four URLs here all at once -> start'.

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

Galaxy Norwave

Analyze Data Workflow Visualize Shared Data Help User

Download from web or upload from disk

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

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

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 shows a dataset editing interface. At the top, there are buttons for "Edit attributes", "Auto-detect", and "Save". The "Save" button is highlighted with an orange box. Below these are sections for "Name" and "Info". The "Name" section contains a text input field with the value "raw_child-1.fq", which is also highlighted with an orange box. The "Info" section contains the text "uploaded fastqsanger file". Below these are sections for "Annotation" and "Database/Build". A note at the bottom says "Add an annotation or notes to a dataset; annotations are available when a history is viewed." To the right of the main interface is a sidebar showing a history of operations:

- 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

Each item in the history has an "Edit attributes" button, which is highlighted with an orange box for the first item.

Let's examine the FastQ file



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

Show all | Save

History



search datasets

Genome

4 shown



85.25 MB

4: <https://zenodo.org/rec>

Click the eye icon to view data

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

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

Label

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

ATCTGGTTCCTACTTCAGGGCCATAAAACCTAAATAGCCCACACGTTCCCCTAAATA

+

BCCCCFFFFFGGGGGGGGGHHHHGHGHHHHHHHGGGGGGHHHHGHHHHHHHHH

Sequence

Quality score as ASCII symbol

Quality scores

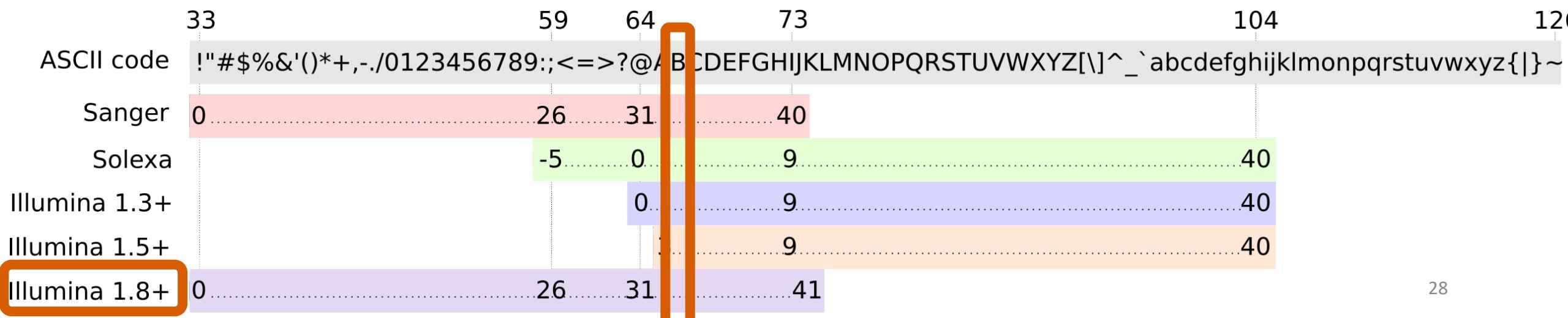
@M01368:8:00000000-A3GHV:1:1101:6911:8255/1

ATCTGGTTCCTACTTCAGGCCATAAAACCTAAATAGCCCACACGTTCCCCTTAAATA

+

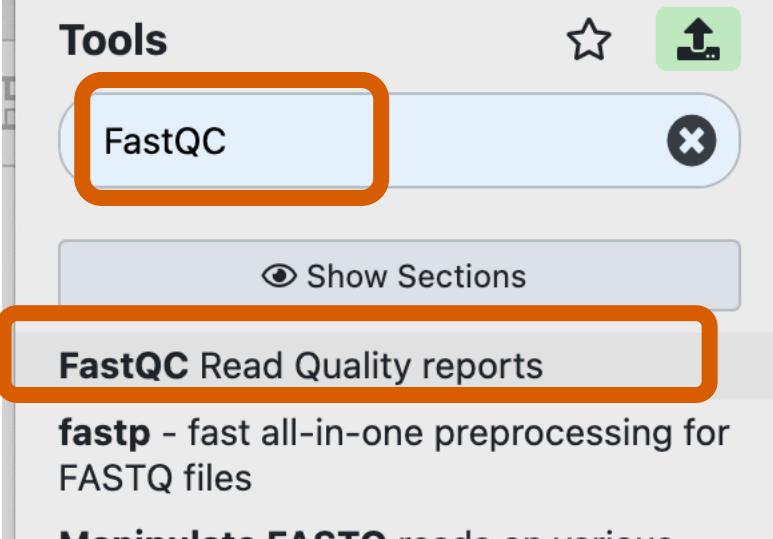
BCCCCFFFFFFGGGGGGGGGGHHHHGHGHHHHHHHGGGGGGHHHHGHHHHHHHHII

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

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

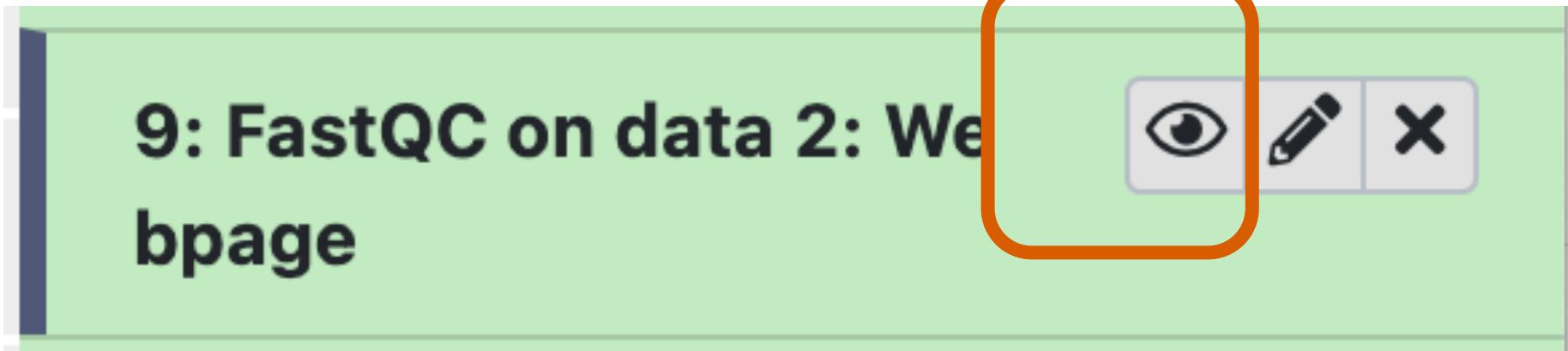
This is a batch mode input field. Separate jobs will be triggered for each dataset selection.



Execute

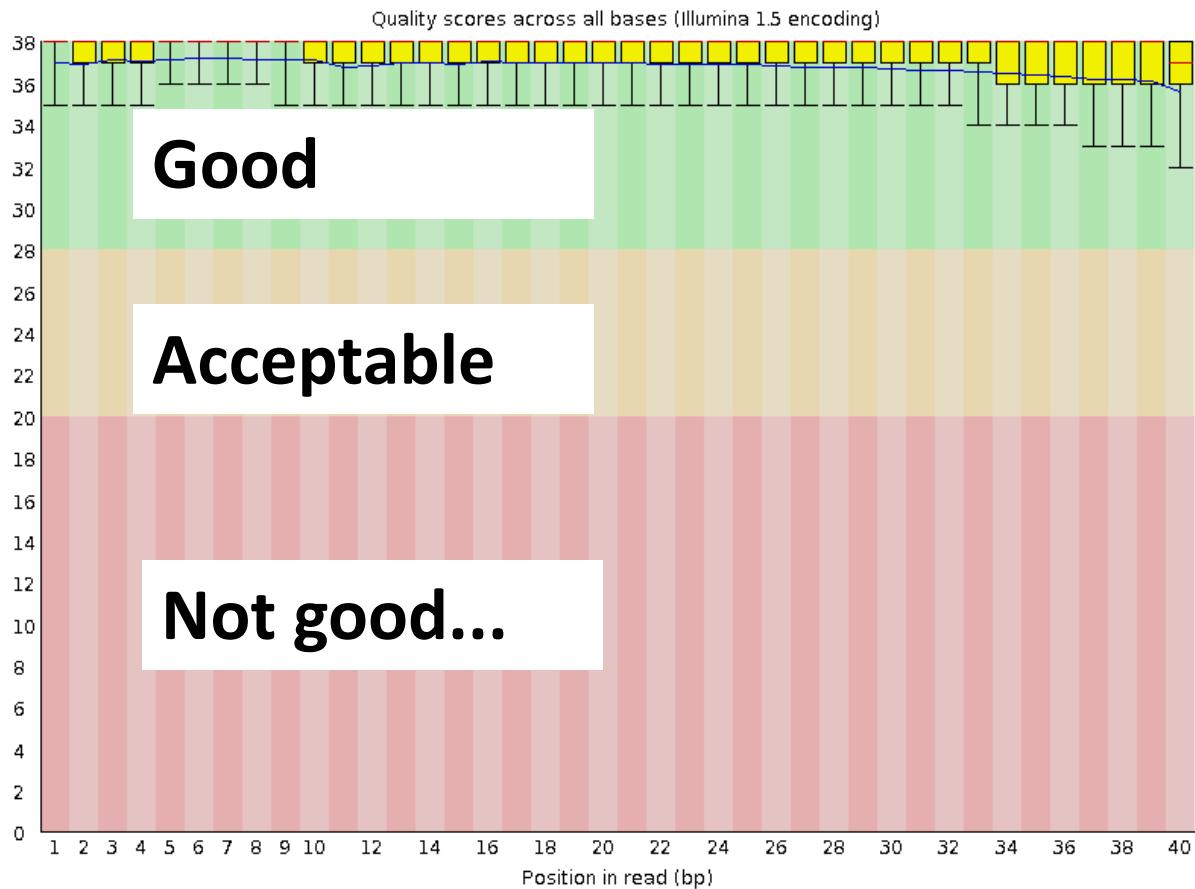
Run FastQC for the quality check

Click the “eye” icon to view data



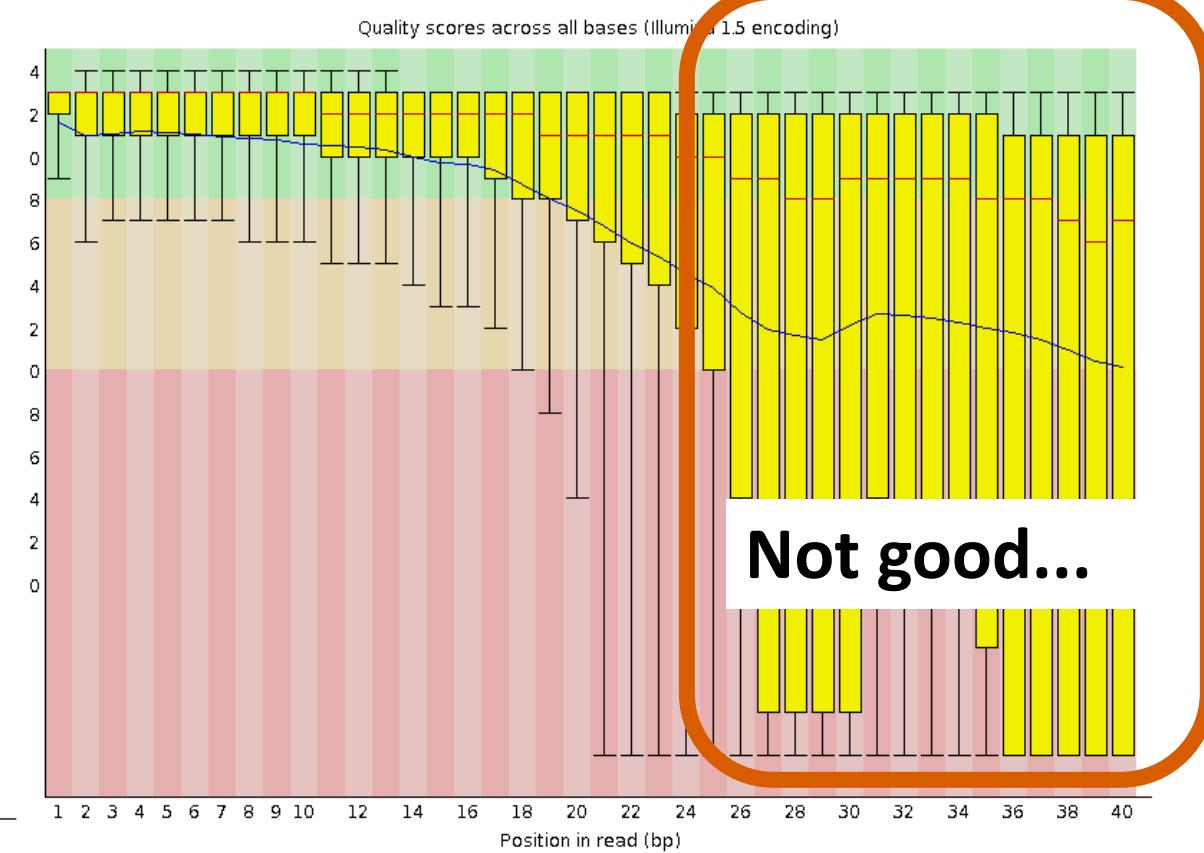
Good Illumina data

Quality



Nucleotide position in read

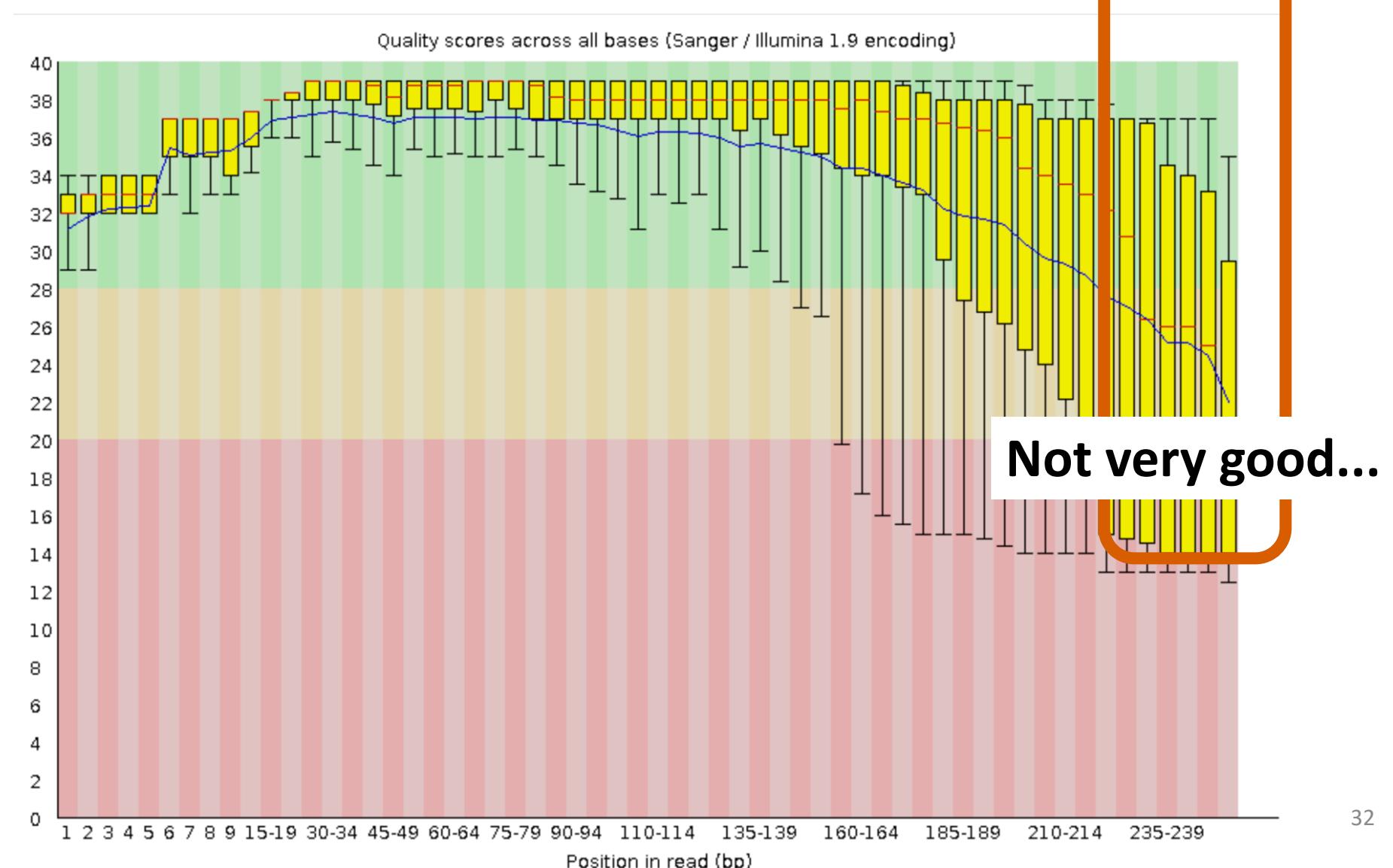
Bad Illumina data



Nucleotide position in read

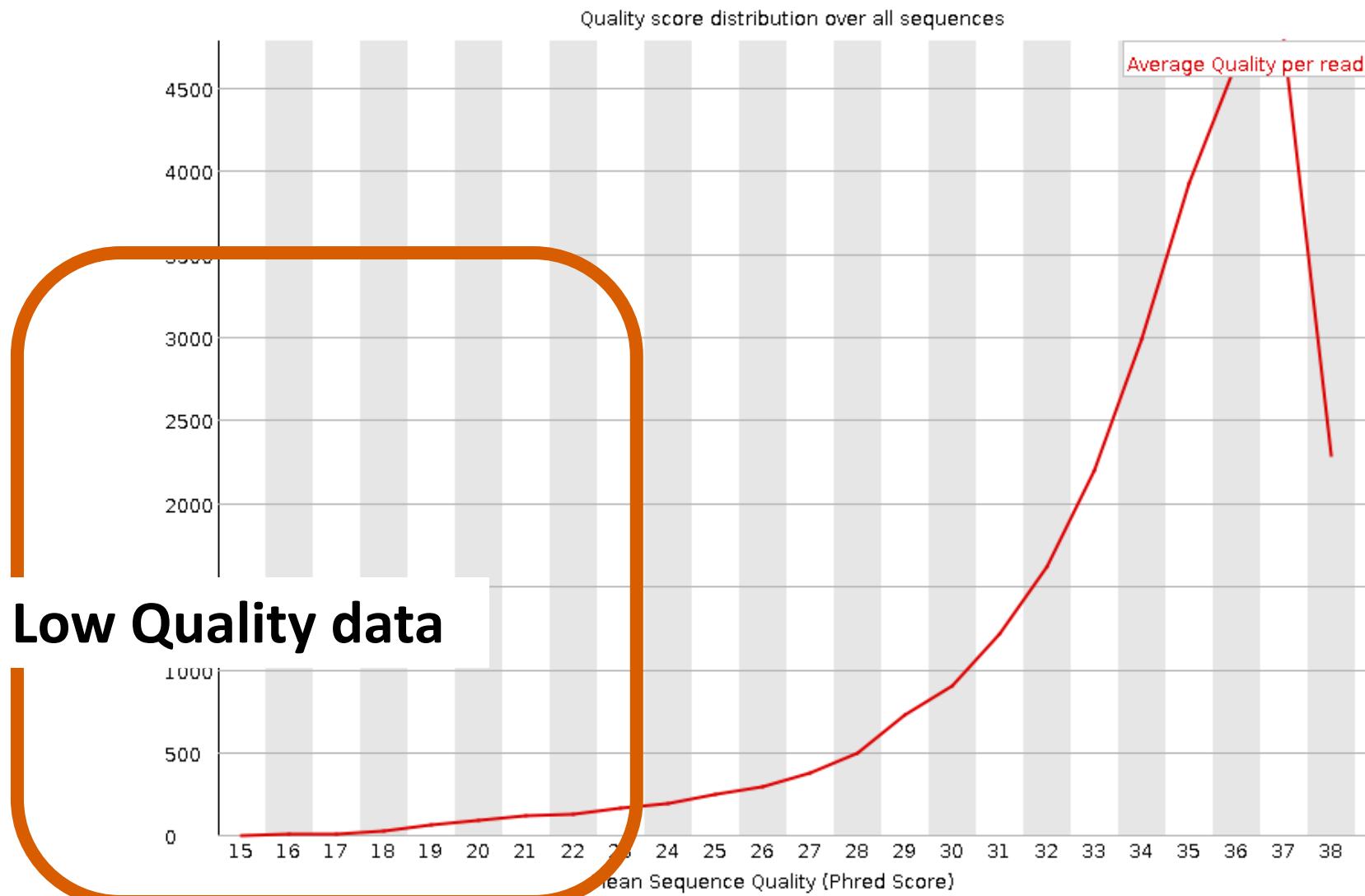
FastQC Data Interpretation

Per base sequence quality

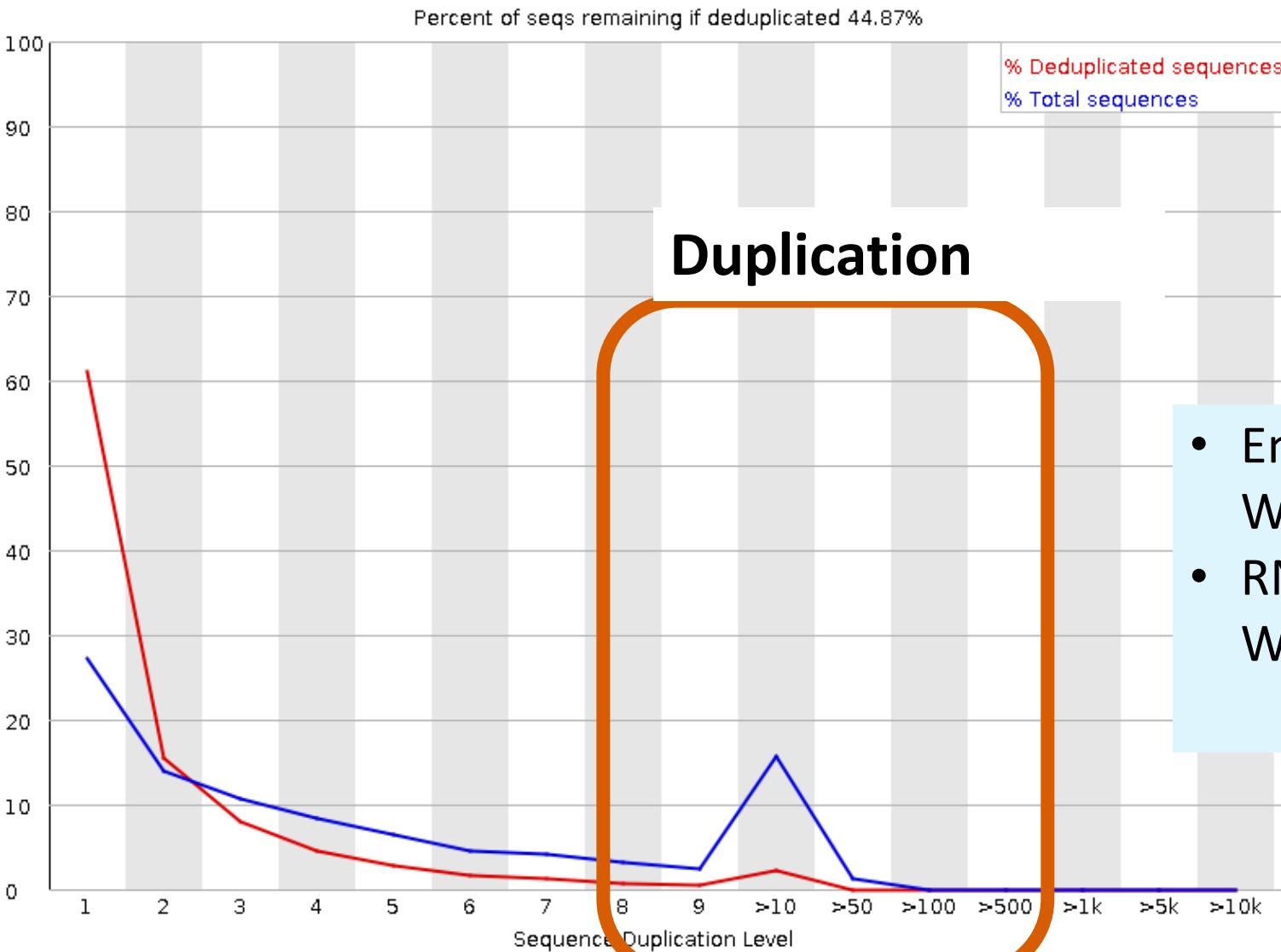


FastQC Data Interpretation

Per sequence quality scores



Sequence Duplication Levels



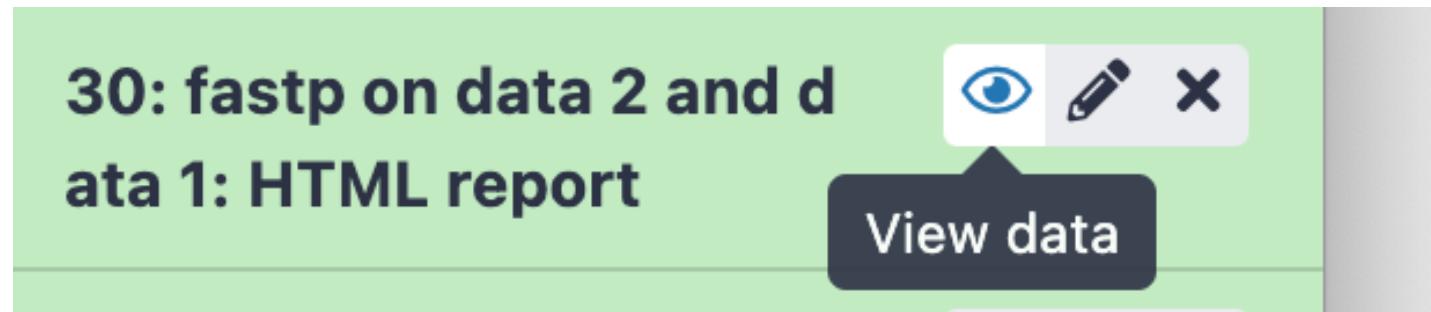
- Enrichment bias (PCR over amplification). Will be removed in a future step
- RNA-sequencing We will handle that data next time

Trimming low quality data with fastp

The screenshot shows the Galaxy web interface with the following elements:

- Tools sidebar:** A list of tools including "fastp" (highlighted with an orange box), "fastpca", "Pedigree", and "WORKFLOWS".
- Tool details panel:** For "fastp - fast all-in-one preprocessing for FASTQ files (Galaxy Version 0.20.1+galaxy0)".
 - Single-end or paired reads:** Set to "Paired" (highlighted with an orange box).
 - Input 1:** "1: child1.fq" (highlighted with an orange box).
 - File selection buttons: folder, clipboard, upload.
 - Input 2:** "2: child2.fq" (highlighted with an orange box).
 - File selection buttons: folder, clipboard, upload.
- Panel buttons:** "Favorite" and "Options".

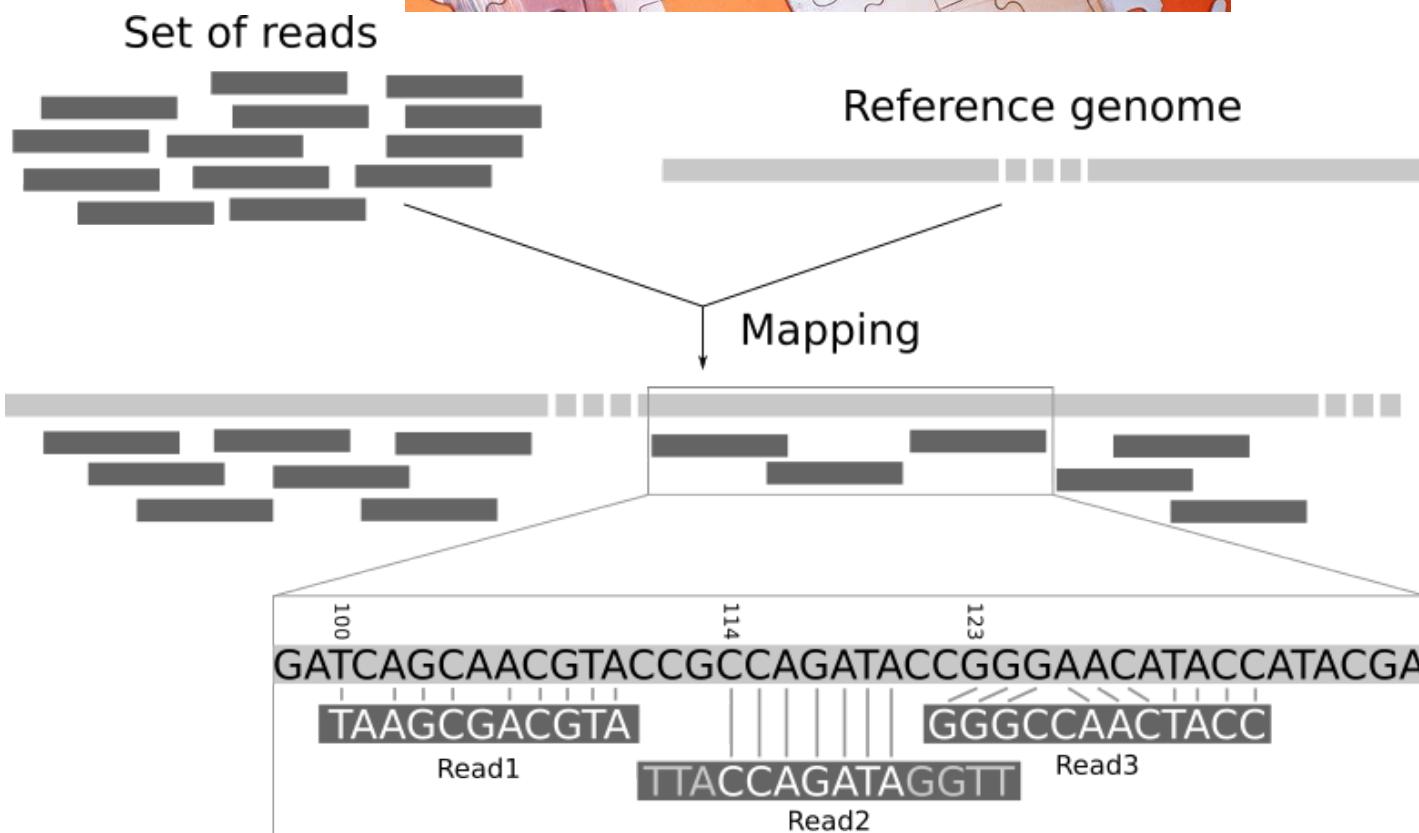
Examine the filtering result of fastp



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 ★  X

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  Options

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

Will 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

Platform/technology used to produce the reads (PL)

ILLUMINA

Merging two BAM data with MergeSamFiles

The screenshot shows the Galaxy web interface. On the left, under the 'Tools' section, the 'MergeSamFiles' tool is highlighted with an orange rectangle. Below it, there's a 'Show Sections' button. In the center, the 'MergeSamFiles' tool card is displayed with its name, a 'Favorite' button, 'Versions' button, and 'Options' dropdown. It also describes the tool: 'merges multiple SAM/BAM datasets into one (Galaxy Version 2.18.2.1)'. On the right, a 'Select SAM/BAM dataset or dataset collection' dropdown menu is open, showing several items. Two specific items are highlighted with an orange rectangle: '64: mapped.mother' and '63: mapped.child'. Other visible items in the list include '52: Filter on data 50. Filtered BAM', '50: BamLeftAlign on data 49 (alignments)', and '49: MarkDuplicates on data 47. MarkDuplicat'. There are also icons for a clipboard and a folder.

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

Removing the PCR duplicates with MarkDuplicates

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)

Select SAM/BAM dataset or dataset collection

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

If empty, upload or

Comment

+ Insert Comment

You can provide mu

If true do not write appropriate flags

Yes No

REMOVE_DUPLICATES

Assume the input is sorted

Yes No

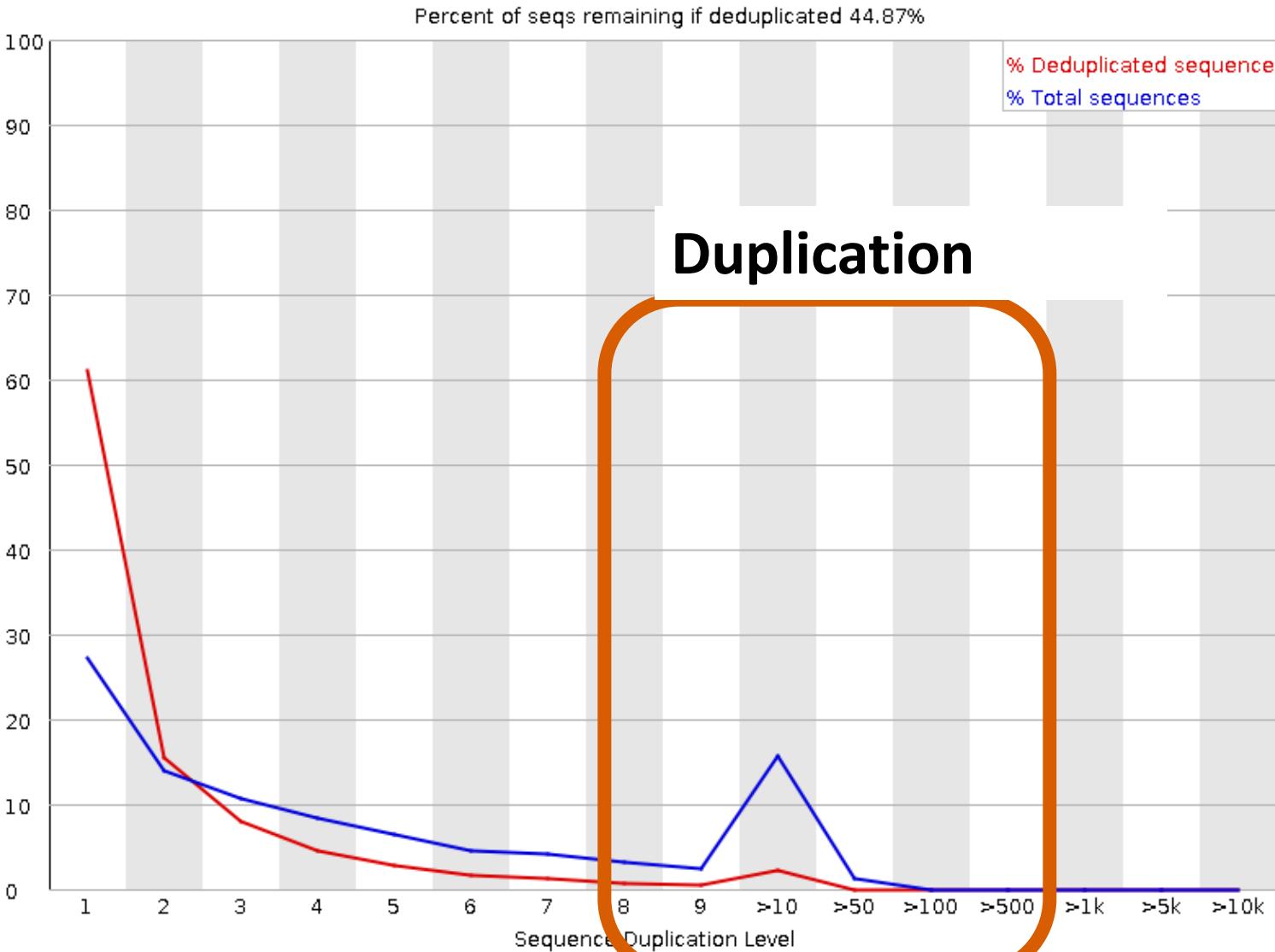
ASSUME_SORTED

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	GGGCACACACACAGGG	Ref: CAC
Your sequence (Alternative sequence)	GGGCAC--ACAGGG	Alt: c
Reference sequence	GGGCACACACACAGGG	Ref: ACA
Your sequence	GGGCA--CACAGGG	Alt: A
Reference sequence	GGGCACACACACAGGG	Ref: GCA
Your sequence	GGG--CACACACAGGG	Alt: G
The deletion is left-aligned		Different Calls!

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

This screenshot shows the configuration interface for the 'BamLeftAlign' workflow. The main title 'BamLeftAlign' is highlighted with an orange box. Below it, there's a sidebar with 'Show Sections' and a list of workflows. The main panel shows the workflow details: 'BamLeftAlign indels in BAM datasets (Galaxy Version 1.3.1)' with 'Favorite', 'Versions', and 'Options' buttons. It asks to 'Choose the source for the reference genome' ('Locally cached') and 'Select alignment file in BAM format' (listing '67: MarkDuplicates on data 65: MarkDuplica...'). Under 'Using reference genome', 'Human (Homo sapiens): hg38' is selected. The 'Maximum number of iterations' is set to '5'. A note says 'Iterate the left-realignment no more than this many times (--max-iterations)'. At the bottom, there's an 'Email notification' section with 'Yes' and 'No' buttons, and a note about sending an email when the job completes. A large orange box highlights the 'Using reference genome' dropdown and the 'Maximum number of iterations' input field.

Filtering BAM file

Tools  

[Filter BAM datasets on a variety of attributes](#) 

 Show Sections

Filter BAM datasets on a variety of attributes 

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 ★  X

FreeBayes ★ Favorite  Options

Show Sections

FreeBayes bayesian genetic variant detector
FreeBayes bayesian genetic variant detector (Galaxy Version 1.3.1)

BamLeftAlign indels in BAM datasets

SnpEff 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

Choose the source for the reference genome

Locally cached

Run in batch mode?

Run individually **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 human hg38

Limit variant calling to a set of regions?

Limit to region

Sets --targets or --region options

Region Chromosome

chrM chrM: 1-16569

(--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   

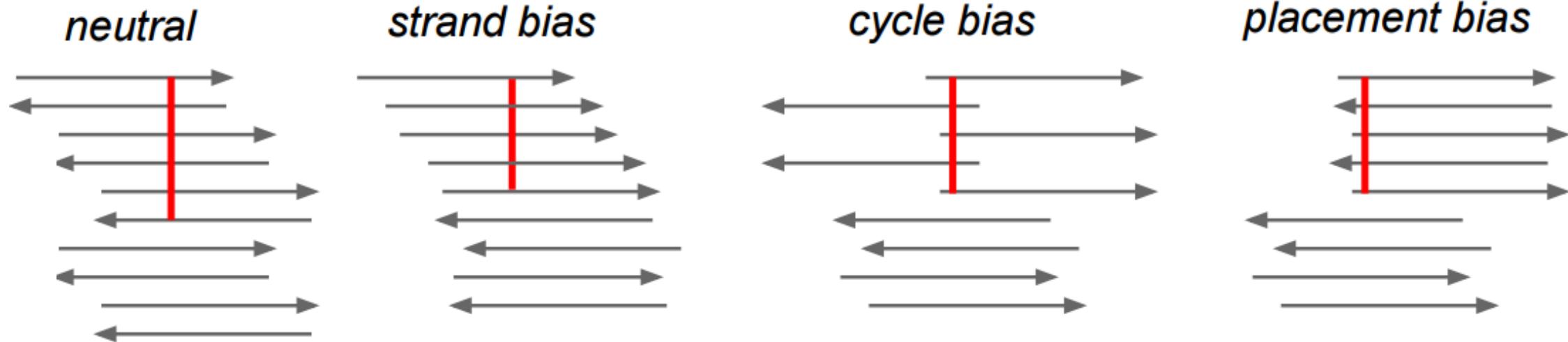
(variants)

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 VCFfilter tool interface in Galaxy. The top navigation bar includes 'Tools' (with a star icon), a file upload icon, and a search bar containing 'VCFfilter' with a delete icon. Below the search bar is a 'Show Sections' button. To the right, there's a summary box for 'VCFfilter': 'filter VCF data in a variety of attributes (Galaxy Version 1.0.0_rc3+galaxy3)' with 'Favorite', 'Versions', and 'Options' buttons. The main workspace is titled 'VCF dataset to filter' and shows a dropdown menu with '53: FreeBayes on data 52 (variants)' selected. Below this is a 'more filters' section with a 'Select the filter type' dropdown showing 'Info filter (-f)' and a 'Specify filtering value' input field containing 'SRP > 20'. A large green callout box on the right side contains the text: 'Set the following filters' followed by five filtering 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.

Set the following filters

- "Specify filtering value": SRP > 20
- "Specify filtering value": SAP > 20
- "Specify filtering value": EPP > 20
- "Specify filtering value": QUAL > 20
- "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?



Reformat the VCF file

The screenshot shows the Galaxy web interface with the following components:

- Tools Panel:** On the left, a list of tools is shown. The "VCFtoTab-delimited" tool is highlighted with an orange box and has a red border around its name.
- Tool Header:** The "VCFtoTab-delimited" tool header includes a "Favorite" button, a "Versions" button, and an "Options" button.
- Dataset Selection:** A section titled "Select VCF dataset to convert" contains three selection buttons (document, clipboard, folder) and a dropdown menu showing "72: VCFfilter: on data 71".
- Report Data:** A "Report data per sample" section with "Yes" and "No" buttons, followed by a "-g option" field.
- Empty Fields:** A "Fill empty fields with" section with a "Nothing" dropdown menu, followed by a "-n option" field.
- Email Notification:** An "Email notification" section with "Yes" and "No" buttons, followed by a descriptive text: "Send an email notification when the job completes."
- Execute Button:** A large blue "✓ Execute" button at the bottom.

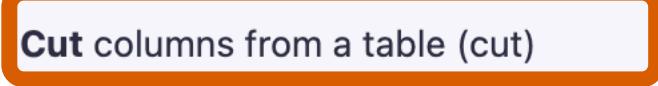
Reformat the VCF file

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

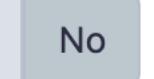


73: VCFtoTab-delimited: on data 72





Email notification

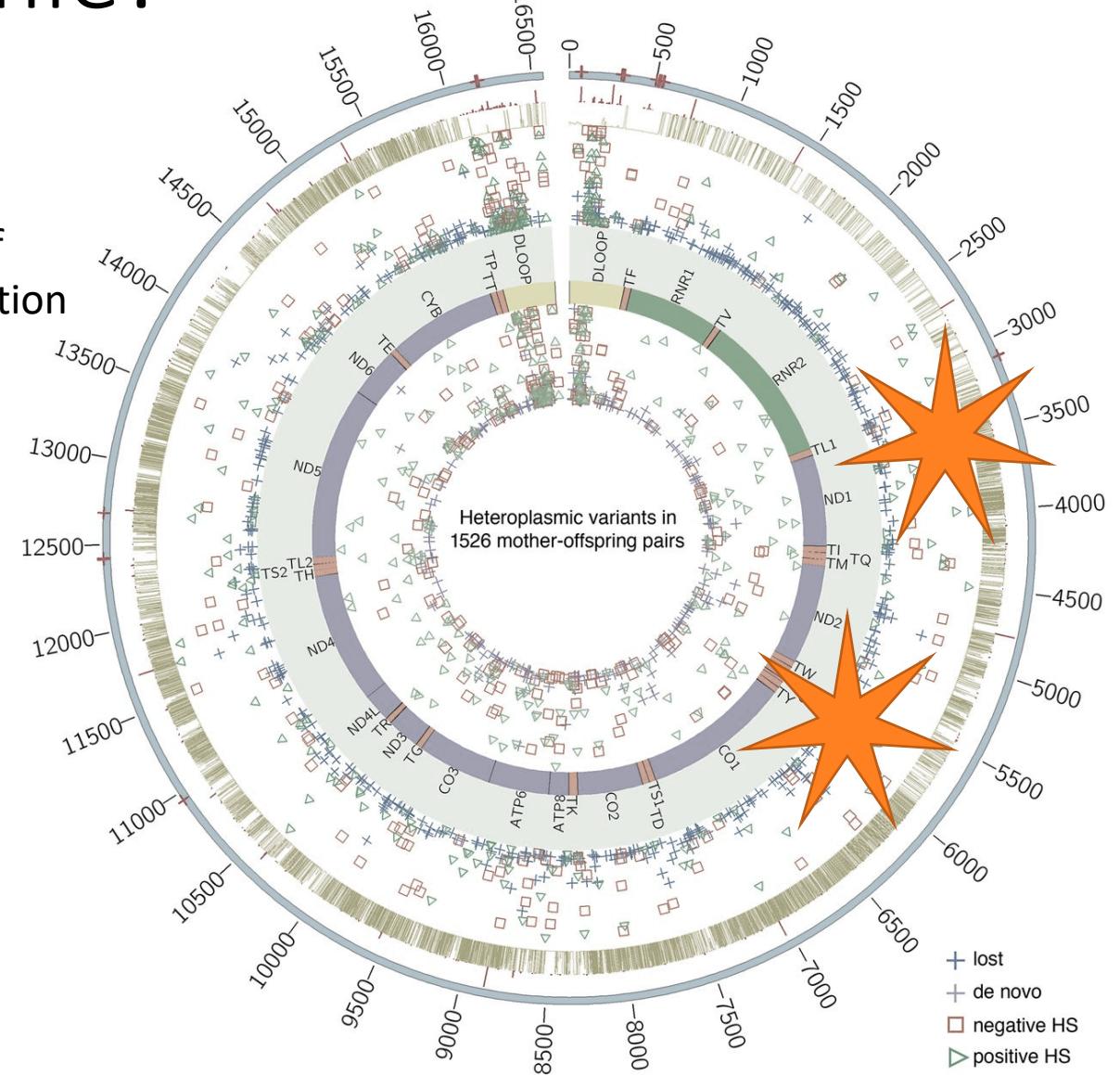
Yes  No

Send an email notification when the job completes.

 Execute

Take a look on the VCF file!

Position	Ref	Alt	Sample	AO	DP
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	DP
1	2	3	4	Number of alternative observations	Depth of this position
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

A
A
A
A

⋮

G
G
G

⋮

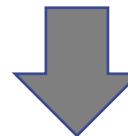
Total Read Counts
995

Alternative Counts:
666

Take a look on the VCF file...

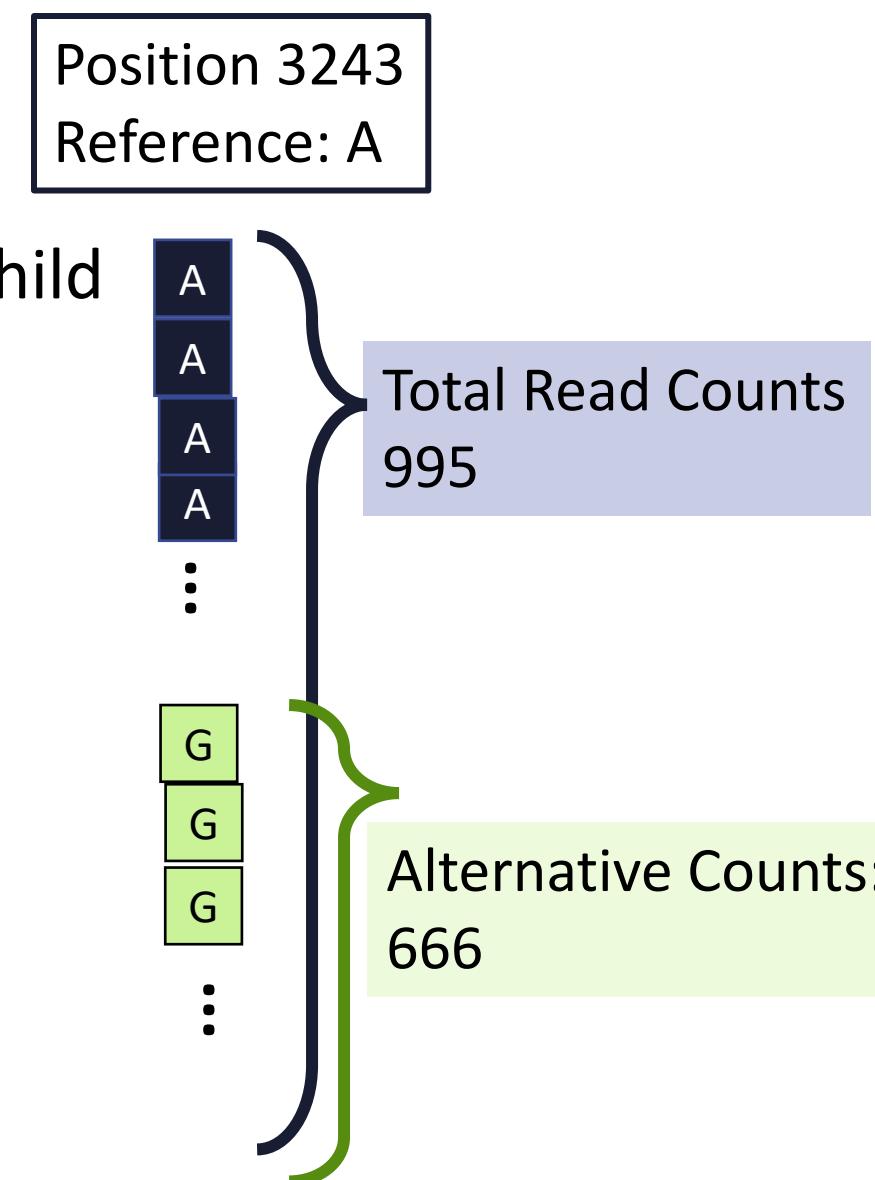
Reference Allele: A, Read counts = $995 - 666 = 329$

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

Posision 3243

Child : G = 0.67

Mother : G = ???

Posision 5538

Child : G = ???

Mother : G = ???

Allele frequency of variants in Mother and Child

Position 3243

Child : G = 0.67

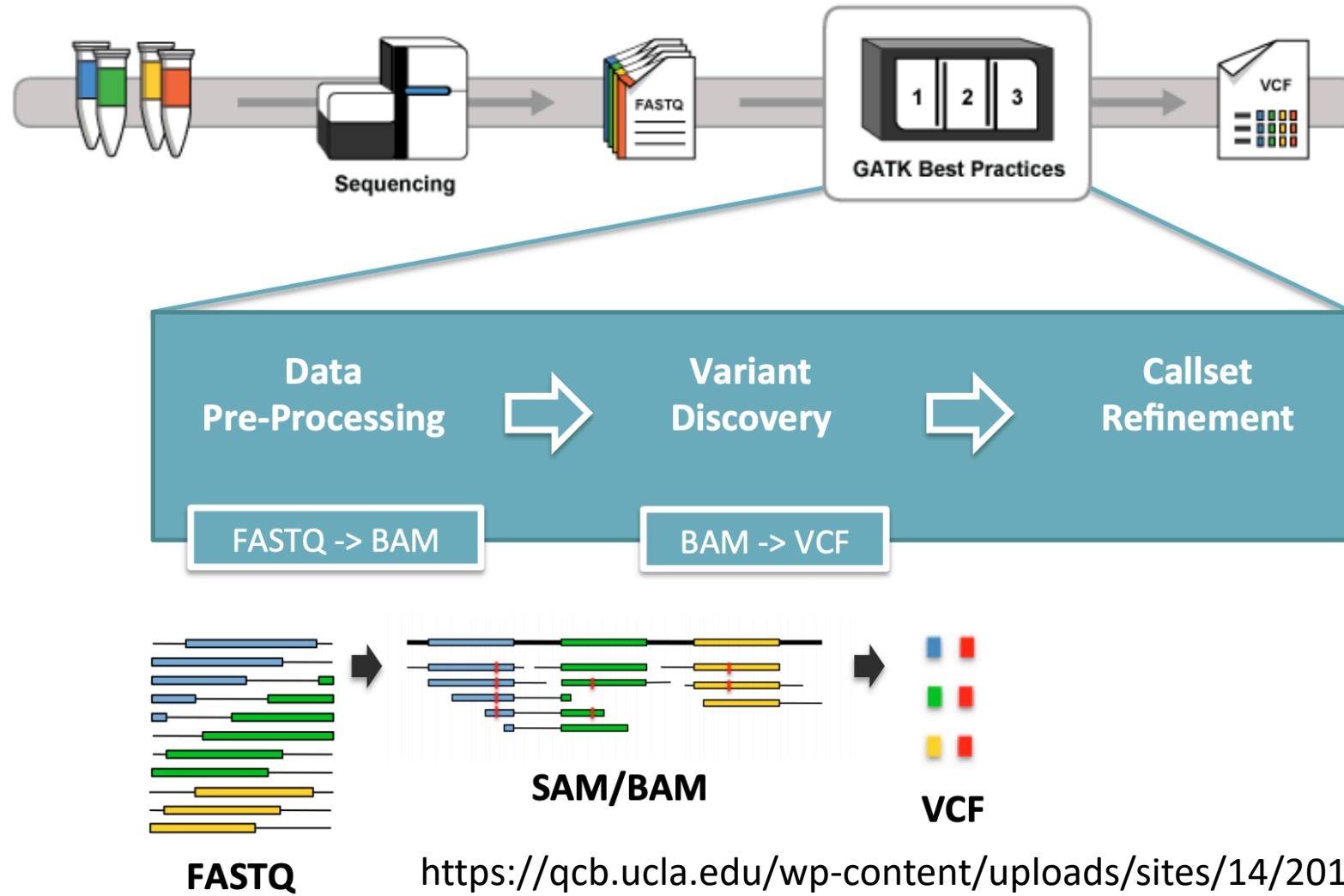
Mother : G = 0.33

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

