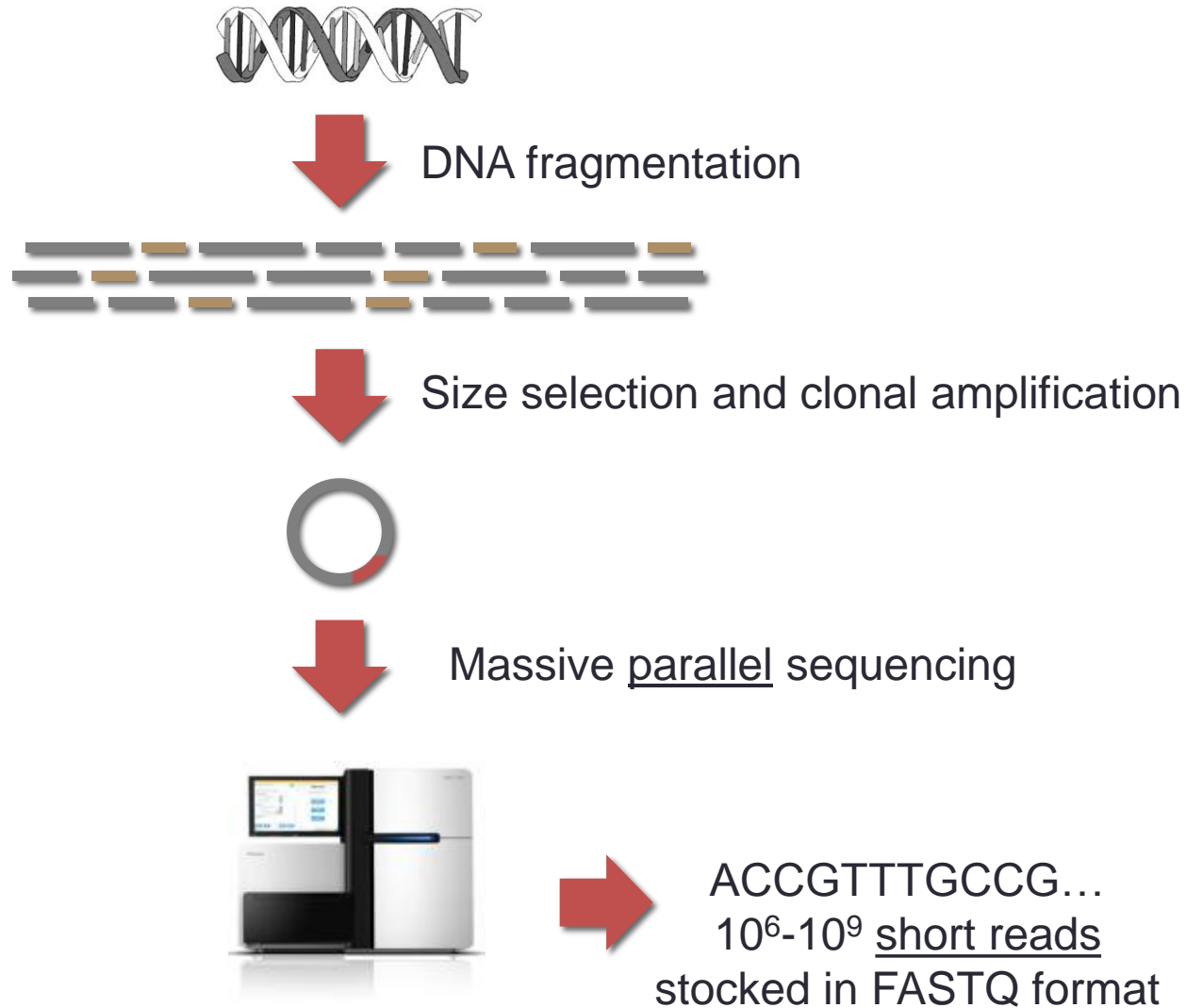


GALAXY INITIATION

A. Lermine

U900 Institut Curie, INSERM, Mines ParisTech

How does Next-Gen sequencing work?

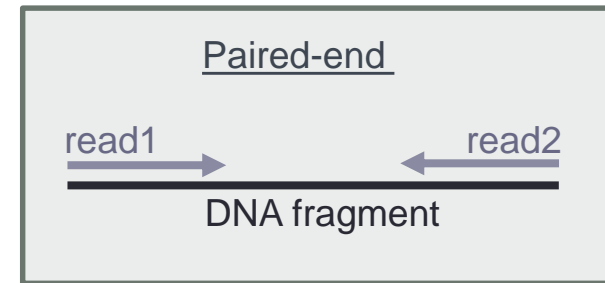


Hi-seq, SOLiD, PGM, ..., What does it mean ?

Platform	Provider	Reads Number (M)	Max Reads Size (bp)	Throughput (Gb)	Time	Space
GS Flex	Roche	1	700	0.7	8d	Base
Hiseq 2000/2500 Normal mode	Illumina	3000	2x100	600	11d	Base
Hiseq 2500 rapid mode	Illumina	600	2x150	120	40h	Base
MiSeq	Illumina	15	2x250	8.5	40h	Base
SOLiD	LifeTech	1400	75-35	150	25d	Color
PGM 314	Ion Torrent	0.5	400	>0.01	2-4h	Base
PGM 316	Ion Torrent	2	400	>0.1	2-4h	Base
PGM 318	Ion Torrent	4	400	>1	2-4h	Base
Proton	Ion Torrent	60-80 M	200	>12	2-4h	Base

Let's start after sequencing ...

- A **raw** data file containing **millions** of reads (A/C/G/T/N sequence + base qualities):
 - **IonTorrent PGM**: FASTQ, SFF, Unmapped BAM of reads of different sizes
 - SFF and Unmapped BAM contain specific IonTorrent flowspace information (helps improve the accuracy during the mapping and variant calling steps)
 - **Illumina**: FASTQ of same size reads
- Reads can be **single-end** (one end of the fragment sequenced) or **paired-end** (both ends sequenced)



ACTGATTAGTCTGAATTAGANNGATAGGAT

ACTAGGCATCGGCATCACGGACNNNNNNNN

GATCGATGCATAGCGATCAGCATCGATACG

ACTAGCTATCGAGCTATCAGCGAGCATCTATC

CGGCGCTCCGCTCTCGAAACTAGCACTGAC

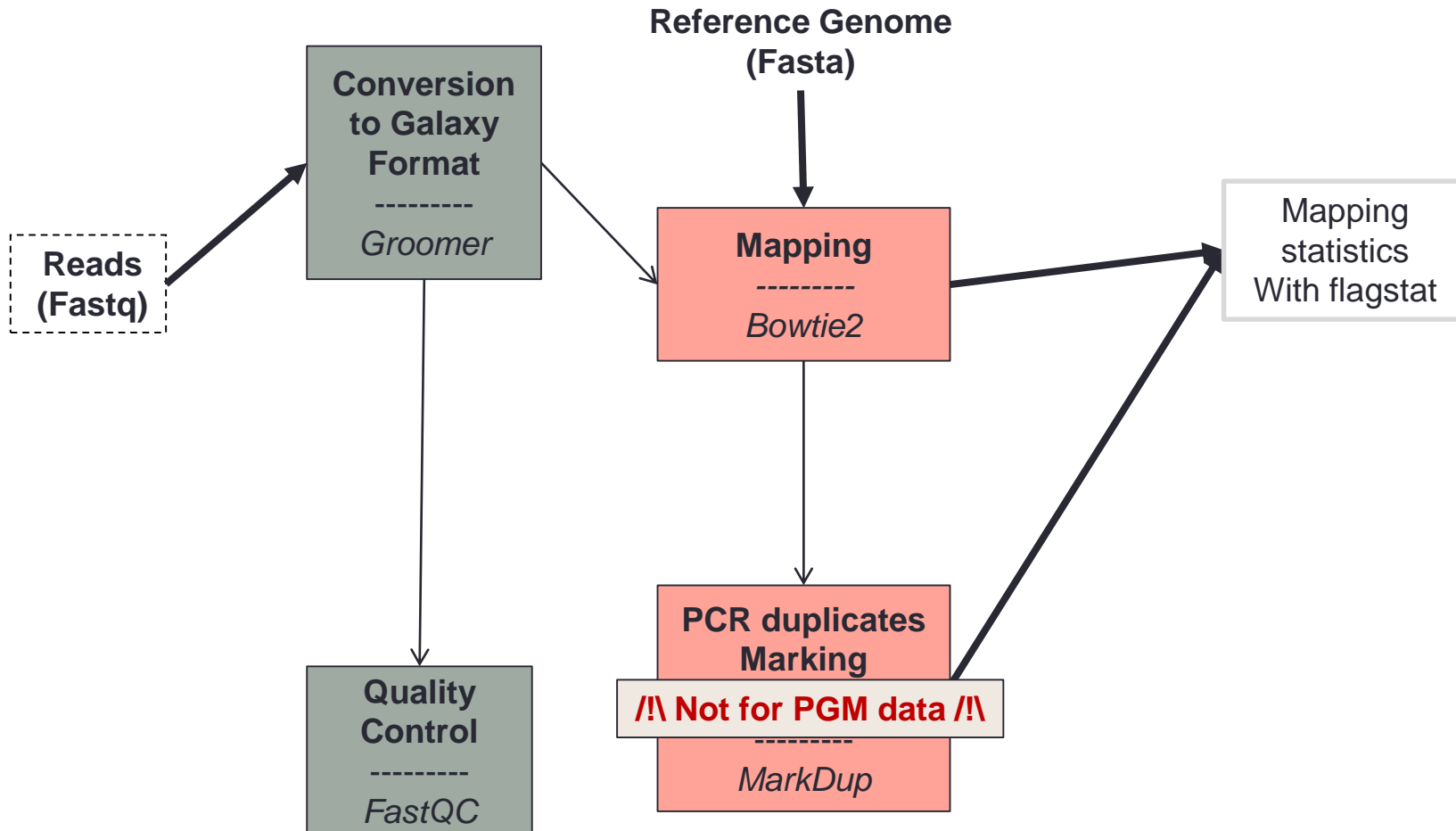
ACTAGCTACTATCGAGCGAGCGATCATCGAC

AGCATCAGGATCTACGATCTAGCGAACTGAC

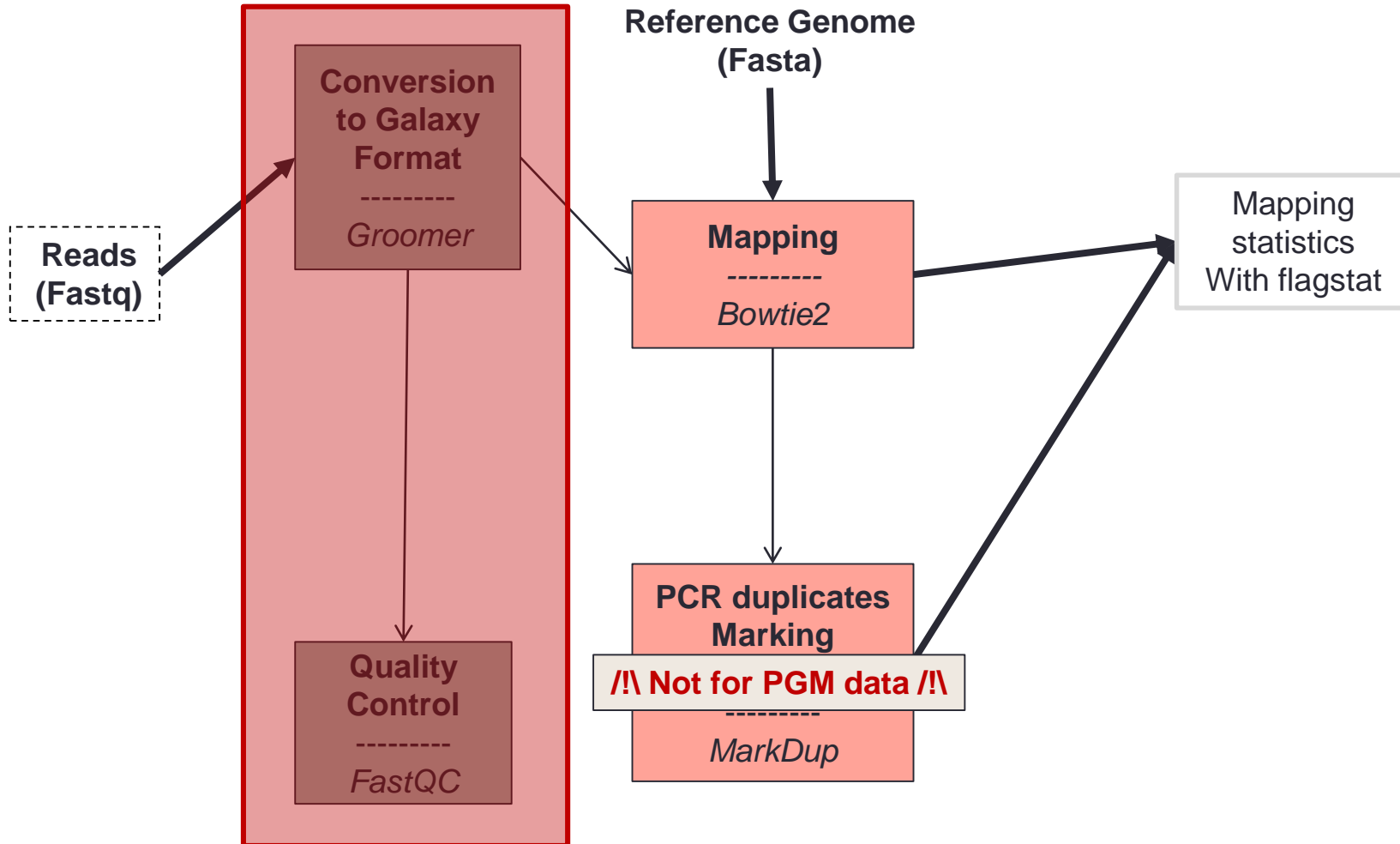
CTGACTACTATCGAGCGAGCTACTAACTGAC

NNNNNNNNNNNNNNNNNNNNNNNNNNNNNN

Galaxy Workflow

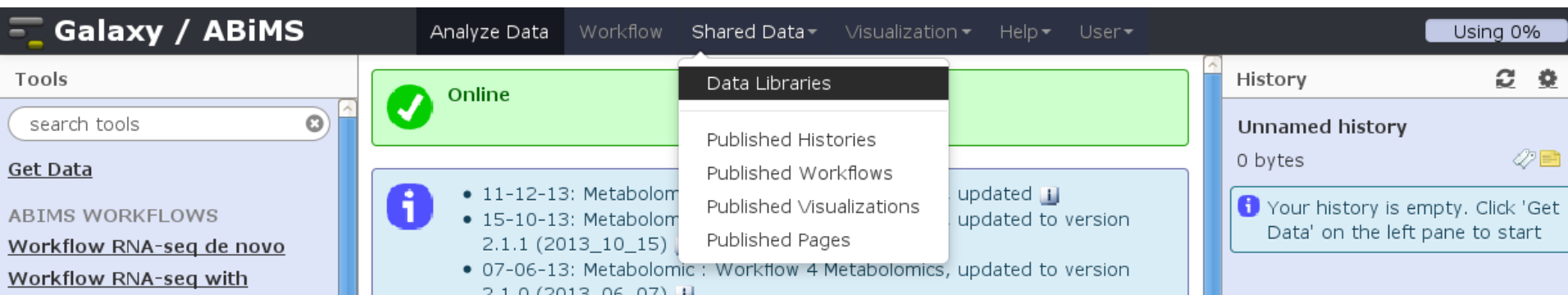


Galaxy Workflow



Two available datasets on Galaxy

1. Open your web browser and go to your IFB Cloud Galaxy instance
2. In the top menu, click on « Shared Data » then « Data librairies »



The screenshot shows the Galaxy / ABiMS web interface. The top navigation bar includes 'Galaxy / ABiMS', 'Analyze Data', 'Workflow', 'Shared Data', 'Visualization', 'Help', and 'User'. A dropdown menu is open under 'Shared Data', with 'Data Libraries' selected. The main content area shows a green 'Online' status bar and a list of datasets with details like dates and versions. The left sidebar has a 'Tools' section with a search box and 'Get Data' links. The right sidebar shows a 'History' section with an 'Unnamed history' and a message: 'Your history is empty. Click 'Get Data' on the left pane to start'.

3. Click on «**TP-INITIATION** »:

→ **Chr4.fastq**: raw reads file

→ **Chr4.fasta**: reference sequence for alignment

Sequence Quality Encoding (FASTQ)

- Extension from traditional FASTA format
- Each block has 4 elements (in 4 lines):
 - Sequence name (read name, group, etc...)
 - Sequence
 - + (optional: sequence name again)
 - Associated quality scores (phred-scaled) : different encoding possible
- Example record:
 - @FCD19MJACXX:2:1101:1735:1993#GTTTCGACA/1
 - NGAGGCTGAGGCGGGCAGAGGTCAGGAGATCGAGACCATC
 - +
 - BP\cccc]ceecheheeZbe_cZbd_dbbdd\!Xab_`b

Sequence Quality Encoding (FASTQ)

- The base calling (A, T, G or C) is performed based on Phred Scores.

Phred Quality Score	Probability of Incorrect Base Call	Base Call Accuracy
10	1 in 10	90%
20	1 in 100	99%
30	1 in 1,000	99.9%
40	1 in 10,000	99.99%
50	1 in 100,000	99.999%

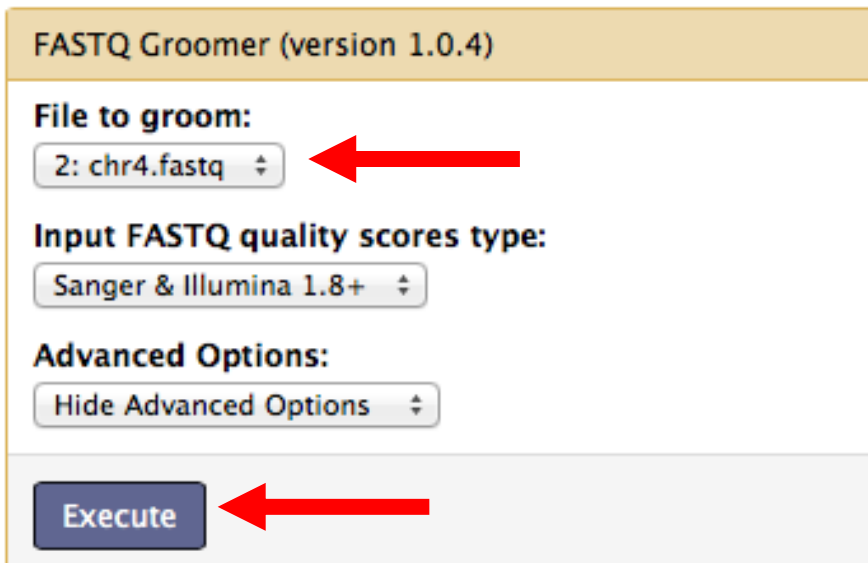
→ 1% error rate

- Phred scores provide \log_{10} -transformed error probability values
→ If p is probability that the base call is wrong the Phred score is

$$Q = -10 \log_{10} (P) \Leftrightarrow P = 10^{-Q/10}$$

FASTQ format conversion

1. Rename your history to « **TP Initiation** » by clicking on « **Unnamed history** »
2. In the left panel, click on « FASTQ Groomer » under the NGS: QC and manipulation section to convert **your FASTQ** into FASTQ Sanger Format
3. Click on « Execute » to launch the conversion



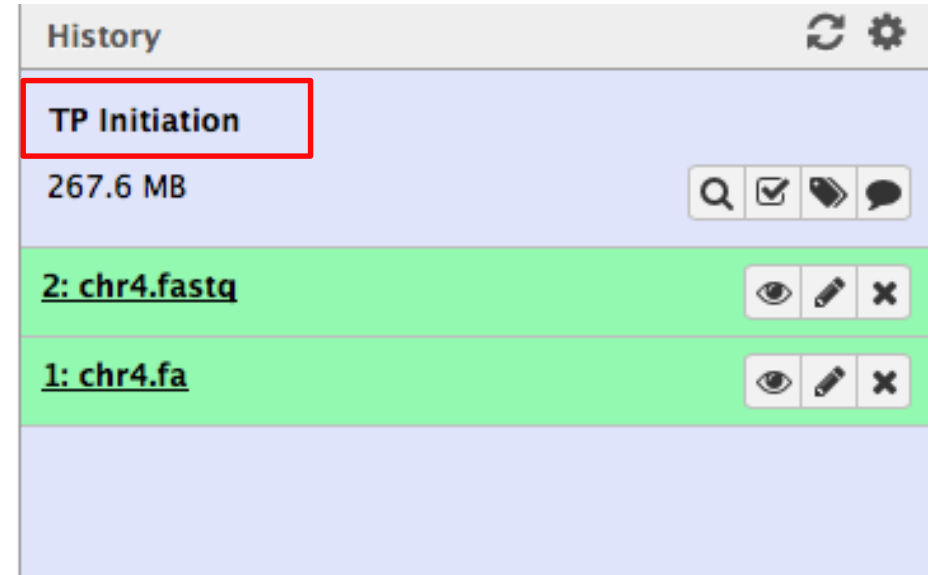
FASTQ Groomer (version 1.0.4)

File to groom:
2: chr4.fastq ←

Input FASTQ quality scores type:
Sanger & Illumina 1.8+ ▾

Advanced Options:
Hide Advanced Options ▾

Execute ←



History

TP Initiation (highlighted with a red box)
267.6 MB

2: chr4.fastq (highlighted with a green background)

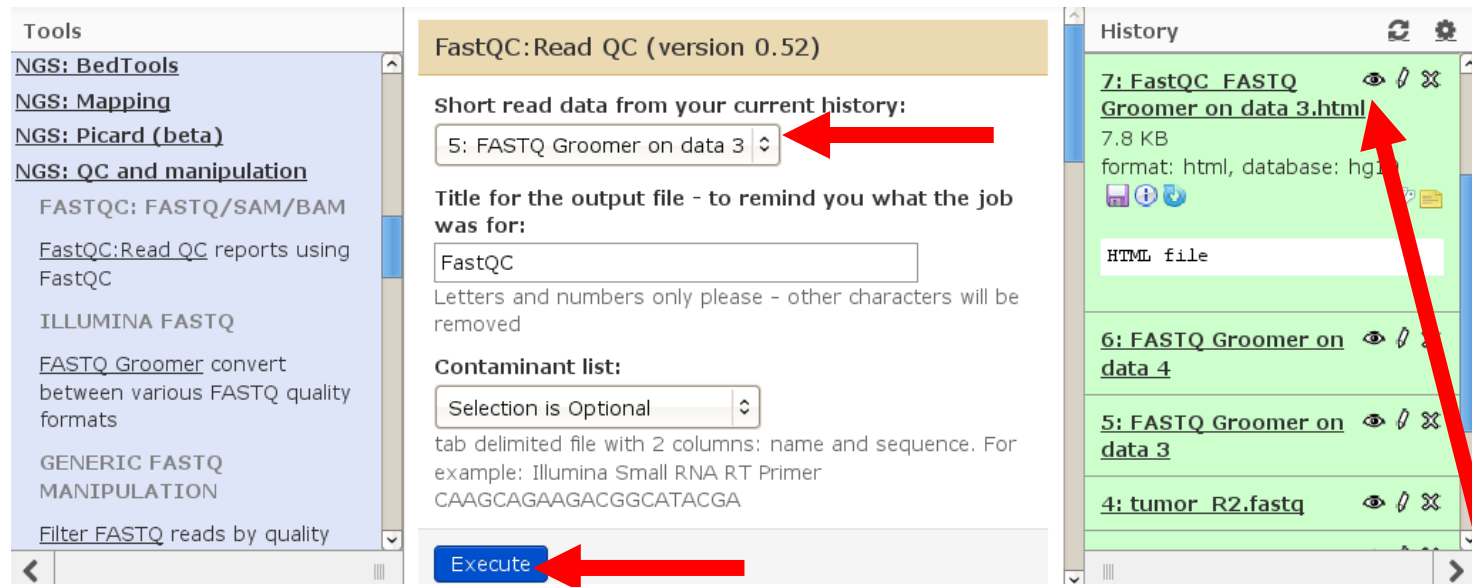
1: chr4.fa (highlighted with a green background)

FASTQC: Quality Control on raw data

- **Sequence Length Distribution**: sequencers produce same (Illumina) or different (PGM) read length. This metric helps identify abnormal read length
- **Sequence content**: % A/C/G/T ; %GC; %N at each position in the read
 - Proportion biased for targeted sequencing
- **Quality score**:
 1. **Per base**: identify base calls with low quality (commonly towards the end of a read)
 2. **Per sequence**: to see if a subset of your sequences have universally low quality values
- **K-mers content**: a k-mer is a motif of length k observed more than once in a sequence (repeats : ACACAC ; spaced occurrences : tccGAGGaaggGAGGaag)
- **Over-represented sequences**: highly duplicated sequence in your library (primer, adapter..)

FASTQC : FASTQ Quality Control

1. In the left panel, click on « FASTQC: Read QC »
2. Select the FASTQ Groomer dataset and click on « Execute »




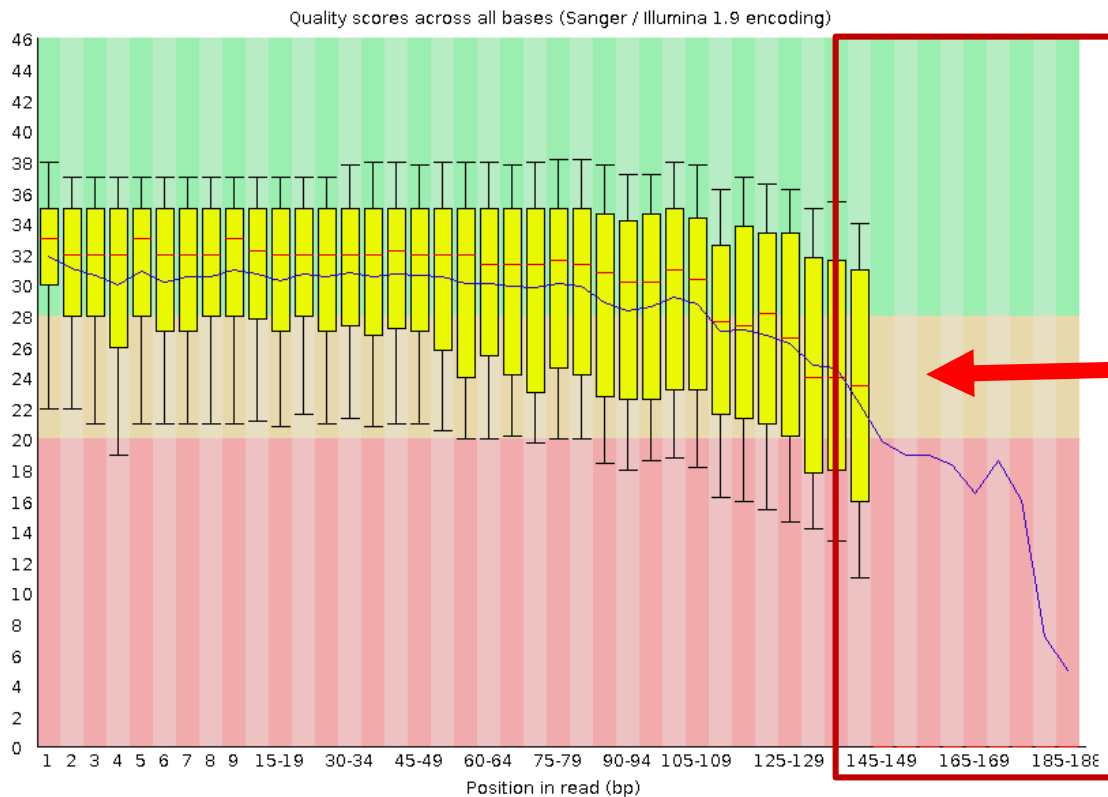
The screenshot displays the FASTQC web interface. On the left, the 'Tools' panel lists various NGS tools, with 'FASTQC: Read QC' selected. The main panel shows the configuration for 'FastQC:Read QC (version 0.52)'. The 'Short read data from your current history' section has a dropdown menu set to '5: FASTQ Groomer on data 3'. The 'Title for the output file' is 'FastQC'. The 'Contaminant list' is set to 'Selection is Optional'. The 'Execute' button is highlighted with a red arrow. The 'History' panel on the right shows a list of jobs, with the top job '7: FastQC FASTQ Groomer on data 3.html' highlighted in green and its eye icon also highlighted with a red arrow.

→The result of FASTQC is an html page that you can view by clicking on the **eye**

FASTQC Metrics

- Look at the different metrics for reads
- **Problem**: the per base sequence quality are quite low towards the end

 **Per base sequence quality**



Solution:


Trim the 50bp from the 3' end of the reads

→ Higher confidence in the sequenced information

FASTQ Trimmer to improve reads quality

1. Use « FASTQ Trimmer » to cut off 50bp from 3' (use the « search tools » object to find the tool)
2. Run « FASTQC » on the trimmed reads

FASTQ Trimmer (version 1.0.0)


FASTQ File:
 

Define Base Offsets as:


Use Absolute for fixed length reads (Illumina, SOLiD)
 Use Percentage for variable length reads (Roche/454)

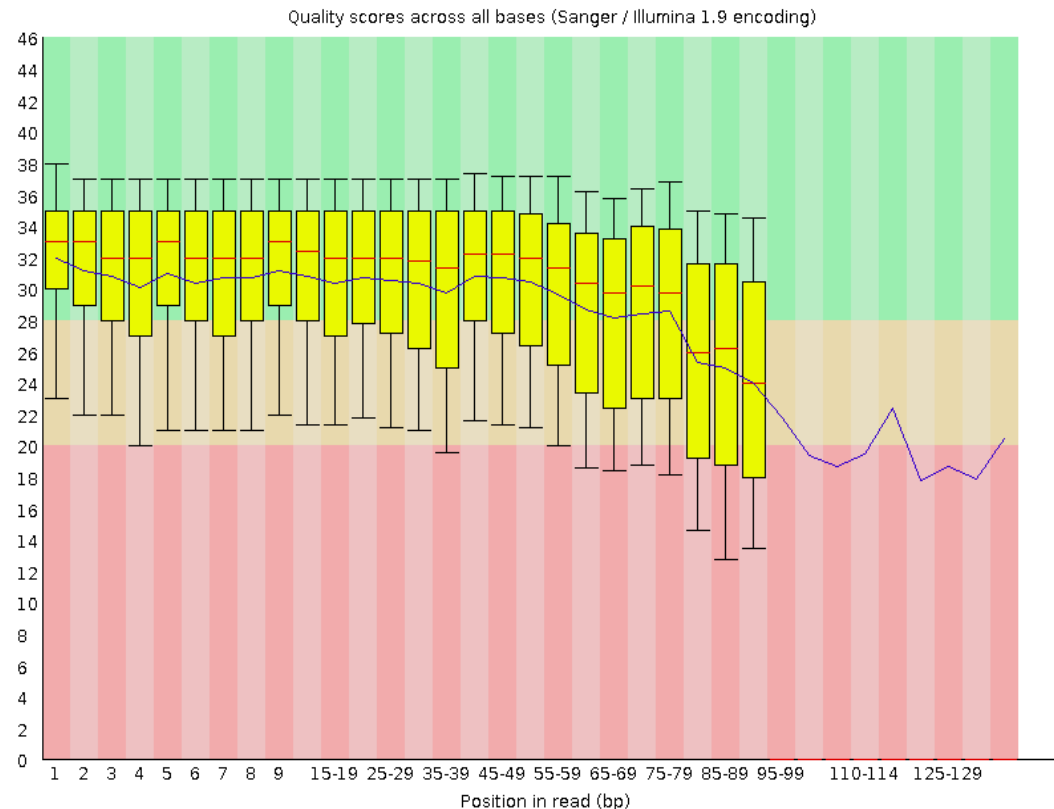
Offset from 5' end:

 Values start at 0, increasing from the left

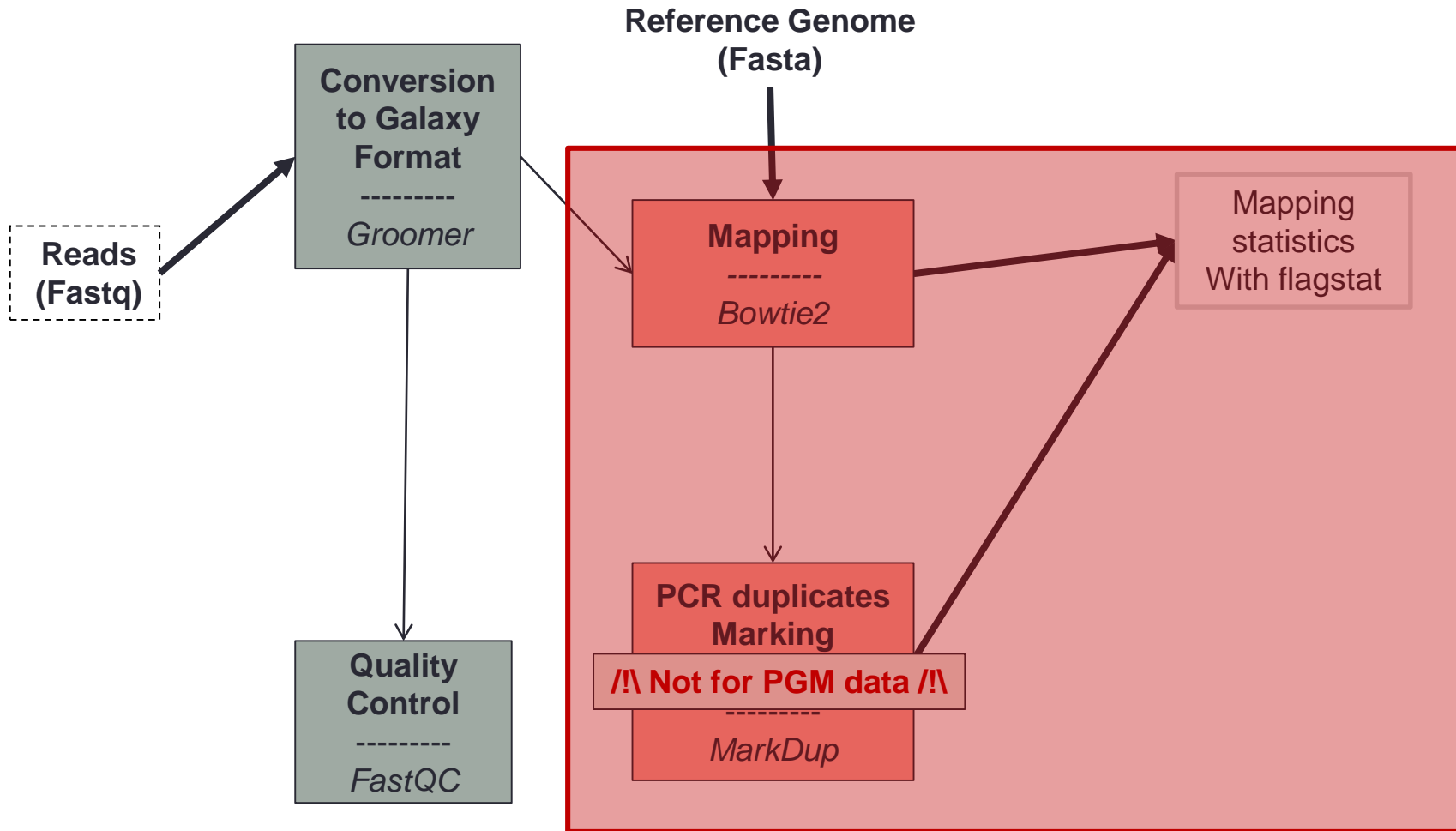
Offset from 3' end:
 
 Values start at 0, increasing from the right

Keep reads with zero length:





Galaxy Workflow



Mapping on a reference Genome

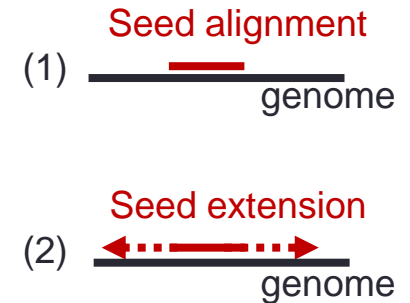
- Reads are aligned ≥ 1 times on the reference genome
- A **mapping quality** is associated to each alignment:
 - Quantify the probability that the alignment is correct
 - Decreases with the number of mismatches (wrong nucleotide) & gaps (small insertions/deletions) & the number of alignments



Reads Alignment - Vocabulary

Mapping method: **seed & extend**

1. Aligning the seed (small part of the read)
2. Extending the seed to align the whole read



Mismatch: Incoherence between two nucleotides

Indels: Insertion/Deletion into the reference genome

Gap: Bridge within the read alignment (*i.e.* small indels)

Mappability: Uniqueness of a region

- repeated region = low mappability
- unique region = good mappability

Multiple Alignments

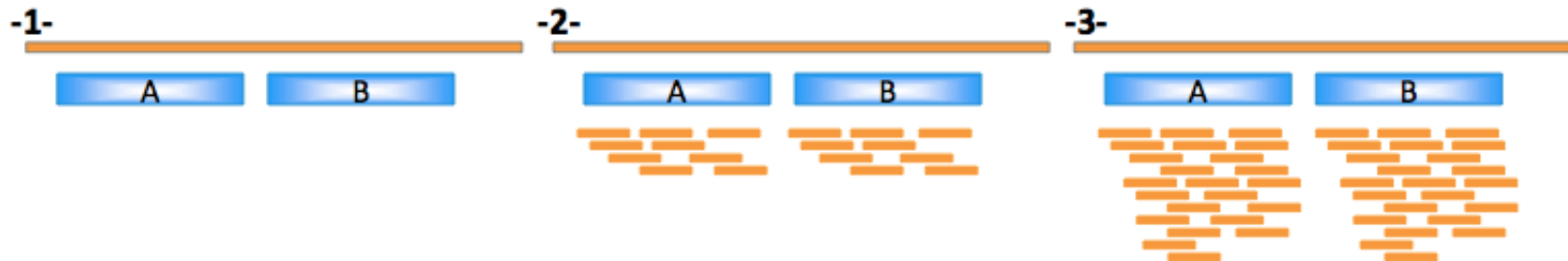
- A read can align **multiple times** on the genome (repeated elements...)
- How to deal with these multiple alignments reads ?

Three strategies:

-1- Report only unique alignment

-2- Report best alignments & randomly assign reads across equally good loci

-3- Report all (best) alignments



- **Mapping Quality:** quantify the probability that a read is misplaced
→ Low if a read has multiple alignments ; several mismatches/gaps

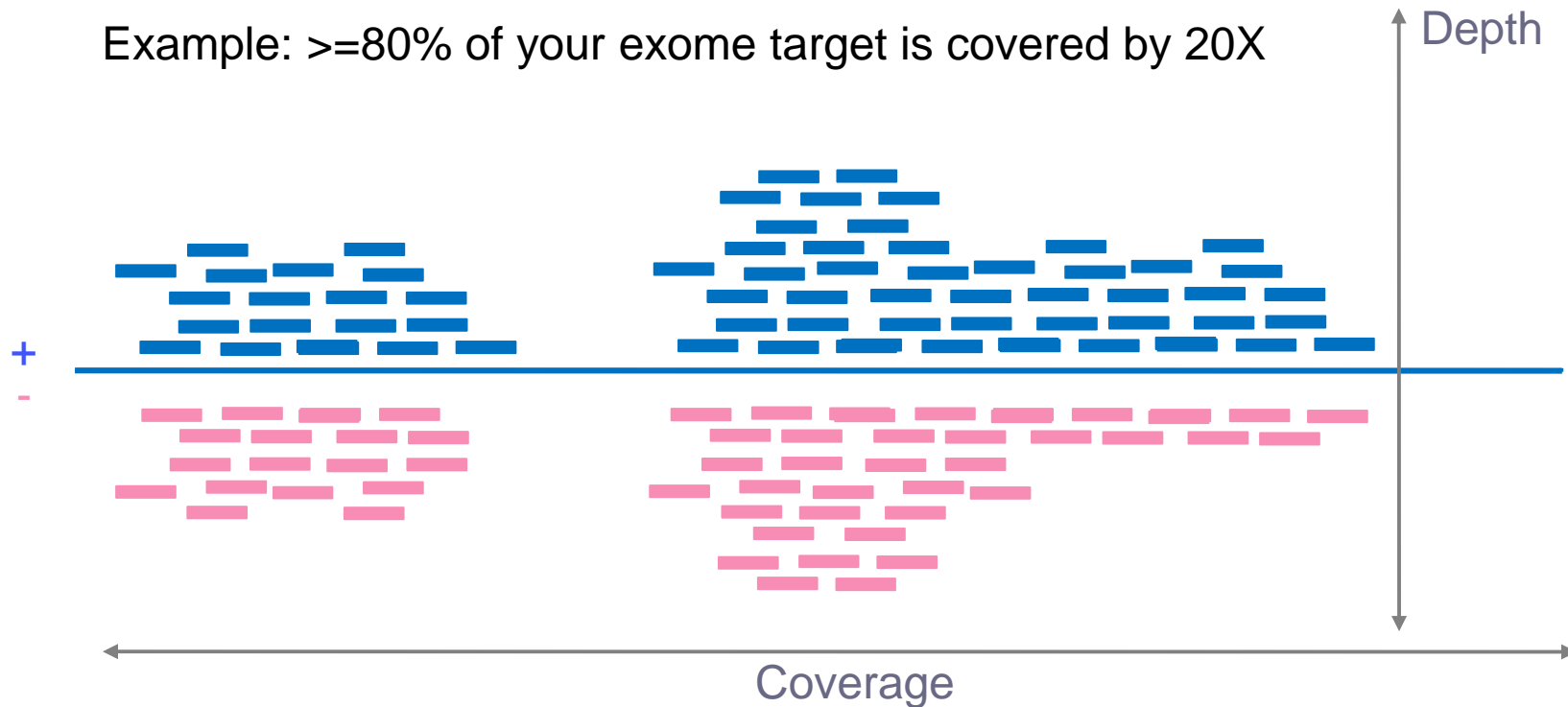
Statistics used as Quality Control

- **Depth of coverage** = mean number of reads covering a base (X)

Example: 30X for normal sample, 100X for tumor sample

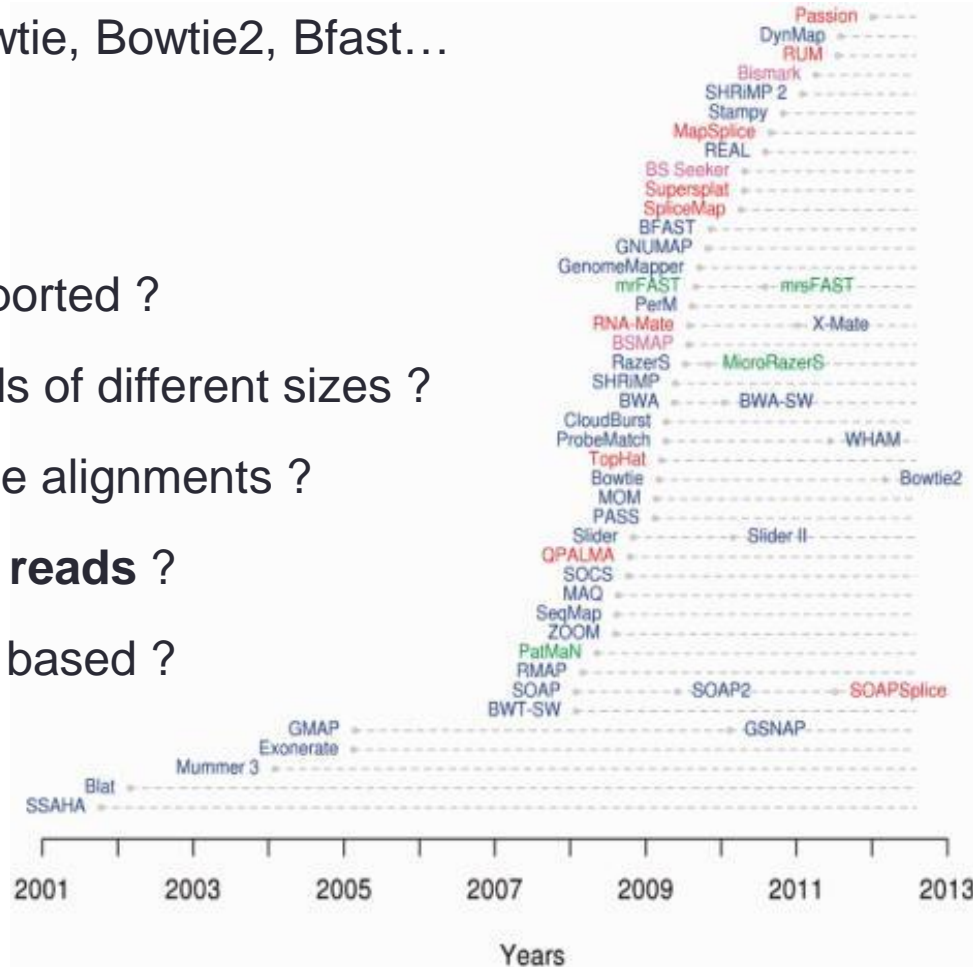
- **Coverage** = part of the reference with at least one read

Example: $\geq 80\%$ of your exome target is covered by 20X



Alignment tools

- Multitude of alignment tools: BWA, Bowtie, Bowtie2, Bfast...
- How to choose the best tool ?
 - Is my sequencing technology supported ?
 - Do I have short/long reads ? Reads of different sizes ?
 - **Do I want to allow gaps ?** Multiple alignments ?
 - Does it support single/**paired-end reads** ?
 - On which alignment algorithm is it based ?
 - Computational issues ?
 - Is it used by the community ?



Mapping with Bowtie2

1. Use « Bowtie2 to align reads on the hg19 genome

Bowtie2 (version 0.2)

Is this library mate-paired?:

Paired-end ↕

FASTQ file:

5: FASTQ Groomer on data 3 ↕

Nucleotide-space: Must have Sanger-scaled quality values

FASTQ file:

9: FASTQ Trimmer on data 6 ↕

Nucleotide-space: Must have Sanger-scaled quality values

Minimum insert size for valid paired-end alignments:

0

Maximum insert size for valid paired-end alignments:

250

Write unaligned reads to separate file(s):

Will you select a reference genome from your history

Use a built-in index ↕

Built-ins were indexed using default options

Select a reference genome:

Homo sapiens hg19 ↕

Specify the read group for this file?:

Yes ↕

Read group identifier (ID). Each @RG I header section.:

chr4

Required if RG specified. Read group ID:

Library name (LB):

chr4

Required if RG specified

Platform/technology used to produce

illumina

Required if RG specified. Valid values : C

Sample (SM):

chr4

Required if RG specified. Use pool name

Full parameter list ↕

You can use the default settings or set custom values

Type of alignment:

End to end ↕

Preset option:

Very sensitive ↕

Preset option:

combination of parameters designed to have a good tradeoff between speed, sensitivity, accuracy

Execute

SAM/BAM aligned format

- SAM Format: aligned format, human readable

@SQ SN:chr12 LN:133851895

@RG ID:Sample_ID LB:Sample_Library PL:ILLUMINA SM:Sample_Name PU:Platform_Unit

Read name	Flag	Chr	5' pos	MAPQ	Cigar	paired	5' pos of the mate	Insert size		
ERR166338.1	99	chr12	82670685	23	101M	=	82670850	266		
GCCCCTGGGGATGTTTTGCACCAAGCCACTGTCTCCAGCTGG							sequence			
BBC@GIIHGCFIEHEAIEIFFGEONDNJFINIONHNGJNNNNKNJN							Base quality			
RG:Z:Sample_ID	XT:A:U	NM:i:0	X0:i:1	X1:i:1	XM:i:0	XO:i:0	XG:i:0	MD:Z:100	XA:Z	tags
Group affiliation										

- BAM Format: Binary SAM Format (not human readable but compressed = smaller)

Mapping Statistics

- Use « Flagstat » from « Samtools » to see some mapping statistics

flagstat (version 1.0.0)

BAM File to Convert:

8: SAM-to-BAM on data 1 and data 4: converted BAM

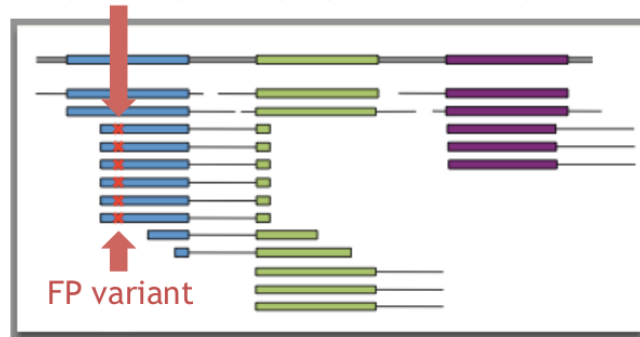
Execute

```
109073 + 0 in total (QC-passed reads + QC-failed reads)
0 + 0 duplicates
107254 + 0 mapped 98.33% % of mapped reads
0 + 0 paired in sequencing
0 + 0 read1
0 + 0 read2
0 + 0 properly paired (-nan%:-nan%)
0 + 0 with itself and mate mapped
0 + 0 singletons (-nan%:-nan%)
0 + 0 with mate mapped to a different chr
0 + 0 with mate mapped to a different chr (mapQ>=5)
```

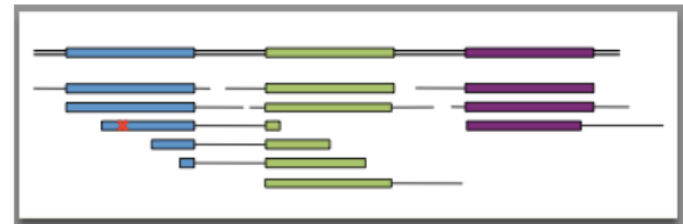
Removing Duplicates

- **Duplicates reads:** different reads having the same sequence caused by PCR amplification during sequencing library preparation
- The removal of the duplicates depends on the application (not suitable for sequencing on small target)

Sequencing error propagated in duplicates



PCRdup
removal




- **Galaxy:** Use “Mark Duplicates reads” from “NGS:Picard” to **mark** duplicates (don’t remove them)
 - If duplicates are marked, samtools and GATK tools will ignore them
- **Galaxy:** Run “Flagstat” on the output BAM to see the number of PCR duplicates

Extract workflow from history

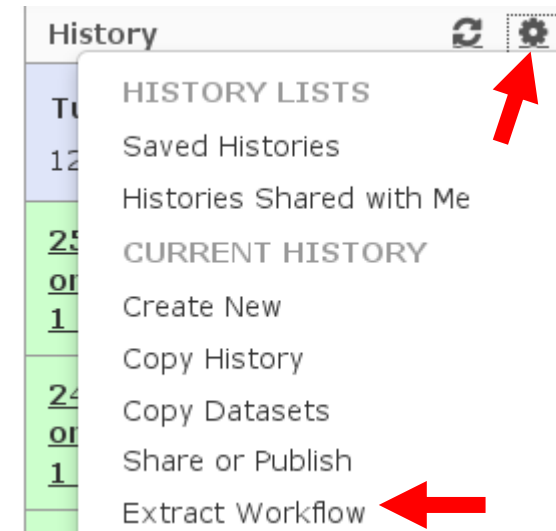
- In the « history » panel, click on the topside wheel then
- on « Extract Workflow »
- Write a name for your workflow then click on

Workflow name

Workflow constructed from history 'TP Initiation'



- Select the steps you want to see
- Check every steps!



Select Libraries on Galaxy

You can only run tools on data that are present in your current history

1. In the Data Libraries
2. Select « **chr4.fastq** » ; « **chr4.fa** »
3. Select « Import to Histories » then click on Go



4. Write a new history name and click on « import library datasets »

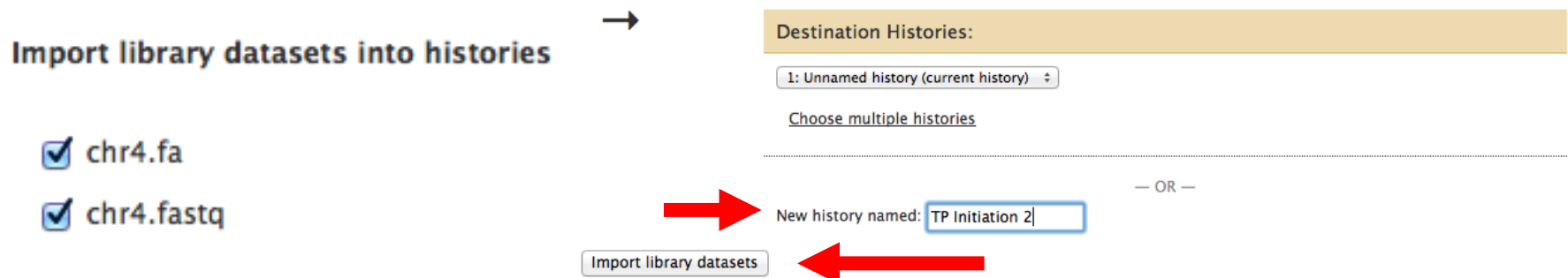
Import library datasets into histories →

chr4.fa
 chr4.fastq

Destination Histories:
1: Unnamed history (current history) ▾
[Choose multiple histories](#)

— OR —

New history named:



Let Galaxy work for you!



Successfully ran workflow "Workflow constructed from history "TP Initiation"". The following datasets have been added to the queue:

- 1: chr4.fa
- 2: chr4.fastq
- 22: FASTQ Groomer on data 2
- 23: FastQC_FASTQ Groomer on data 2.html
- 24: FASTQ Trimmer on data 22
- 25: FastQC_FASTQ Trimmer on data 22.html
- 26: Bowtie2 on data 1 and data 24: aligned reads
- 27: flagstat on data 26
- 28: MarkDups_Dupes Marked.bam
- 29: MarkDups_Dupes Marked.html

History ↻ ⚙

TP Initiation
306.8 MB 🔍 ✓ 🗑 🗨

- 🕒 29: [MarkDups_Dupes Marked.html](#) 👁 🖋 ✕
- 🕒 28: [MarkDups_Dupes Marked.bam](#) 👁 🖋 ✕
- 🕒 27: [flagstat on data 26](#) 👁 🖋 ✕
- 🕒 26: [Bowtie2 on data 1 and data 24: aligned reads](#) 👁 🖋 ✕
- 🕒 25: [FastQC FASTQ Trimmer on data 22.html](#) 👁 🖋 ✕
- 🕒 24: [FASTQ Trimmer on data 22](#) 👁 🖋 ✕
- 🕒 23: [FastQC FASTQ Groomer on data 2.html](#) 👁 🖋 ✕
- 🌟 22: [FASTQ Groomer on data 2](#) 👁 🖋 ✕
- 21: [MarkDups_Dupes Marked.html](#) 👁 🖋 ✕
- 20: [MarkDups_Dupes Marked.bam](#) 👁 🖋 ✕
- 19: [flagstat on data 18](#) 👁 🖋 ✕
- 18: [Bowtie2 on data 1 and data 13: aligned reads](#) 👁 🖋 ✕
- 15: [FastQC FASTQ Trimmer on data 11.html](#) 👁 🖋 ✕