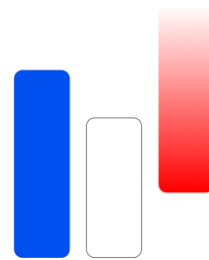


Module 1/6: Analyses ADN

- 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*)

Download 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

The screenshot shows the Galaxy web interface. The top navigation bar includes 'Galaxy', 'Analyze Data', 'Workflow', 'Shared Data', 'Visualization', 'Help', and 'User'. The right corner indicates 'Using 77.7 MB'. On the left, there is a 'Tools' panel with a search bar and 'Get Data'/'Send Data' buttons. The main workspace is empty. On the right, the 'History' panel is visible, containing a search bar and a list of datasets. A red box highlights the entry 'Unnamed history' with '0 bytes' and a 'Click to rename history' button. A red callout box points to this entry with the text 'Default name = « Unnamed history »'. Below the main workspace, red text reads '1 Rename history => TP1' with three vertical lines underneath.

Default name = « Unnamed history »

1 Rename history => TP1

The screenshot shows the Galaxy web interface with the 'History' menu open. The top navigation bar is the same as in the previous screenshot. The 'History' panel on the right now shows a search bar and a list with the entry 'TP1' and '0 bytes'. A red box highlights the gear icon in the top right of the 'History' panel. A red box also highlights the 'Create New' option in the 'CURRENT HISTORY' section of the menu. Red text in the center reads '2 Explore history menu' and '3 Create new history'. A blue arrow points from the 'Create New' option down to the main workspace.

2 Explore history menu
3 Create new history

List histories, go back to TP1

Galaxy Analyze Data Workflow Shared Data Visualization Help User Using 77.7 MB

Tools search tools

History search datasets View all histories

1 List all histories

Galaxy Analyze Data Workflow Shared Data Visualization Help User Using 77.7 MB

Tools search tools

Get Data Send Data Lift-Over Text Manipulation Convert Formats Filter and Sort Join, Subtract and Group Extract Features Fetch Sequences Fetch Alignments Get Genomic Scores Statistics Graph/Display Data Phenotype Association TEST DNASEQ ANALYSIS Quality Control

Saved Histories

search history names and tags

Advanced Search

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	

Switch View Share or Publish Copy Rename Delete Delete Permanently Undelete

2 Go back to TP1 history

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

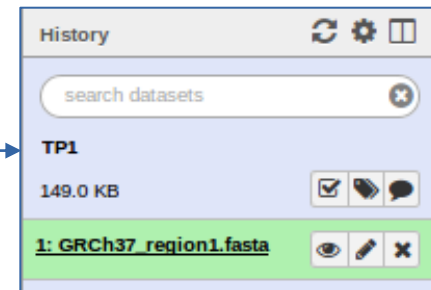
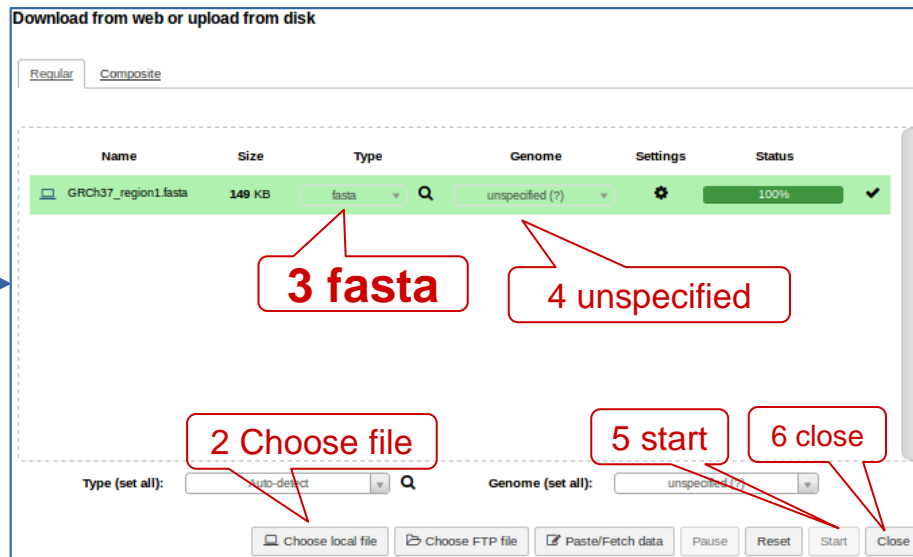
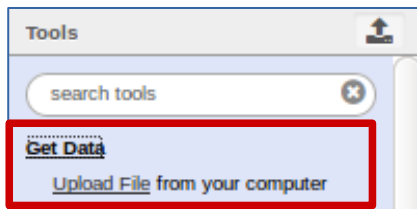
OTHER ACTIONS

Import from File

Dataset ~ « Data file »

Upload reference in a dataset

1 Tools
Get Data / Upload File



- 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

Galaxy interface showing dataset summary for **1: GRCh37_region1.fasta** (149.0 KB). The main panel displays a sequence of DNA bases: `>chr20 20:380000-530000` followed by a long string of nucleotide characters.

Galaxy interface showing dataset summary for **1: GRCh37_region1.fasta** (148.96 KB). It indicates **1 sequences** in **fasta** format. The main panel displays the same DNA sequence as the top screenshot.

Galaxy interface showing the **Edit Attributes** and **Convert Format** panels. The **Edit Attributes** panel shows the dataset name `GRCh37_region1.fasta` and options to **Save** or **Auto-detect**. The **Convert Format** panel shows a dropdown menu set to **Convert FASTA to Tabular** and a **Convert** button. The **Change data type** panel 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 [Len File](#) [Len Entry](#)

1: GRCh37_region1.fa

Submit

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

Current Custom Builds:

Name	Key	Number of chroms/contigs	
TP_ref	TP_ref	Processing	Delete

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 »

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

>chr20 20:380000-530000
CAGACAGGGACAAGTCAGTCTCATCTCTGTGCACCA
TTGTGTCTAGGGTCTCATG666CAGCCCCGACCTC
CAGACCCCTGTCCAGCTCCCTCCAAGCTGAGTGTG
GGGGAACACGAGGACTGCCAAGGGCAGGTACCGTGC

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

This will inspect the dataset and attempt to correct the above column values if they are not accurate.

History

search datasets

TP1

1 shown, 1 deleted, 2 hidden

148.97 KB

1: GRCh37_region1.fasta

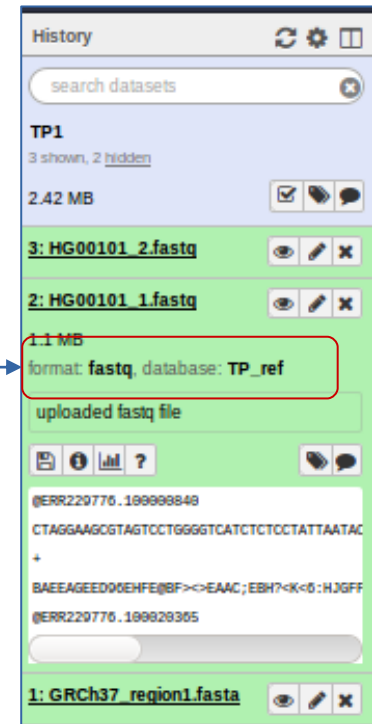
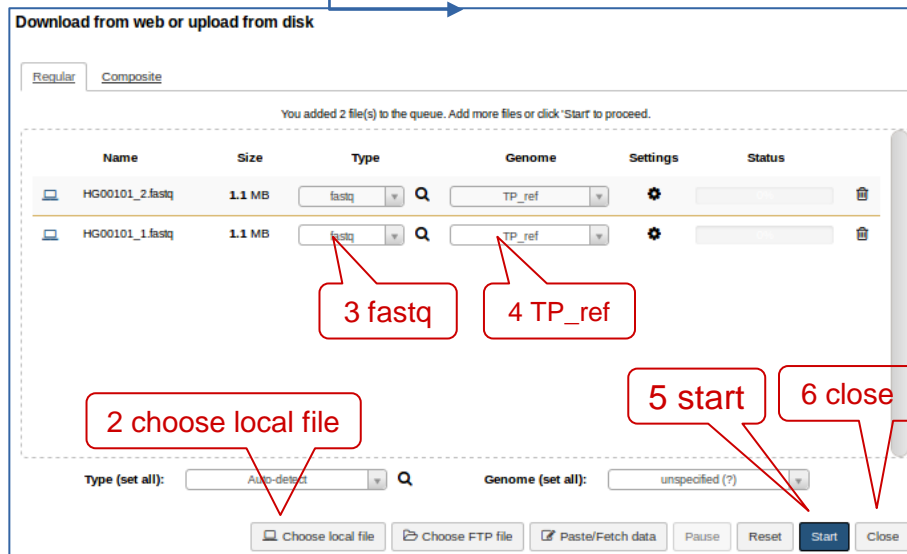
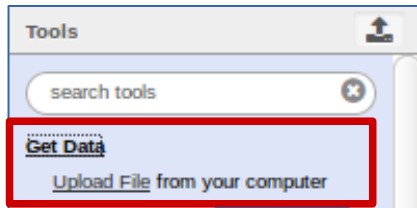
1 sequences
format: **fasta**, database: **TP_ref**

uploaded fasta file

>chr20 20:380000-530000
CAGACAGGGACAAGTCAGTCTCATCTCTGTGCACCA
TTGTGTCTAGGGTCTCATG666CAGCCCCGACCTC
CAGACCCCTGTCCAGCTCCCTCCAAGCTGAGTGTG
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



Convert to Sanger format : *groomer* tool

Galaxy

Analyze Data Workflow Shared Data Visualization Help User

Tools

groom

Preprocessing

FASTQ Groomer convert between various FASTQ quality formats

Alignment

Bowtie2 - map reads against reference genome

Workflows

All workflows

FASTQ Groomer convert between various FASTQ quality formats (Galaxy Version 1.0.4)

Options

File to groom

3: HG00101_2.fastq
2: HG00101_1.fastq

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

Input FASTQ quality scores type

Sanger & illumina 1.8+

Advanced Options

Hide Advanced Options

Execute

History

search datasets

TP1

0 shown, 2 hidden

4.7 MB

5: FASTQ Groomer on data 3

1.1 MB

format: fastqsanger, database: TP_ref

Groomed 5283 sanger reads into sanger reads.
Based upon quality and sequence, the input data is valid for: sanger
Input ASCII range: ""(34) - 'P'(80)
Input decimal range: 1 - 47

4: FASTQ Groomer on data 2

1.1 MB

format: fastqsanger, database: TP_ref

Groomed 5283 sanger reads into sanger reads.
Based upon quality and sequence, the input data is valid for: sanger
Input ASCII range: ""(34) - 'P'(80)
Input decimal range: 1 - 47

- 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*

Attributes Convert Format Datatype Permissions

Edit Attributes

Name: FASTQ Groomer on data 2 2 => HG0101_OK_1.fastq

Info: Groomed 5283 sanger reads into sanger reads. Based upon quality and sequence.

Annotation / Notes:

Database/Build: TP_ref (TP_ref) [Custom]

Save 3

Auto-detect

History

search datasets

TP1

5 shown, 2 hidden

4.7 MB

5: HG0101_OK_2.fastq

4: HG0101_OK_1.fastq

3: HG00101_2.fastq

2: HG00101_1.fastq

1: GRCh37_region1.fasta

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 retrieve *dataset* origin ?

Retreive *dataset* origin

History

search datasets

TP1

5 shown, 2 hidden

4.7 MB

5: HG0101_OK_2.fastq

4: HG0101_OK_1.fastq

1.1 MB

format: **fastqsanger**, database: TP_ref

Groomed 5283 sanger reads into sanger reads.
Based upon quality and sequence, the input data is valid for: sanger
Input ASCII range: ""(34) - "P"(80)
Input decimal range: 1 - 47

Information icon circled in red

@ERR220776.169999848
CTAGGAAGCGTAGTCTCTGGGGTCATCTCTCTATTAATA
+
BAEEAGEED0EHFE@BF<>>EAAC;EBH7<K<5:HJGF
@ERR220776.169920305

Tool: FASTQ Groomer

Number: 4

Name: HG0101_OK_1.fastq

Created: Tue 24 Jan 2017 10:59:29 AM (UTC)

Filesize: 1.1 MB

Dbkey: TP_ref

Format: fastqsanger

Galaxy Tool ID: toolshed.g2.bx.psu.edu/repos/devteam/fastq_groomer/fastq_groomer/1.0.4

Galaxy Tool Version: 1.0.4

Tool Version:

Tool Standard Output: [stdout](#)

Tool Standard Error: [stderr](#)

Tool Exit Code: 0

History Content API ID: 9ad713a5aaf99f6f

Job API ID: b3a78854daef4a5a

History API ID: 5114a2a207b7caff

UUID: 02f01336-3810-4a82-a42-4dc9d62cd9a5

Input Parameter	Value	Note for rerun
File to groom	2: HG00101_1.fastq	
Input FASTQ quality scores type	Sanger & Illumina 1.8+	
Advanced Options	basic	

Dependency	Dependency Type	Version
galaxy_sequence_utils	tool_shed_package	1.0.0

Inheritance Chain

HG0101_OK_1.fastq

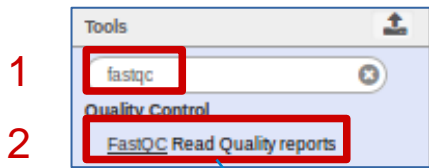
This *dataset* results from *groomer* tool, applied on *dataset 2* (HG101_1.fastq)

Reads quality control

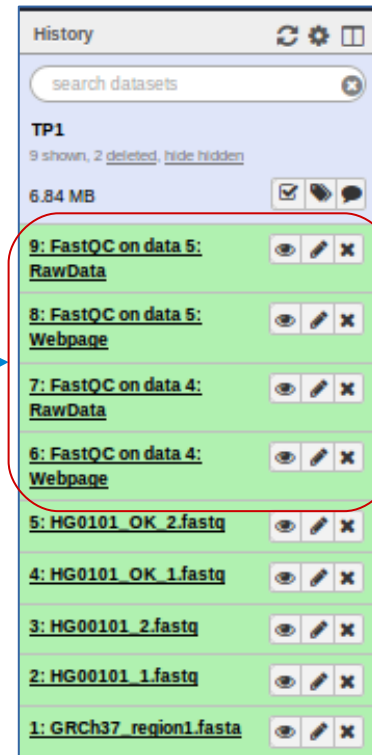
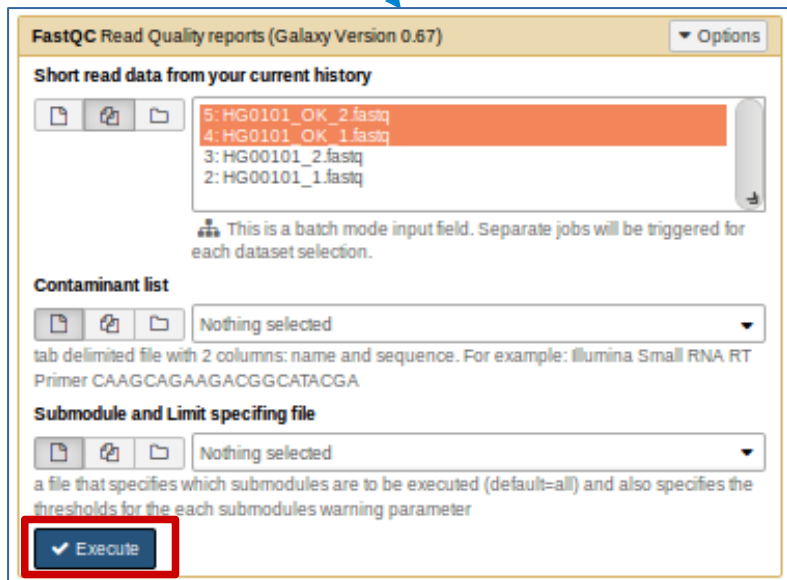
- Per base quality
- Per read mean quality
- Read size
- Adaptators
- Duplicated reads

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

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

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

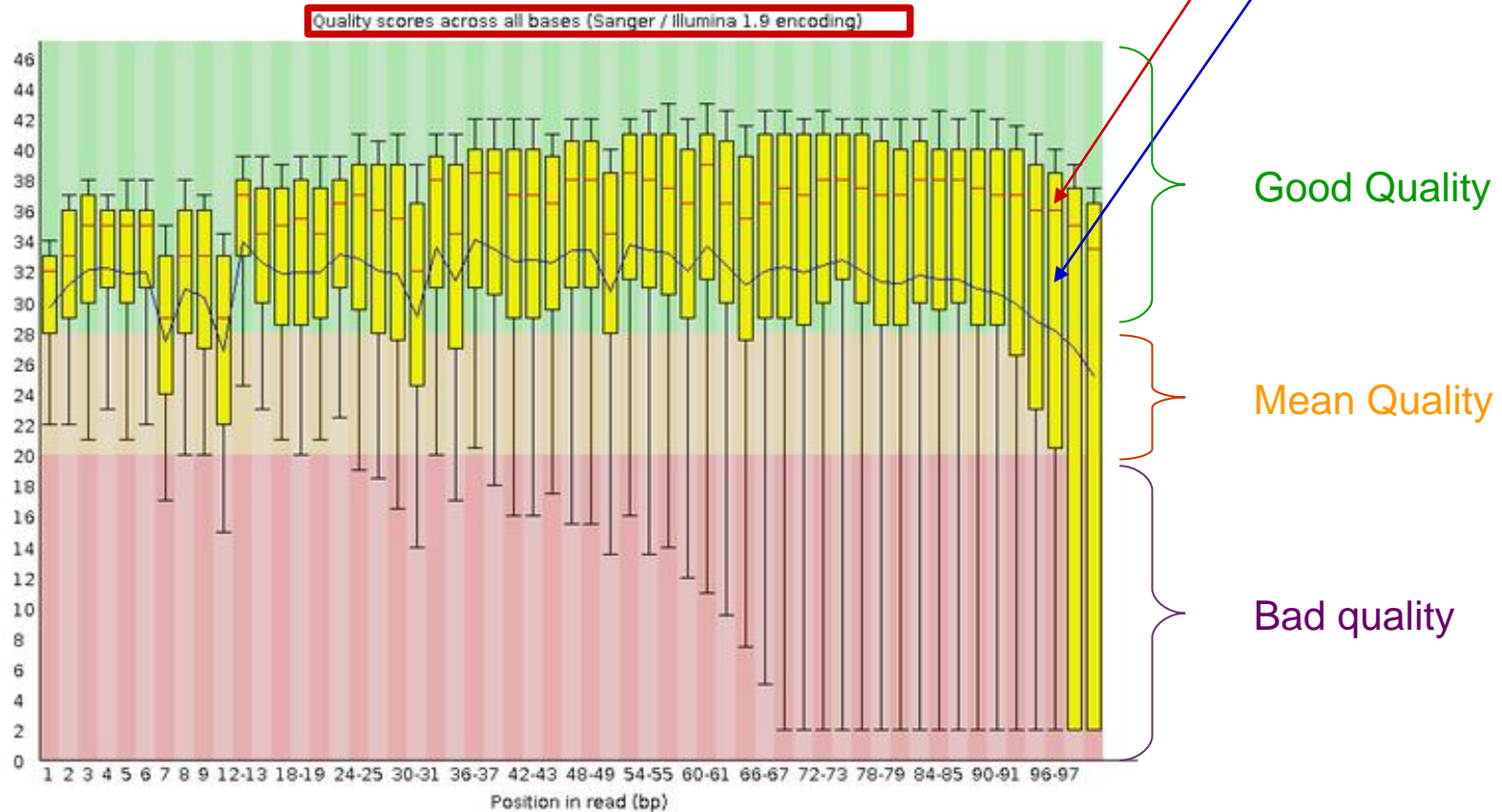
-  [Basic Statistics](#)
-  [Per base sequence quality](#)
-  [Per sequence quality scores](#)
-  [Per base sequence content](#)
-  [Per sequence GC content](#)
-  [Per base N content](#)
-  [Sequence Length Distribution](#)
-  [Sequence Duplication Levels](#)
-  [Overrepresented sequences](#)
-  [Adapter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

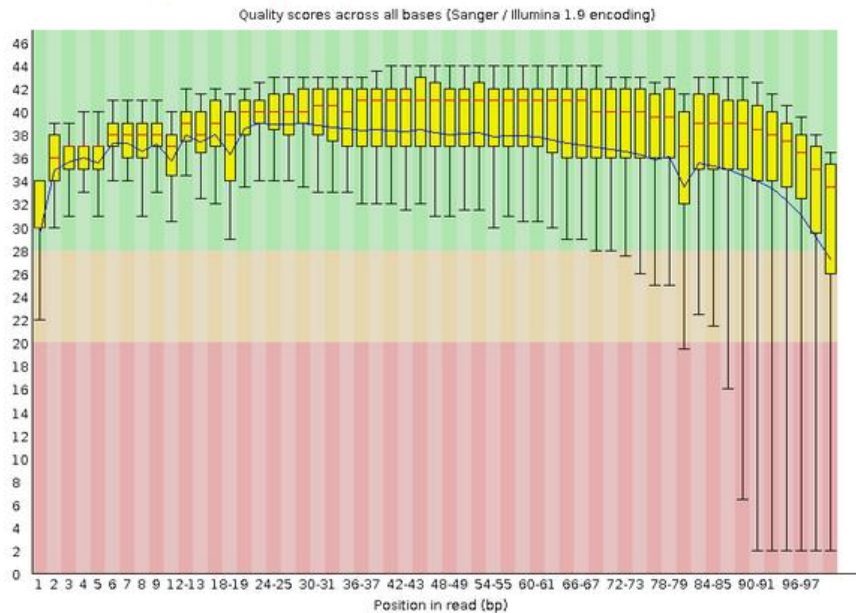
✖ Per base sequence quality



Fasqc : Per base sequence quality

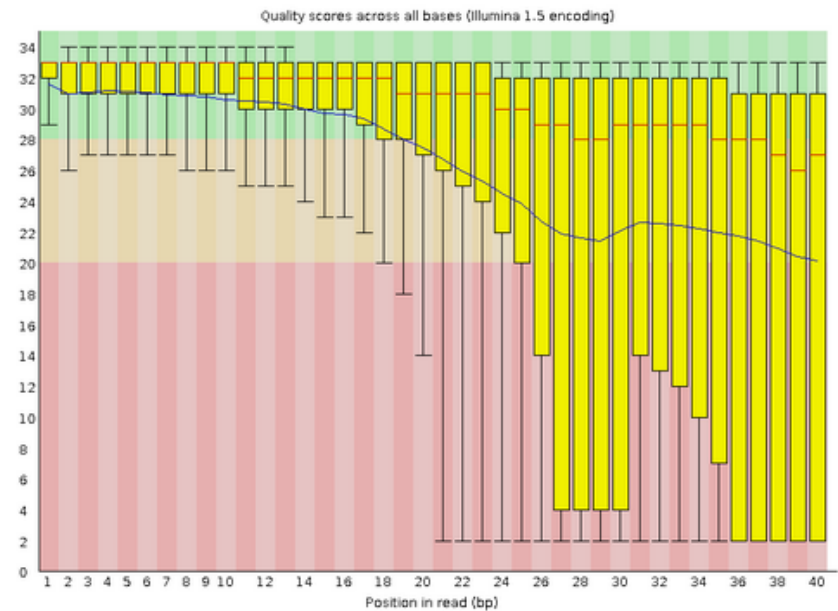
Example OK

✔ Per base sequence quality



Example KO

✘ Per base sequence quality



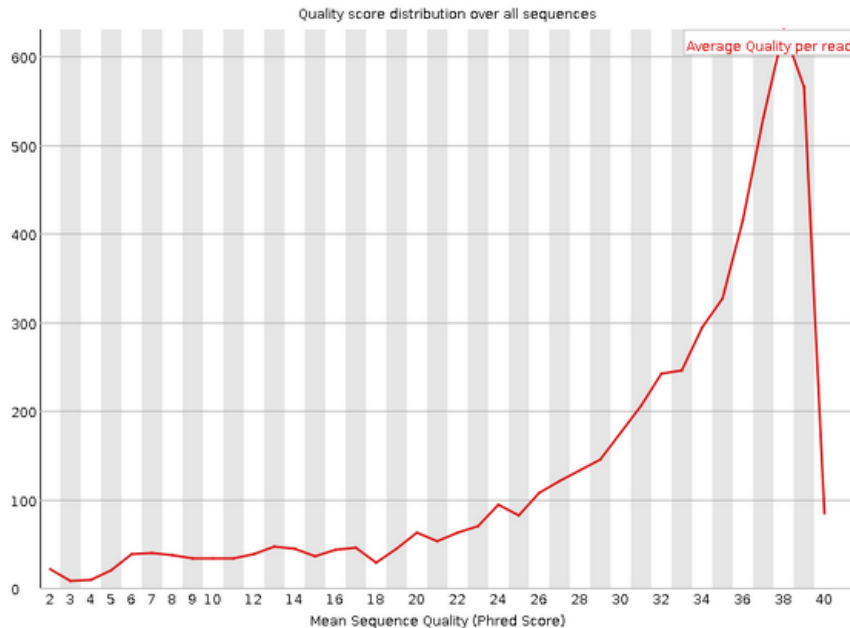
Source :

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

Fastqc : Per sequence quality score

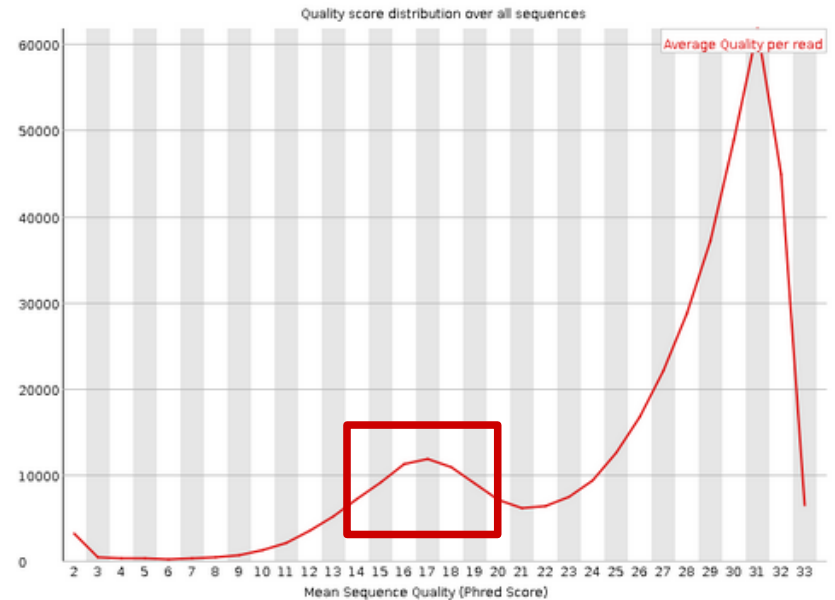
Example OK

✔ Per sequence quality scores



Example KO

✔ Per sequence quality scores

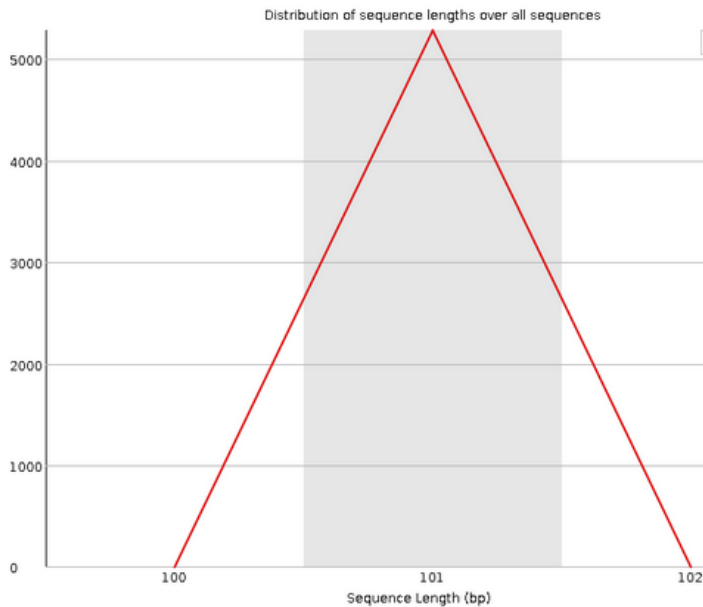


Source :

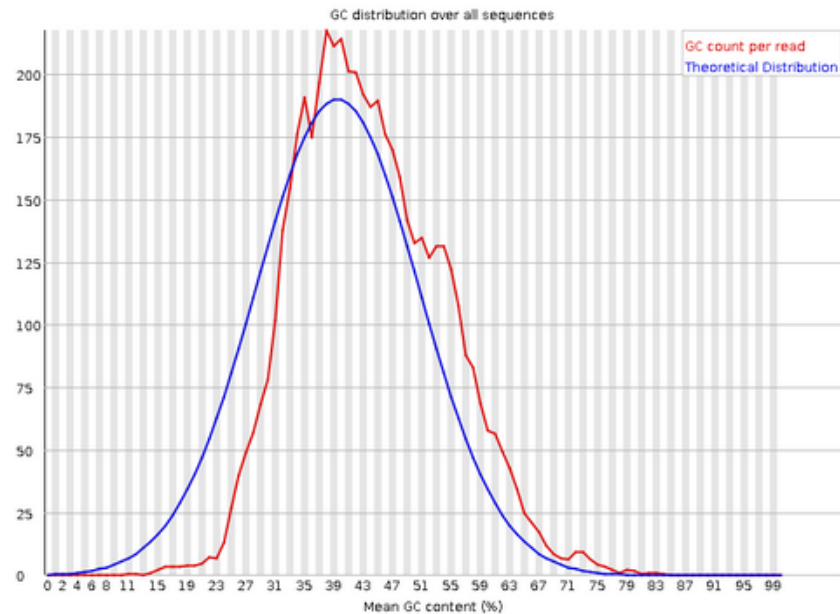
http://www.bioinformatics.babraham.ac.uk/projects/fastqc/bad_sequence_fastqc.html

FastQC : Sequence Length Distribution & Per sequence GC content

✔ Sequence Length Distribution



⚠ Per sequence GC content

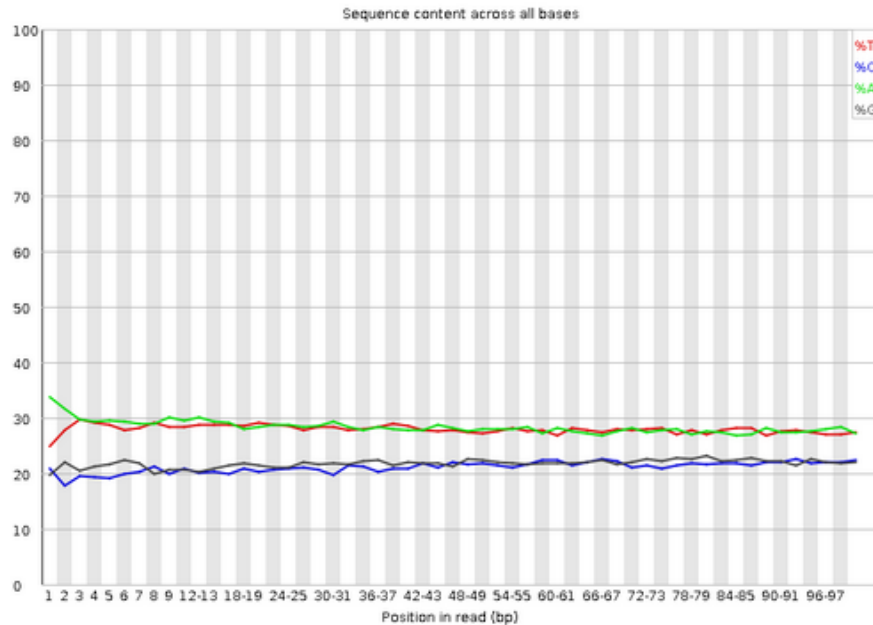


FastQC : Per base sequence content

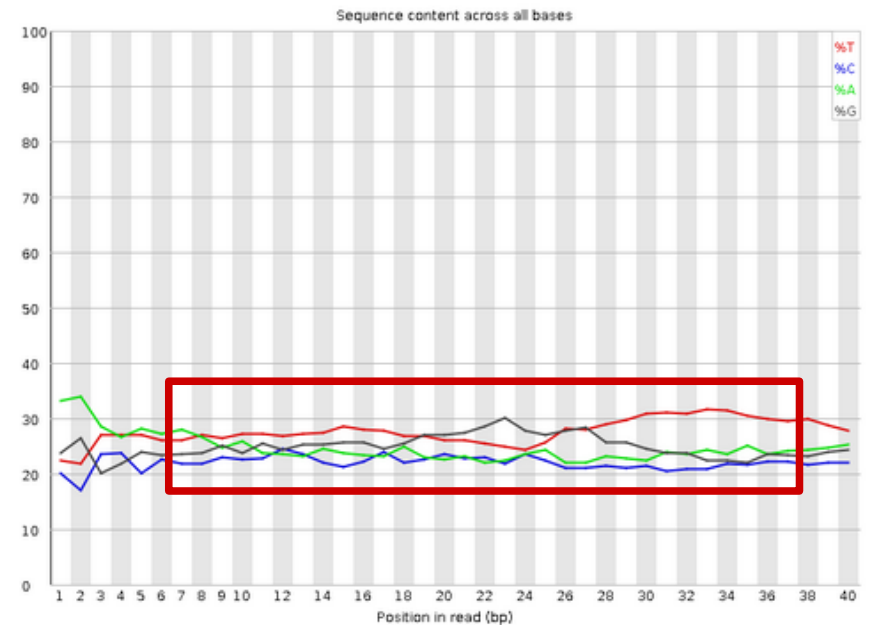
Example OK

Example KO

✔ Per base sequence content



⚠ Per base sequence content



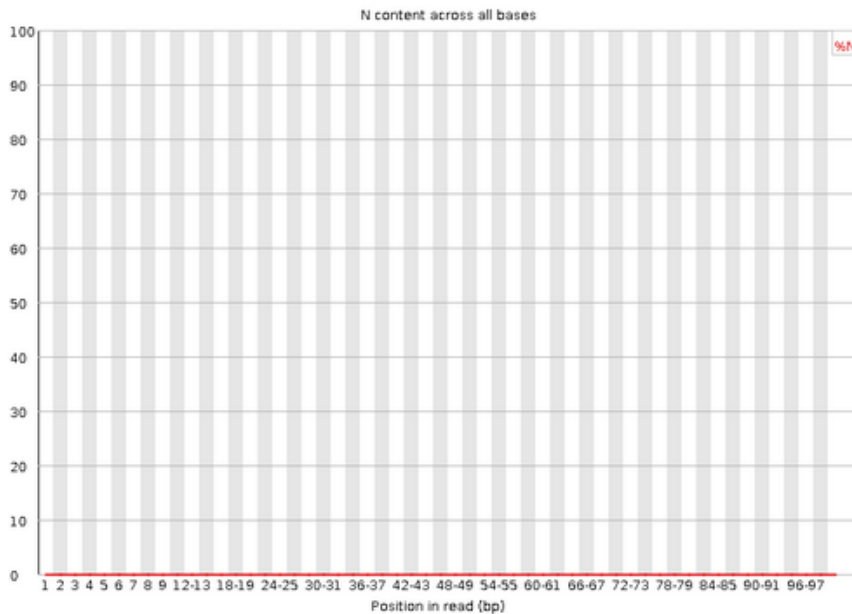
Source :

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

FastQC : Per base N content

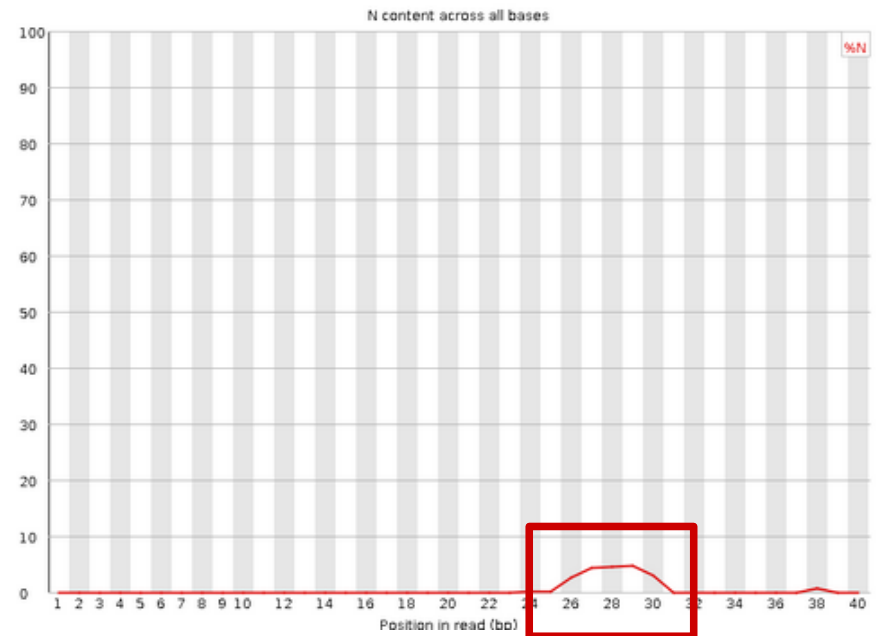
Example OK

✔ 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

 **Overrepresented sequences**
No overrepresented sequences

Example KO

 **Overrepresented sequences**

Sequence	Count	Percentage	Possible Source
AGAGTTTTATCGCTTCCATGACGCAGAAGTTAACACTTTC	2065	0.5224039181558763	No Hit
GATTGGCGTATCCAACCTGCAGAGTTTTATCGCTTCCATG	2047	0.5178502762542754	No Hit
ATTGGCGTATCCAACCTGCAGAGTTTTATCGCTTCCATGA	2014	0.5095019327680071	No Hit
CGATAAAAATGATTGGCGTATCCAACCTGCAGAGTTTTAT	1913	0.4839509420979134	No Hit

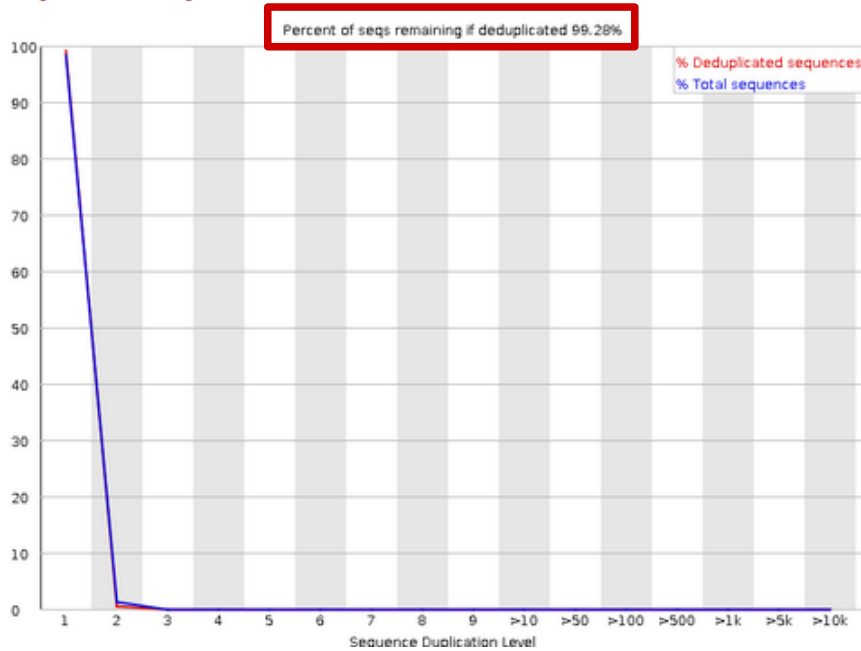
CCTGCAGAGTTTTATCGCTTCCATGACGCAGAAGTTAACA	613	0.15507600476007366	No Hit
CGGTCAGCAGGAATGCCGAGATCGGAAGACGGTTCAGC	599	0.15153508328105078	Illumina Paired End PCR Primer 2 (96% over 25bp)
TCTGCAGGTTGGATACGCCAATCATTTTTATCGAAGCGCG	585	0.1479933618020279	No Hit

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

FastQC : Sequence Duplication Levels

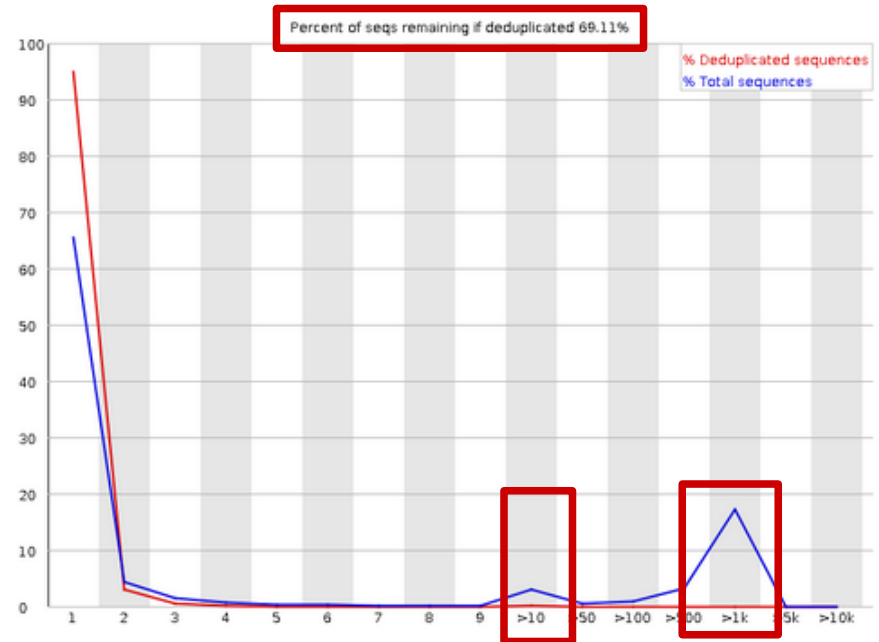
Example OK

✔ Sequence Duplication Levels



Example KO

⚠ Sequence Duplication Levels



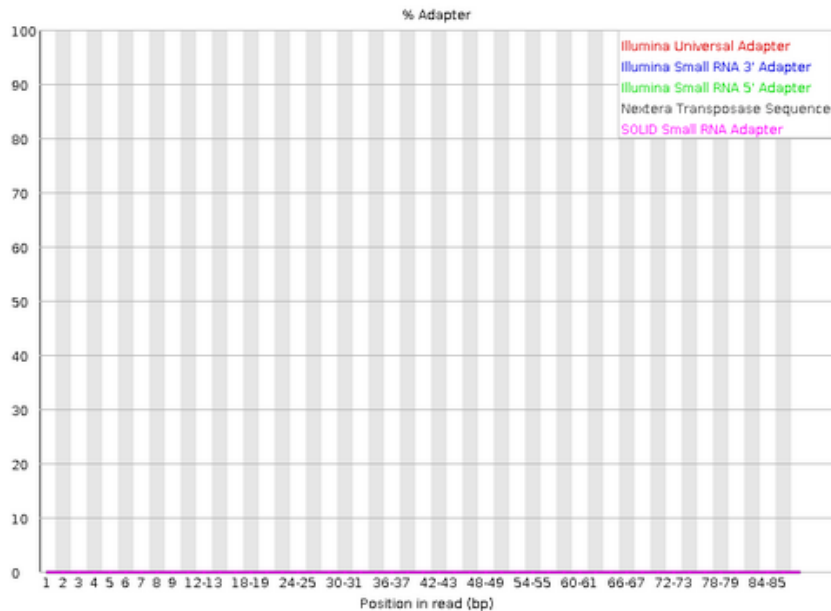
Source :

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

FastQC Adapter Content

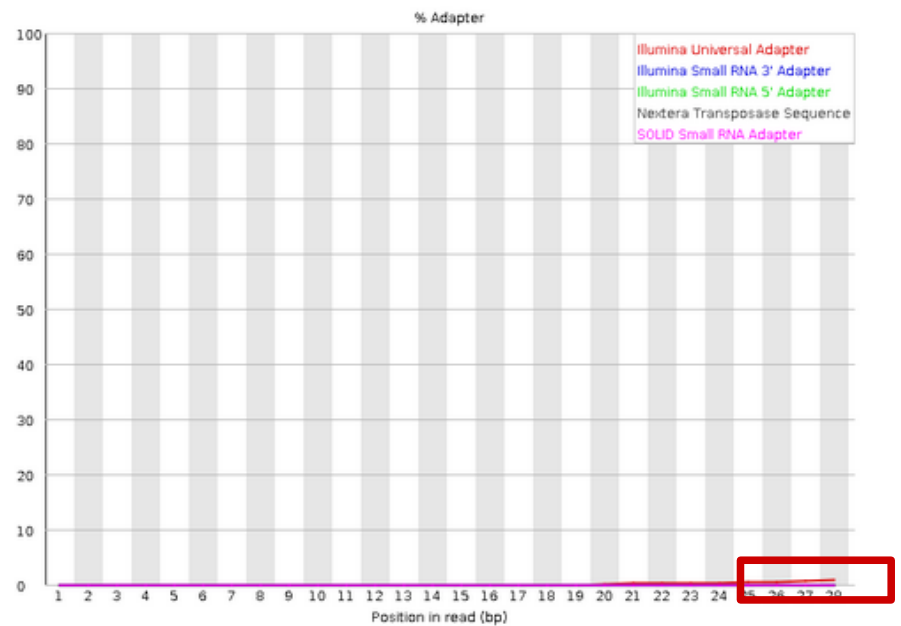
Example OK

Adapter Content



Example KO

Adapter Content

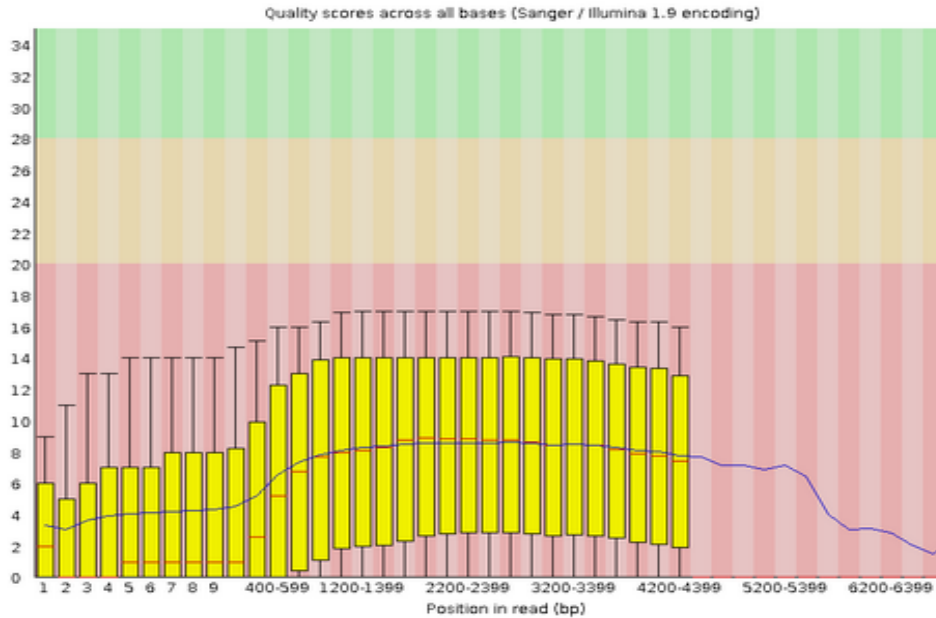


Source :

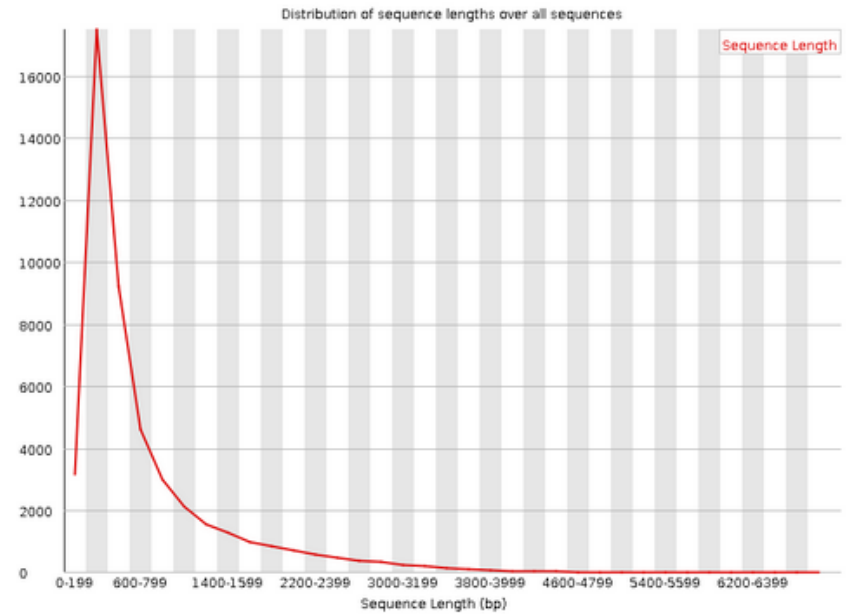
http://www.bioinformatics.babraham.ac.uk/projects/fastqc/bad_sequence_fastqc.html

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

Filtering & trimming

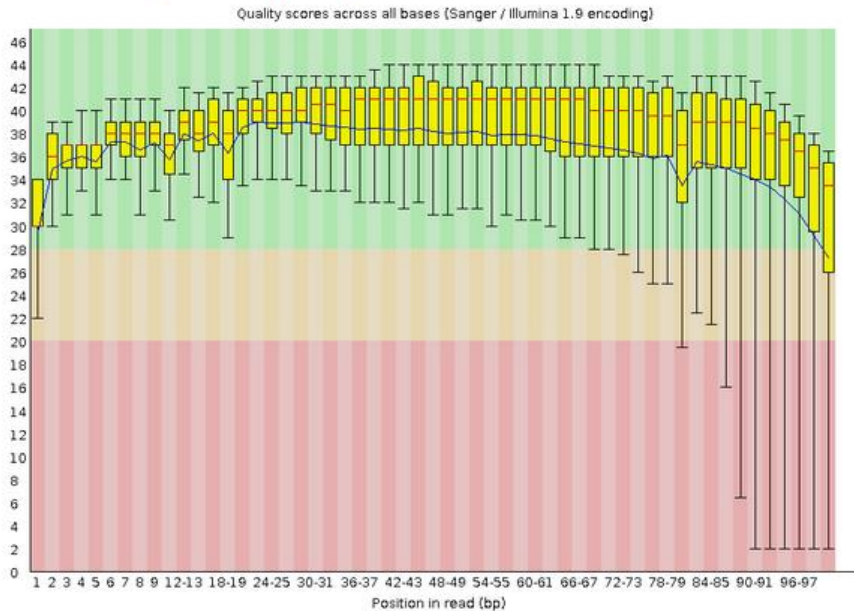
- Filtering = remove reads
 - Based on quality or size criteria
- Trimming = remove read ends
 - Fixed number of bases
 - Bases < quality

Trimming

Cut bad quality bases at the end of reads

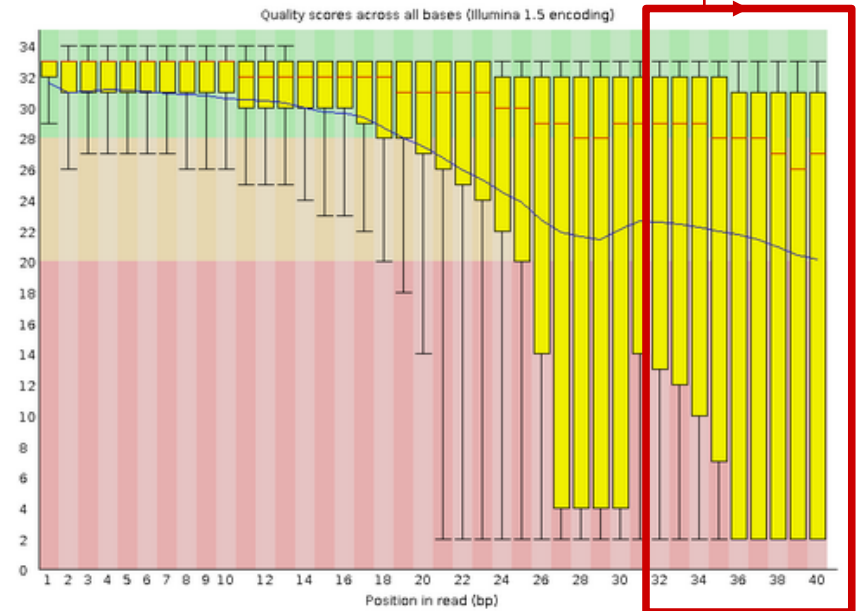
Exemple OK

✔ Per base sequence quality



Exemple KO

✘ Per base sequence quality



Source :

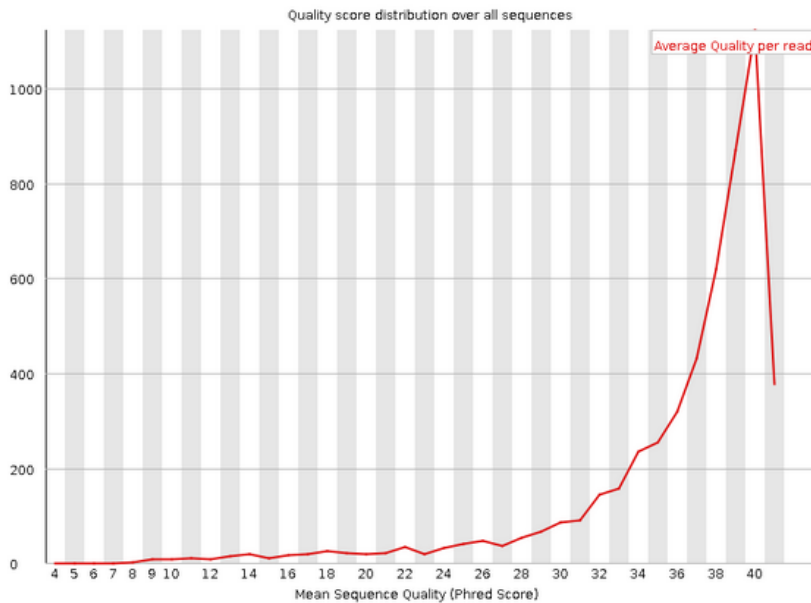
http://www.bioinformatics.babraham.ac.uk/projects/fastqc/bad_sequence_fastqc.html

Filtering

Remove reads with bad mean quality

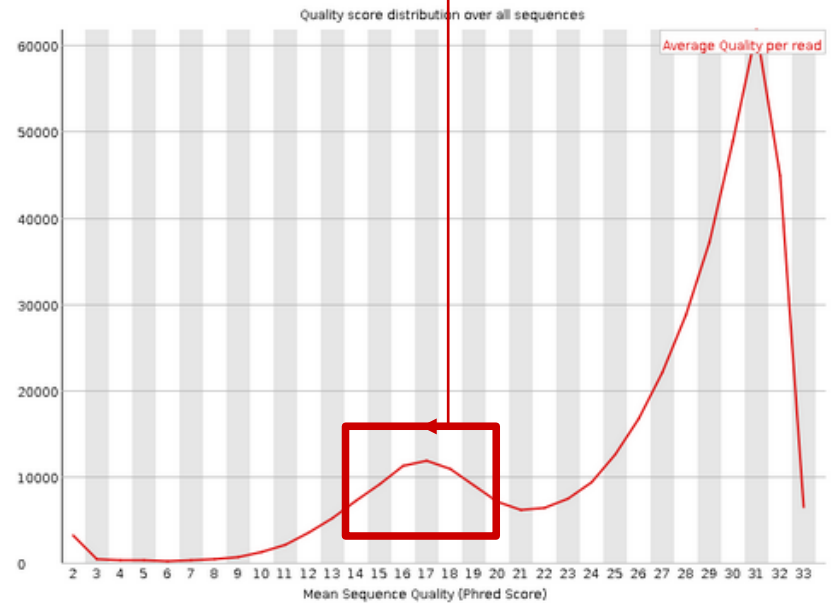
Exemple OK

✔ Per sequence quality scores



Exemple KO

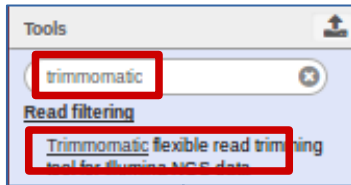
✔ Per sequence quality scores



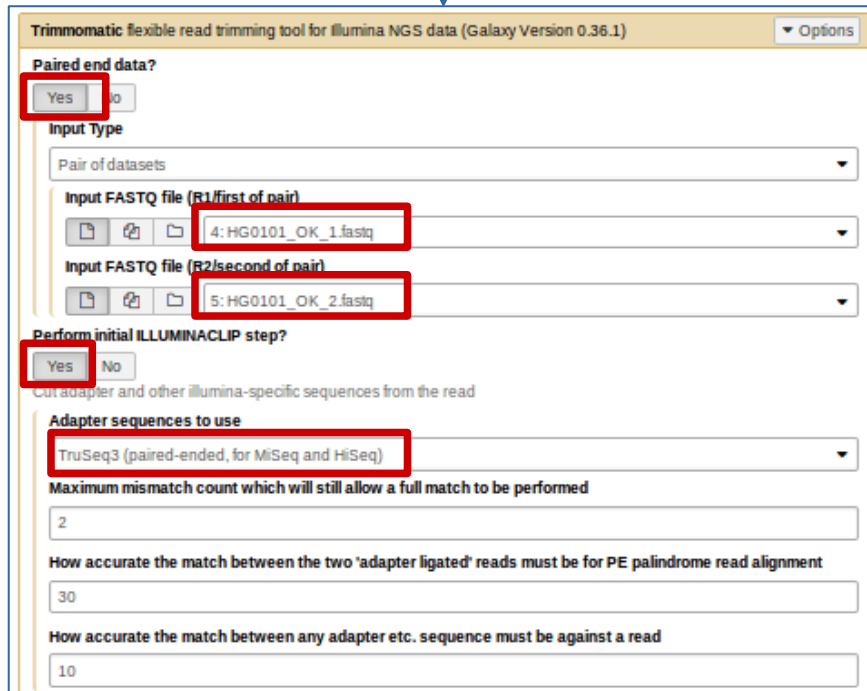
Source :

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



1 Choose files

2 Parameters for adaptators

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



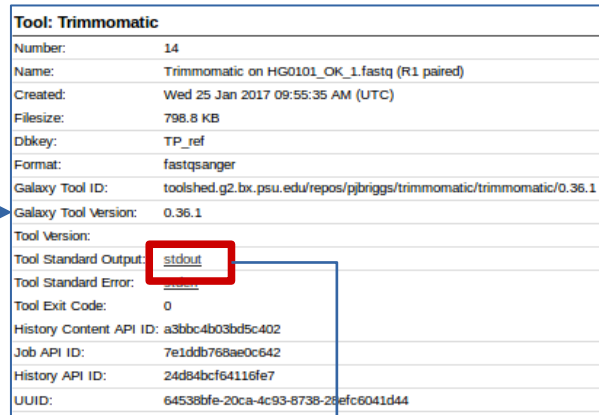
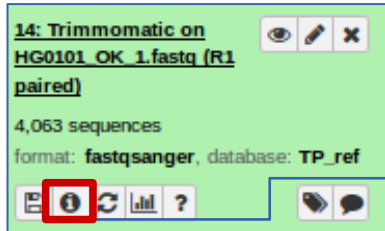
Unpaired reads2 (corresponding reads1 has been removed during cleaning)

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.

```
TrimmomaticPE: Started with arguments:
.-threads 1 -phred33 /ifb/galaxy-dist/database/files/000/dataset_387.dat /ifb/galaxy-dist/database/files/000/dataset_388.dat /ifb/galaxy-dist/data
Using PrefixPair: 'TACACTCTTTCCCTACACGACGCTCTTCCGATCT' and 'GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT'
ILLUMINACLI: Using 1 prefix pairs, 0 forward/reverse sequences, 0 forward only sequences, 0 reverse only sequences
Input Read Pairs: 5283 Both Surviving: 4063 (76.91%) Forward Only Surviving: 28 (0.53%) Reverse Only Surviving: 1108 (20.97%) Dropped: 84 (1.59%)
TrimmomaticPE: Completed successfully
```

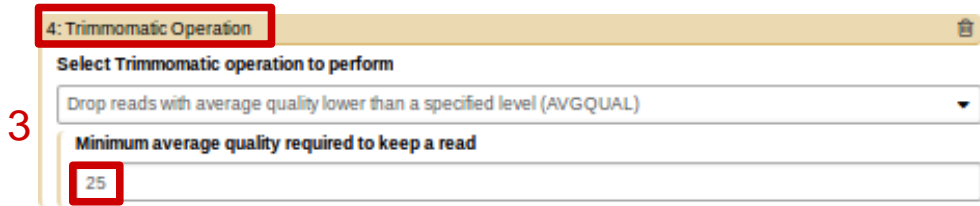
Trimmomatic : 2nd try

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

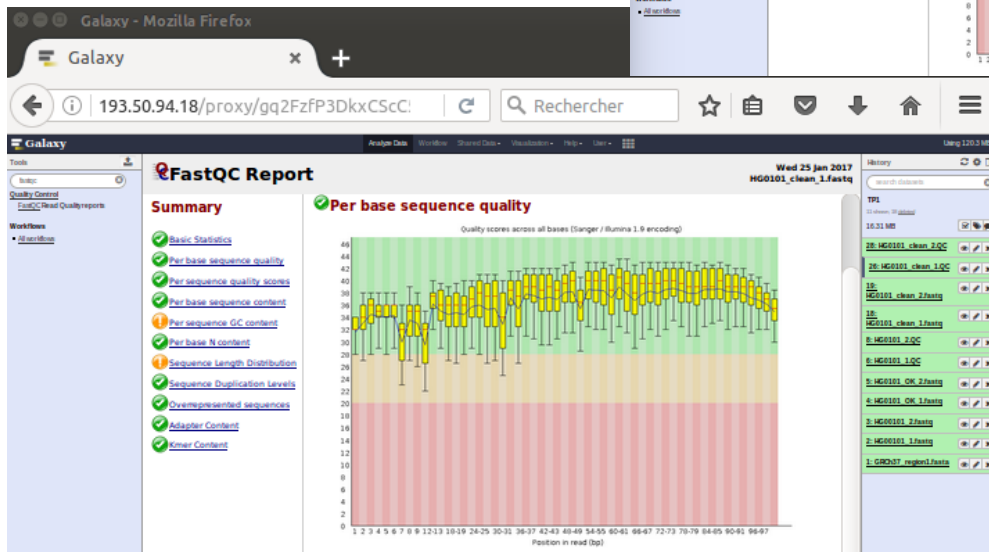
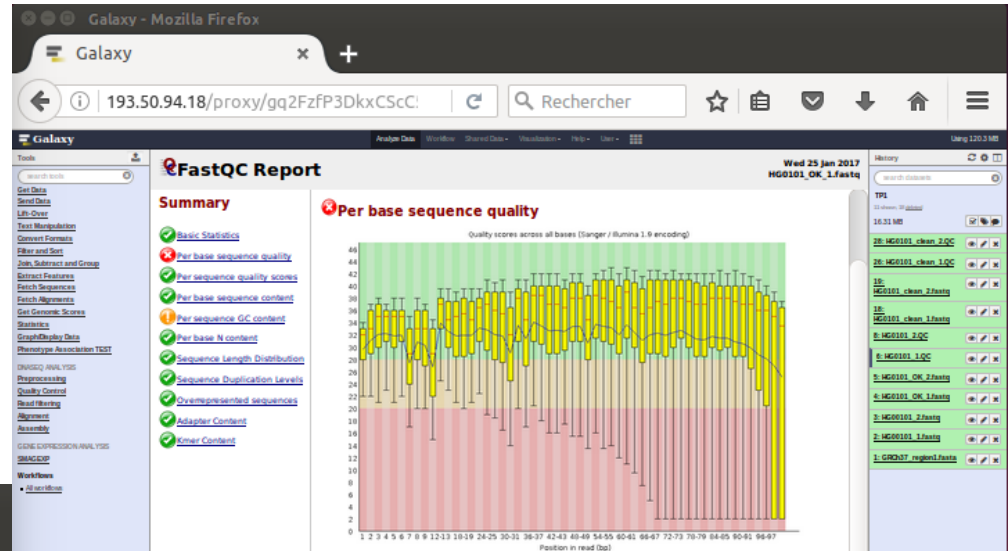
Compare quality control before / after cleaning

Solution 1 :

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

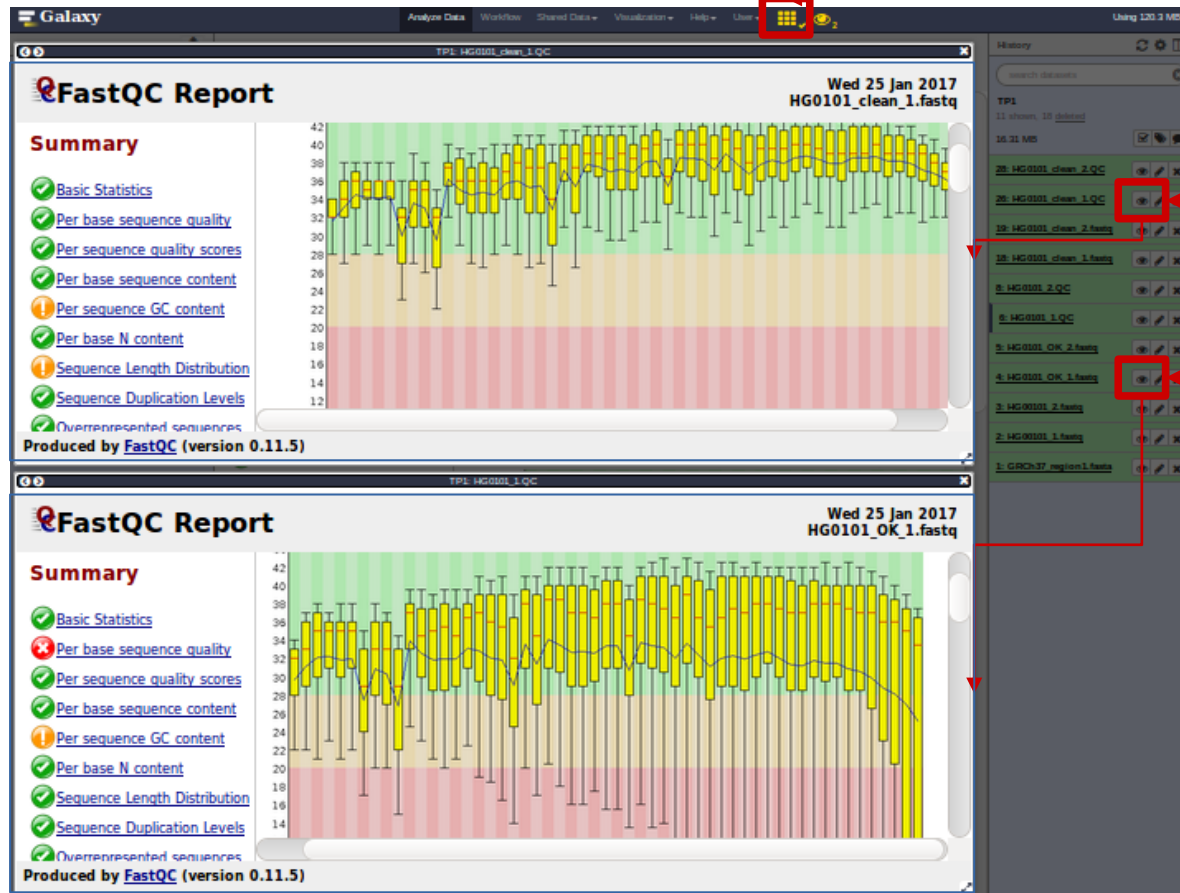
Visualize datasets

- 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

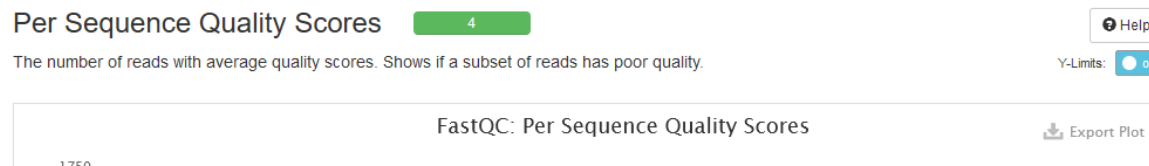
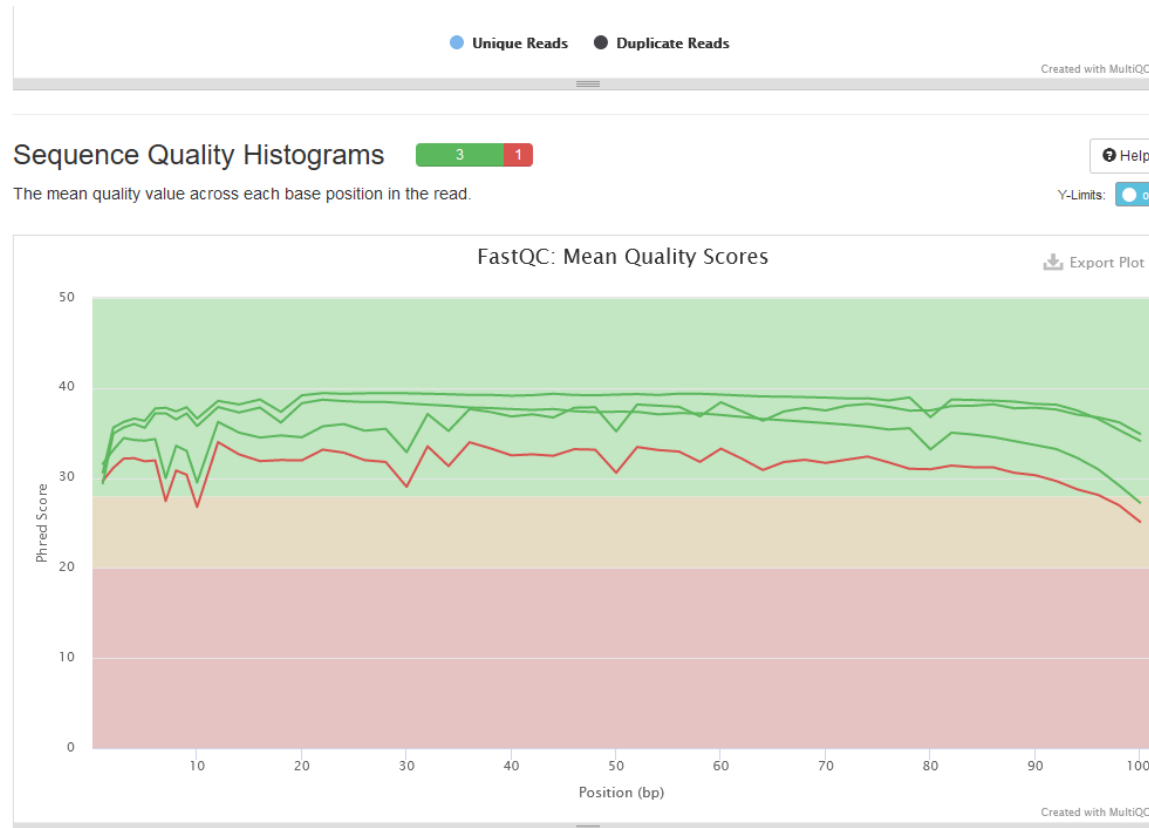
3 Visualize dataset
HG0101_OK_1.fast
q

Compare quality control before / after cleaning

Solution 3 : Use MultiQC

MultiQC
v1.9

- General Stats
- Trimmomatic
- FastQC
- Sequence Counts
- Sequence Quality Histograms
- 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
- Status Checks



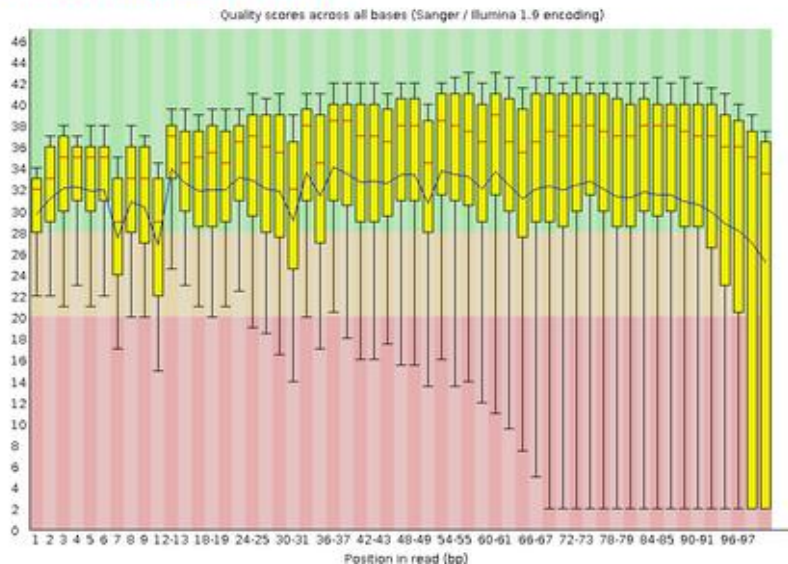
Quality control before & after cleaning

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

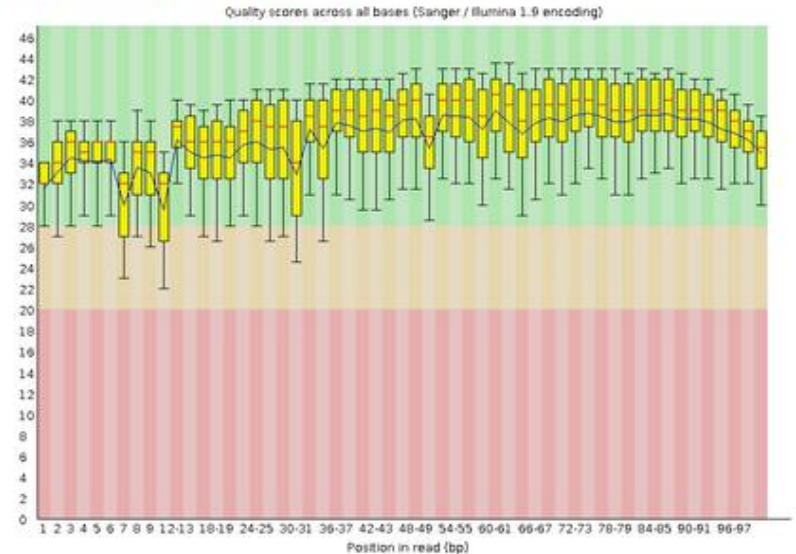


HG0101_clean_1.QC

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

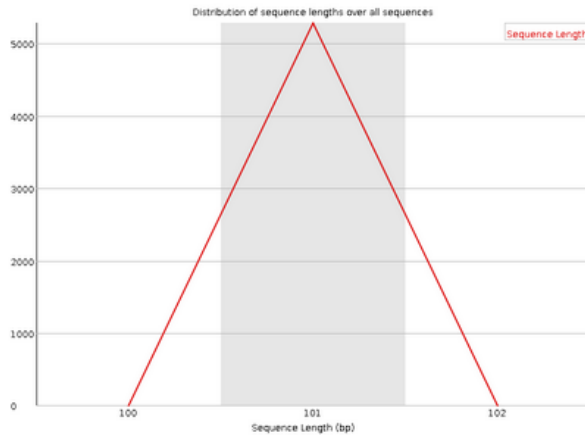


Quality control before & after cleaning

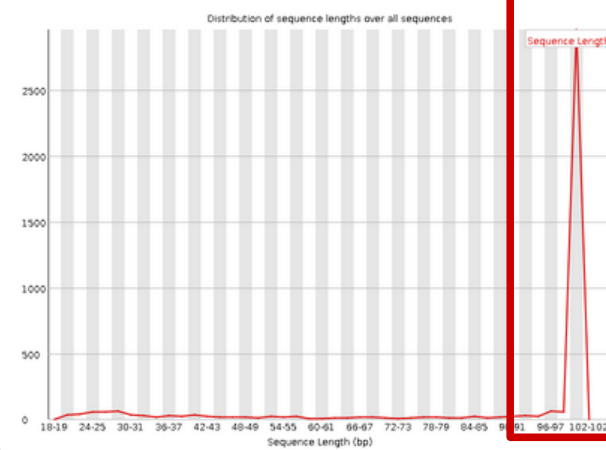
HG0101_OK_1.QC

HG0101_clean_1.QC

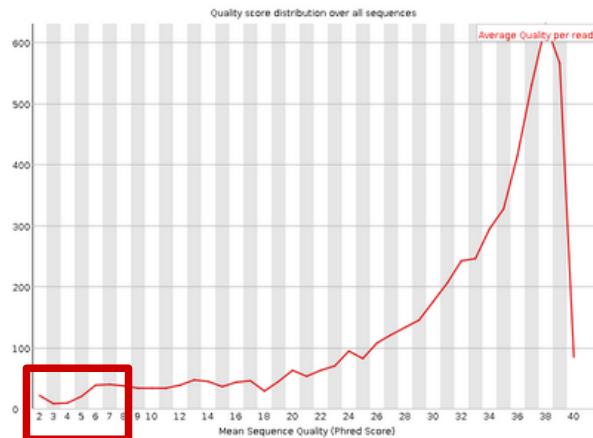
✔ Sequence Length Distribution



⚠ Sequence Length Distribution



✔ Per sequence quality scores



✔ Per sequence quality scores

