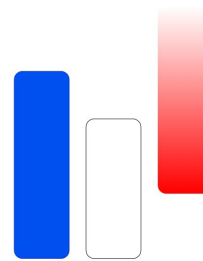


Module 1/5: Analyses ADN

- NGS Introduction
- Reads Quality Control
- Reads Cleaning

→ **Practical #1**



usegalaxy.fr

Galaxy

- Connection
- TIAAS: <https://usegalaxy.fr/join-training/bilille-2022-dnaseq>
- 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*)

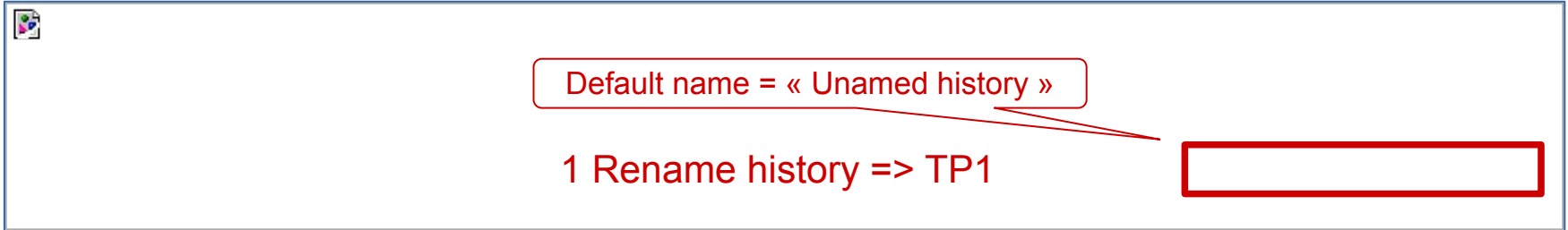
Download files on bilille wiki : <https://wikis.univ-lille.fr/bilille/formation>

Main goals for this first part of tutorial

- Upload reference and reads for one sample (HG0101)

- Work with histories, datasets and tools

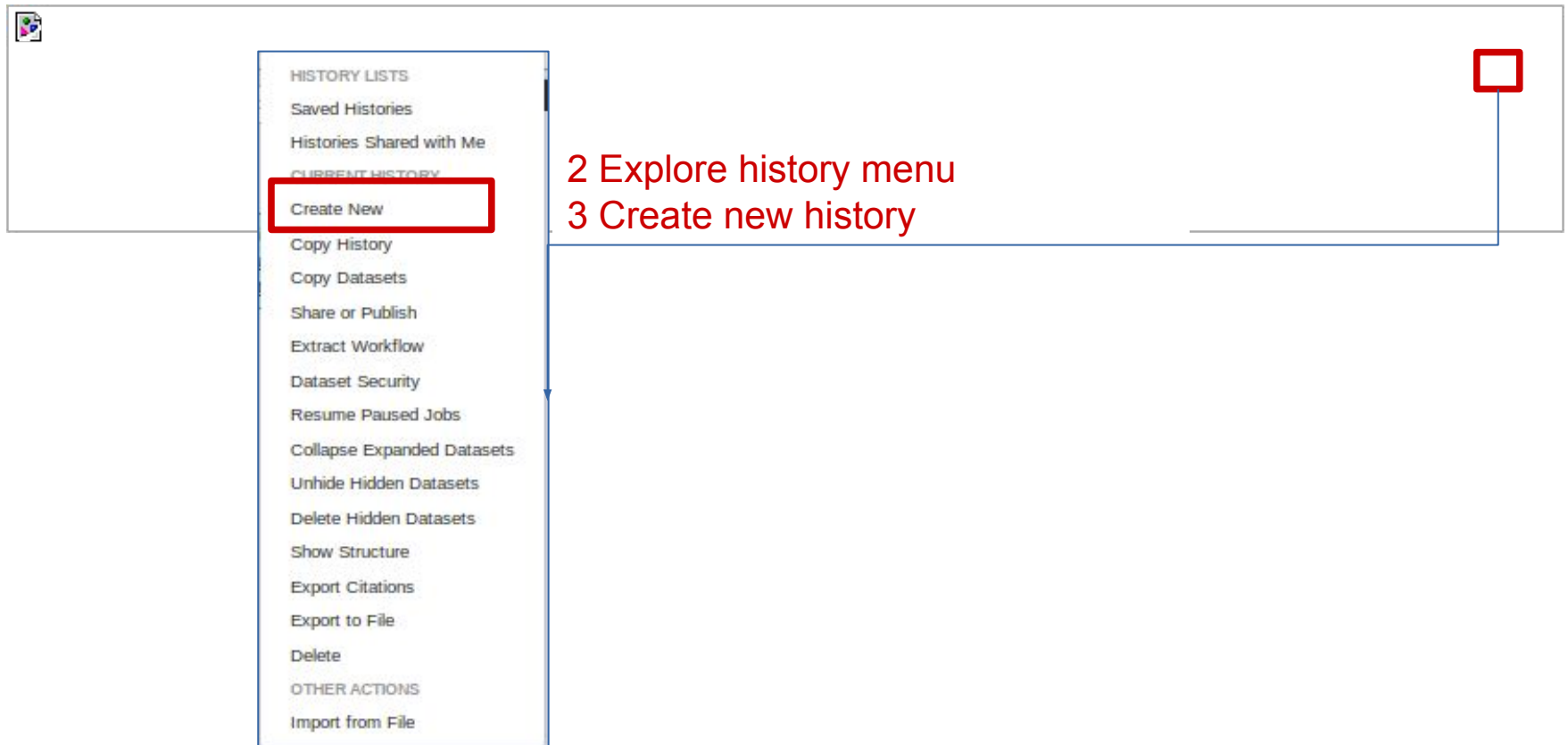
History : « Folder » containing a set of data



Default name = « Unnamed history »

1 Rename history => TP1

[Red box]



HISTORY LISTS

- Saved Histories
- Histories Shared with Me

CURRENT HISTORY

- Create New**
- Copy History
- Copy Datasets
- Share or Publish
- Extract Workflow
- Dataset Security
- Resume Paused Jobs
- Collapse Expanded Datasets
- Unhide Hidden Datasets
- Delete Hidden Datasets
- Show Structure
- Export Citations
- Export to File
- Delete

OTHER ACTIONS

- Import from File

2 Explore history menu

3 Create new history

[Red box]

List histories, go back to TP1

1 List all histories

The screenshot shows the Galaxy web interface. The main content area is titled 'Saved Histories' and contains a table with the following data:

Name	Datasets	Tags	Sharing	Size on Disk	Created	Last Updated	Status
Unnamed history	0 Datasets	0 Tags		0 bytes	~4 seconds ago	~4 seconds ago	current history
TP1	0 Datasets	0 Tags		0 bytes	Nov 17, 2016	~49 seconds ago	

A context menu is open over the 'TP1' history, with the following options: Switch, View, Share or Publish, Copy, Rename, Delete, and Delete Permanently. The 'Switch' option is highlighted. A red box highlights the 'TP1' dropdown in the table. A red box in the top right of the screenshot points to a 'History Lists' sidebar on the right.

2 Go back to TP1 history

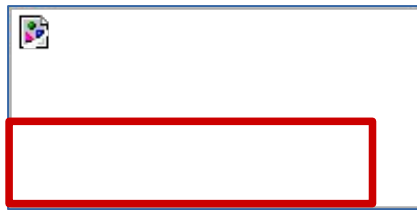
The 'History Lists' sidebar contains the following items:

- HISTORY LISTS
- Saved Histories**
- Histories Shared with Me
- CURRENT HISTORY
- Create New
- Copy History
- Copy Datasets
- Share or Publish
- Extract Workflow
- Dataset Security
- Resume Paused Jobs
- Collapse Expanded Datasets
- Unhide Hidden Datasets
- Delete Hidden Datasets
- Show Structure
- Export Citations
- Export to File
- Delete
- OTHER ACTIONS
- Import from File

Dataset ~ « Data file »

Upload reference in a *dataset*

1 Tools
Get Data / Upload File



Download from web or upload from disk

Regular Composite

Name	Size	Type	Genome	Settings	Status
GRCh37_region1.fasta	149 KB	fasta	unspecified (?)		100% ✓

2 Choose file

3 fasta

4 unspecified

5 start

6 close

Type (set all): Auto-detect

Genome (set all): unspecified (?)

Choose local file Choose FTP file Paste/Fetch data Pause Reset Start Close

History

search datasets

TP1

149.0 KB

1: GRCh37_region1.fasta

- 2 Choose file GRCh37_region1.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

The screenshot shows the Galaxy interface. On the left, the 'History' panel lists a dataset 'TP1' (149.0 KB) with a red box around the dataset name '1: GRCh37_region1.fasta' and another red box around the eye and pencil icons. The main panel shows a code viewer with a DNA sequence: `>chr20 20:380000-530000
CAGACAGGGACAAGTCAGTCTCATCTCTGTGCACCCAGCATTGCCAGAACAGGGCCCTAG
TTGTGTCTAGGGTCTCATGGGGCAGCCCTGACCTCTATCTTGCCCTCCAGCTTCCAA
CAGACCTGTCCAGCTCCCTCCAAGCTGAGTGTGGCCTGATACCTACCAGTGGAGCGA
GGGGAACACGAGGACTGCCAAGGGCAGGTACCGTGCCAAACCCTTCACTCCATTCACAAA
GAGACTCATTTACTCTCATGACAATCAGTGAGGCAGATGTTCTGTCACTTTACAGAT
AAGACAAATGAGCTTTAGAGAGAGCAGAAGACTCATCCAAGACCTCAACACAGAGAAAA
GACTCCTCTGGTTTTTGCCTTGGAGCAGGAACCTTTGACAAGGCTGCCAGATGCAGCCA
CATGAAAACCGCCATTAAAAATGTAATATCGGCCCGGAGCGGGGGCTCACGCTGTAAT
CCCAGCACTTTGGGAGGCCGAGGCAGGCAGATCACTTGAGGTCAGGAGTTGAAGACAGC
CTGGCCAACATGGTGAACCCCGTCTCTACTAAAAATACAAAAATAGCCAGGCGTGGT
GCATGGGCTGTAATCCAGCTACTCGGGAGGCTGAGGCAGGAGAATCACTTGAACCCAG`. On the right, the 'History' panel shows the dataset 'TP1' (148.95 KB) with a red box around the dataset name '1: GRCh37_region1.fasta' and another red box around the eye, pencil, and delete icons.

The screenshot shows the 'Summary' view of the dataset '1: GRCh37_region1.fasta'. It displays '1 sequences' in 'fasta' format, uploaded from a 'fasta file'. The 'History' panel on the left shows 'TP1' (148.96 KB) with a red box around the dataset name '1: GRCh37_region1.fasta' and another red box around the eye, pencil, and delete icons. The main panel shows the dataset details and a code viewer with a DNA sequence: `>chr20 20:389999-530000
CAGACAGGGACAAGTCAGTCTCATCTCTGTGCACCCAGCA
TTGTGTCTAGGGTCTCATGGGGCAGCCCTGACCTCTATC
CAGACCTGTCCAGCTCCCTCCAAGCTGAGTGTGGCCT
GGGGAACACGAGGACTGCCAAGGGCAGGTACCGTCCAAAC`

The screenshot shows the 'Attributes' and 'Convert Format' panels. The 'Attributes' panel has tabs for 'Attributes', 'Convert Format', 'Datatype', and 'Permissions'. The 'Edit Attributes' section shows the dataset name 'GRCh37_region1.fasta', the info 'uploaded fasta file', and the 'Database/Build' set to 'unspecified (?)'. The 'Convert Format' panel has tabs for 'Attributes', 'Convert Format', 'Datatype', and 'Permissions'. The 'Convert to new format' section shows a dropdown menu set to 'Convert FASTA to Tabular' and a 'Convert' button. The 'Change data type' section shows a dropdown menu set to 'fasta' and a 'Save' button.

Add a local reference (TP_ref)



1 Menu User / Custom Builds

Add a Custom Build

New Build

Name (eg: Hamster):
TP_ref

Key (eg: hamster_v1):
TP_ref

Definition:
FASTA

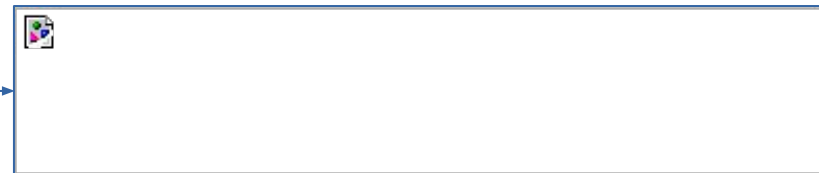
1: GRCh37_region1.fa

2 Choose name TP_ref

3 Choose *fasta* format

4 Choose *dataset n° 1* : GRCh37_region1.fasta

5 Submit



Reference is now available

Check / Change *database* attribute

1 Analyse Data

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 »

2

3

4

5

6

History

search datasets

TP1

149.0 KB

1: GRCh37_region1.fasta

History

search datasets

TP1

1 shown

148.95 KB

1: GRCh37_region1.fasta

1 sequences
format: fasta, database: ?

uploaded fasta file

Attributes

Convert Format

Datatype

Permissions

Edit Attributes

Name:
GRCh37_region1.fasta

Info:
uploaded fasta file

Annotation / Notes:

Database/Build:
TP_ref (TP_ref) [Custom]

Save

Auto-detect

History

search datasets

TP1

1 shown, 1 deleted, 2 hidden

148.97 KB

1: GRCh37_region1.fasta

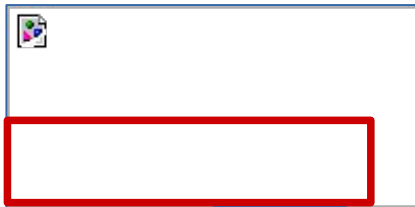
1 sequences
format: fasta, database: TP_ref

uploaded fasta file

>chr20 20:380000-530000
CAGACAGGGACAAGTCAGTCTCATCTCTGTGCACCCA
TTGTGTCTAGGGTCTCATG666CAGCCCCGTGACCTC
CAGACCCCTGTCCCAGCTCCCTCCAAGCTGAGTGTG
GGGGAACACGAGGACTGCCAAGGGCAGGTACCGTGC

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



Download from web or upload from disk

Regular Composite

You added 2 file(s) to the queue. Add more files or click 'Start' to proceed.

Name	Size	Type	Genome	Settings	Status
HG00101_2.fastq	1.1 MB	fastq	TP_ref		
HG00101_1.fastq	1.1 MB	fastq	TP_ref		

Type (set all): Auto-detect Genome (set all): unspecified (?)

Choose local file Choose FTP file Paste/Fetch data Pause Reset Start Close

3 fastq 4 TP_ref 5 start 6 close 2 choose local file

History

search datasets

TP1
3 shown, 2 hidden
2.42 MB

3: HG00101_2.fastq
1.1 MB
format: fastq, database: TP_ref
uploaded fastq file

@ERR229776.100000000
CTAGGAAGCGTAGTCTG6596TCATCTCTCTATTAATAC
+
BAEEAGEEDD0EHFE@BF><<EAAC;EBH7<K<0:HJGFF
@ERR229776.100000005

1: GRCh37_region1.fasta

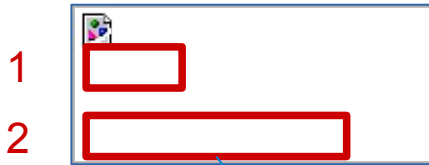
7

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° 2 & 3
 - 4 Execute ⇒ Create 4 new datasets
- For each fastq file :
1 « raw data » and 1 « Webpage »

3

FastQC Read Quality reports (Galaxy Version 0.67)

Short read data from your current history

- 5: HG0101_OK_2.fastq
- 4: HG0101_OK_1.fastq
- 3: HG00101_2.fastq
- 2: HG00101_1.fastq

Contaminant list

Submodule and Limit specifying file

Execute

4

History

search datasets

TP1

9 shown, 2 deleted, hide hidden

6.84 MB

- 9: FastQC on data 5: RawData
- 8: FastQC on data 5: Webpage
- 7: FastQC on data 4: RawData
- 6: FastQC on data 4: Webpage
- 5: HG0101_OK_2.fastq
- 4: HG0101_OK_1.fastq
- 3: HG00101_2.fastq
- 2: HG00101_1.fastq
- 1: GRCh37_region1.fasta

Manage FastQC result *datasets*

1

2

```
##FastQC 0.11.5
>>Basic Statistics pass
#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
%END_MODULE
>>Per base sequence quality fail
#Base Mean Median Lower Quartile Upper Quartile 10th Percentile 90th
1 29.691084611016468 32.0 28.0 33.0 22.0 34.0
2 31.118114707552525 33.0 29.0 36.0 22.0 37.0
3 32.15332197614992 35.0 30.0 37.0 21.0 38.0
4 32.19080068143101 35.0 31.0 36.0 23.0 37.0
5 31.8536816202915 35.0 30.0 36.0 21.0 38.0
6 31.951164111300397 35.0 31.0 36.0 22.0 38.0
7 27.434033692977476 29.0 24.0 33.0 17.0 35.0
8 30.830967253454478 33.0 28.0 36.0 20.0 38.0
9 30.349001249290177 33.0 27.0 36.0 20.0 37.0
10-11 26.774654552337687 29.0 22.0 33.0 15.0 34.5
```

1

3

FastQC Report

Tue 24 Jan 2017
HG0101_OK_1.fastq

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
- Kmer 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

Per base sequence quality

Quality scores across all bases (Sanger / Ill)

46

History

search datasets

TP1
7 shown, 4 deleted, hide hidden

6.84 MB

- 8: HG0101_2.QC
- 6: HG0101_1.QC
- 5: HG0101_OK_2.fastq
- 4: HG0101_OK_1.fastq
- 3: HG00101_2.fastq
- 2: HG00101_1.fastq
- 1: GRCh37_region1.fasta










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

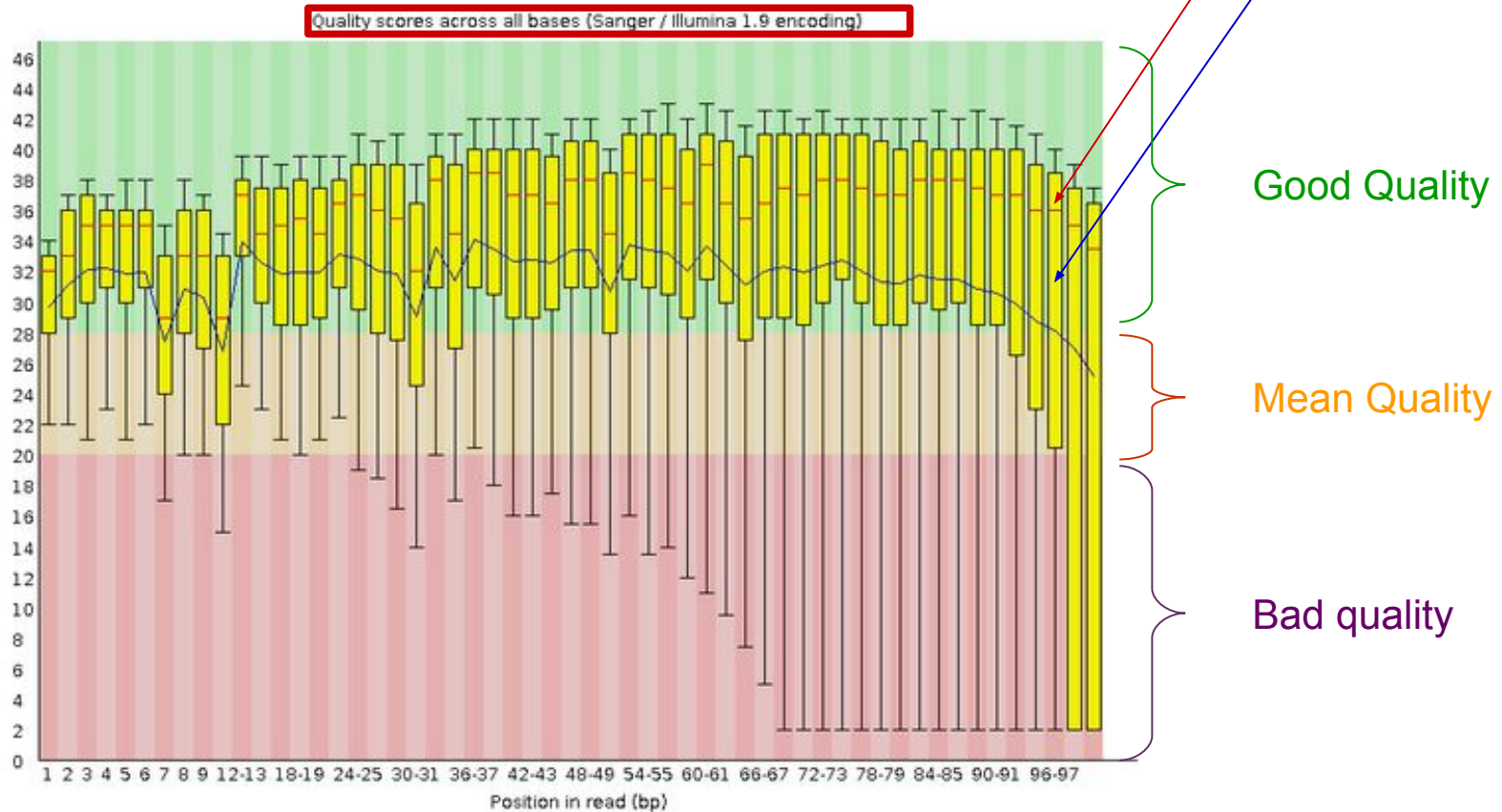
-  [Basic Statistics](#)
-  [Per base sequence quality](#)
-  [Per sequence quality scores](#)
-  [Per base sequence content](#)
-  [Per sequence GC content](#)
-  [Per base N content](#)
-  [Sequence Length Distribution](#)
-  [Sequence Duplication Levels](#)
-  [Overrepresented sequences](#)
-  [Adapter 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

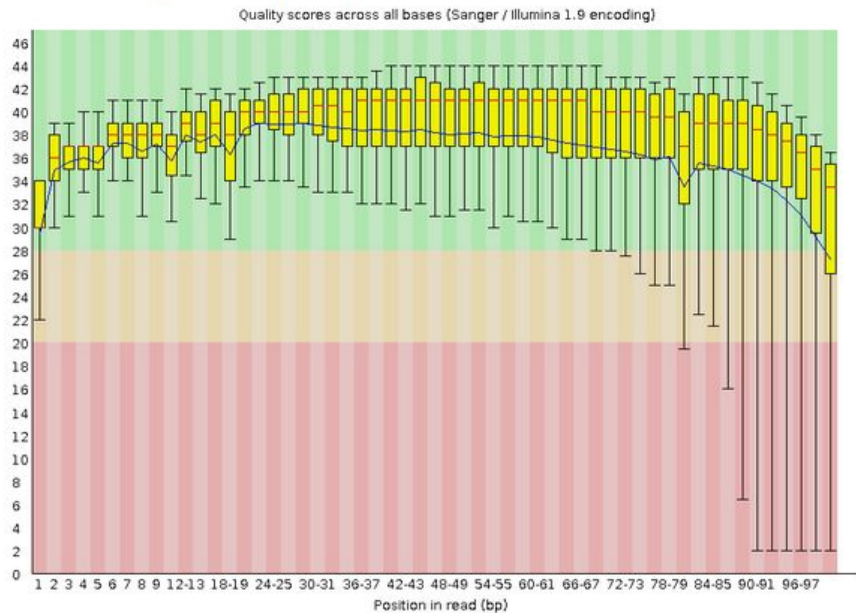
❌ Per base sequence quality



Fasqc : Per base sequence quality

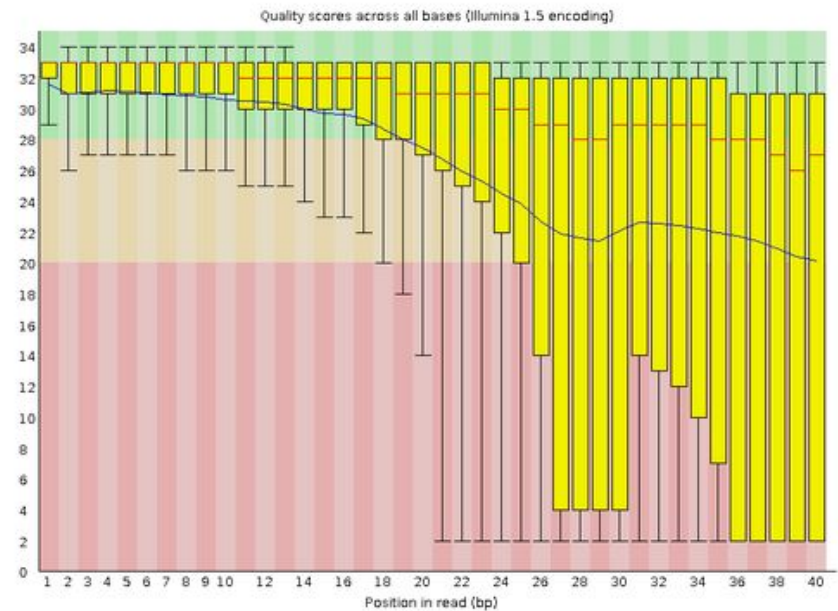
Example OK

✔ Per base sequence quality



Example KO

✘ Per base sequence quality



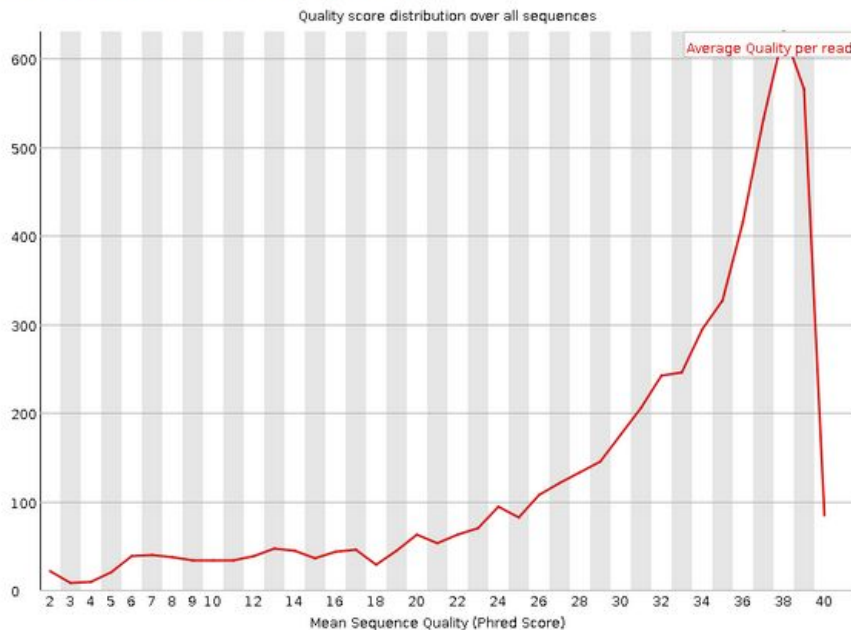
Source :

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

Fastqc : Per sequence quality score

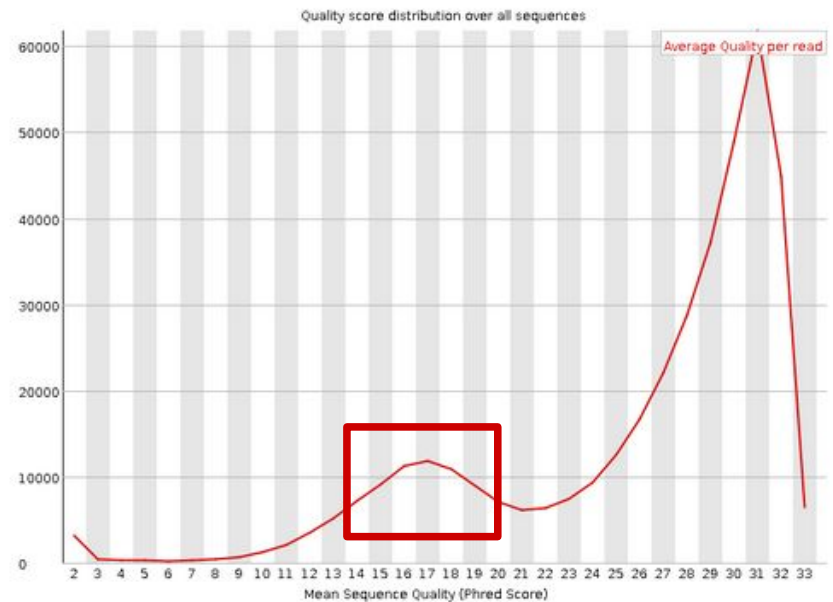
Example OK

✔ Per sequence quality scores



Example KO

✔ Per sequence quality scores

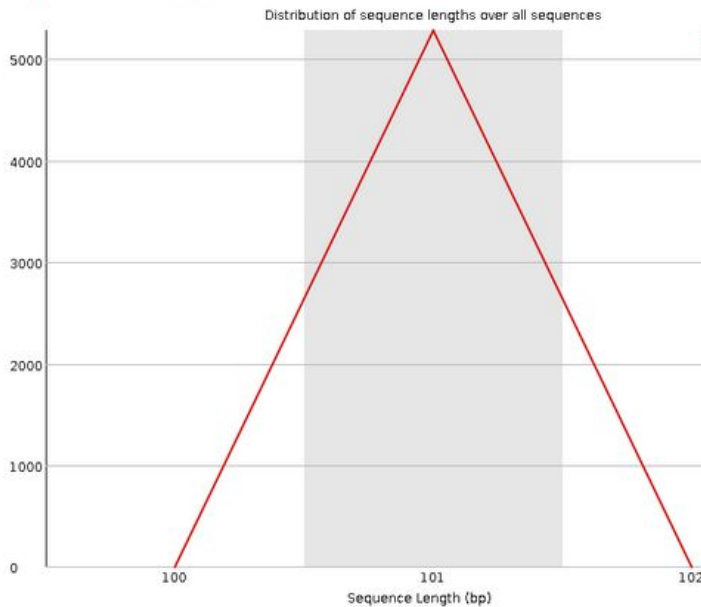


Source :

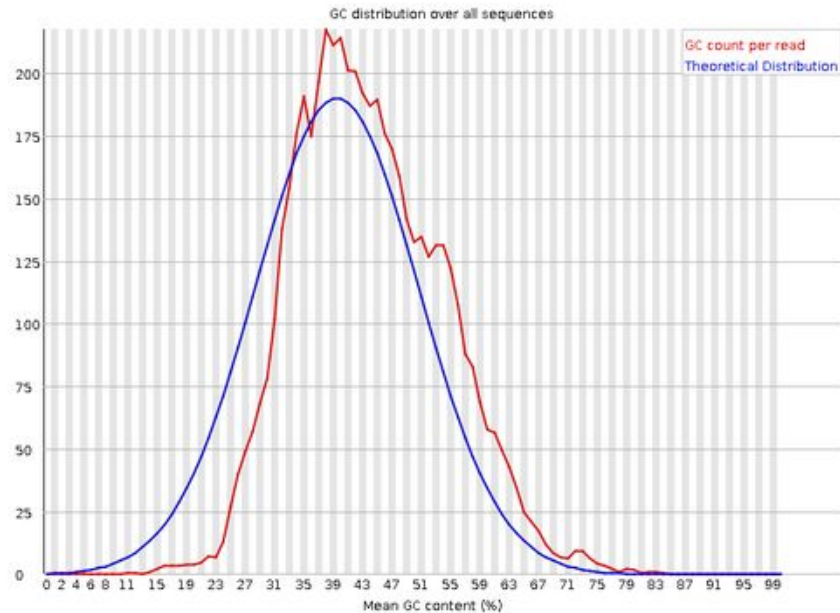
http://www.bioinformatics.babraham.ac.uk/projects/fastqc/bad_sequence_fastqc.html

FastQC : Sequence Length Distribution & Per sequence GC content

✔ Sequence Length Distribution



⚠ Per sequence GC content

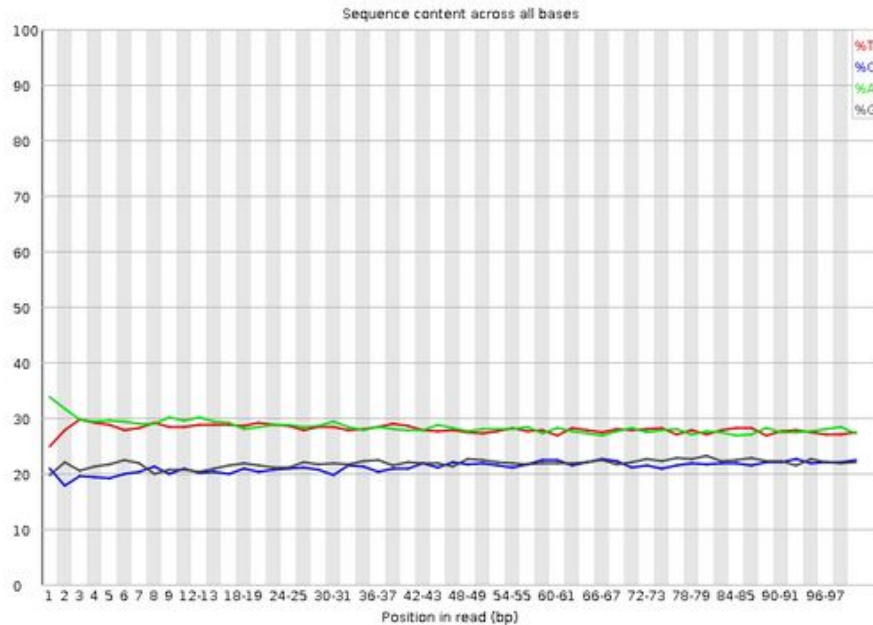


FastQC : Per base sequence content

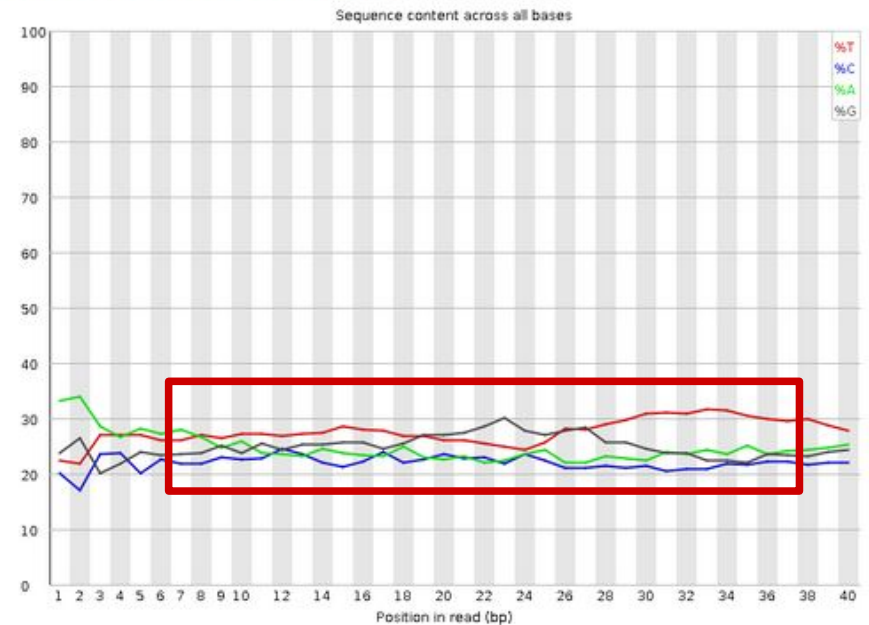
Example OK

Example KO

✔ Per base sequence content



⚠ Per base sequence content



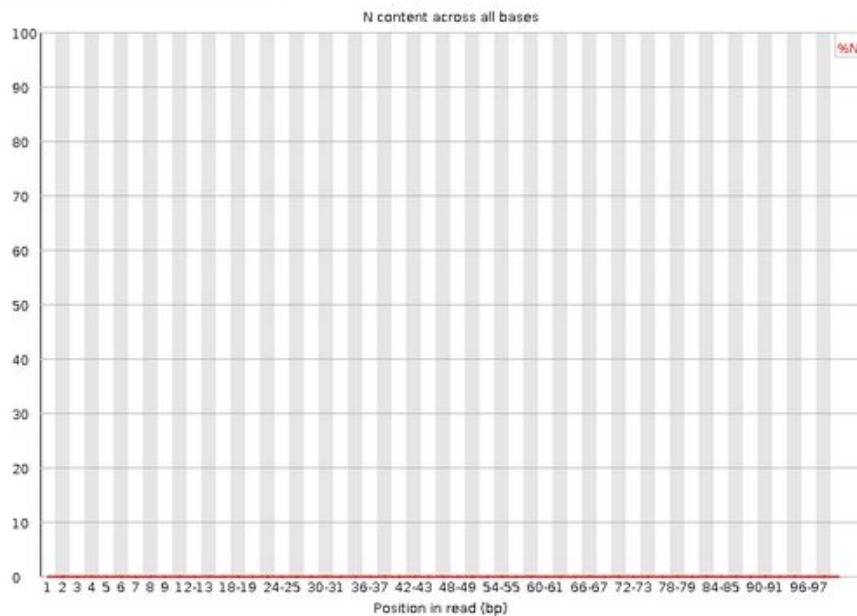
Source :

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

FastQC : Per base N content

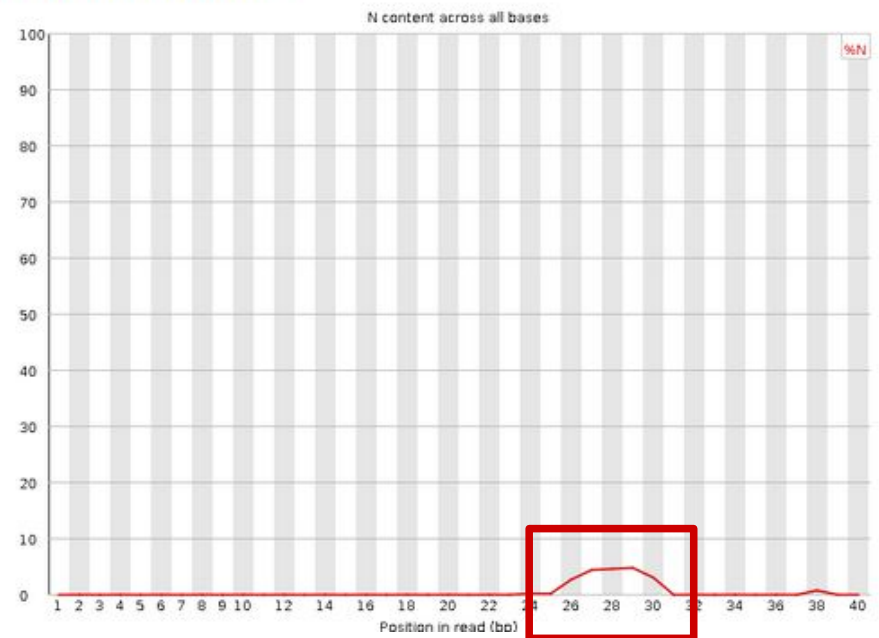
Example OK

✔ Per base N content



Example KO

✔ Per base N content



Source :

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

FastQC : Overrepresented sequences

Example OK

Example KO



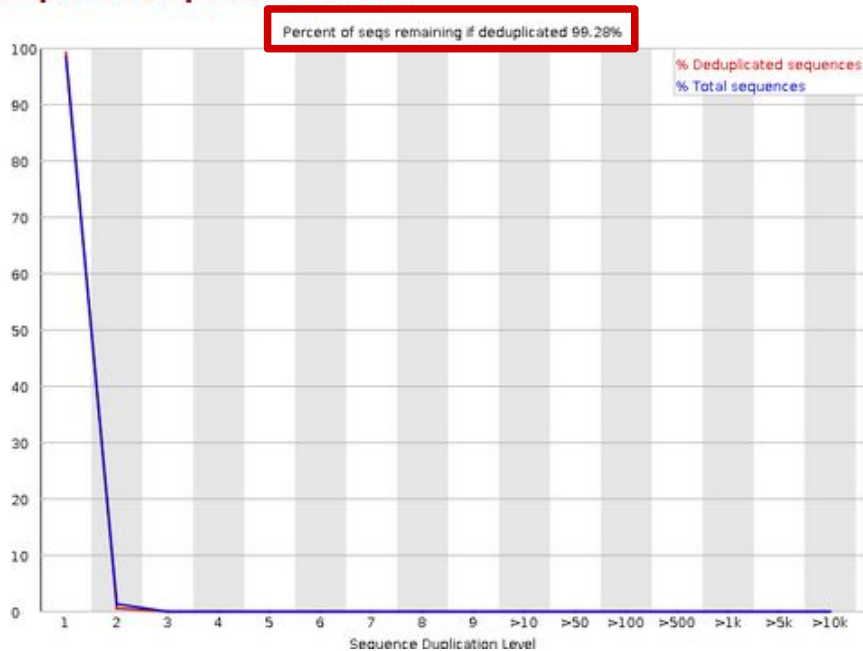
Source :
http://www.bioinformatics.babraham.ac.uk/projects/fastqc/bad_sequence_fastqc.html

FastQC : Sequence Duplication Levels

Example OK



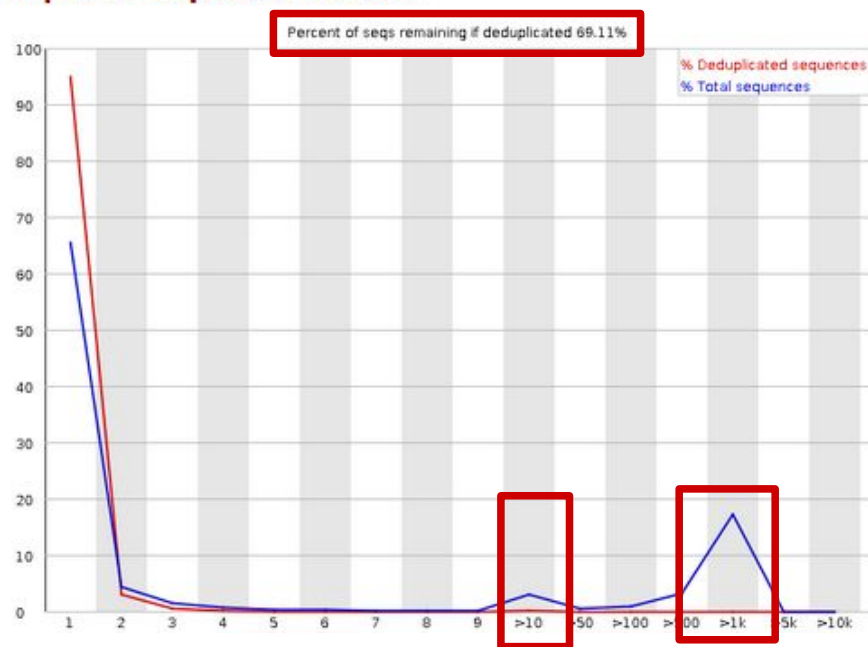
Sequence Duplication Levels



Example KO



Sequence Duplication Levels



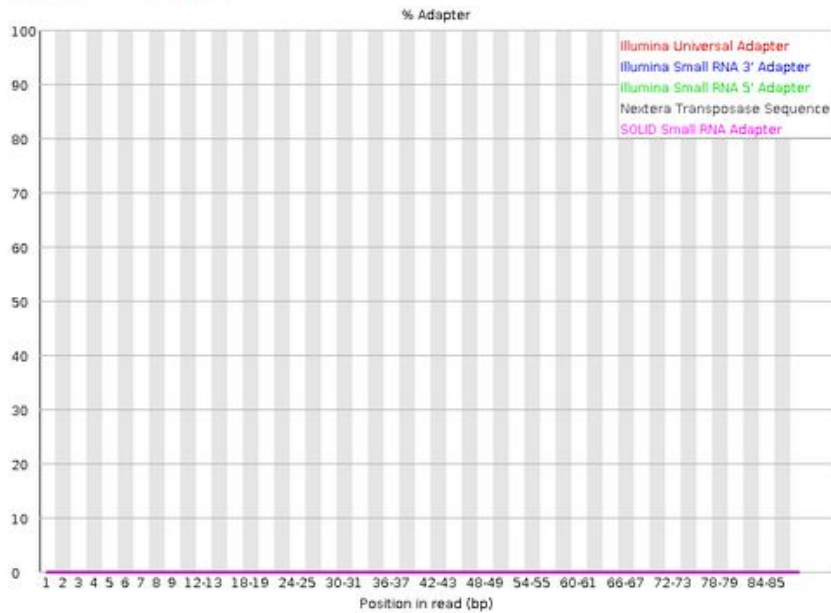
Source :

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

FastQC Adapter Content

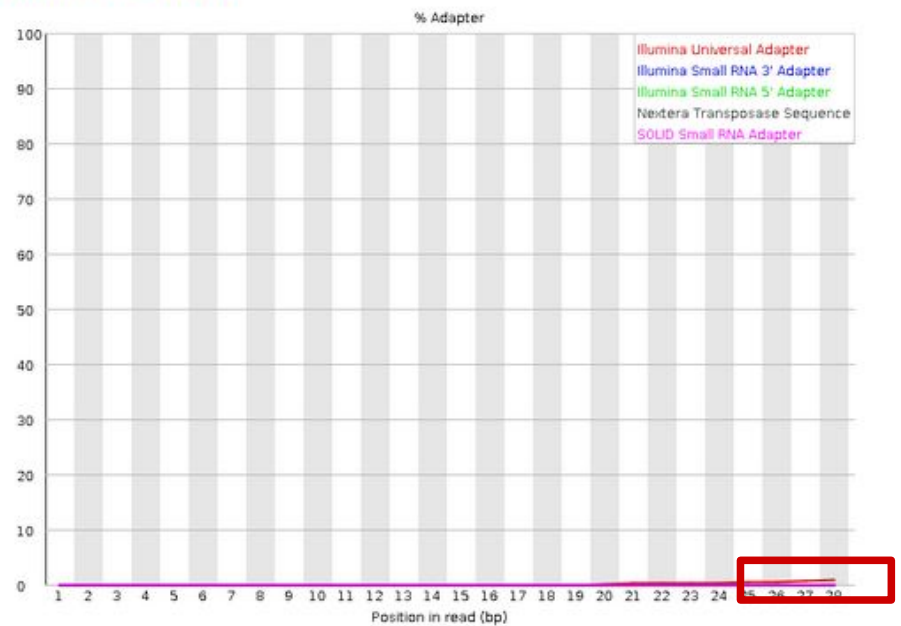
Example OK

Adapter Content



Example KO

Adapter Content

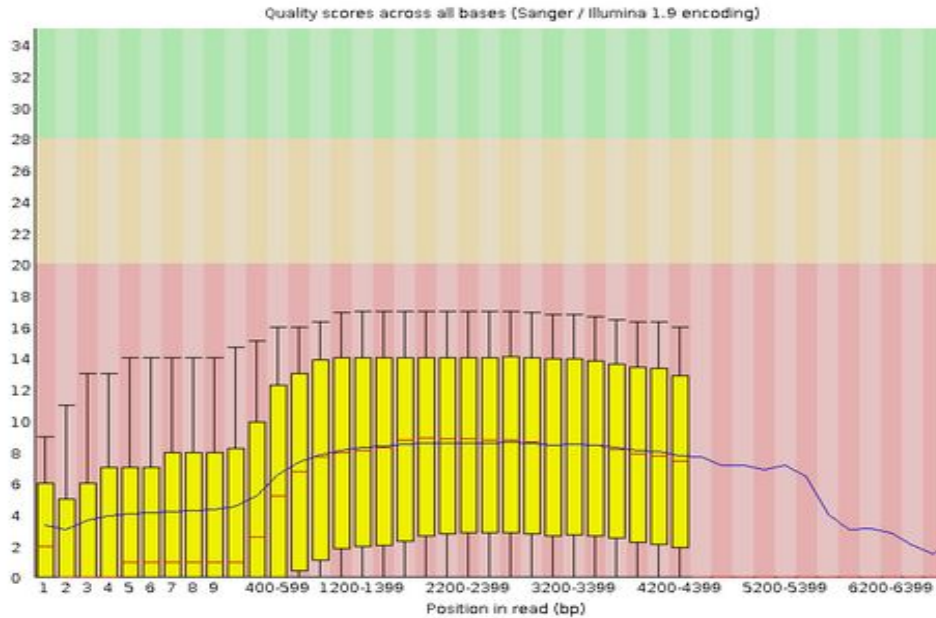


Source :

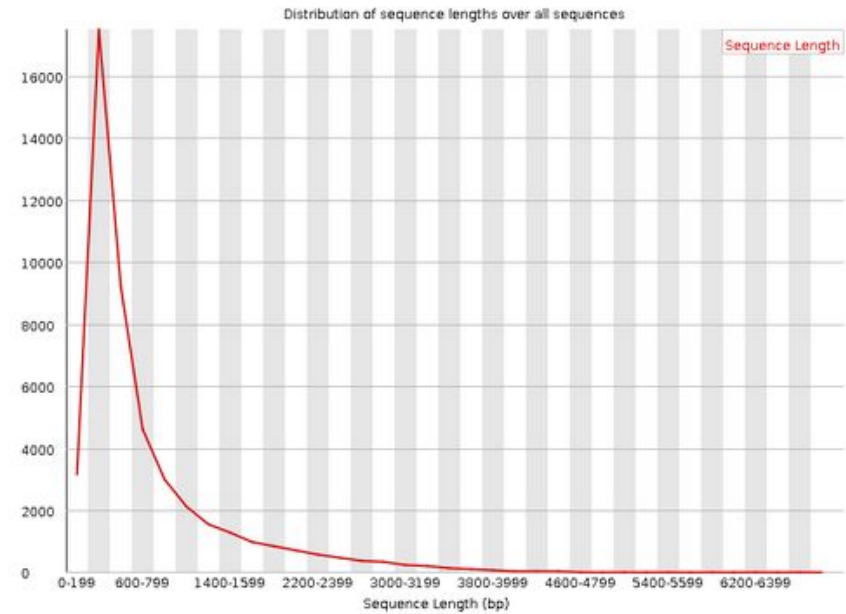
http://www.bioinformatics.babraham.ac.uk/projects/fastqc/bad_sequence_fastqc.html

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

- Filtering adapters
- Filtering & trimming reads
- Comparing quality before and after cleaning

Filtering & trimming

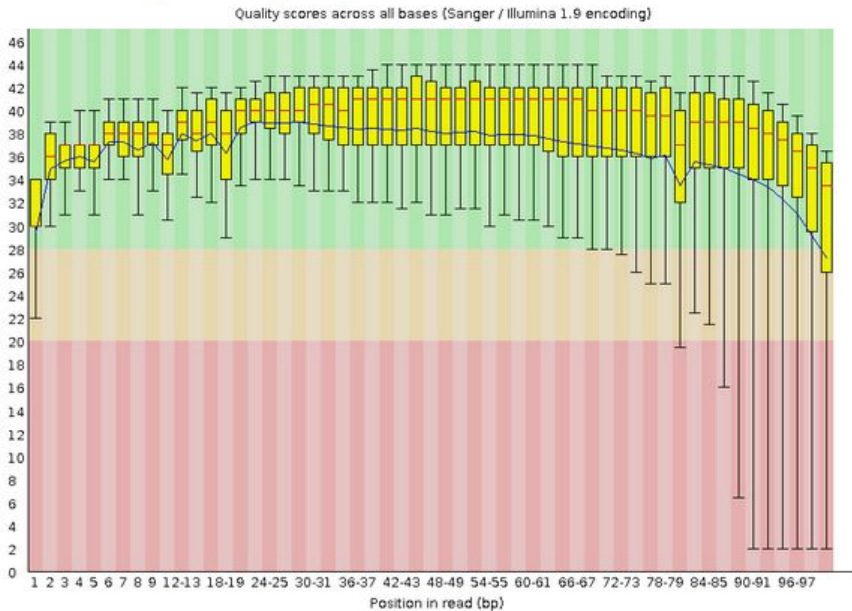
- Filtering = remove reads
 - Based on quality or size criteria
- Trimming = remove read ends
 - Fixed number of bases
 - Bases < quality

Trimming

Cut bad quality bases at the end of reads

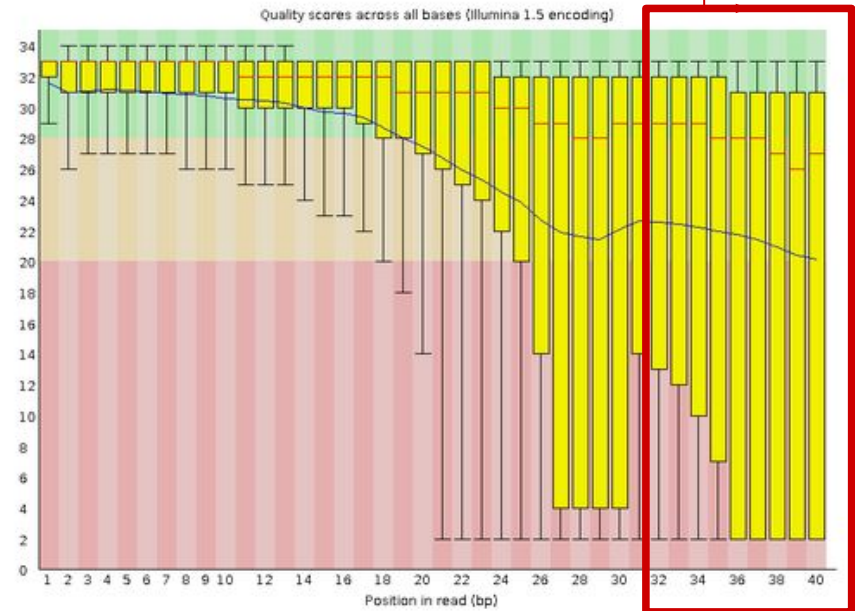
Exemple OK

✔ Per base sequence quality



Exemple KO

✘ Per base sequence quality



Source :

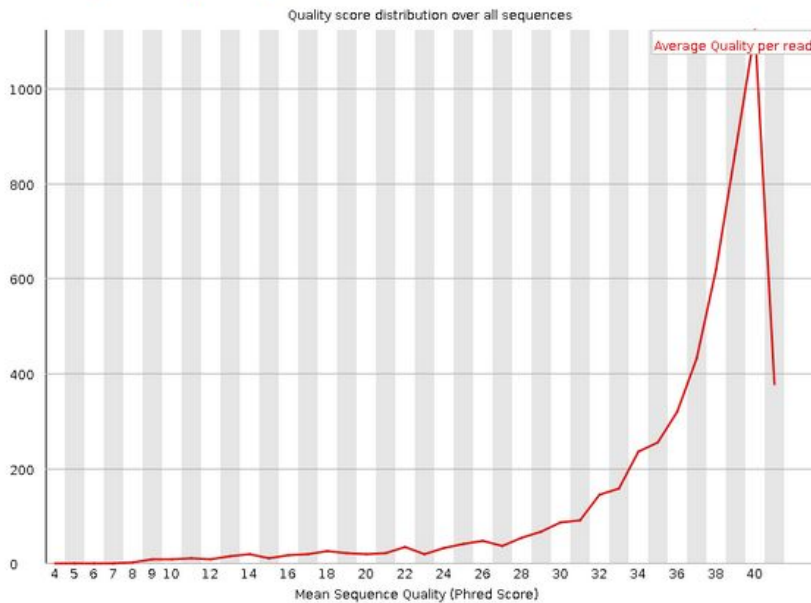
http://www.bioinformatics.babraham.ac.uk/projects/fastqc/bad_sequence_fastqc.html

Filtering

Remove reads with bad mean quality

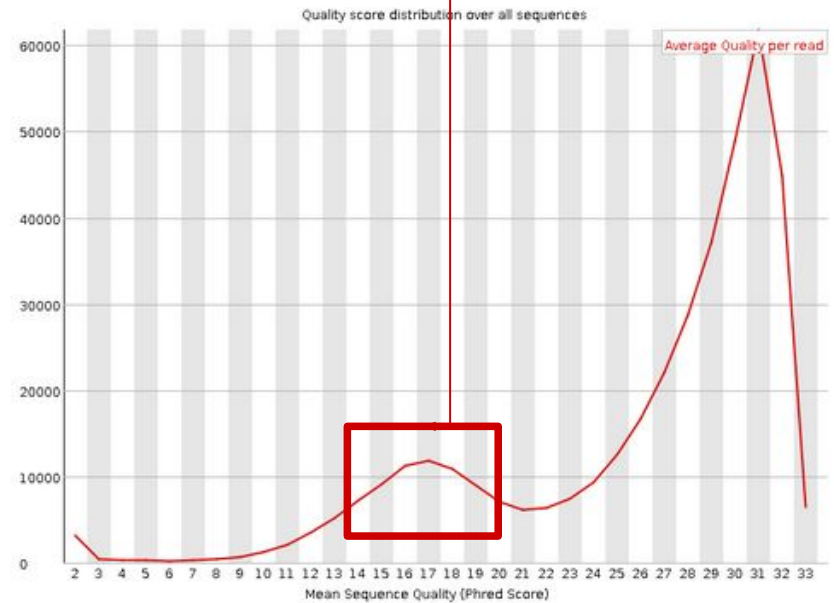
Exemple OK

✔ Per sequence quality scores



Exemple KO

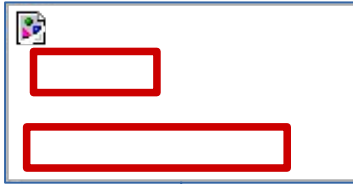
✔ Per sequence quality scores



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

The screenshot shows the Trimmomatic Galaxy interface. The following elements are highlighted with red boxes:

- Paired end data?**: The **Yes** radio button is selected.
- Input Type**: The dropdown menu is set to **Pair of datasets**.
- Input FASTQ file (R1/first of pair)**: The file path **4:HG0101_OK_1.fastq** is entered.
- Input FASTQ file (R2/second of pair)**: The file path **5:HG0101_OK_2.fastq** is entered.
- Perform initial ILLUMINACLIP step?**: The **Yes** radio button is selected.
- Adapter sequences to use**: The dropdown menu is set to **TruSeq3 (paired-ended, for MiSeq and HiSeq)**.

1 Choose files

2 Parameters for adapters

Reads cleaning (*Trimmomatic*) 2/2

+ Insert Trimmomatic Operation

Trimmomatic Operation

1: Trimmomatic Operation

Select Trimmomatic operation to perform
Cut bases off the start of a read, if below a threshold quality (LEADING)

Minimum quality required to keep a base
3
Bases at the start of the read with quality below the threshold will be removed

2: Trimmomatic Operation

Select Trimmomatic operation to perform
Cut bases off the end of a read, if below a threshold quality (TRAILING)

Minimum quality required to keep a base
3
Bases at the end of the read with quality below the threshold will be removed

3: Trimmomatic Operation

Select Trimmomatic operation to perform
Sliding window trimming (SLIDINGWINDOW)

Number of bases to average across
4

Average quality required
20

4: Trimmomatic Operation

Select Trimmomatic operation to perform
Drop reads with average quality lower than a specified level (AVGQUAL)

Minimum average quality required to keep a read
20

5: Trimmomatic Operation

Select Trimmomatic operation to perform
Drop reads below a specified length (MINLEN)

Minimum length of reads to be kept
20

Add operations (cleaning steps) :

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

2 : *TRAILING*: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



17: Trimmomatic on HG0101_OK_2.fastq (R2 unpaired)

16: Trimmomatic on HG0101_OK_1.fastq (R1 unpaired)

15: Trimmomatic on HG0101_OK_2.fastq (R2 paired)

14: Trimmomatic on HG0101_OK_1.fastq (R1 paired)

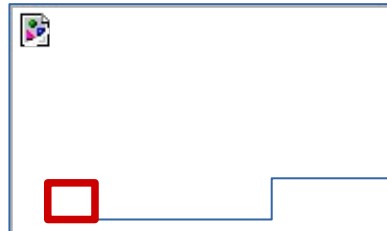
Unpaired reads2 (corresponding reads1 has been removed during cleaning)

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 ?



Tool: Trimmomatic	
Number:	14
Name:	Trimmomatic on HG0101_OK_1.fastq (R1 paired)
Created:	Wed 25 Jan 2017 09:55:35 AM (UTC)
Filesize:	798.8 KB
Dbkey:	TP_ref
Format:	fastqsanger
Galaxy Tool ID:	toolshed.g2.bx.psu.edu/repos/pjbriggs/trimmomatic/trimmomatic/0.36.1
Galaxy Tool Version:	0.36.1
Tool Version:	
Tool Standard Output:	stdout
Tool Standard Error:	stderr
Tool Exit Code:	0
History Content API ID:	a3bbc4b03bd5c402
Job API ID:	7e1ddb768ae0c642
History API ID:	24d84bcf64116fe7
UUID:	64538bfe-20ca-4c93-8738-29efc6041d44

From summary, click on « i » icon

Look content of « *stdout* »

This contains all messages send by this tool during execution.



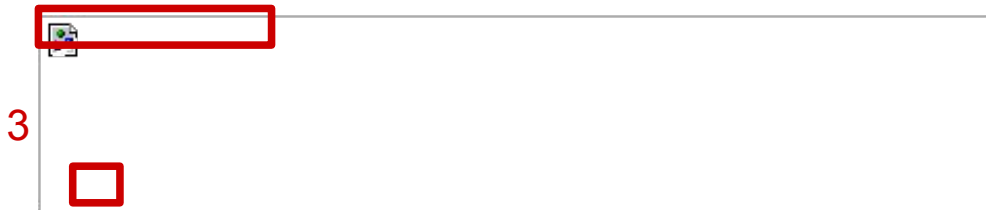
Trimmomatic : 2nd try

Run this tool again after changing AVGQUAL parameter value to 25



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

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



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 *datasets* HG0101_clean_1.fastq and HG0101_clean_2.fastq
- Run quality control on these 2 datasets.
- Compare quality control before and after cleaning

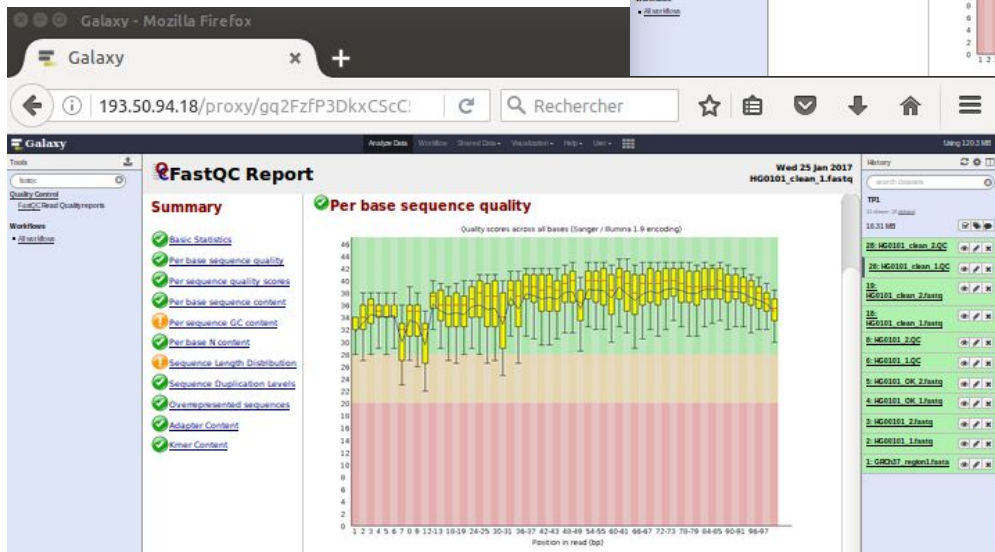
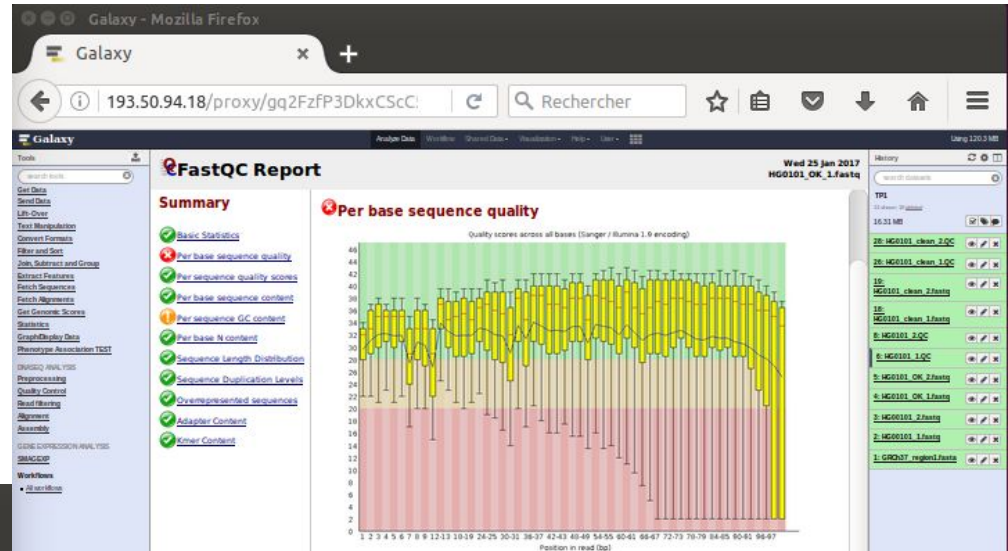
Compare quality control before / after cleaning

Solution 1 :

Open a second web browser window
Connect to access history and datasets

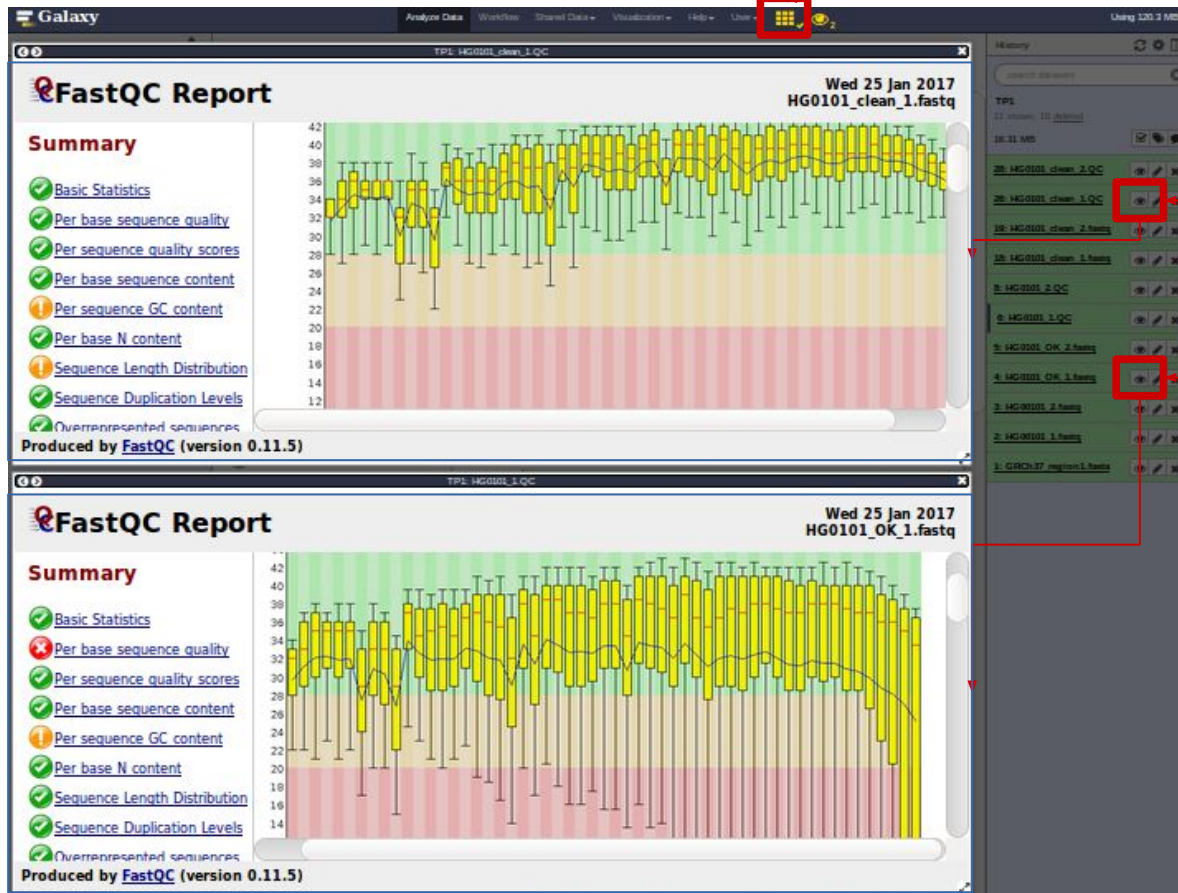
Visualize datasets

- HG101_OK_1.CQ
- HG101_clean_1.QC



Compare quality control before / after cleaning

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



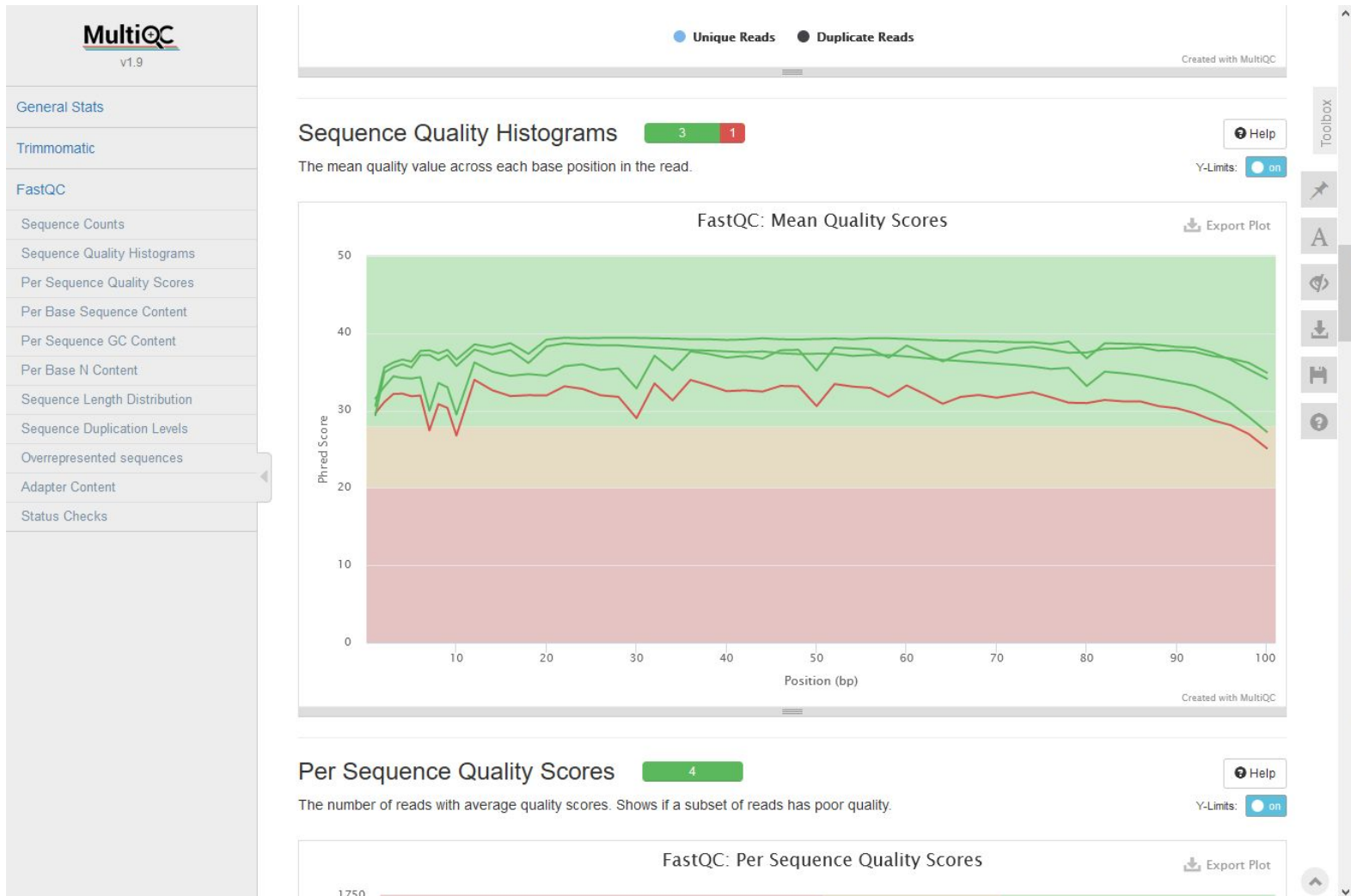
1 « Enable Scratchbook »

2 Visualize dataset
HG0101_clean_1.fastq

3 Visualize dataset
HG0101_OK_1.fast
q

Compare quality control before / after cleaning

Solution 3 : Use MultiQC



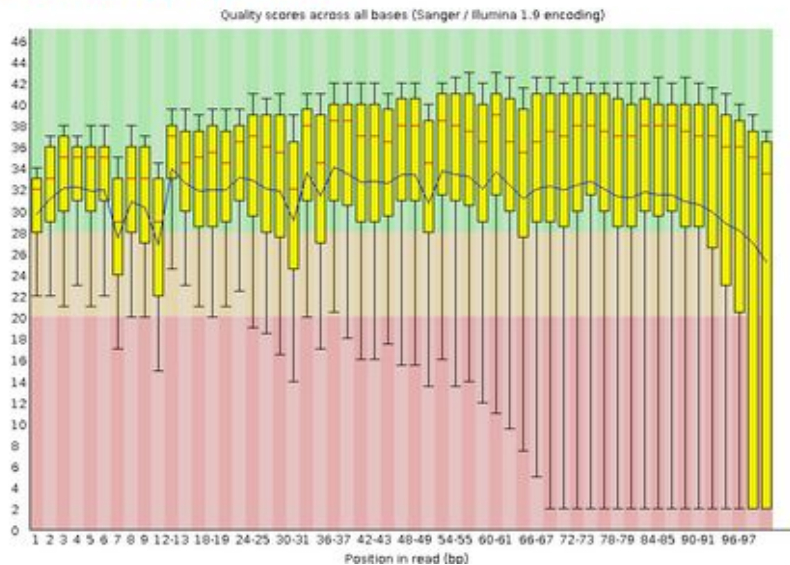
Quality control before & after cleaning

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

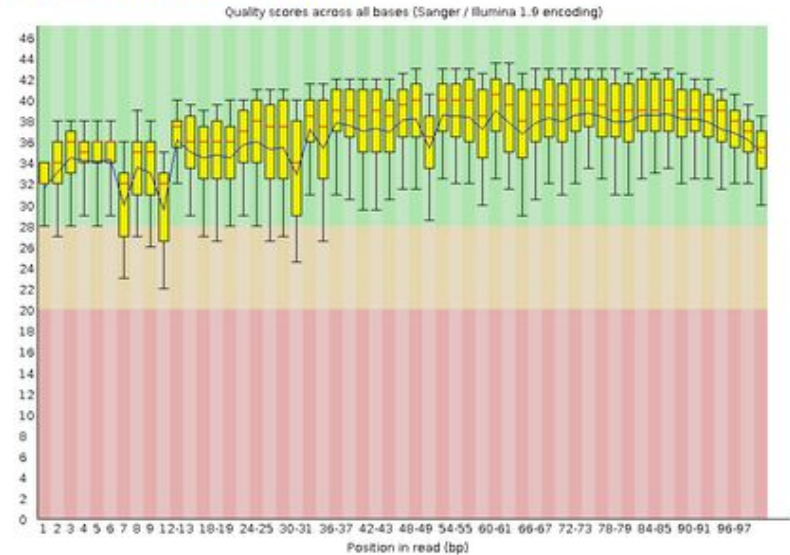


HG0101_clean_1.QC

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

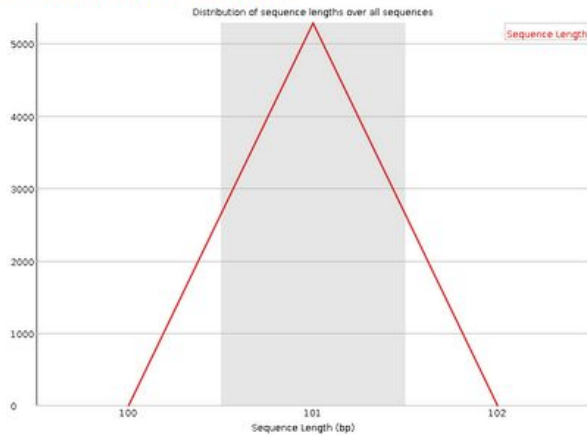


Quality control before & after cleaning

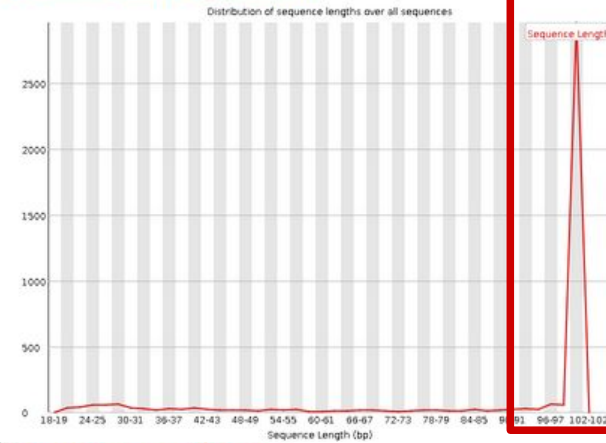
HG0101_OK_1.QC

HG0101_clean_1.QC

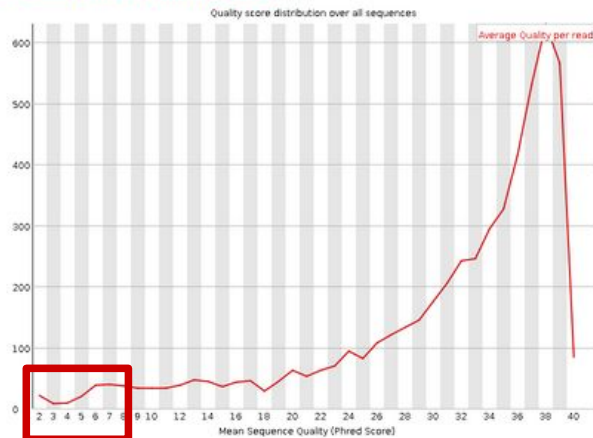
✔ Sequence Length Distribution



⚠ Sequence Length Distribution



✔ Per sequence quality scores



✔ Per sequence quality scores

