Module 1/6: Analyses ADN

- NGS Introduction
- Reads Quality Control
- Reads Cleaning

→ **Practical #1**

usegalaxy.fr

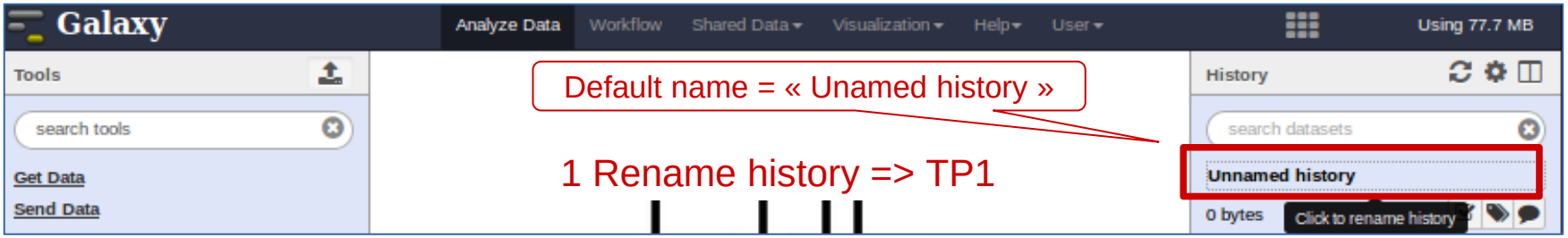
Galaxy

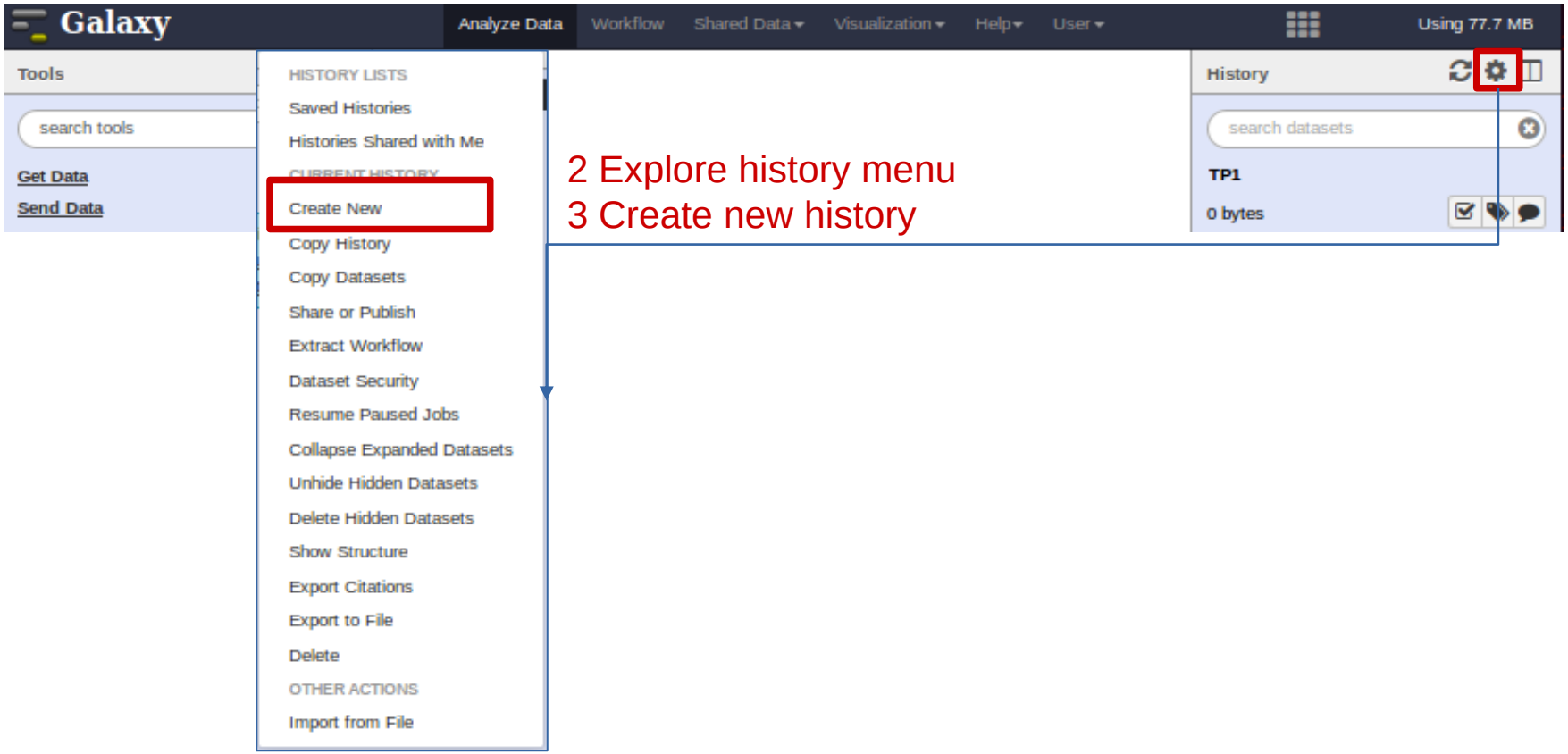
- Connection
- Upload data
- Working with datasets and histories
- Adding local reference
- Converting to *fastqsanger* format

Data for this tutorial

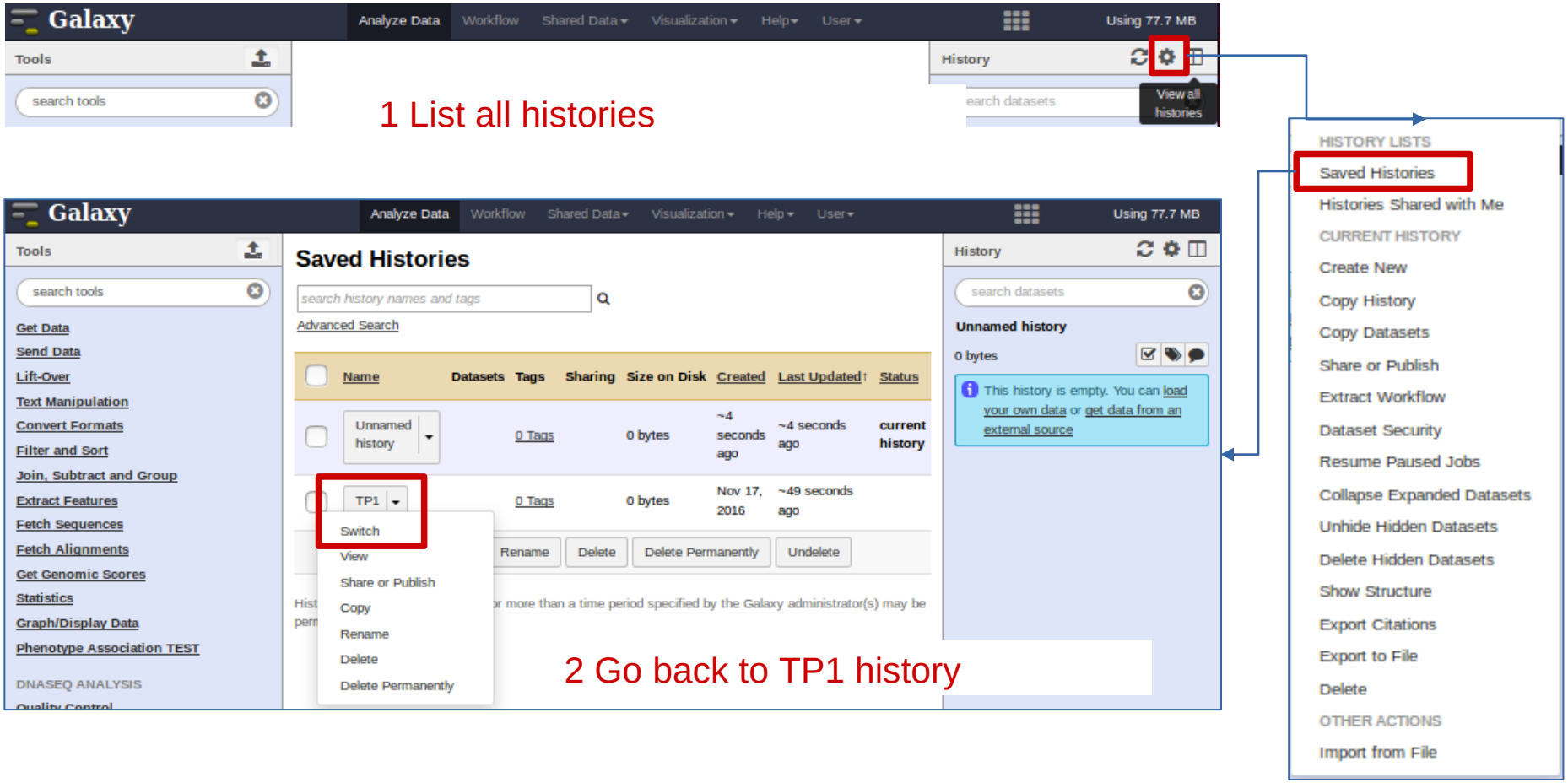
- Data from Human genome from Hapmap project
  https://www.ncbi.nlm.nih.gov/variation/news/NCBI_retiring_HapMap/
- Reference : small region from chromosome 20 20:380000-530000 (assembly GRCh37)
  - ⇒ file *GRCh37_region1.fasta*
- Reads: Illumina paired-end (2x100bp) for 3 samples (HG0096, HG0101 and HG0103)
  - ⇒ files *HG0XXX_1.fastq*, *HG0XXX_2.fastq*
  - *(only reads for this small region, for reasons of speed)*
- Dowload files on billile *wiki : https://wikis.univ-lille.fr/bilille/formation*
- Main goals for this first part of tutorial
  - Upload reference and reads for one sample (HG0101)
  - Work with histories, datasets and tools

History : « Folder » containing a set of data



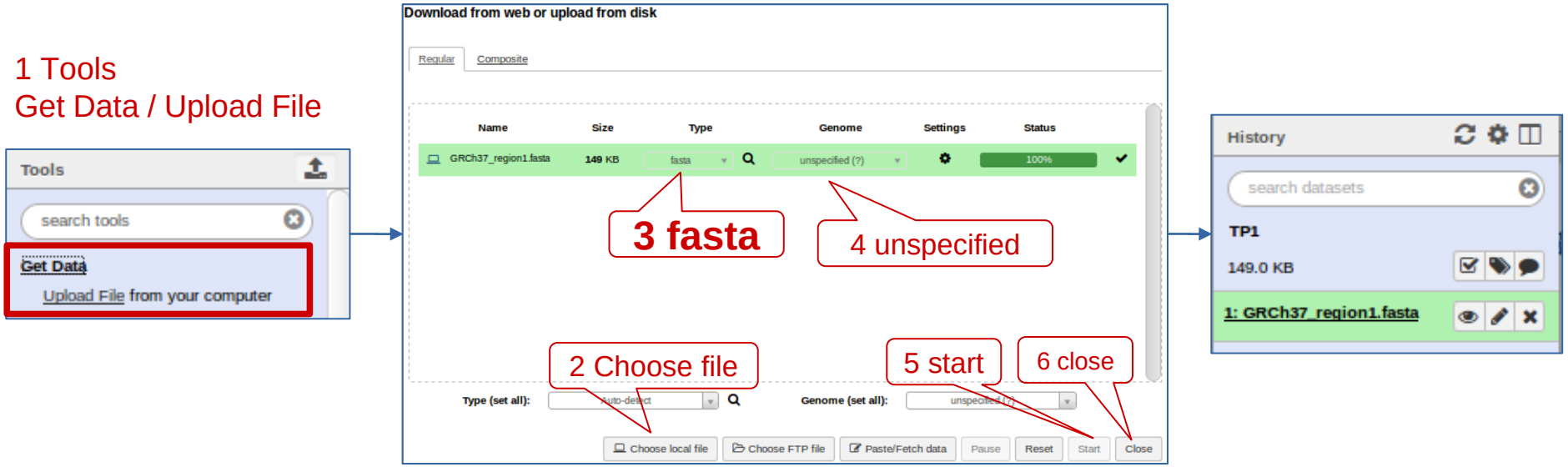


List histories, go back to TP1



Dataset ~ « Data file »
Upload reference in a *dataset*

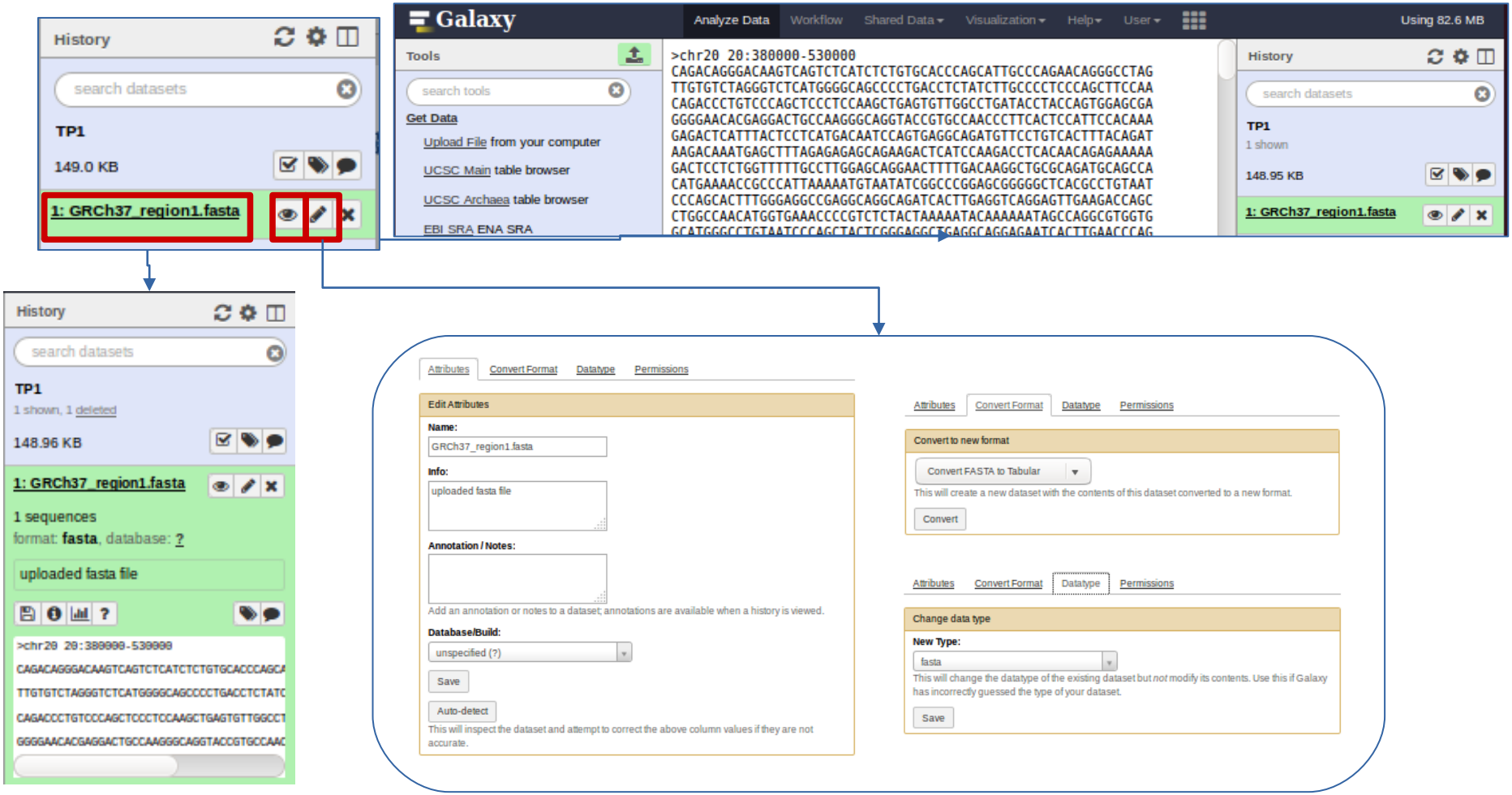
1 Tools
Get Data / Upload File



2 Choose file GRCh37_regions1.fasta
3 Choose fasta format (*! not csfasta*)
4 Keep « Unspecified » as genome
5 Run with start
6 Close

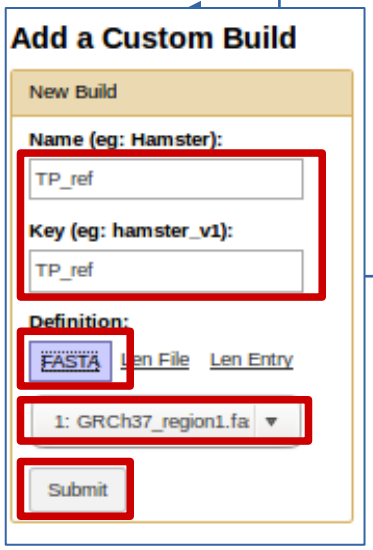
Dataset : summary, attributes, full data

1 Click on dataset name ⇒ show summary of attributes and data
2 Click on the eye ⇒ show data
3 Click on pencil ⇒ show attributes



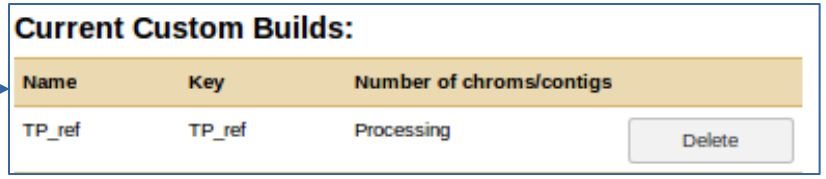
Add a local reference (TP_ref)

1 Menu User / Custom Builds



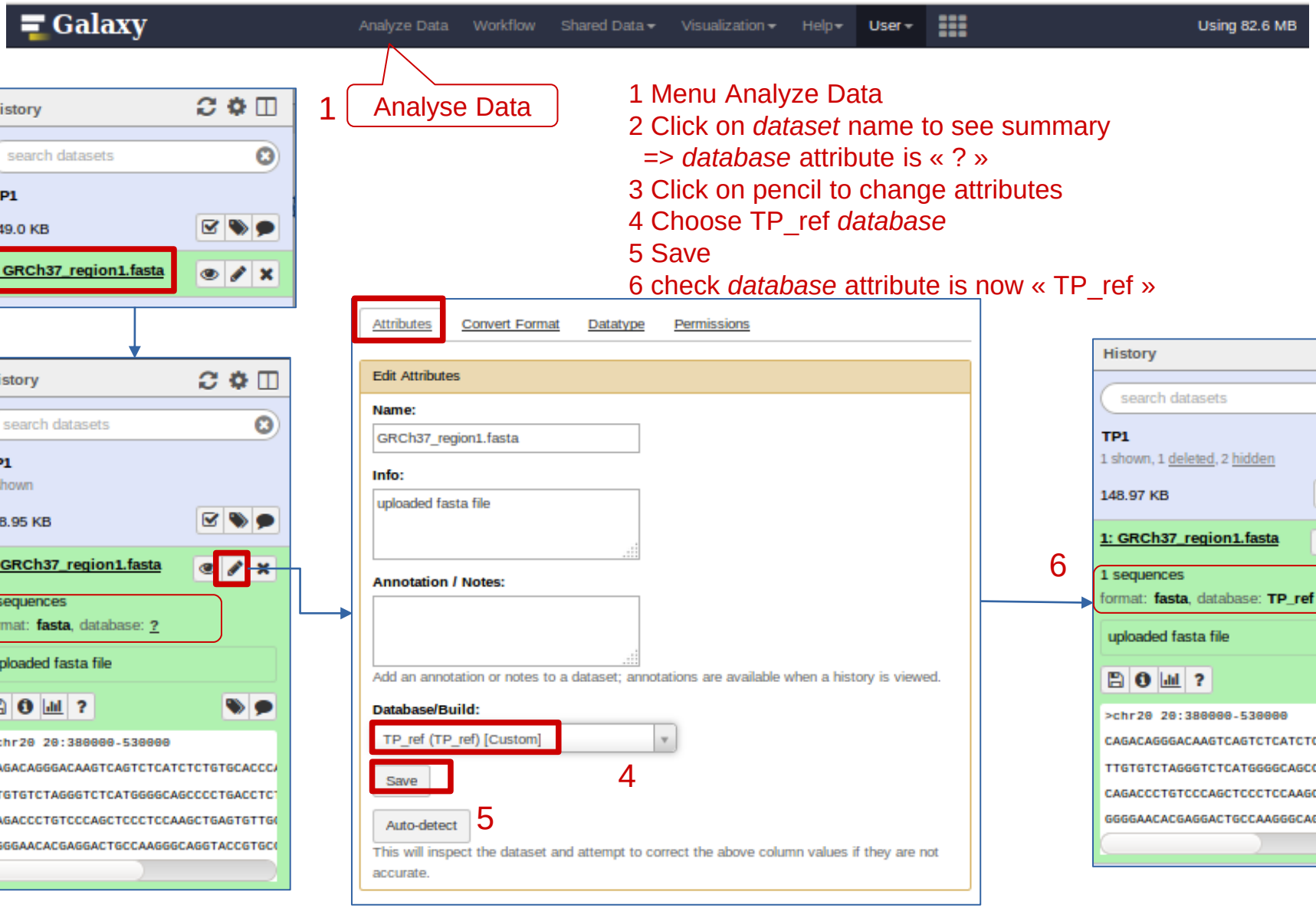
2
3
4
5

2 Choose name TP_ref
3 Choose *fasta* format
4 Choose *dataset* n° 1 : GRCh37_regions1.fasta
5 Submit



Reference is now available

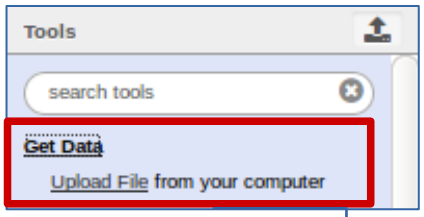
Check / Change *database* attribute



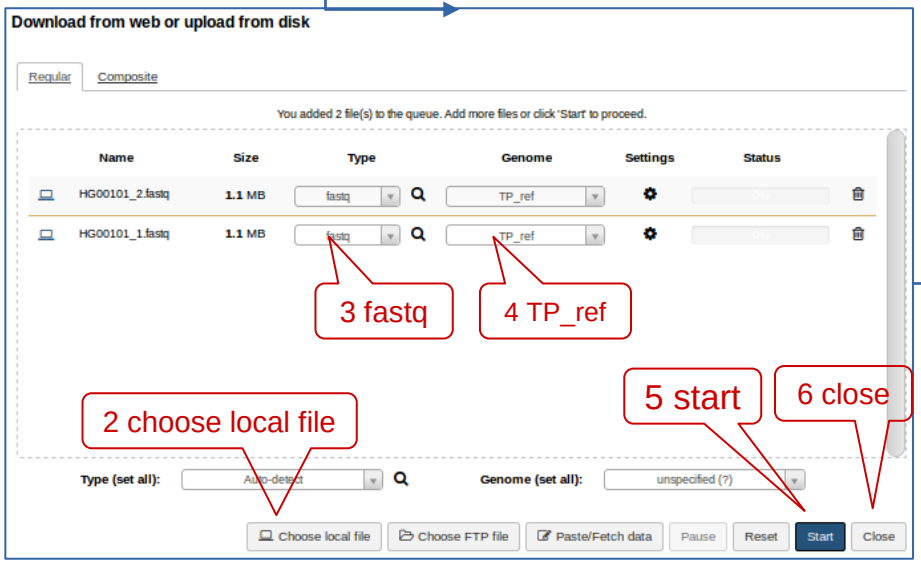
1 Menu Analyze Data
2 Click on *dataset* name to see summary
=> *database* attribute is « ? »
3 Click on pencil to change attributes
4 Choose TP_ref *database*
5 Save
6 check *database* attribute is now « TP_ref »

Upload reads (*fastq*) for sample HG0101

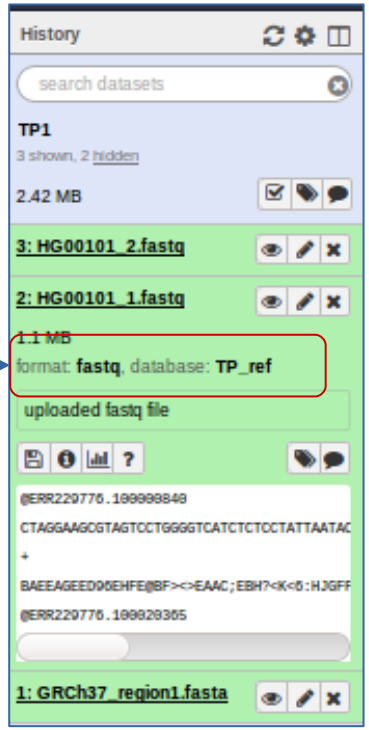
1 Tools: Get Data / Upload File
2 Choose files HG0101_1.fastq and HG0101_2.fastq
3 Choose « fastq » format
4 choose « TP_ref » genome
5 Run with start
6 Close
7 Check attributes



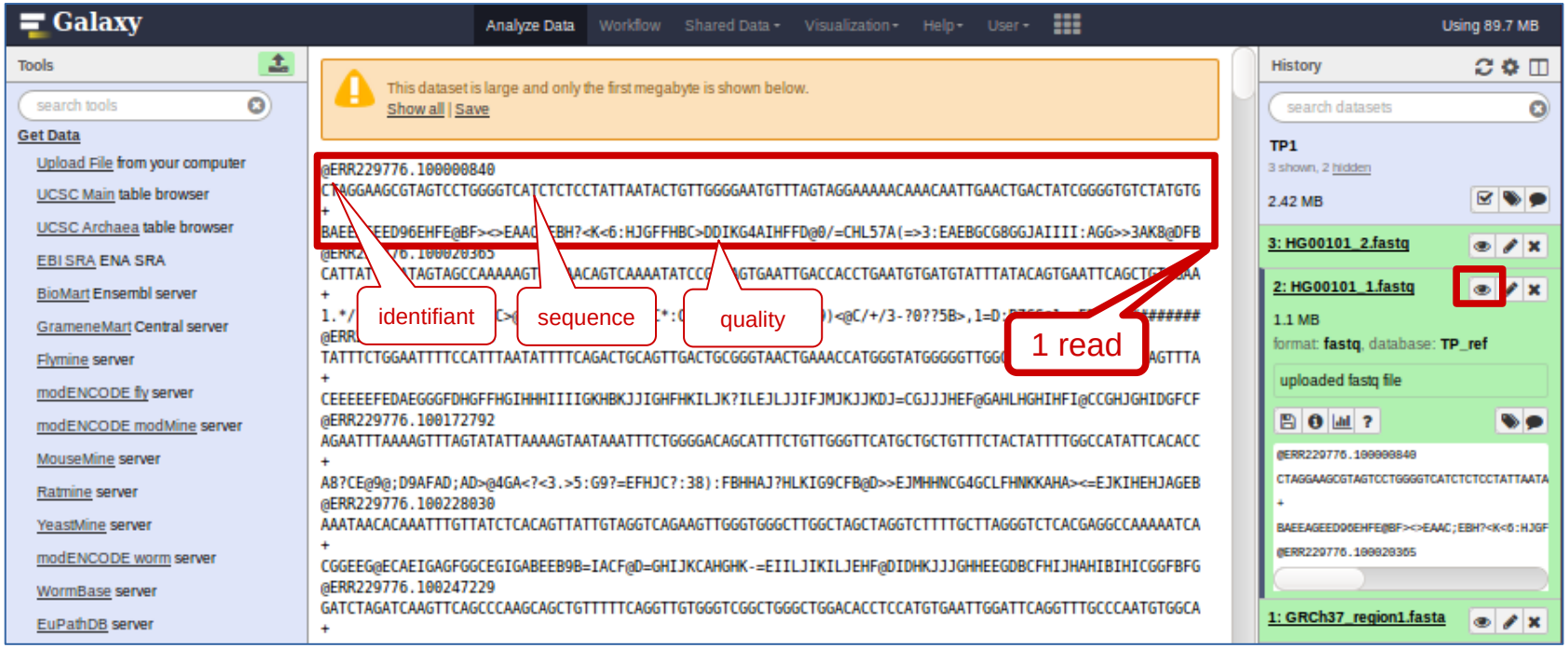
1



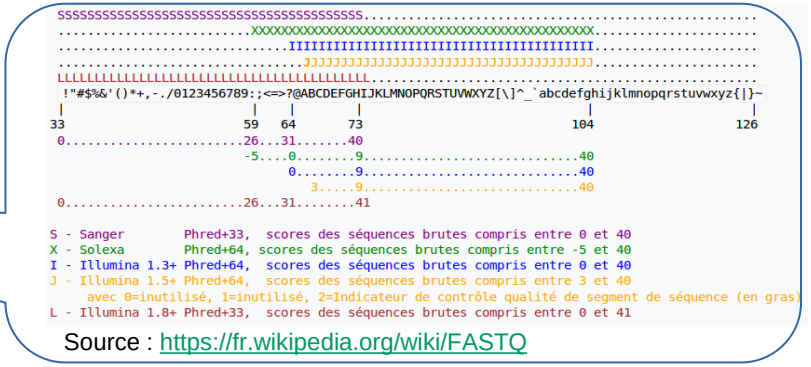
7



Look at a *fastq* file

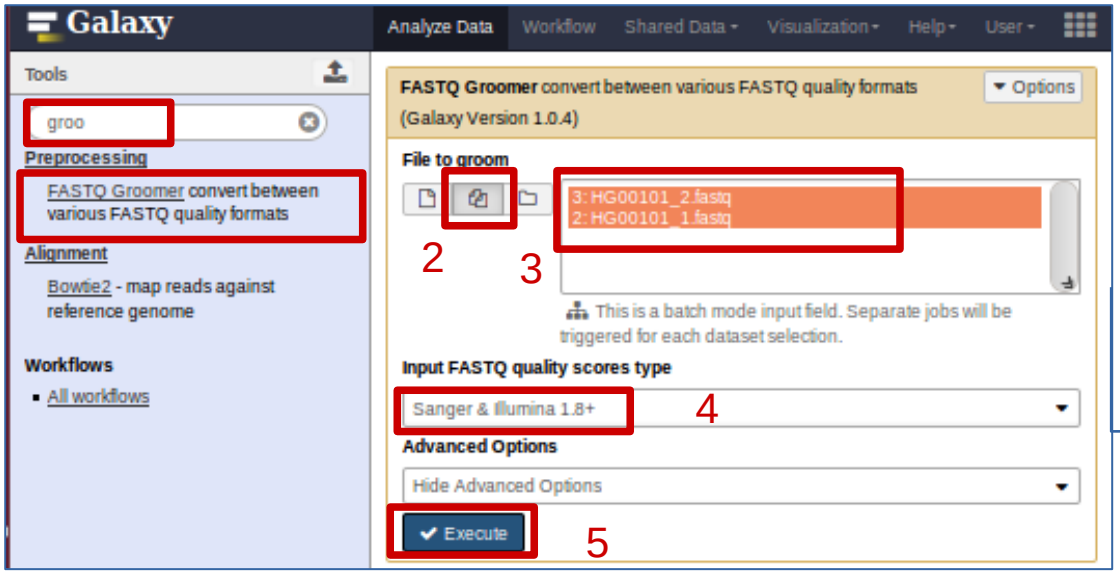


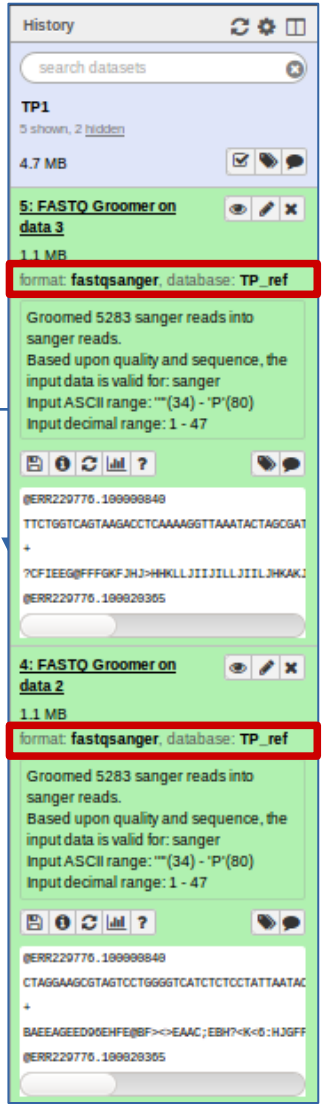
What is the file size ?
How many reads ?
Sizes of reads ?
Which is the quality coding ?



Source : https://fr.wikipedia.org/wiki/FASTQ

Convert to Sanger format : *groomer* tool

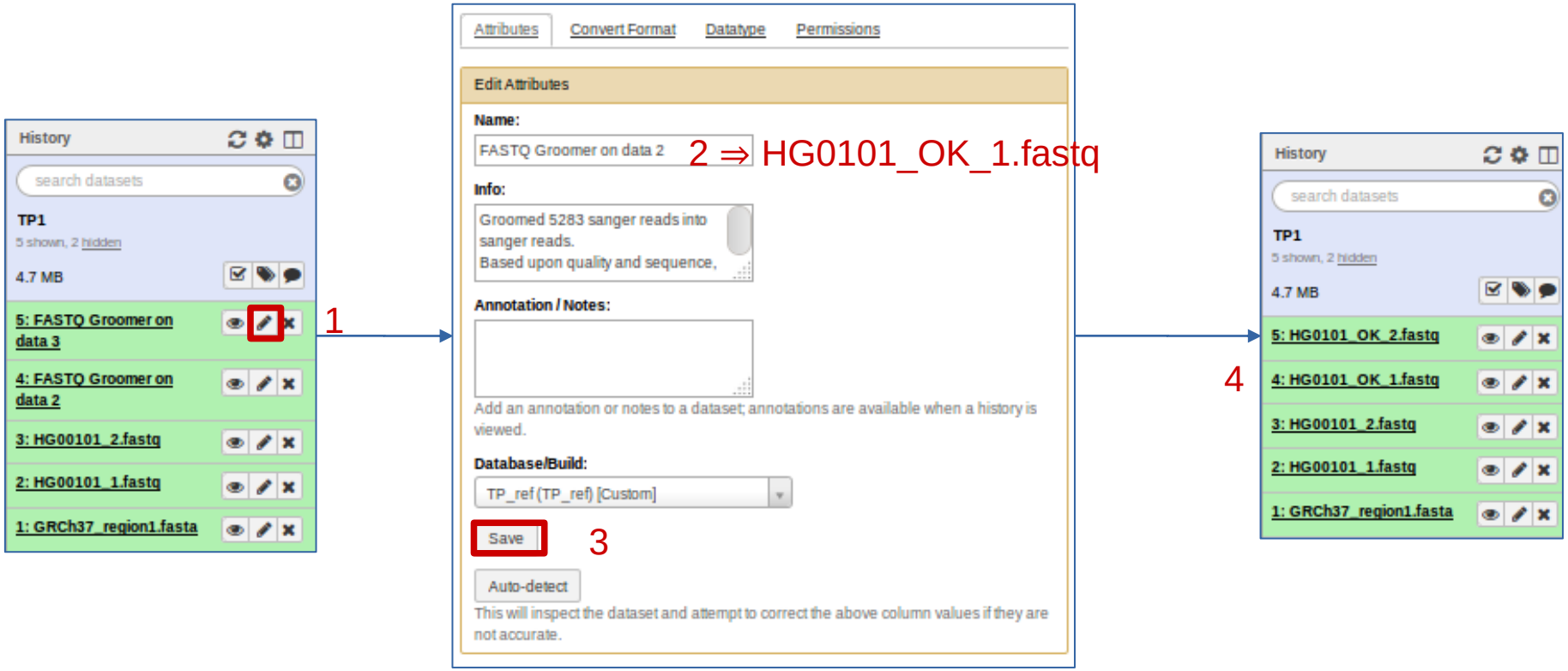




1 Tools: FASTQ Groomer
2 Choose to « groom » many files
3 Choose files HG0101_1_fastq and HG0101_2.fastq
4 Choose « Sanger & Illumina 1.8+ » format
5 Execute ⇒ Create 2 new *datasets* (N° 4 et 5)
6 Check new *datasets* attributes

What are the sizes of new *datasets* ?
How many reads ?
Which is the quality coding ?
What are the names of new *datasets* ?

Rename *datasets*



For each new *datasets* :

1 Click on pencil to change attributes

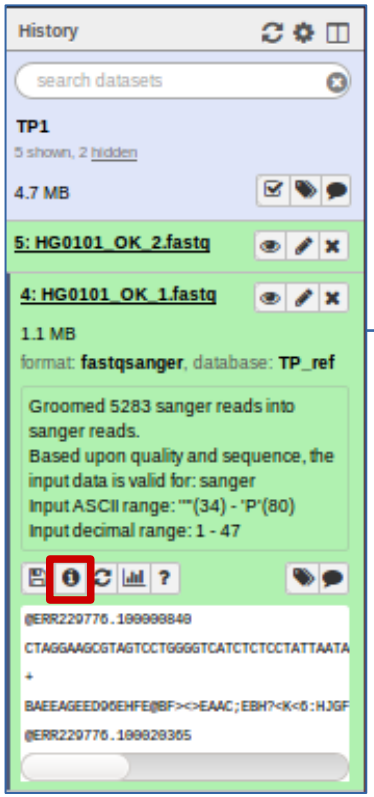
2 Change the name

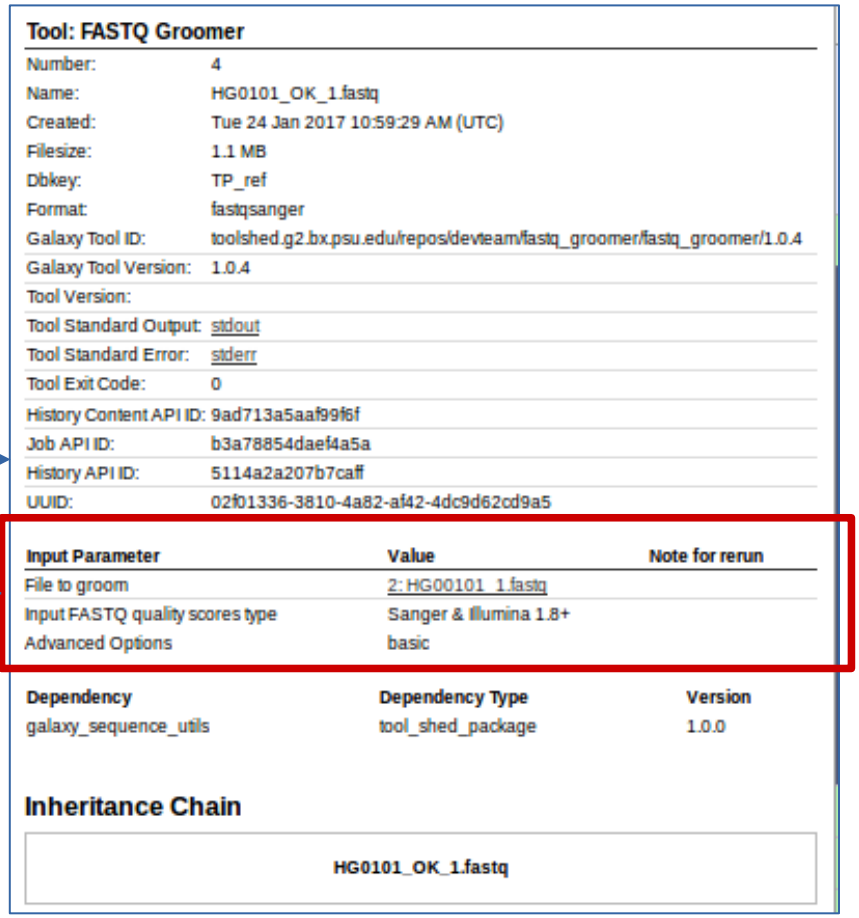
3 Save

4 Check new *datasets* names

After changin a *dataset* name, how can we retreive *dataset* origin ?

Retreive *dataset* origin





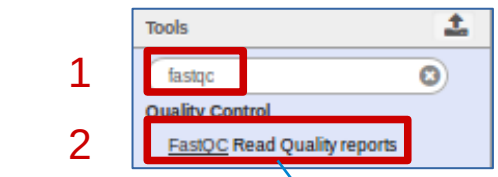
This *dataset* results from *groomer* tool, applied on *dataset* 2 (HG101_1.fastq)

Reads quality control

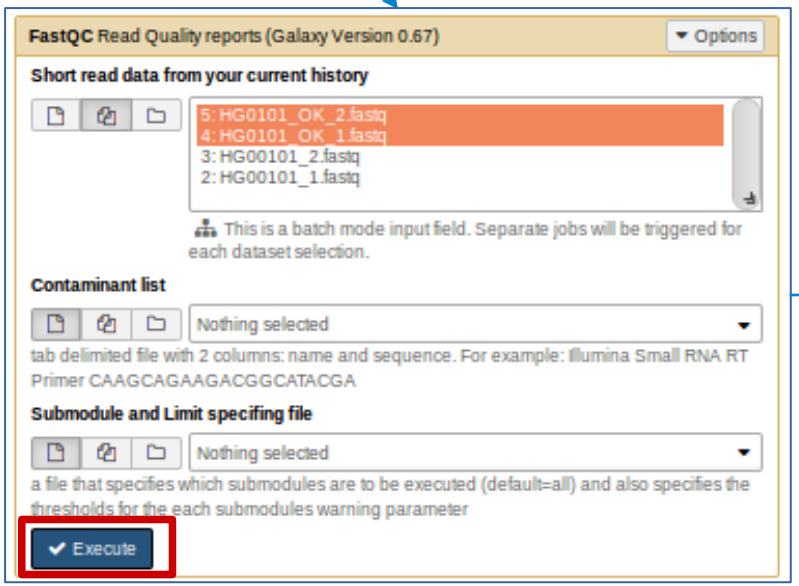
- Per base quality
- Per read mean quality
- Read size
- Adaptators
- Duplicated reads

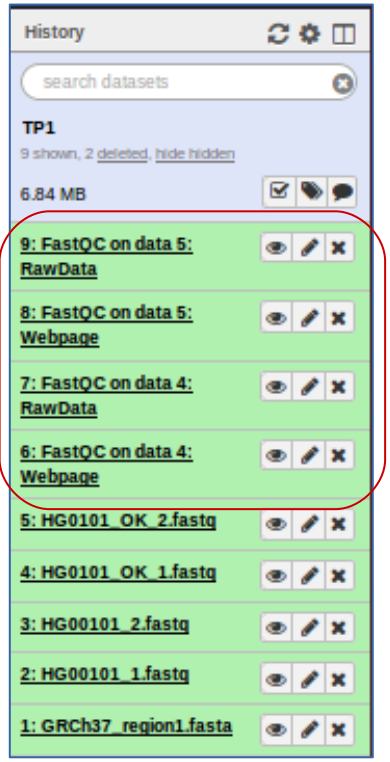
Reads quality control (*fastqc*)

- *Andrews, S. FastQC A Quality Control tool for High Throughput Sequence Data.*

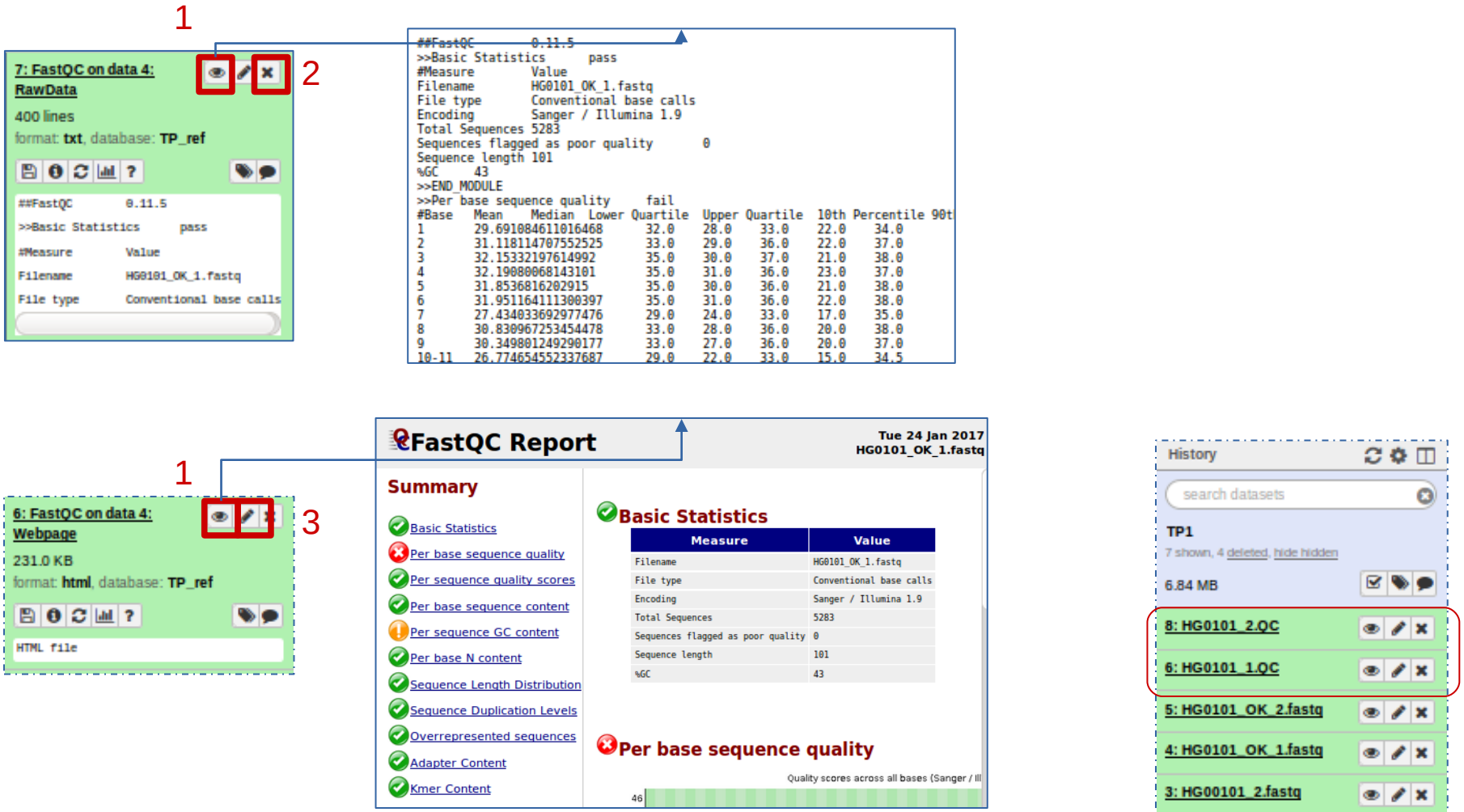


1,2 Choose tool : FastQC
3 Choose datasets n° 4 et 5
4 Execute ⇒ Create 4 new datasets
For each fastq file :
1 « raw data » and 1 « Webpage »





Manage FastQC result *datasets*



1 Look quickly at dataset content (we will deeply look at that later)

2 Rename « Webpage » datasets ⇒ HG0101_1.QC et HG0101_2.QC

3 Rename « RawData » datasets ⇒ HG0101_1.QC_raw et HG0101_2.QC_raw

FastQC : Summary & Basic Statistics

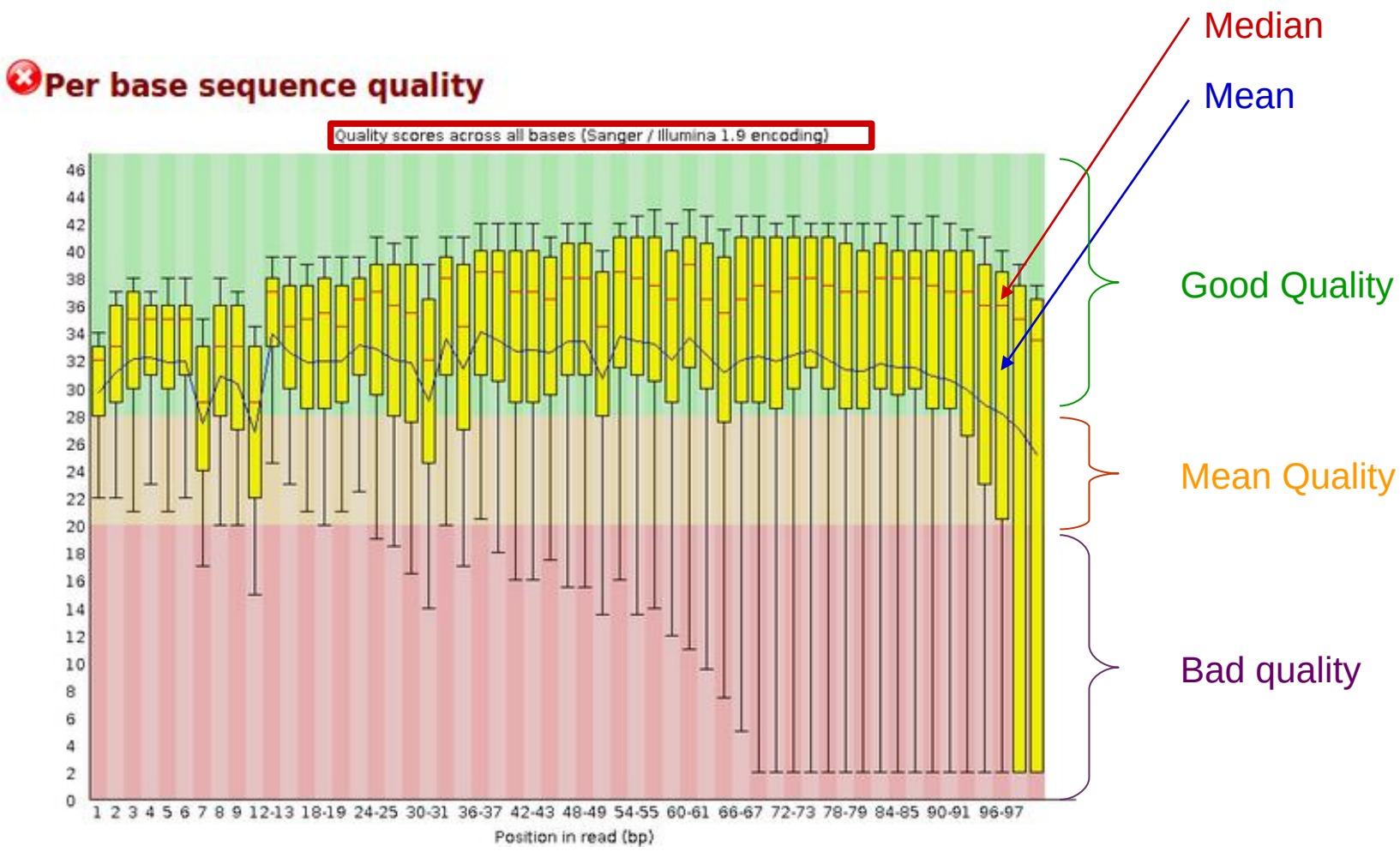
Summary

- Basic Statistics
- Per base sequence quality
- Per sequence quality scores
- Per base sequence content
- Per sequence GC content
- Per base N content
- Sequence Length Distribution
- Sequence Duplication Levels
- Overrepresented sequences
- Adapter Content

Basic Statistics

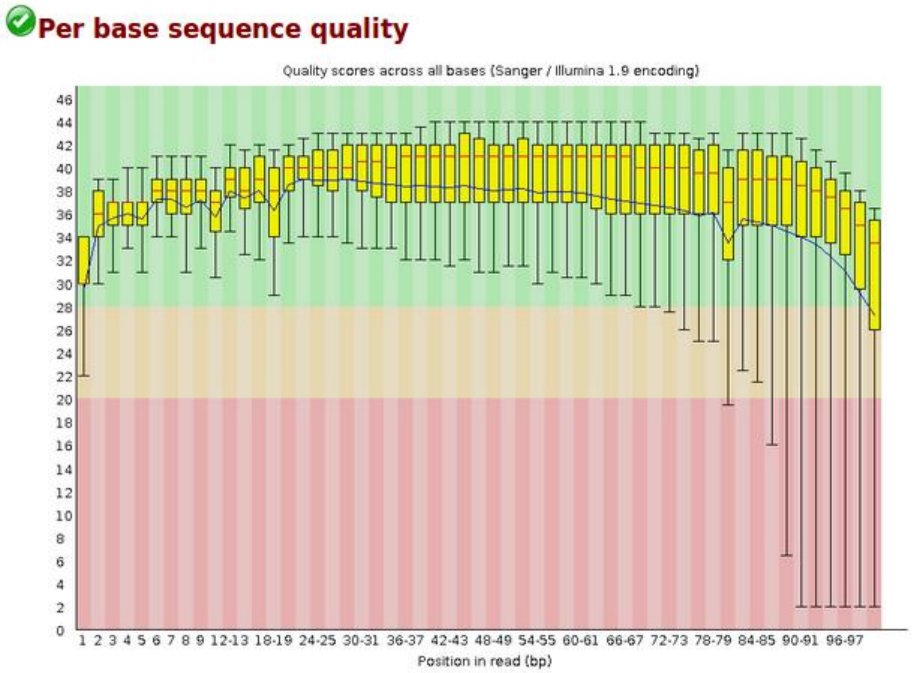
Measure	Value
Filename	HG0101_OK_1.fastq
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	5283
Sequences flagged as poor quality	0
Sequence length	101
%GC	43

FastQC : Per base sequence quality

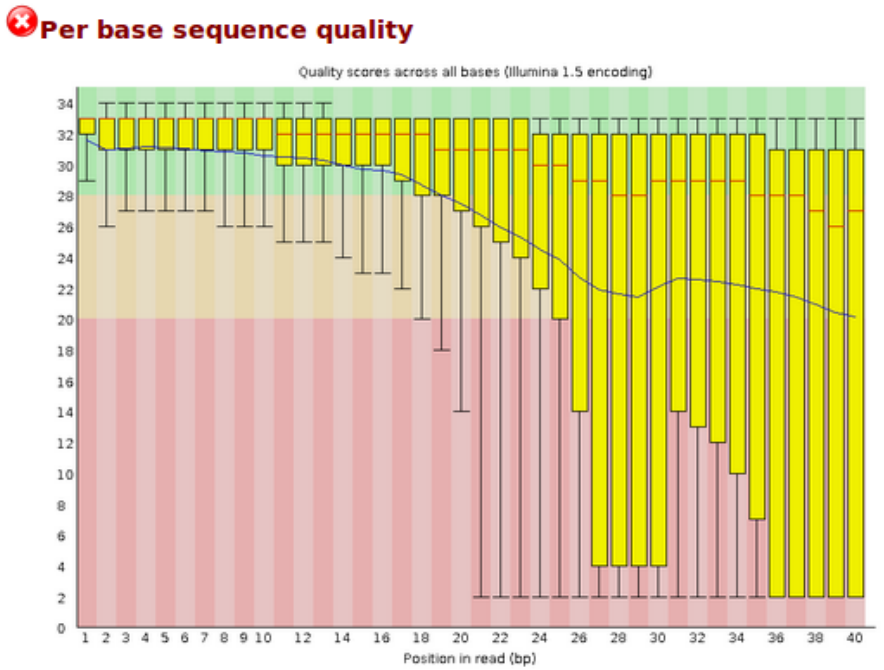


Fasqc : Per base sequence quality

Example OK



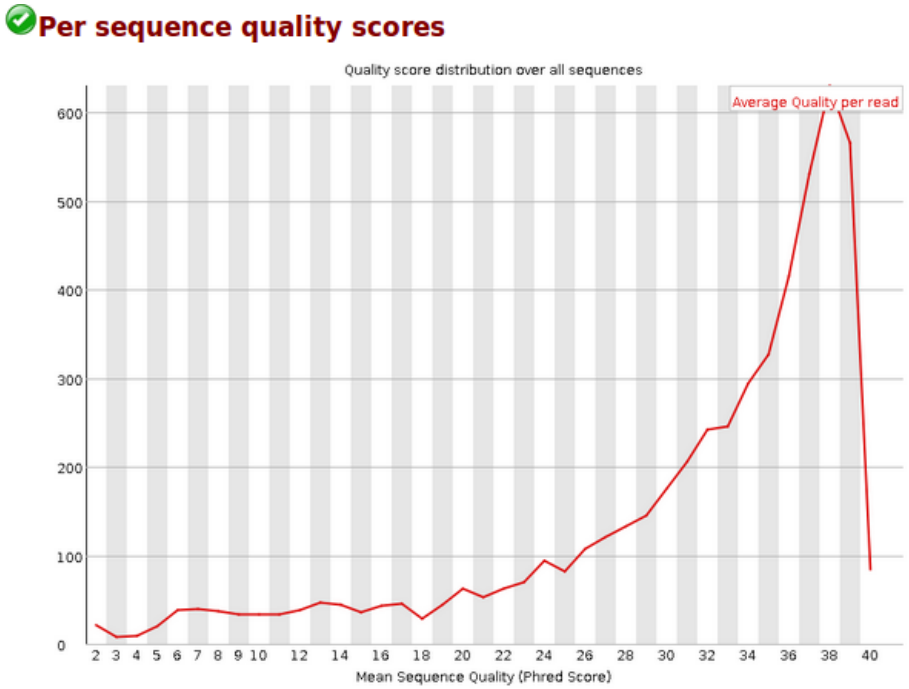
Example KO



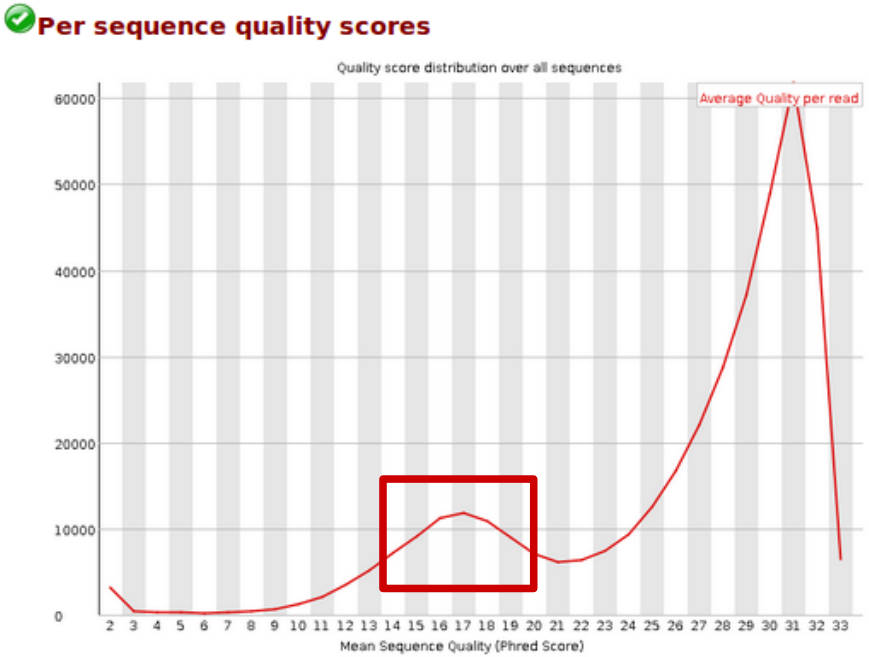
Source :
http://www.bioinformatics.babraham.ac.uk/projects/fastqc/bad_sequence_fastqc.html

Fastqc : Per sequence quality score

Example OK

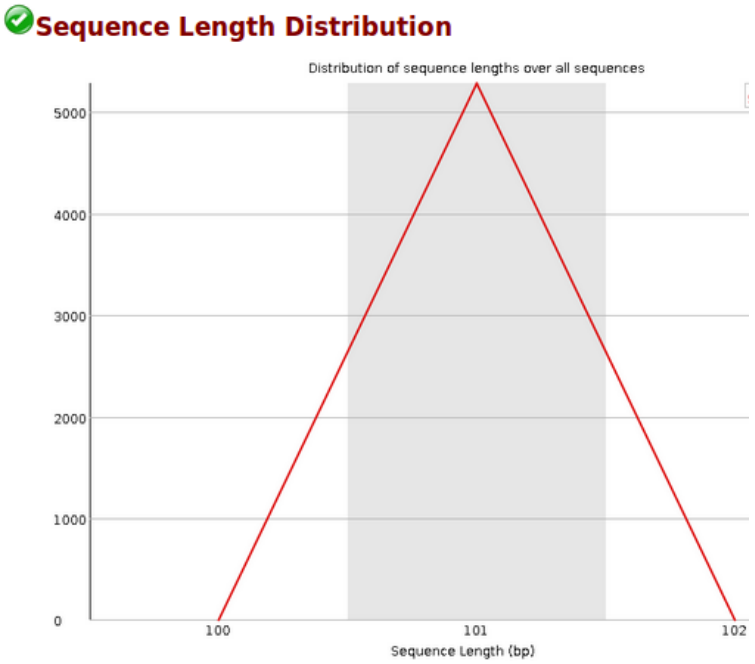


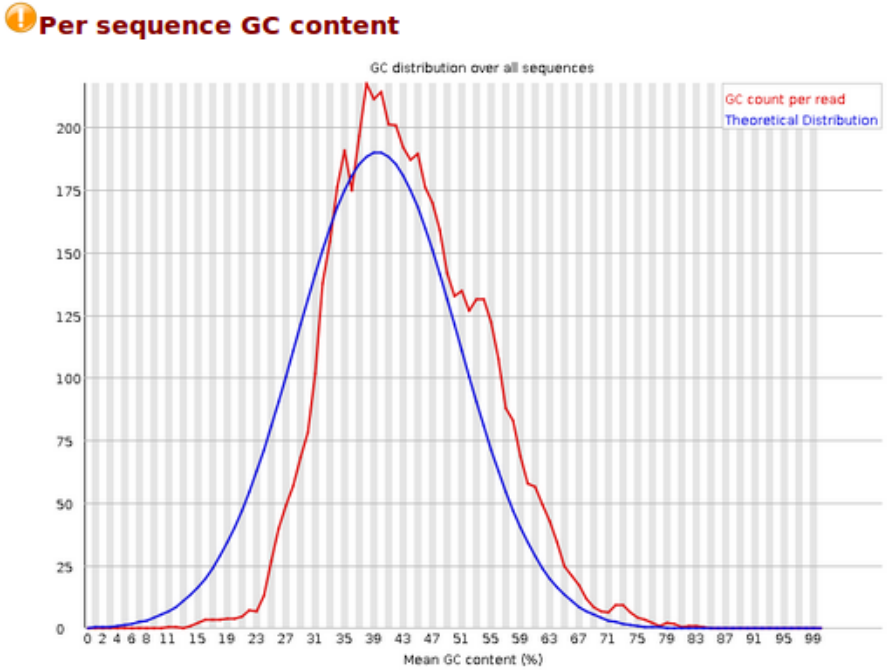
Example KO



Source :
http://www.bioinformatics.babraham.ac.uk/projects/fastqc/bad_sequence_fastqc.html

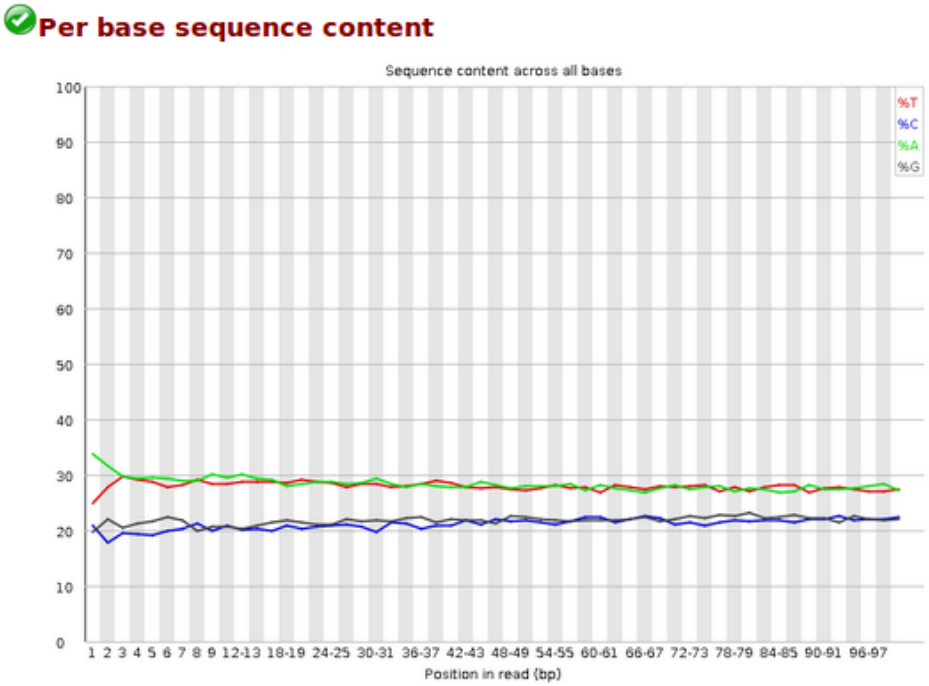
FastQC : Sequence Length Distribution & Per sequence GC content



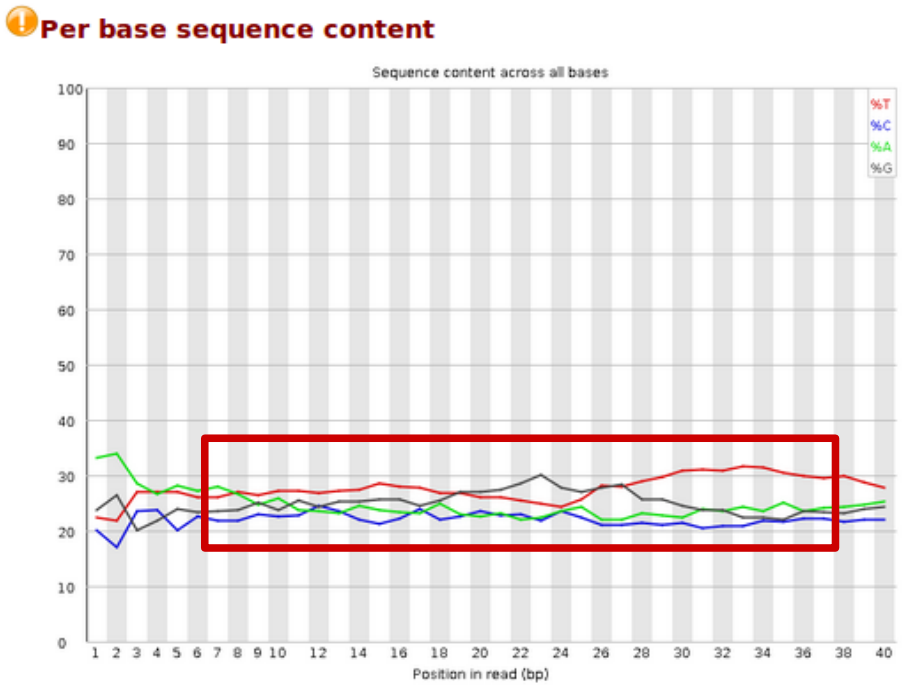


FastQC : Per base sequence content

Example OK



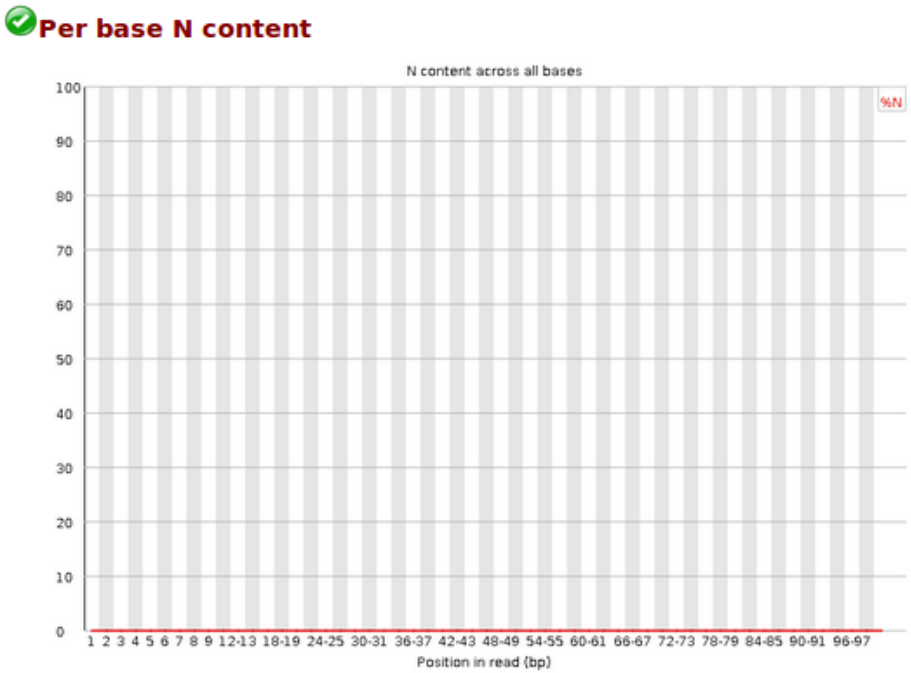
Example KO



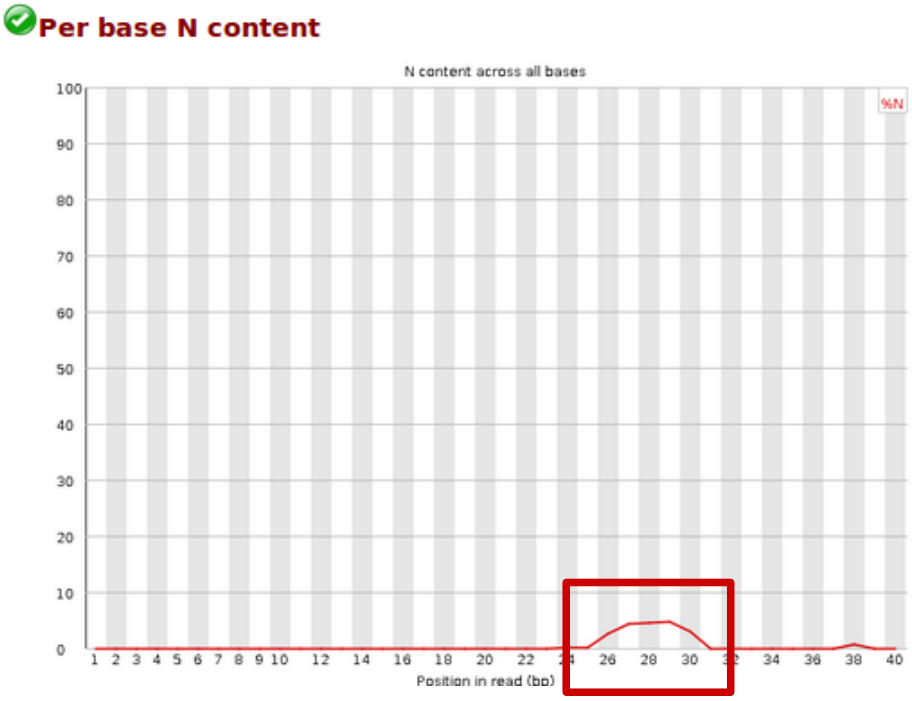
Source :
http://www.bioinformatics.babraham.ac.uk/projects/fastqc/bad_sequence_fastqc.html

FastQC : Per base N content

Example OK



Example KO



Source :
http://www.bioinformatics.babraham.ac.uk/projects/fastqc/bad_sequence_fastqc.html

FastQC : Overrepresented sequences

Example OK

Overrepresented sequences

No overrepresented sequences

Example KO

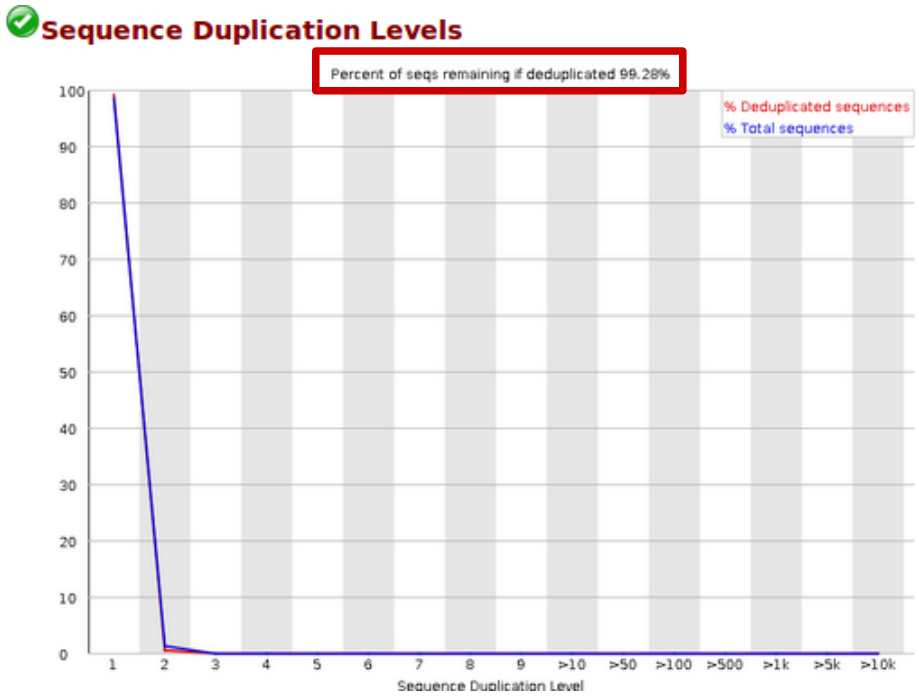
Overrepresented sequences

Sequence	Count	Percentage	Possible Source
AGAGTTTTATCGCTTCCATGACGCAGAAGTTAACACTTTC	2065	0.5224039181558763	No Hit
GATTGGCGTATCCAACCTGCAGAGTTTTATCGCTTCCATG	2047	0.5178582762542754	No Hit
ATTGGCGTATCCAACCTGCAGAGTTTTATCGCTTCCATGA	2014	0.5095019327680871	No Hit
CGATAAAAATGATTGGCGTATCCAACCTGCAGAGTTTTAT	1913	0.4839509420979134	No Hit
CCTGCAGAGTTTTATCGCTTCCATGACGCAGAAGTTAACA	613	0.15507680476007366	No Hit
CGGTTCAGCAGGAATGCCGAGATCGGAAGAGCGGTTCAGC	599	0.15153508328105078	Illumina Paired End PCR Primer 2 (96% over 25bp)
TCTGCAGGTTGGATACGCCAATCATTTTTATCGAAGCGCG	585	0.1479933618020279	No Hit

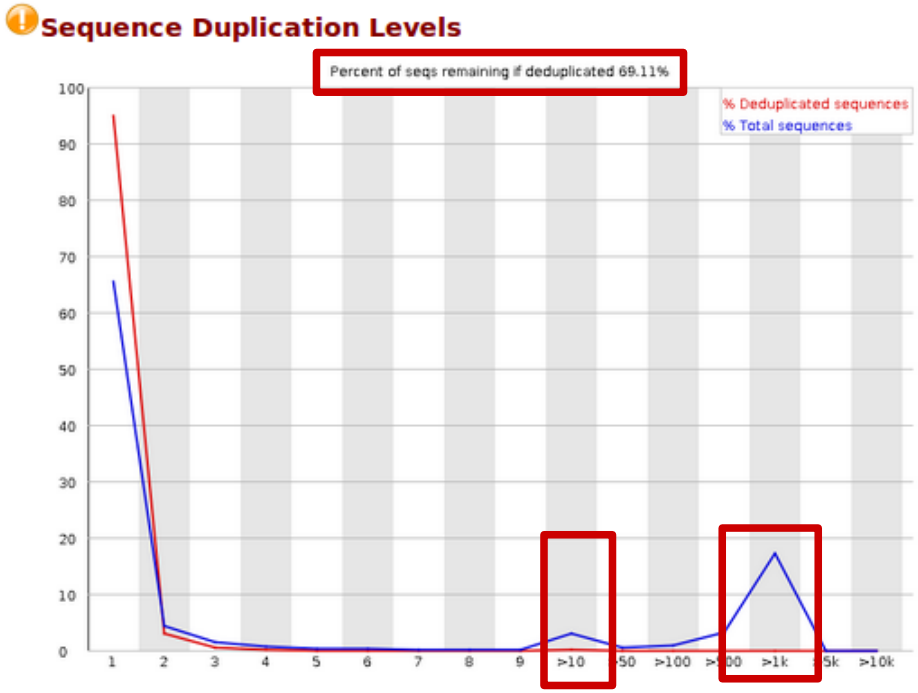
Source :
http://www.bioinformatics.babraham.ac.uk/projects/fastqc/bad_sequence_fastqc.html

FastQC : Sequence Duplication Levels

Example OK



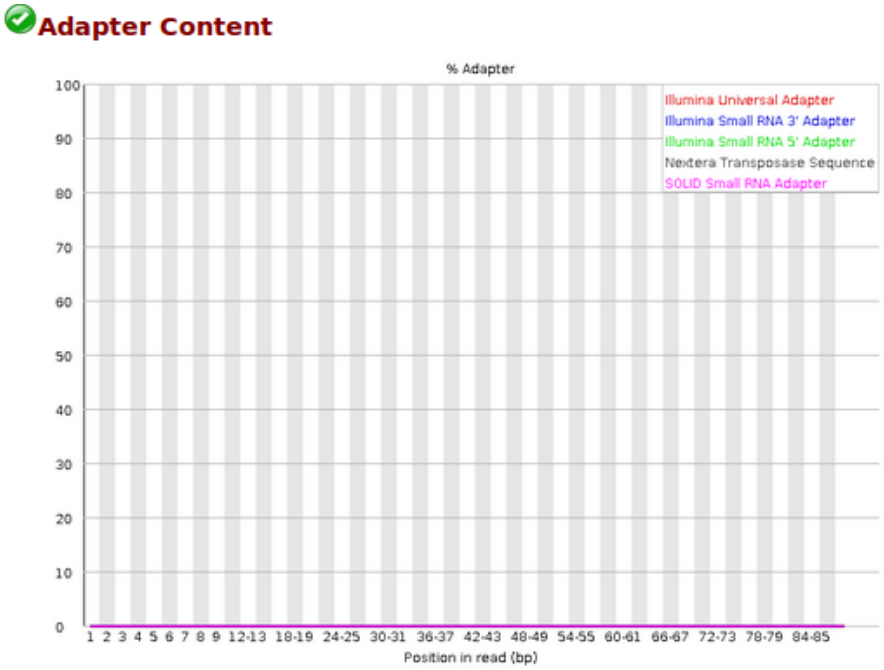
Example KO



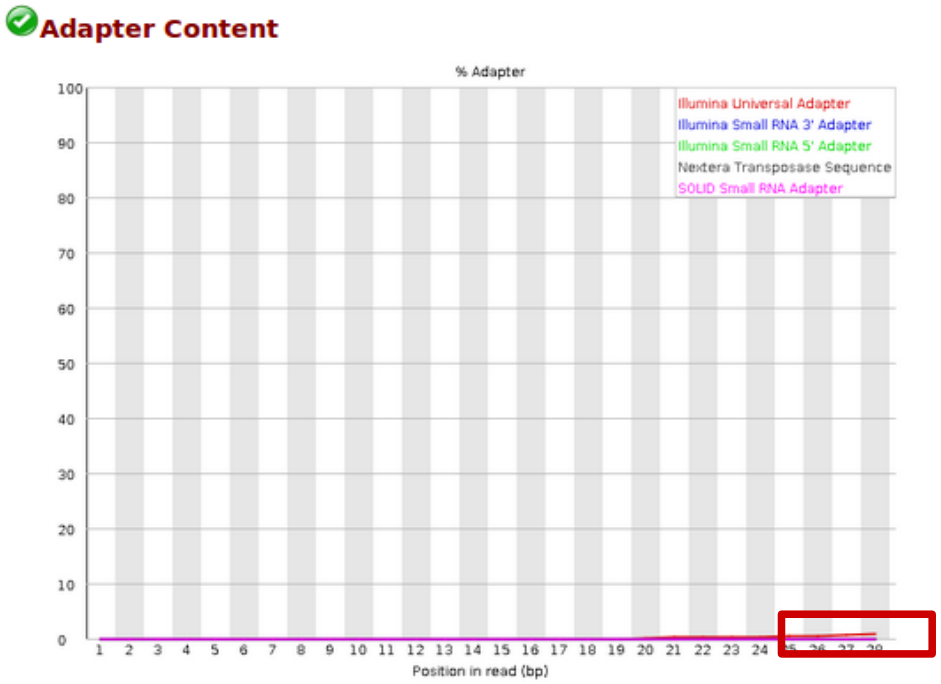
Source :
http://www.bioinformatics.babraham.ac.uk/projects/fastqc/bad_sequence_fastqc.html

FastQC Adapter Content

Example OK



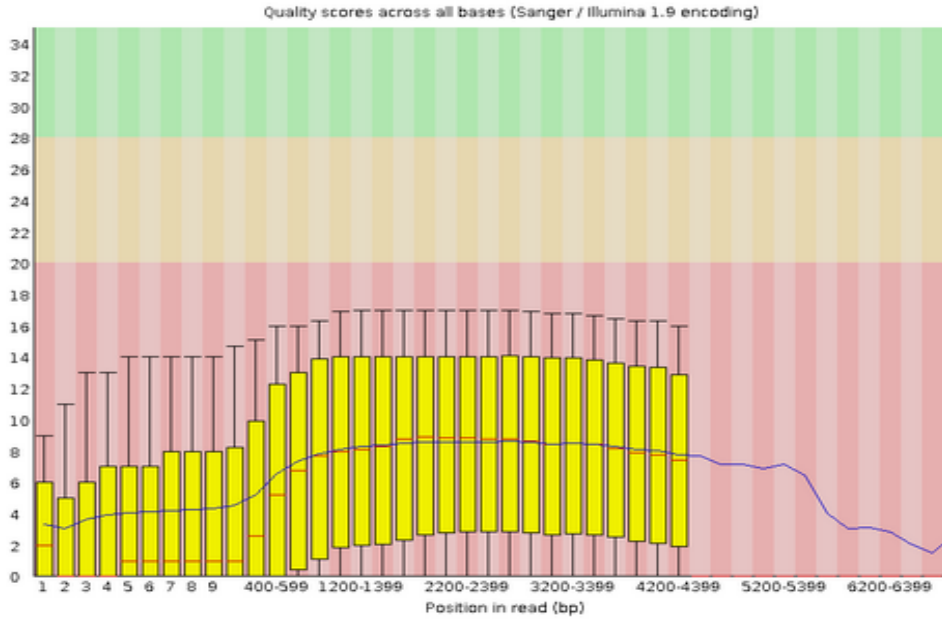
Example KO



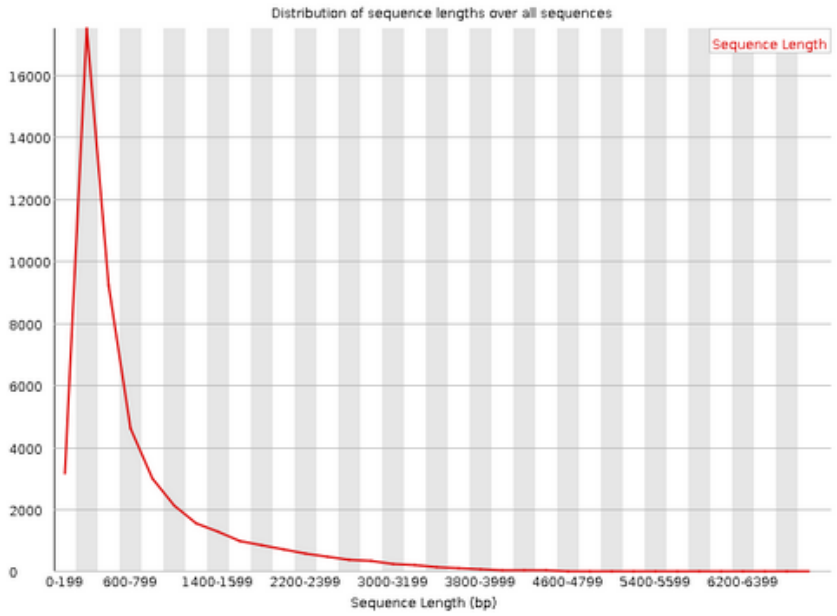
Source :
http://www.bioinformatics.babraham.ac.uk/projects/fastqc/bad_sequence_fastqc.html

FastQC – Example with PacBio

Per base sequence quality



Sequence Length Distribution



Source :
http://www.bioinformatics.babraham.ac.uk/projects/fastqc/pacbio_srr075104_fastqc.html

Cleaning Reads

- Filtrering adaptators
- Filtering & trimming reads
- Comparing quality before and after cleaning

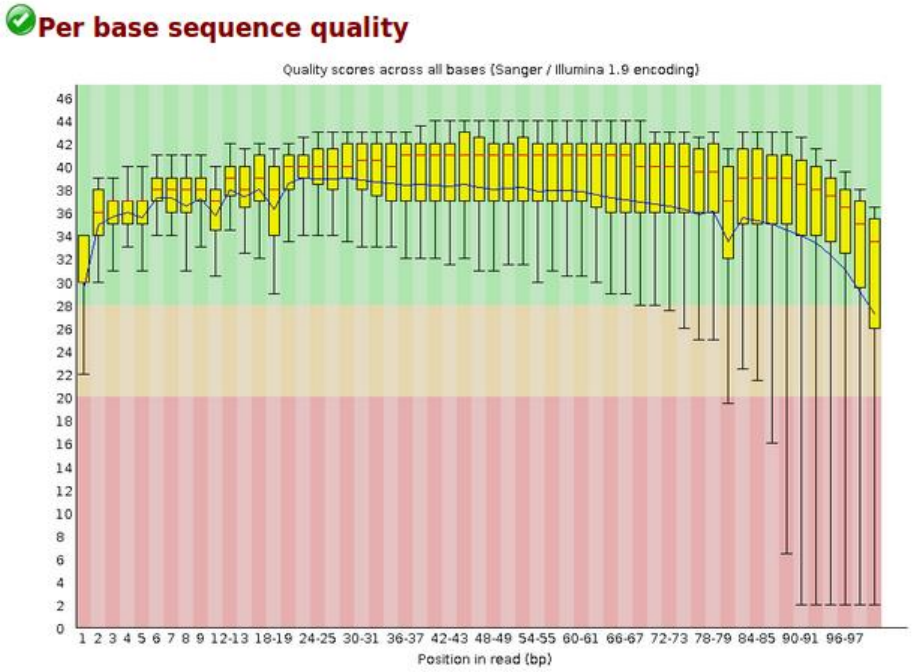
Filtering & trimming

- Filtering = remove reads
  - Based on quality or size criteria
- Trimming = remove read ends
  - Fixed number of bases
  - Bases < quality

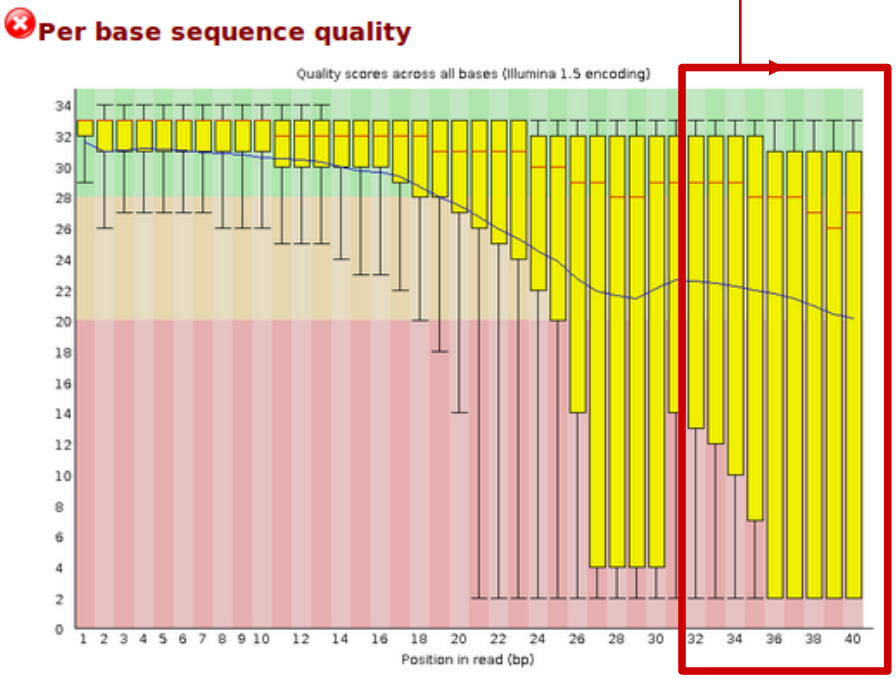
Trimming

Cut bad quality bases at the end of reads

Exemple OK



Exemple KO

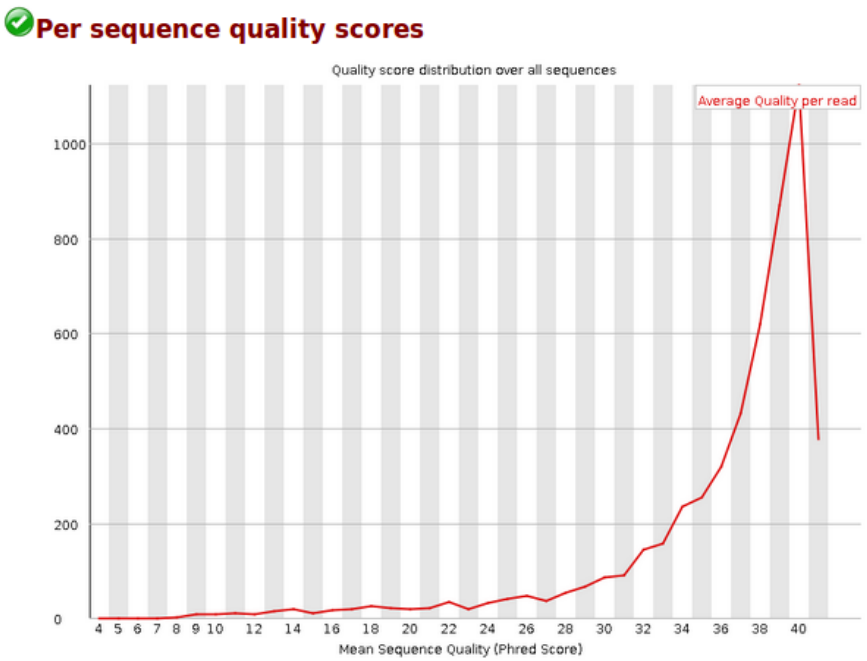


Source :
http://www.bioinformatics.babraham.ac.uk/projects/fastqc/bad_sequence_fastqc.html

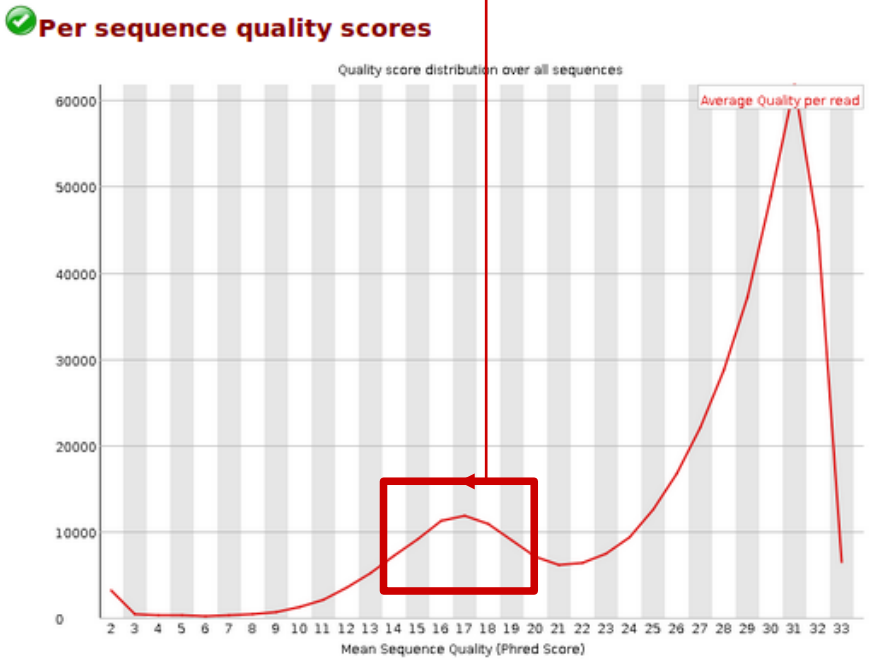
Filtering

Remove reads with bad mean quality

Exemple OK

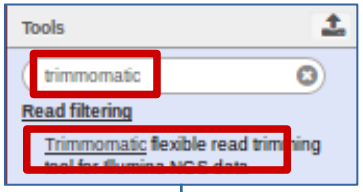


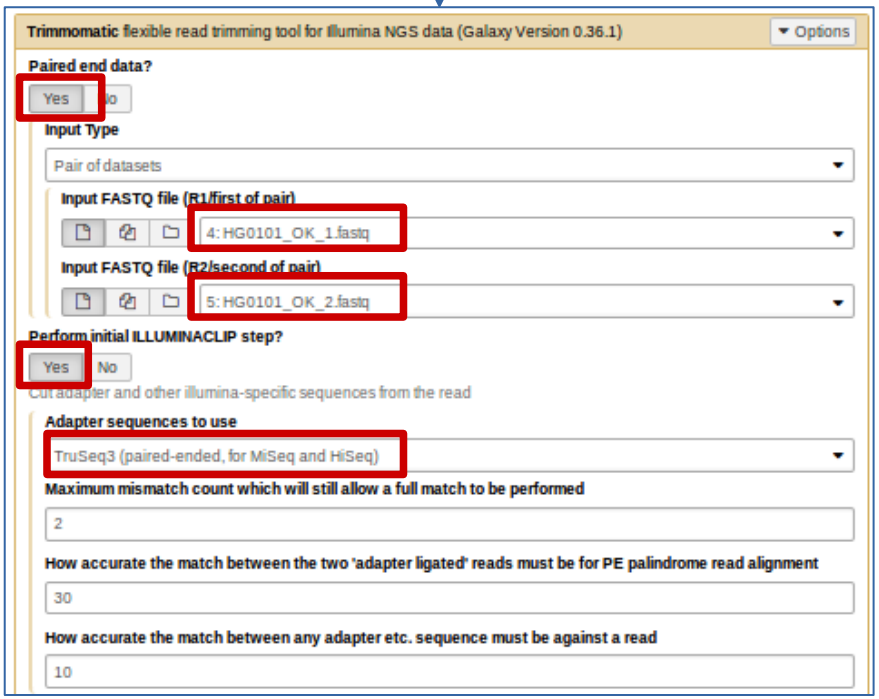
Exemple KO



Source : http://www.bioinformatics.babraham.ac.uk/projects/fastqc/bad_sequence_fastqc.html

Reads cleaning (*Trimmomatic*) 1/2



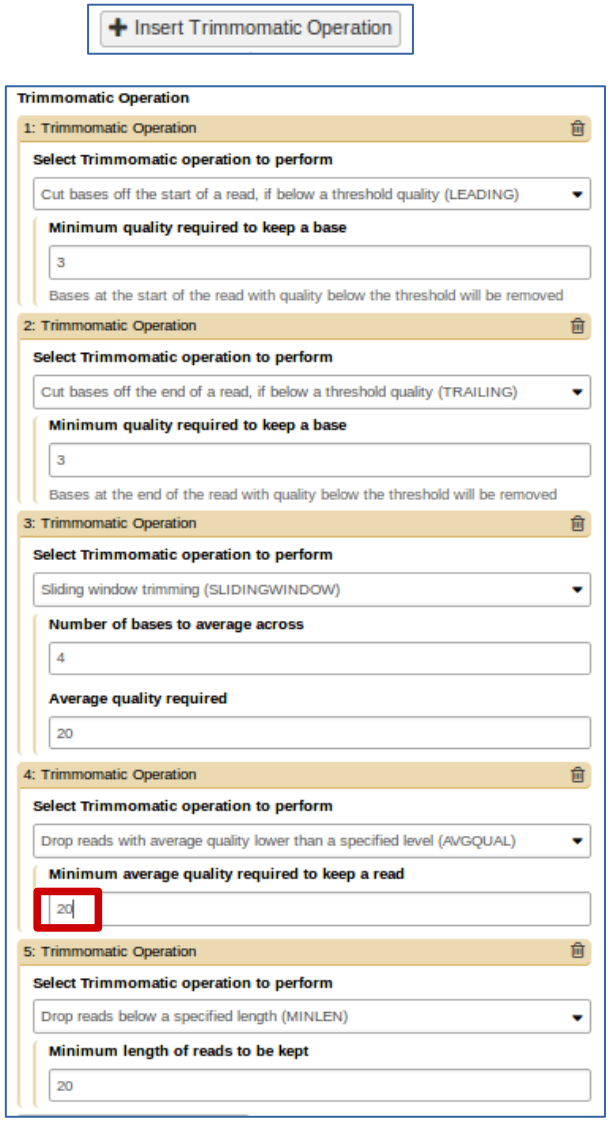


Bolger, A. M. and Lohse, M. and Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. In Bioinformatics, 30 (15), pp. 2114–2120

1 Choose files

2 Parameters for adaptators

Reads cleaning (*Trimmomatic*) 2/2



Add operations (cleaning steps) :

1 : *LEADING* : Cut bad quality 5' bases

2 : *TRAILNG*:Cut bad quality 3' bases

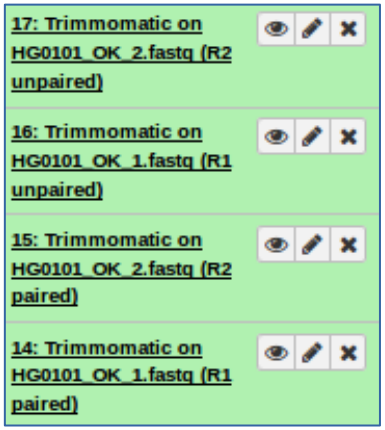
3 : *SLIDINGWINDOW* : Cut bases with bad mean quality in a sliding window

4 : *AVGQUAL* : remove reads with bad mean quality

5 : *MINLEN* : Remove small size reads

+ Output trimmomatic log messages?: Yes

Trimmomatic : Results



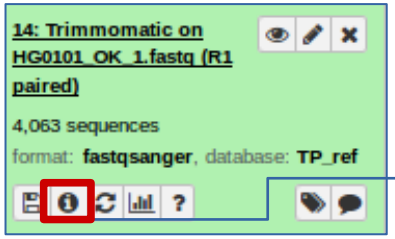
Unpaired reads2 (corresponding reads1 has been removed during cleaning)

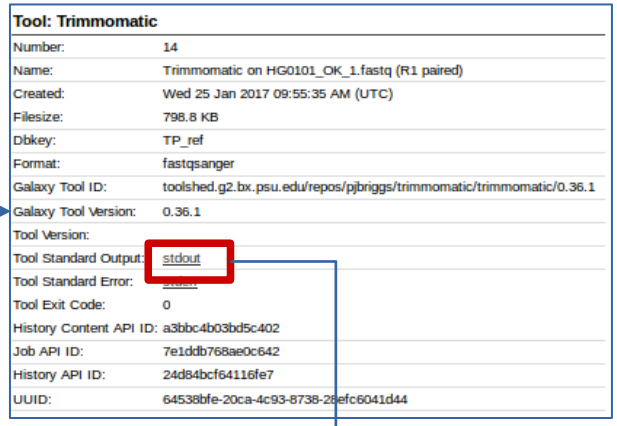
Unpaired reads1 (corresponding reads2 has been removed during cleaning)

Reads2 after cleaning

Reads1 after cleaning

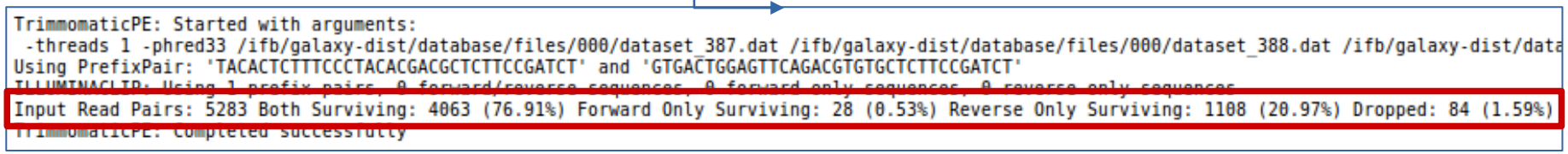
How many paired reads after cleaning ?
Are there any trace or log of what happens during cleaning ?





From summary, click on « i » icon

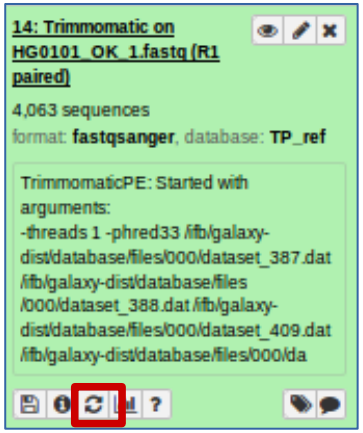
Look content of « *stdout* »
This contains all messages send by this tool during execution.



Trimmomatic : 2nd try

Run this tool again after changing AVGQUAL parameter value to 25

1

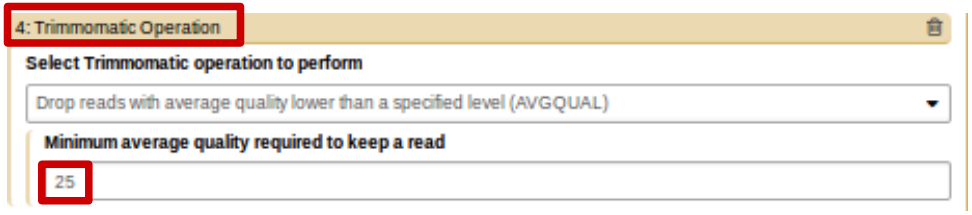


2

1 Use one of the *datasets* produced by previous analysis

2 Click on « *Run this job again* » icon

3



All parameters are pre-sets with values used in the previous execution

3 Change only the parameter in step n° 4 AVGQUAL ⇒ value 25

- How many paired reads after new cleaning ?
- Work only with paired *dataset* produced by first cleaning (remove second try)
- Rename the *datatsets* HG0101_clean_1.fastq and HG0101_clean_2.fastq
- Run quality control on these 2 datasets.
- Compare quality control before and after cleaning

Compare quality control before / after cleaning

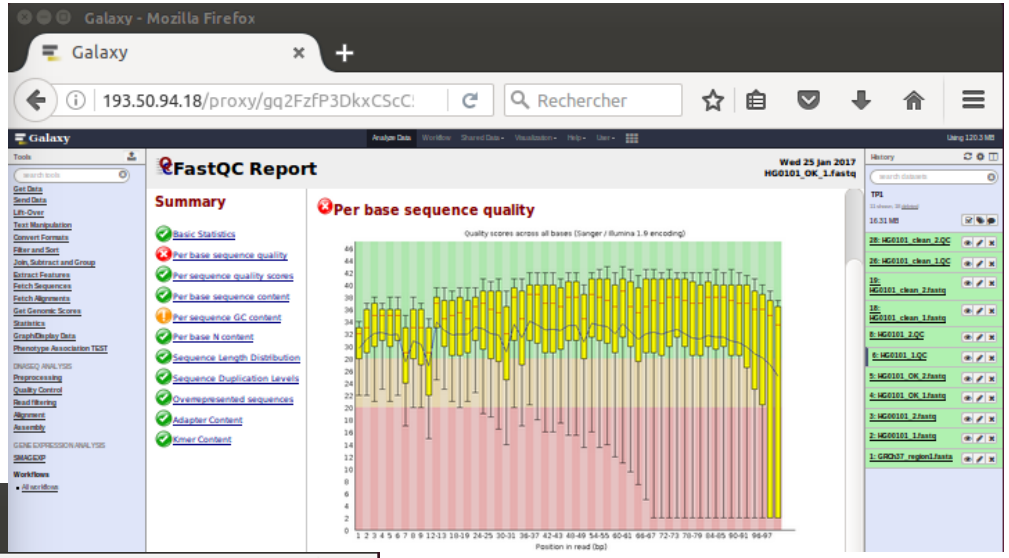
Solution 1 :

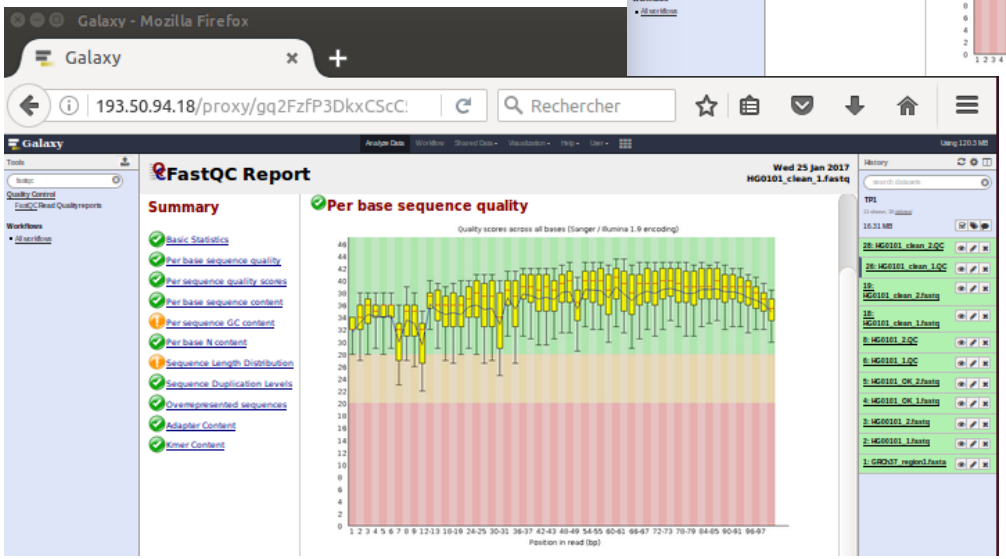
Open a second web browser window

Connect to access history and datasets

Visualize datasets

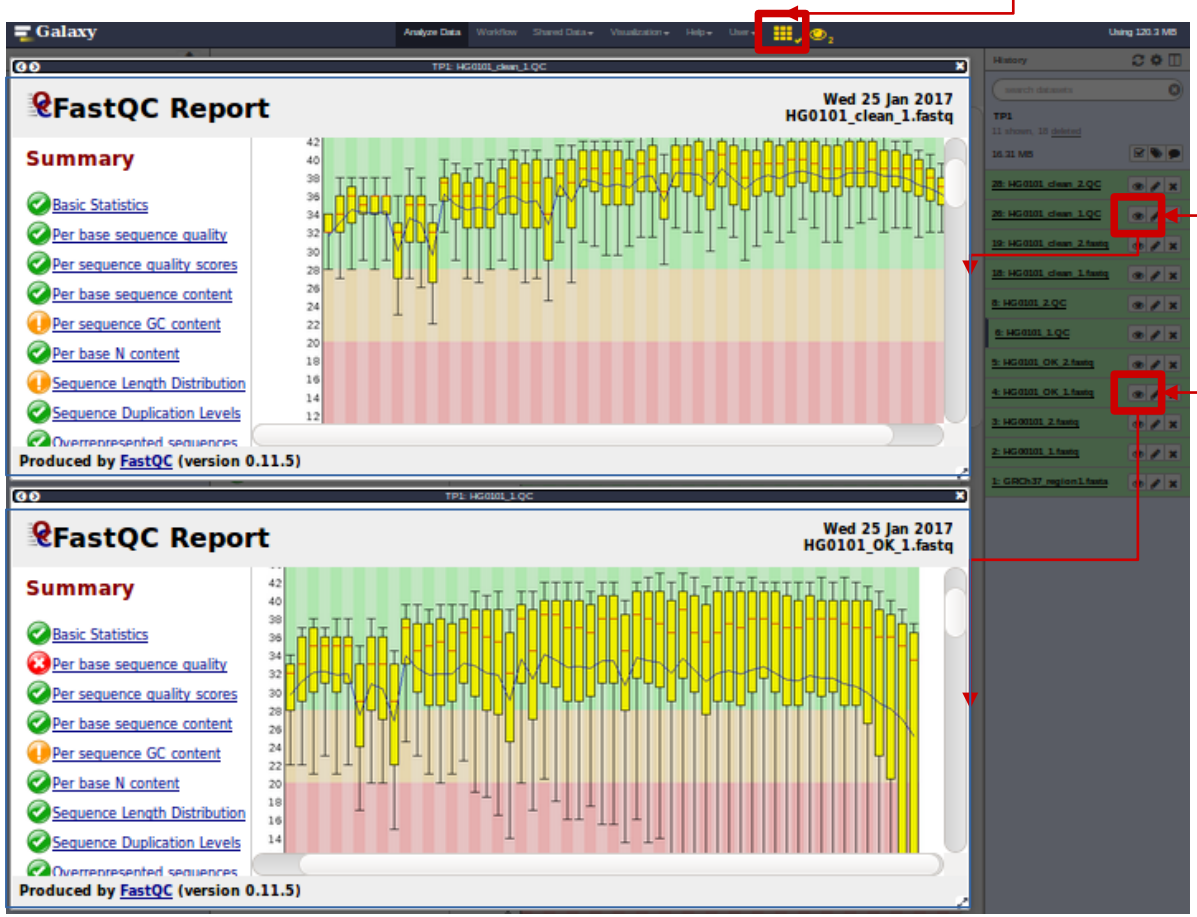
- HG101_OK_1.CQ
- HG101_clean_1.QC





Compare quality control before / after cleaning

Solution 2 : Use Galaxy « Scratchbook » to manage Galaxy windows



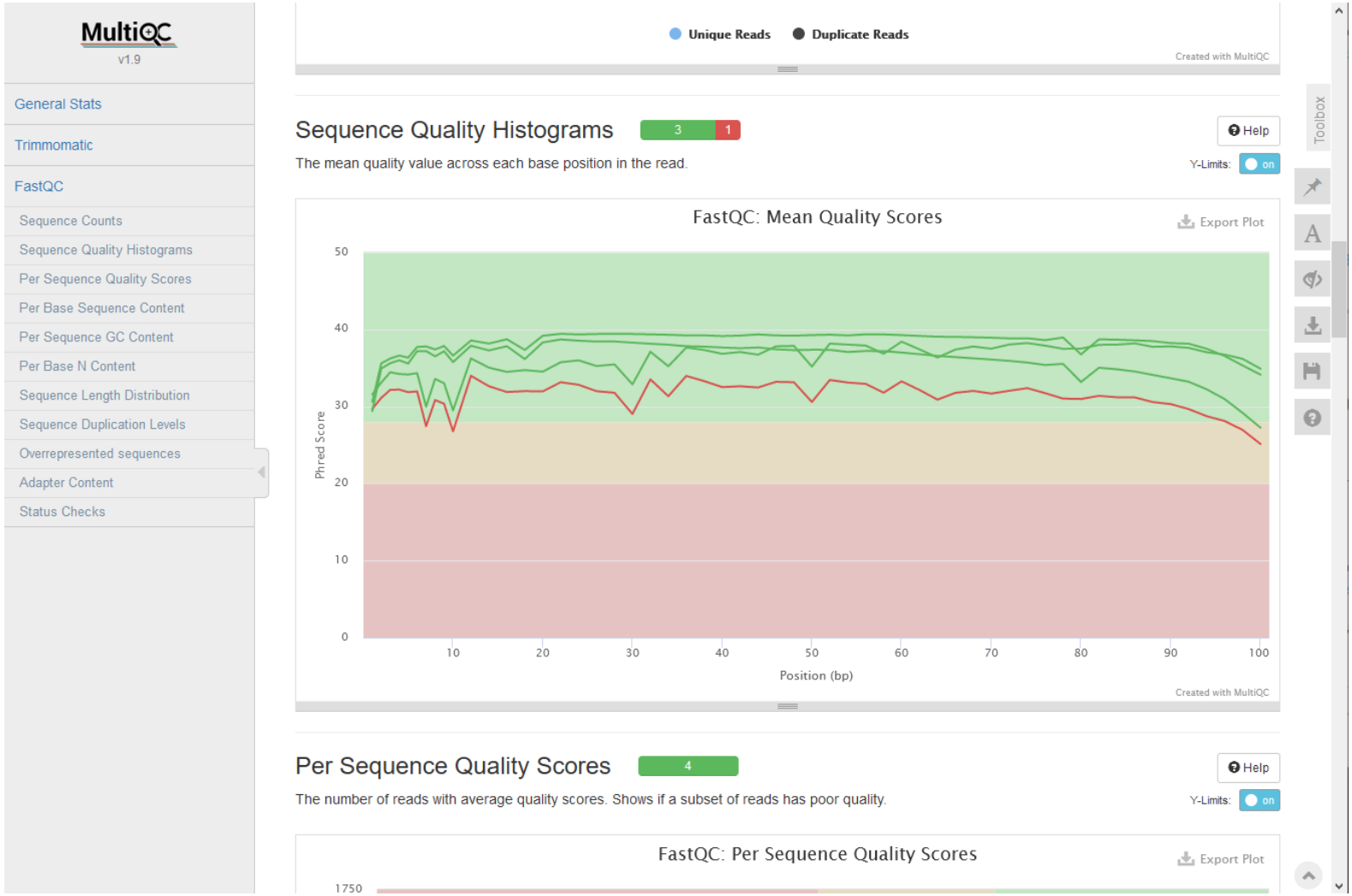
1 « Enable Scratchbook »

2 Visualize dataset HG0101_clean_1.fastq

3 Visualize dataset HG0101_OK_1.fastq

Compare quality control before / after cleaning

Solution 3 : Use MultiQC



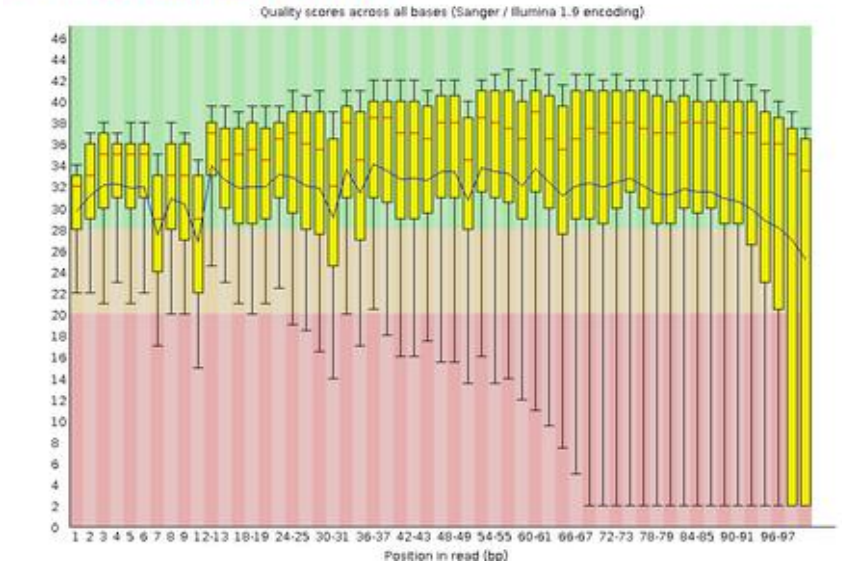
Quality control before & after cleaning

HG0101_OK_1.QC

Basic Statistics

Measure	Value
Filename	HG0101_OK_1.fastq
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	5283
Sequences flagged as poor quality	0
Sequence length	101
%GC	43

Per base sequence quality

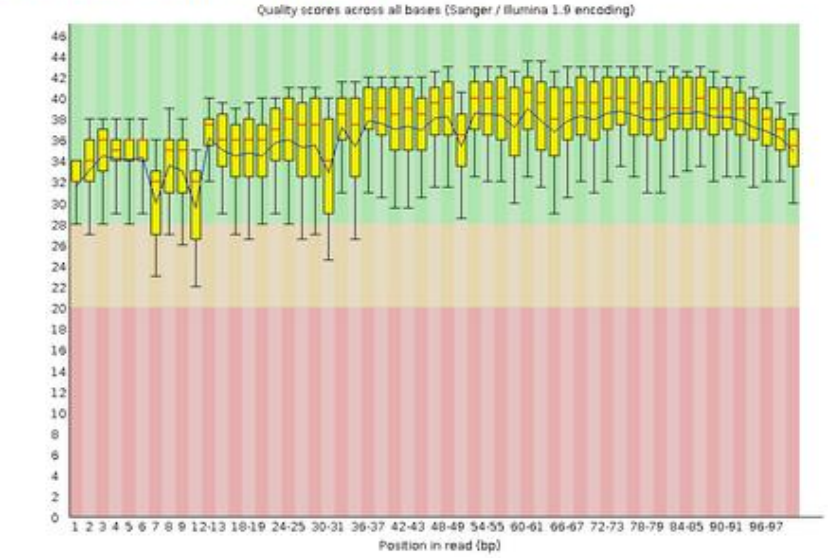


HG0101_clean_1.QC

Basic Statistics

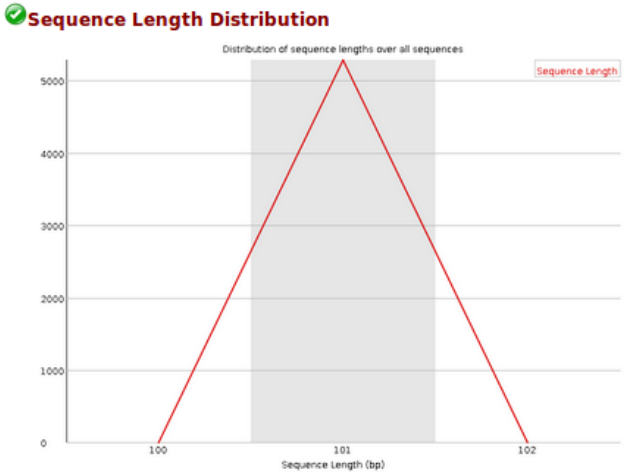
Measure	Value
Filename	HG0101_clean_1.fastq
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	4060
Sequences flagged as poor quality	0
Sequence length	20-101
%GC	42

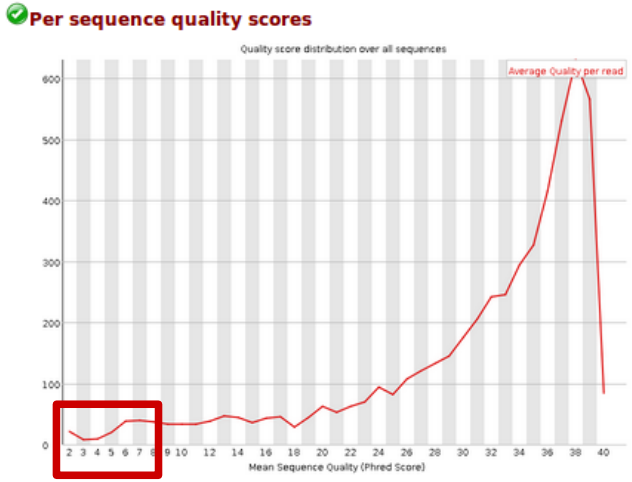
Per base sequence quality



Quality control before & after cleaning

HG0101_OK_1.QC





HG0101_clean_1.QC