Galaxy for virologist training

Exercise 2: Quality control and trimming

Table of Contents

Despite the improvement of sequencing methods, there is no error-free technique. A correct measuring of the sequencing quality is essential for identifying problems in the sequencing, thus, this must be the first step in every sequencing analysis. Once the quality control is finished, it's important to remove those low quality reads, or short reads, for which a trimming step is mandatory. After the trimming step it is recommended to perform a new quality control step to be sure that trimming worked.

1. Illumina Quality control and trimming

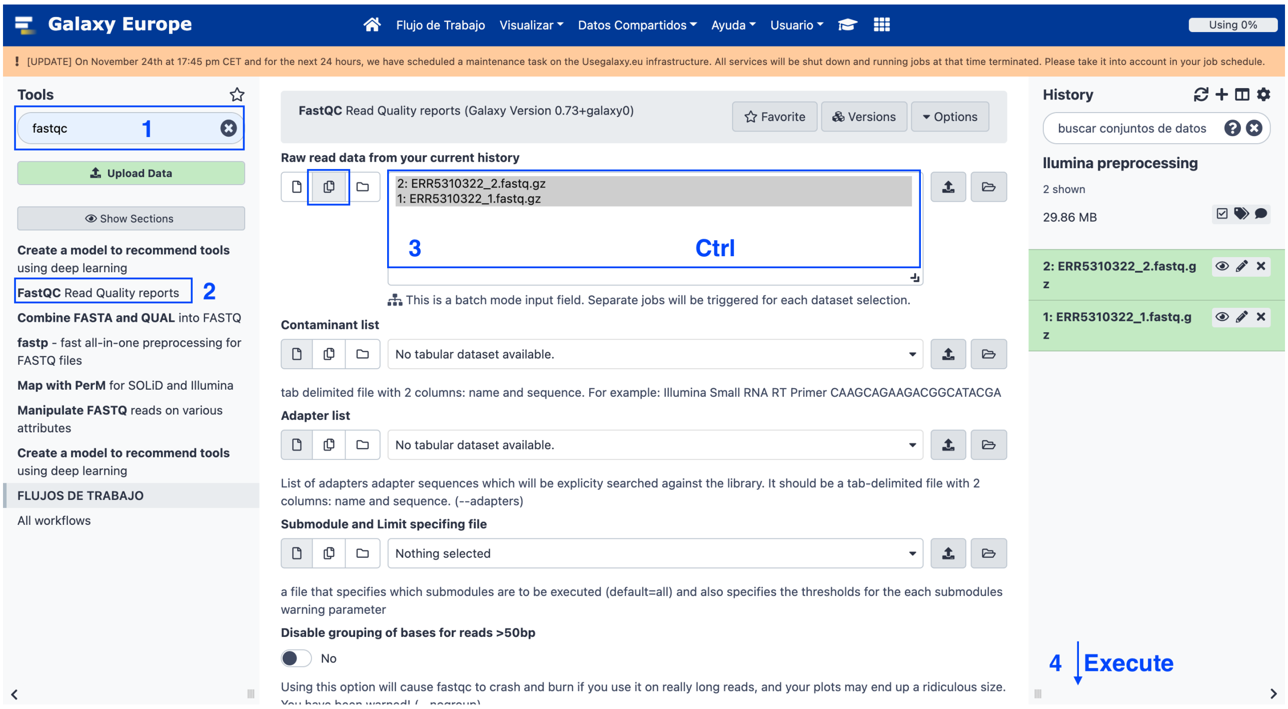
Title	**Pre-processing**
Training dataset:	PRJEB43037 - In August 2020, an outbreak of West Nile Virus affected 71 people with meningoencephalitis in Andalusia and 6 more cases in Extremadura (south-west of Spain), causing a total of eight deaths. The virus belonged to the lineage 1 and was relatively similar to previous outbreaks occurred in the Mediterranean region. Here, we present a detailed analysis of the outbreak, including an extensive phylogenetic study. This is one of the outbreak samples.
Questions:	• How do I check whether my Illumina data was correctly sequenced? • How can I improve the quality of my data?
Objectives:	• Perform a quality control in raw Illumina reads • Perform a quality trimming in raw Illumina reads • Perform a quality control in trimmed Illumina reads
Estimated time:	25 min

1. Quality control

To run the quality control over the samples, follow these steps: 1. Create a new history, as we explained yesterday named **Illumina preprocessing** 2. Upload data as seen yesterday, copy and paste the following URLs:

```
ftp://ftp.sra.ebi.ac.uk/vol1/fastq/ERR531/002/ERR5310322/ERR5310
ftp://ftp.sra.ebi.ac.uk/vol1/fastq/ERR531/002/ERR5310322/ERR5310
```

1. Search for the **fastqc** tool and select **FastQC Read Quality reports** and set the following parameters:
   - Select multiple file data set in Raw read data from your current history
   - With the *Ctrl* key pressed, select the two datasets
   - Then go down and select **Execute**



Galaxy Europe
Flujo de Trabajo
Visualizar
Datos Compartidos
Ayuda
Usuario
Using 0%

! [UPDATE] On November 24th at 17:45 pm CET and for the next 24 hours, we have scheduled a maintenance task on the Usegalaxy.eu infrastructure. All services will be shut down and running jobs at that time terminated. Please take it into account in your job schedule.

Tools
fastqc
1
Upload Data
Show Sections
Create a model to recommend tools using deep learning
FastQC Read Quality reports
2
Combine FASTA and QUAL into FASTQ
fastp - fast all-in-one preprocessing for FASTQ files
Map with PerM for SOLiD and Illumina
Manipulate FASTQ reads on various attributes
Create a model to recommend tools using deep learning
FLUJOS DE TRABAJO
All workflows

FastQC Read Quality reports (Galaxy Version 0.73+galaxy0)

Favorite
Versions
Options

Raw read data from your current history

2: ERR5310322_2.fastq.gz
1: ERR5310322_1.fastq.gz

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

Contaminant list

No tabular dataset available.

tab delimited file with 2 columns: name and sequence. For example: Illumina Small RNA RT Primer CAAGCAGAAGACGGCATACGA

Adapter list

No tabular dataset available.

List of adapters adapter sequences which will be explicity searched against the library. It should be a tab-delimited file with 2 columns: name and sequence. (--adapters)

Submodule and Limit specifing file

Nothing selected

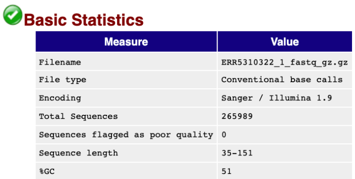
a file that specifies which submodules are to be executed (default=all) and also specifies the thresholds for the each submodules warning parameter

Disable grouping of bases for reads >50bp

No

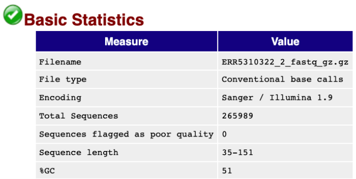
Using this option will cause fastqc to crash and burn if you use it on really long reads, and your plots may end up a ridiculous size. You have been warned! (--nogroup)

To see the results we are going to open the jobs with **Web page** in their name for both data 1 and data 2.



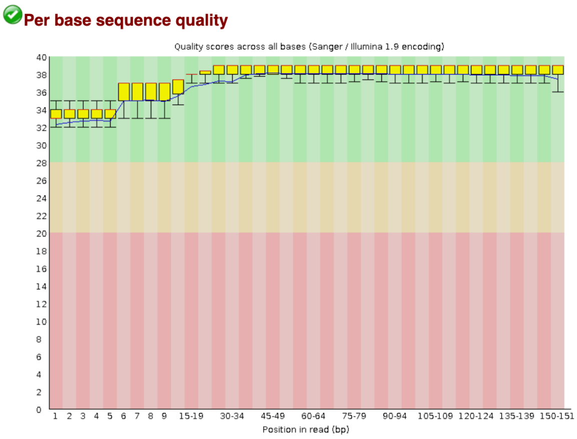
Basic Statistics

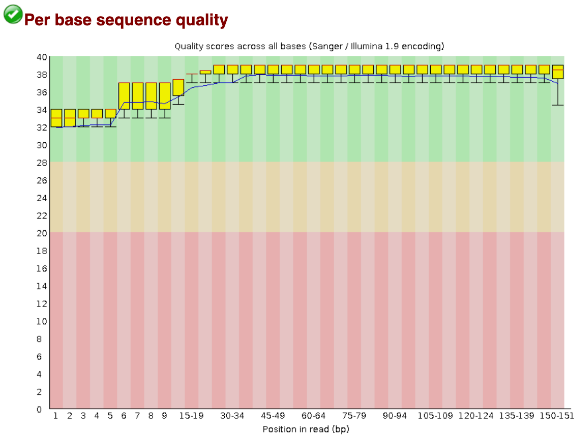
Measure	Value
Filename	ERR5310322_1_fastq_gz.gz
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	265989
Sequences flagged as poor quality	0
Sequence length	35-151
%GC	51



Basic Statistics

Measure	Value
Filename	ERR5310322_2_fastq_gz.gz
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	265989
Sequences flagged as poor quality	0
Sequence length	35-151
%GC	51





Here, you can see the number of reads in each file, the maximum and minimum length of all reads in the sample, and the quality plots for both R1 and R2. They look quite good, but we are going to run trimming over the samples.

▶ How many reads do the samples have?

First question

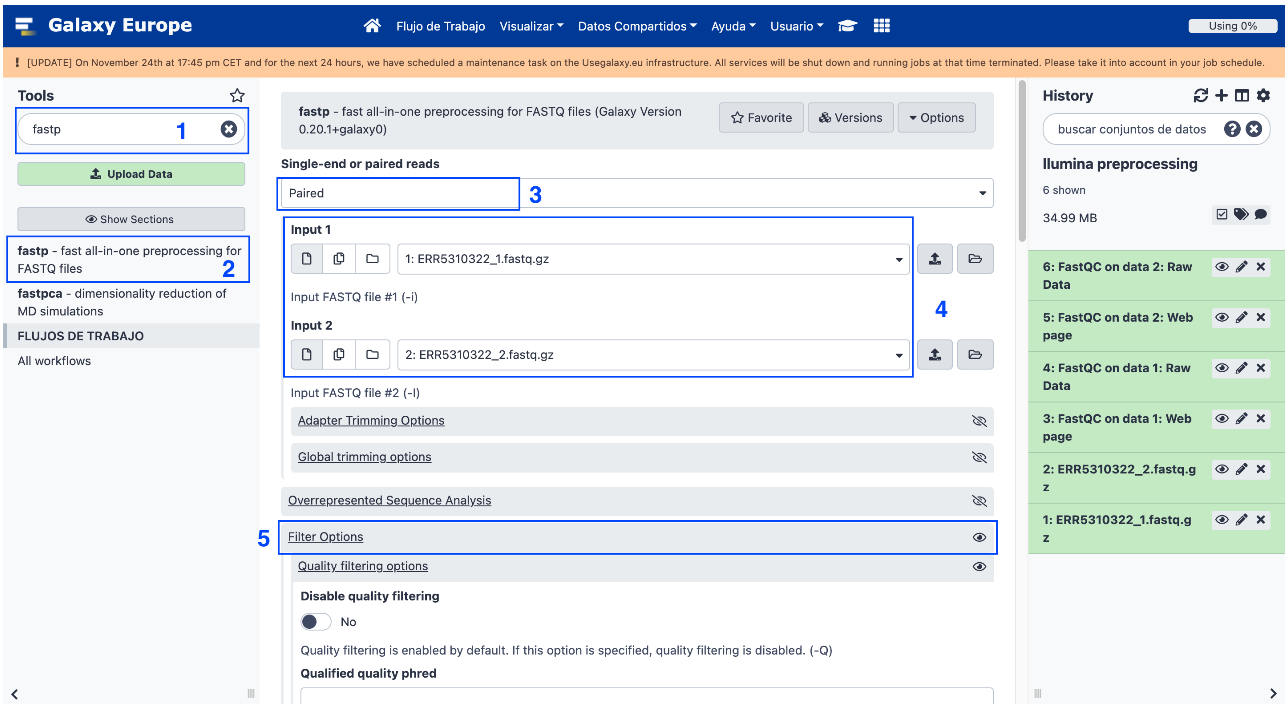
▶ How do I check whether my Illumina data was correctly sequenced?

2. Trimming

Once we have performed the quality control, we have to perform the quality and read length trimming:

1. Search for **fastp** in the tools and select **fastp - fast all-in-one preprocessing for FASTQ files**

2. Select custom parameters:
    - Single-end or paired reads > Paired
        - Input 1 > Browse datasets (right folder icon) > Select ERR5310322_1.fastq.gz
        - Input 2 > Browse datasets > Select ERR5310322_2.fastq.gz
    - Display Filter Options
        - Quality Filtering options
            - Qualified Quality Phred = 30
            - Unqualified percent limit = 10
        - Length Filtering Options
            - Length required = 50
    - Read modification options
        - PoliX tail trimming > Enable polyX tail trimming
        - Per read cutting by quality options
            - Cut by quality in front (5') > Yes
            - Cut by quality in tail (3') > Yes
            - Cutting mean quality = 30
3. Finally, click on **Execute**



! [UPDATE] On November 24th at 17:45 pm CET and for the next 24 hours, we have scheduled a maintenance task on the Usegalaxy.eu infrastructure. All services will be shut down and running jobs at that time terminated. Please take it into account in your job schedule.

fastp - fast all-in-one preprocessing for FASTQ files (Galaxy Version 0.20.1+galaxy0)

Single-end or paired reads

Paired

Input 1

1: ERR5310322_1.fastq.gz

Input FASTQ file #1 (-i)

Input 2

2: ERR5310322_2.fastq.gz

Input FASTQ file #2 (-I)

Adapter Trimming Options

Global trimming options

Overrepresented Sequence Analysis

Filter Options

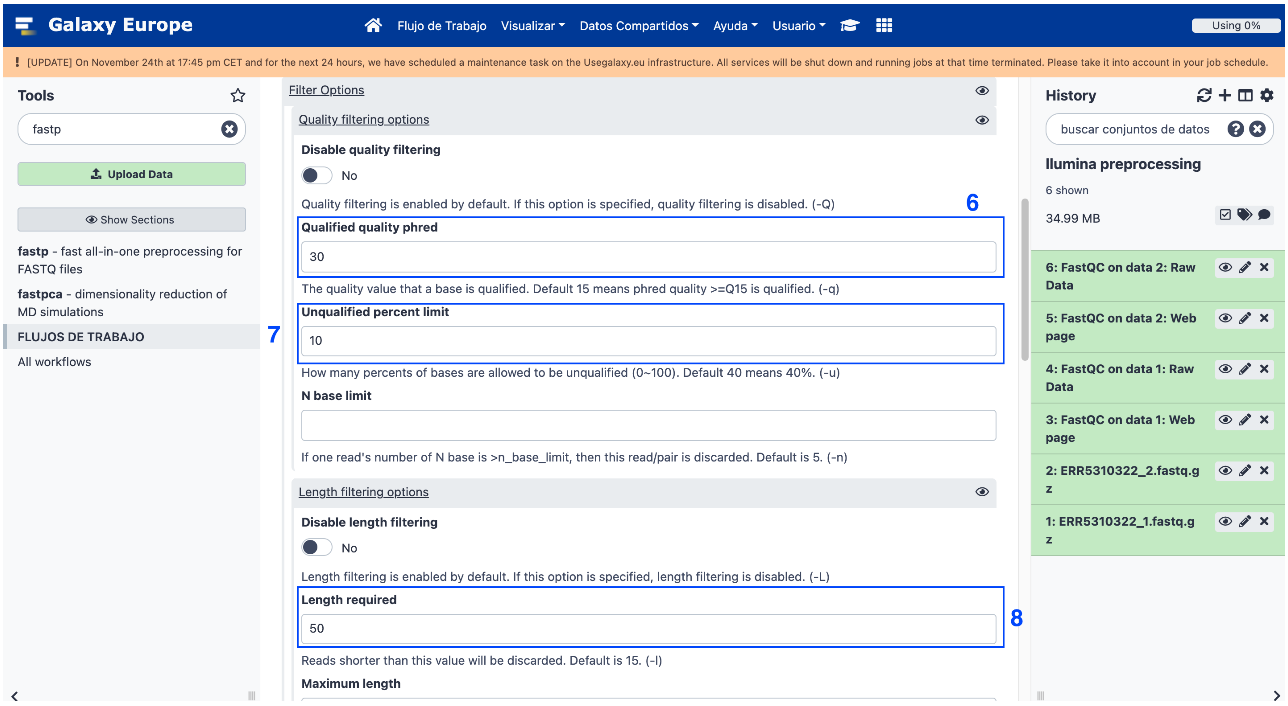
Quality filtering options

Disable quality filtering

No

Quality filtering is enabled by default. If this option is specified, quality filtering is disabled. (-Q)

Qualified quality phred


Galaxy Europe
Flujo de Trabajo
Visualizar
Datos Compartidos
Ayuda
Usuario
Using 0%
Tools
fastp
Upload Data
Show Sections
fastp - fast all-in-one preprocessing for FASTQ files
fastpca - dimensionality reduction of MD simulations
FLUJOS DE TRABAJO
All workflows
6
7
8
History
buscar conjuntos de datos
Ilumina preprocessing
6 shown
34.99 MB
6: FastQC on data 2: Raw Data
5: FastQC on data 2: Web page
4: FastQC on data 1: Raw Data
3: FastQC on data 1: Web page
2: ERR5310322_2.fastq.gz
1: ERR5310322_1.fastq.gz

! [UPDATE] On November 24th at 17:45 pm CET and for the next 24 hours, we have scheduled a maintenance task on the Usegalaxy.eu infrastructure. All services will be shut down and running jobs at that time terminated. Please take it into account in your job schedule.

Filter Options

Quality filtering options

Disable quality filtering

No

Quality filtering is enabled by default. If this option is specified, quality filtering is disabled. (-Q)

Qualified quality phred

30

The quality value that a base is qualified. Default 15 means phred quality >=Q15 is qualified. (-q)

Unqualified percent limit

10

How many percents of bases are allowed to be unqualified (0~100). Default 40 means 40%. (-u)

N base limit

If one read's number of N base is >n_base_limit, then this read/pair is discarded. Default is 5. (-n)

Length filtering options

Disable length filtering

No

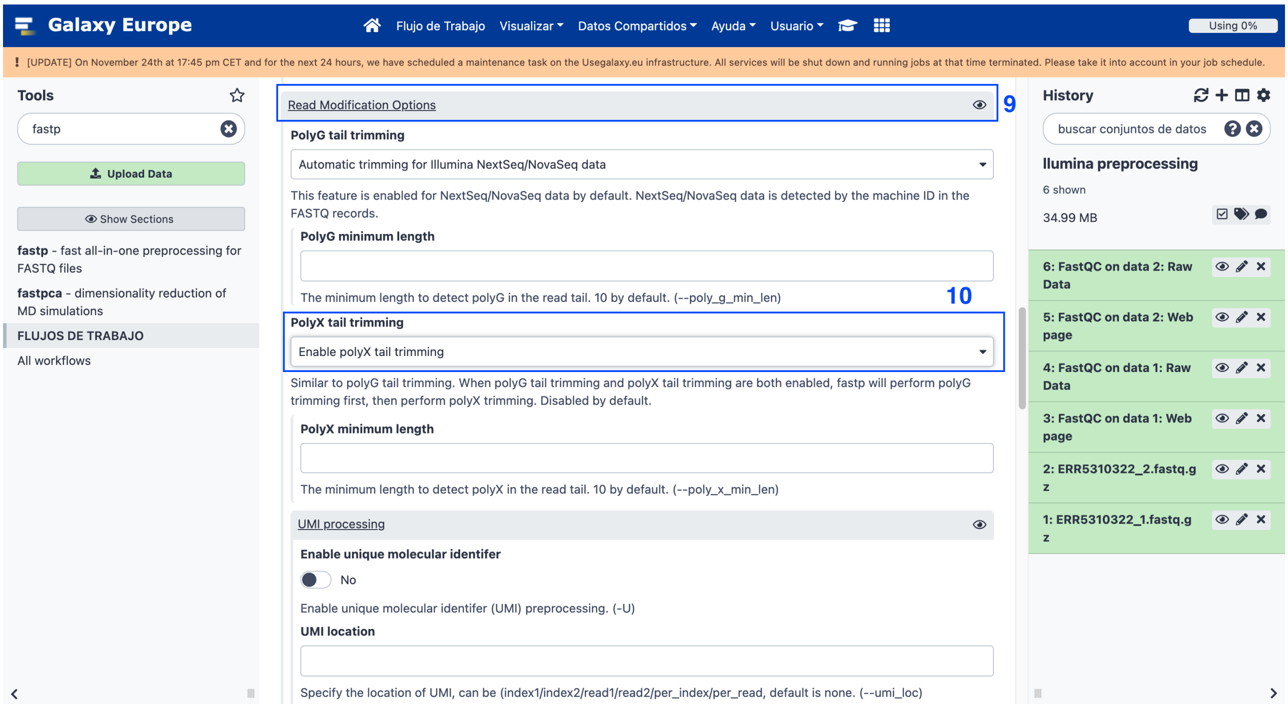
Length filtering is enabled by default. If this option is specified, length filtering is disabled. (-L)

Length required

50

Reads shorter than this value will be discarded. Default is 15. (-l)

Maximum length



! [UPDATE] On November 24th at 17:45 pm CET and for the next 24 hours, we have scheduled a maintenance task on the Usegalaxy.eu infrastructure. All services will be shut down and running jobs at that time terminated. Please take it into account in your job schedule.

fastp - fast all-in-one preprocessing for FASTQ files

fastpca - dimensionality reduction of MD simulations

Read Modification Options

PolyG tail trimming

Automatic trimming for Illumina NextSeq/NovaSeq data

This feature is enabled for NextSeq/NovaSeq data by default. NextSeq/NovaSeq data is detected by the machine ID in the FASTQ records.

PolyG minimum length

The minimum length to detect polyG in the read tail. 10 by default. (--poly_g_min_len)

PolyX tail trimming

Enable polyX tail trimming

Similar to polyG tail trimming. When polyG tail trimming and polyX tail trimming are both enabled, fastp will perform polyG trimming first, then perform polyX trimming. Disabled by default.

PolyX minimum length

The minimum length to detect polyX in the read tail. 10 by default. (--poly_x_min_len)

UMI processing

Enable unique molecular identifer

No

Enable unique molecular identifer (UMI) preprocessing. (-U)

UMI location

Specify the location of UMI, can be (index1/index2/read1/read2/per_index/per_read, default is none. (--umi_loc)

llumina preprocessing

6 shown

34.99 MB

6: FastQC on data 2: Raw Data

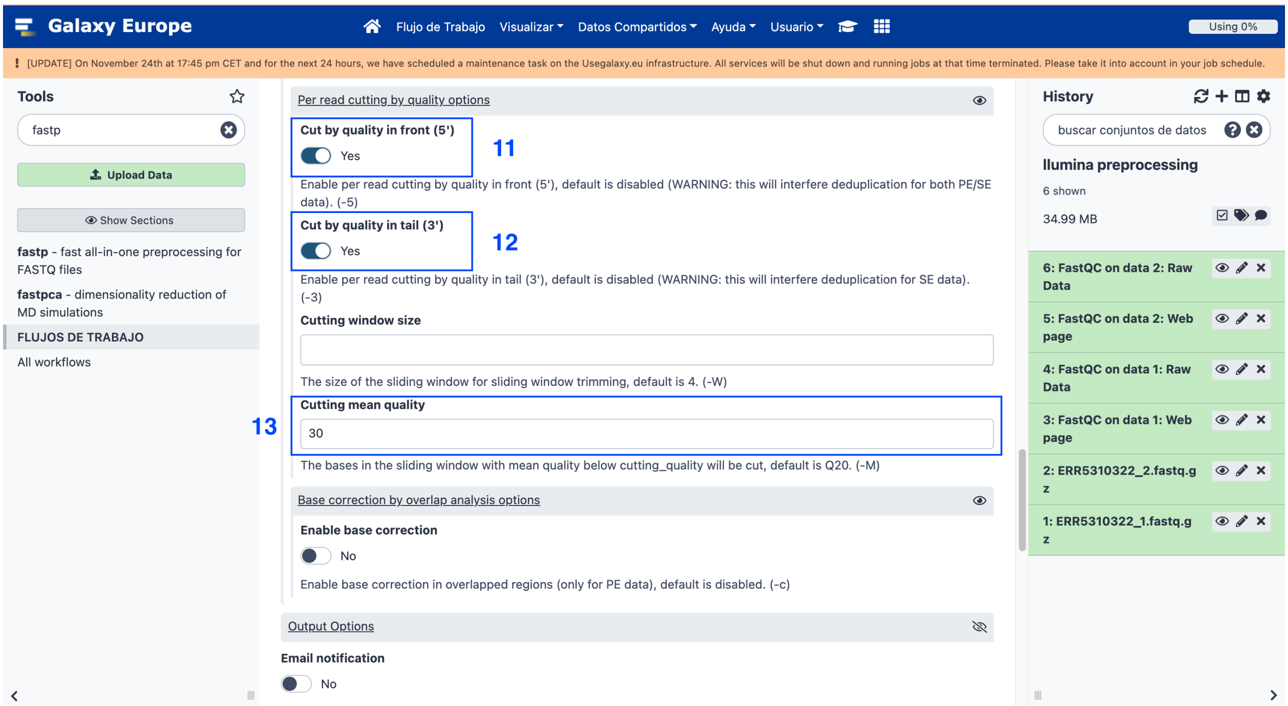
5: FastQC on data 2: Web page

4: FastQC on data 1: Raw Data

3: FastQC on data 1: Web page

2: ERR5310322_2.fastq.gz

1: ERR5310322_1.fastq.gz



! [UPDATE] On November 24th at 17:45 pm CET and for the next 24 hours, we have scheduled a maintenance task on the Usegalaxy.eu infrastructure. All services will be shut down and running jobs at that time terminated. Please take it into account in your job schedule.

fastp - fast all-in-one preprocessing for FASTQ files

fastpca - dimensionality reduction of MD simulations

Per read cutting by quality options

Cut by quality in front (5')

Yes

Enable per read cutting by quality in front (5'), default is disabled (WARNING: this will interfere deduplication for both PE/SE data). (-5)

Cut by quality in tail (3')

Yes

Enable per read cutting by quality in tail (3'), default is disabled (WARNING: this will interfere deduplication for SE data). (-3)

Cutting window size

The size of the sliding window for sliding window trimming, default is 4. (-W)

Cutting mean quality

30

The bases in the sliding window with mean quality below cutting_quality will be cut, default is Q20. (-M)

Base correction by overlap analysis options

Enable base correction

No

Enable base correction in overlapped regions (only for PE data), default is disabled. (-c)

Output Options

Email notification

No

llumina preprocessing

6 shown

34.99 MB

6: FastQC on data 2: Raw Data

5: FastQC on data 2: Web page

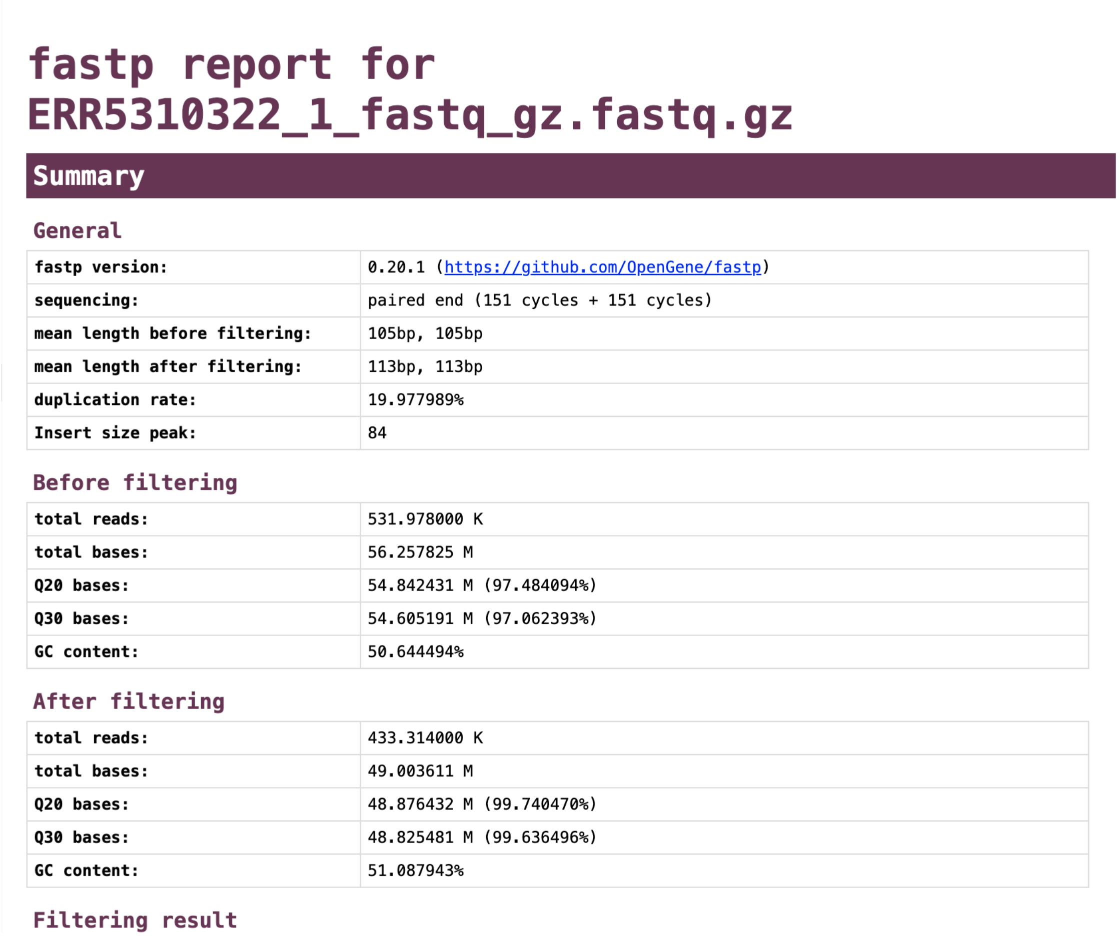
4: FastQC on data 1: Raw Data

3: FastQC on data 1: Web page

2: ERR5310322_2.fastq.gz

1: ERR5310322_1.fastq.gz

To see the trimming stats, have a look at the **fastp on data 2 and data 1: HTML report** file. You should see something like that.



fastp report for ERR5310322_1_fastq_gz.fastq.gz

Summary

General

fastp version:	0.20.1 (https://github.com/OpenGene/fastp)
sequencing:	paired end (151 cycles + 151 cycles)
mean length before filtering:	105bp, 105bp
mean length after filtering:	113bp, 113bp
duplication rate:	19.977989%
Insert size peak:	84

Before filtering

total reads:	531.978000 K
total bases:	56.257825 M
Q20 bases:	54.842431 M (97.484094%)
Q30 bases:	54.605191 M (97.062393%)
GC content:	50.644494%

After filtering

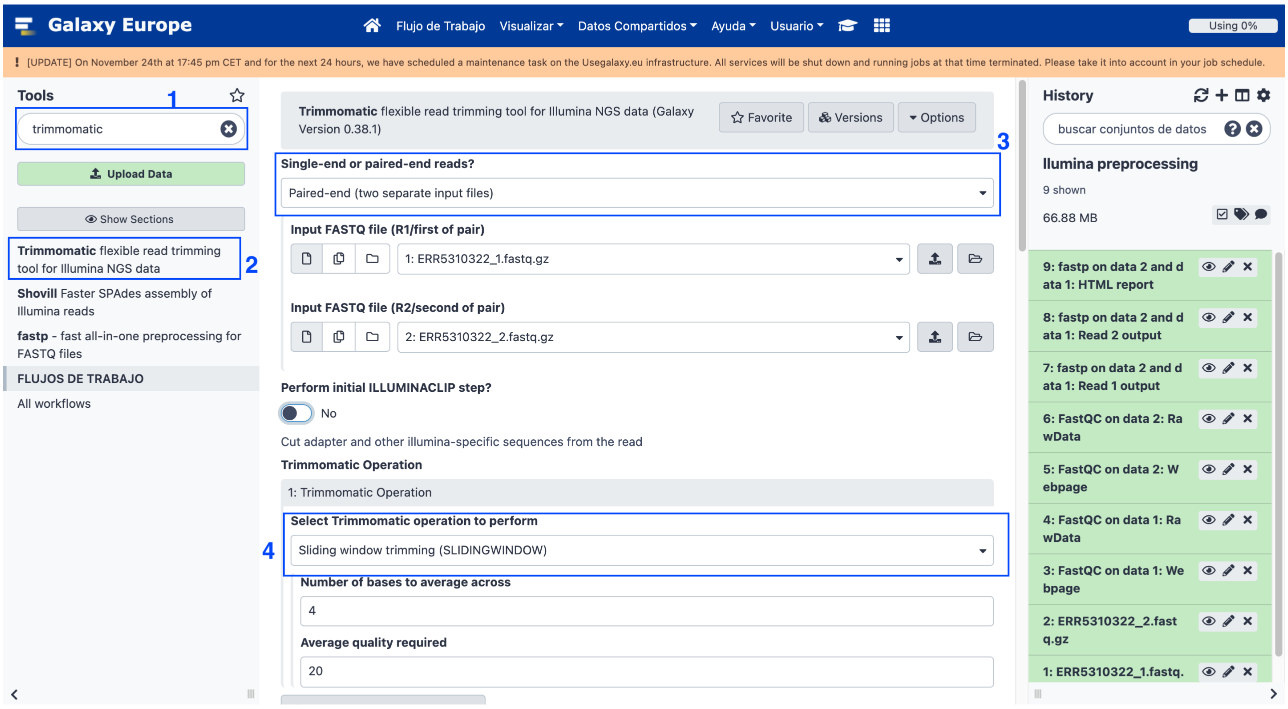
total reads:	433.314000 K
total bases:	49.003611 M
Q20 bases:	48.876432 M (99.740470%)
Q30 bases:	48.825481 M (99.636496%)
GC content:	51.087943%

Filtering result

▶ How many reads have we lost?

Other trimming tools

1. Search for **trimmomatic** in the tools and select **Trimmomatic flexible read trimming tool for Illumina NGS data**
2. Select custom parameters:
   - Single-end or paired-end reads? = Paired-end (two separated files)
   - Input FASTQ file (R1/first of pair) = ERR5310322_1.fastq.gz
   - Input FASTQ file (R2/second of pair) = ERR5310322_2.fastq.gz
   - Insert Trimmomatic Operation:
     - Select Trimmomatic operation to perform: **MINLEN**
     - Minimum length of reads to be kept = 50
3. Select **Execute**



! [UPDATE] On November 24th at 17:45 pm CET and for the next 24 hours, we have scheduled a maintenance task on the Usegalaxy.eu infrastructure. All services will be shut down and running jobs at that time terminated. Please take it into account in your job schedule.

Tools

trimmomatic

Upload Data

Show Sections

Trimmomatic flexible read trimming tool for Illumina NGS data

Shovill Faster SPAdes assembly of Illumina reads

fastp - fast all-in-one preprocessing for FASTQ files

FLUJOS DE TRABAJO

All workflows

Trimmomatic flexible read trimming tool for Illumina NGS data (Galaxy Version 0.38.1)

Single-end or paired-end reads?
Paired-end (two separate input files)

Input FASTQ file (R1/first of pair)
1: ERR5310322_1.fastq.gz

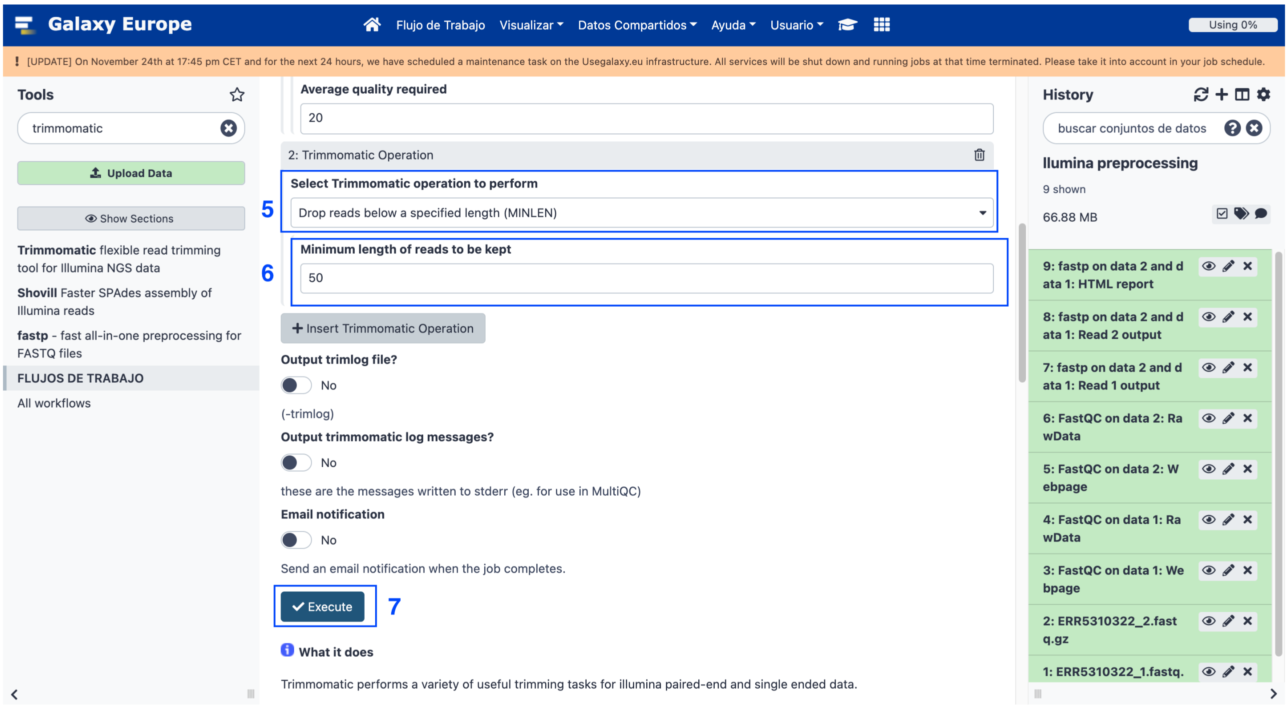
Input FASTQ file (R2/second of pair)
2: ERR5310322_2.fastq.gz

Perform initial ILLUMINACLIP step?
No
Cut adapter and other illumina-specific sequences from the read

Trimmomatic Operation
1: Trimmomatic Operation
Select Trimmomatic operation to perform
Sliding window trimming (SLIDINGWINDOW)
Number of bases to average across
4
Average quality required
20

llumina preprocessing
9 shown
66.88 MB

9: fastp on data 2 and data 1: HTML report
8: fastp on data 2 and data 1: Read 2 output
7: fastp on data 2 and data 1: Read 1 output
6: FastQC on data 2: RawData
5: FastQC on data 2: Webpage
4: FastQC on data 1: RawData
3: FastQC on data 1: Webpage
2: ERR5310322_2.fastq.gz
1: ERR5310322_1.fastq.



! [UPDATE] On November 24th at 17:45 pm CET and for the next 24 hours, we have scheduled a maintenance task on the Usegalaxy.eu infrastructure. All services will be shut down and running jobs at that time terminated. Please take it into account in your job schedule.

Tools

trimmomatic

Upload Data

Show Sections

Trimmomatic flexible read trimming tool for Illumina NGS data

Shovill Faster SPAdes assembly of Illumina reads

fastp - fast all-in-one preprocessing for FASTQ files

FLUJOS DE TRABAJO

All workflows

Average quality required
20

2: Trimmomatic Operation
Select Trimmomatic operation to perform
Drop reads below a specified length (MINLEN)
Minimum length of reads to be kept
50

+ Insert Trimmomatic Operation

Output trimlog file?
No
(-trimlog)

Output trimmomatic log messages?
No
these are the messages written to stderr (eg. for use in MultiQC)

Email notification
No
Send an email notification when the job completes.

✓ Execute

What it does

Trimmomatic performs a variety of useful trimming tasks for illumina paired-end and single ended data.

llumina preprocessing
9 shown
66.88 MB

9: fastp on data 2 and data 1: HTML report
8: fastp on data 2 and data 1: Read 2 output
7: fastp on data 2 and data 1: Read 1 output
6: FastQC on data 2: RawData
5: FastQC on data 2: Webpage
4: FastQC on data 1: RawData
3: FastQC on data 1: Webpage
2: ERR5310322_2.fastq.gz
1: ERR5310322_1.fastq.

Trimmomatic does not perform statistics over trimmed reads, so we need to perform FastQC again over the Trimmomatic results.

▶ Try to do it on your own.

Second question

▶ How can I improve the quality of my data?

- This hands-on history URL: https://usegalaxy.eu/u/svarona/h/llumina-preprocessing

2. Nanopore Quality control and trimming

Title	Galaxy

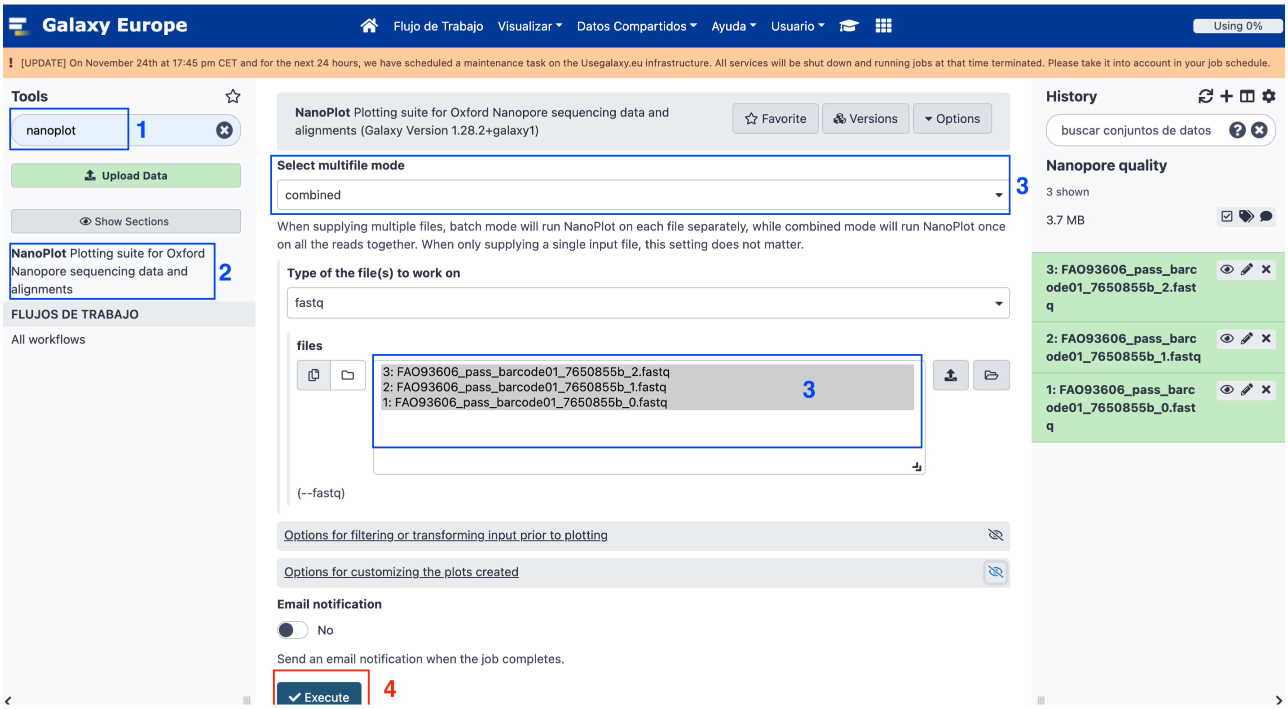
Title	**Galaxy**
Training dataset:	The data we are going to manage corresponds to Nanopore amplicon sequencing data using ARTIC network primers por SARS-CoV-2 genome. From the Fast5 files generated by the ONT software, we are going to select the pass reads, so they are already filtered by quality.
Questions:	• How do I know if my Nanopore data was correctly sequenced?
Objectives:	• Perform a quality control in raw Illumina reads • Perform a quality trimming in raw Nanopore reads • Perform a quality control in trimmed Nanopore reads
Estimated time:	15 min

1. Quality control

To run the quality control over the samples, follow these steps: 1. [Create a new history has explained yesterday] named **Nanopore quality** 2. [Upload data as seen yesterday], copy and paste the following URLs:

```
https://raw.githubusercontent.com/nf-core/test-datasets/viralrec
https://raw.githubusercontent.com/nf-core/test-datasets/viralrec
https://raw.githubusercontent.com/nf-core/test-datasets/viralrec
```

1. Search for the **Nanoplot** tool and select **NanoPlot Plotting suite for Oxford Nanopore sequencing data and alignments**
2. Run the tool as follows:
   - In *Select multifile mode*: **Combined** (as we are working with 3 different fastq files for the same sample, we can analyze them in batch)
   - In the *files* part, use *Ctrl* to select the three fastq files.
   - Select **Execute**



! [UPDATE] On November 24th at 17:45 pm CET and for the next 24 hours, we have scheduled a maintenance task on the Usegalaxy.eu infrastructure. All services will be shut down and running jobs at that time terminated. Please take it into account in your job schedule.

nanoplot

NanoPlot Plotting suite for Oxford Nanopore sequencing data and alignments

NanoPlot Plotting suite for Oxford Nanopore sequencing data and alignments (Galaxy Version 1.28.2+galaxy1)

Select multifile mode
combined
When supplying multiple files, batch mode will run NanoPlot on each file separately, while combined mode will run NanoPlot once on all the reads together. When only supplying a single input file, this setting does not matter.

Type of the file(s) to work on
fastq

files
3: FAO93606_pass_barcode01_7650855b_2.fastq
2: FAO93606_pass_barcode01_7650855b_1.fastq
1: FAO93606_pass_barcode01_7650855b_0.fastq
(--fastq)

Options for filtering or transforming input prior to plotting

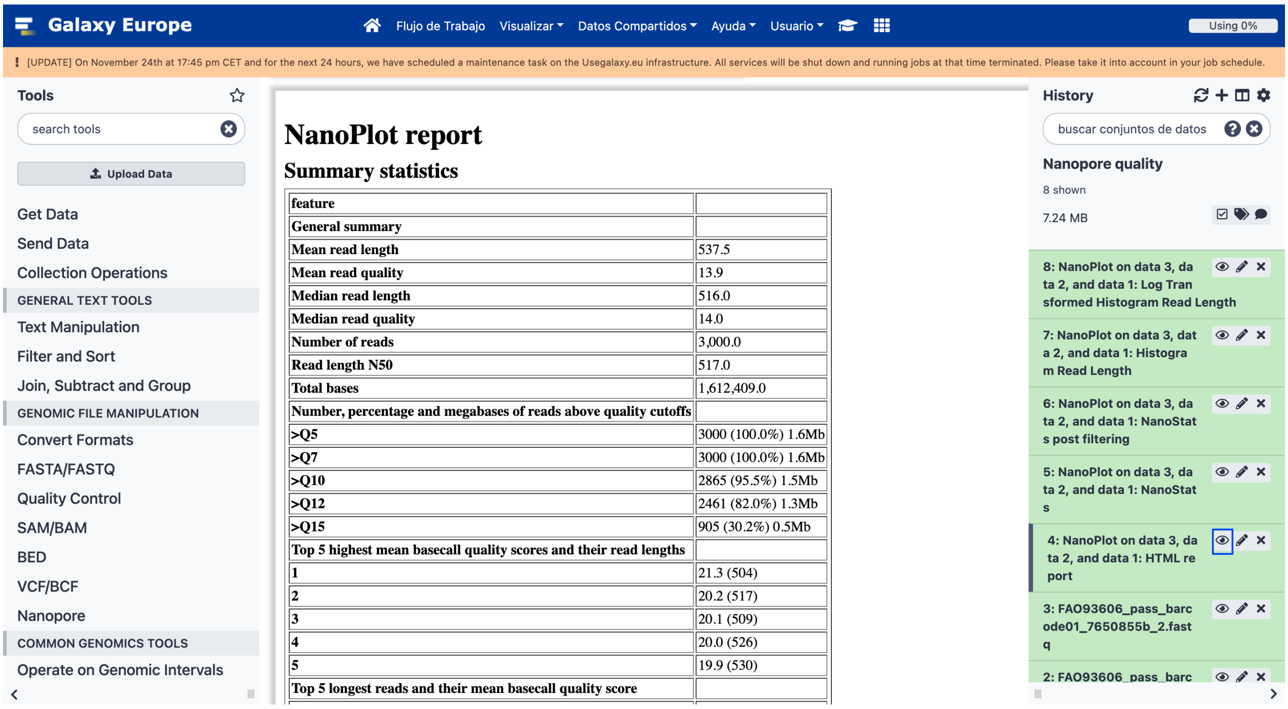
Options for customizing the plots created

Email notification
No
Send an email notification when the job completes.

Nanopore quality
3 shown
3.7 MB
3: FAO93606_pass_barcode01_7650855b_2.fastq
2: FAO93606_pass_barcode01_7650855b_1.fastq
1: FAO93606_pass_barcode01_7650855b_0.fastq

Now we are going to have a look to the results.

1. Select the :eye: icon in the **NanoPlot on data 3, data 2, and data 1: HTML report** result.
2. Have a look to the stats.


Galaxy Europe
Flujo de Trabajo
Visualizar
Datos Compartidos
Ayuda
Usuario
Using 0%
Tools
search tools
Upload Data
Get Data
Send Data
Collection Operations
GENERAL TEXT TOOLS
Text Manipulation
Filter and Sort
Join, Subtract and Group
GENOMIC FILE MANIPULATION
Convert Formats
FASTA/FASTQ
Quality Control
SAM/BAM
BED
VCF/BCF
Nanopore
COMMON GENOMICS TOOLS
Operate on Genomic Intervals
History
buscar conjuntos de datos
Nanopore quality
8 shown
7.24 MB
8: NanoPlot on data 3, data 2, and data 1: Log Transformed Histogram Read Length
7: NanoPlot on data 3, data 2, and data 1: Histogram Read Length
6: NanoPlot on data 3, data 2, and data 1: NanoStats post filtering
5: NanoPlot on data 3, data 2, and data 1: NanoStats
4: NanoPlot on data 3, data 2, and data 1: HTML report
3: FAO93606_pass_barcode01_7650855b_2.fastq
2: FAO93606_pass_barc

! [UPDATE] On November 24th at 17:45 pm CET and for the next 24 hours, we have scheduled a maintenance task on the Usegalaxy.eu infrastructure. All services will be shut down and running jobs at that time terminated. Please take it into account in your job schedule.

NanoPlot report

Summary statistics

feature	
General summary	
Mean read length	537.5
Mean read quality	13.9
Median read length	516.0
Median read quality	14.0
Number of reads	3,000.0
Read length N50	517.0
Total bases	1,612,409.0
Number, percentage and megabases of reads above quality cutoffs	
>Q5	3000 (100.0%) 1.6Mb
>Q7	3000 (100.0%) 1.6Mb
>Q10	2865 (95.5%) 1.5Mb
>Q12	2461 (82.0%) 1.3Mb
>Q15	905 (30.2%) 0.5Mb
Top 5 highest mean basecall quality scores and their read lengths	
1	21.3 (504)
2	20.2 (517)
3	20.1 (509)
4	20.0 (526)
5	19.9 (530)
Top 5 longest reads and their mean basecall quality score	

As you can see, the Mean read length is around 500 nt, which makes sense because we are using amplicon sequencing data.

▶ How many reads do the samples have?

First question

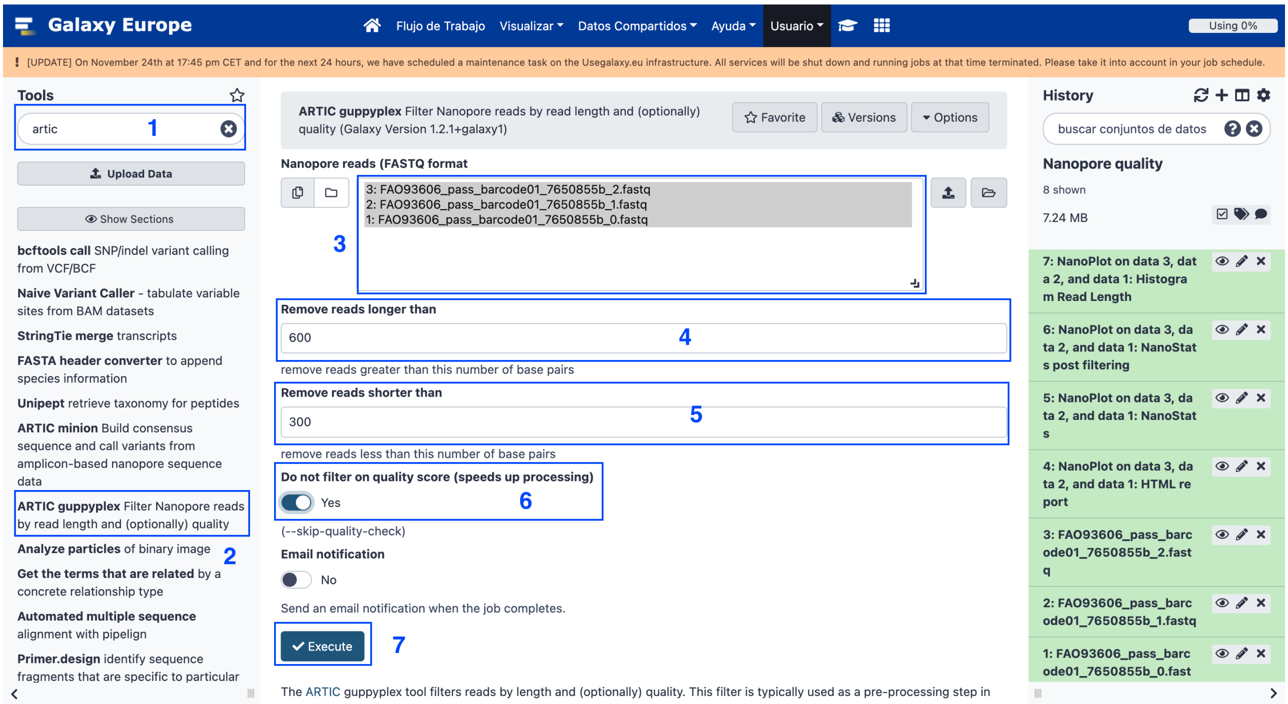
▶ How do I check whether my Nanopore data was correctly sequenced?

2. Trimming

When Nanopore reads are being sequenced, the MinKnown software splits Fast5 reads into quality pass and quality fail. As we will select only Fast5 pass reads, we won't need to perform a quality trimming, so even if we see that the reads have a bad Phred score, we know that the ONT software considered the reads as "good quality".

Then we will only be performing a read length trimming. As we are using amplicon sequencing data, we won't be expecting reads smaller than 400 nucleotides, nor higher than 600, which would obviously correspond to chimeric reads.

1. Search for **artic** tool
2. Select **ARTIC guppyplex Filter Nanopore reads by read length and (optionally) quality**
3. While pressing the *Ctrl* key, select the three samples
4. Remove reads longer than = 600
5. Remove reads shorter than = 400
6. Do not filter on quality score (speeds up processing) = Yes (we had already select pass reads)



! [UPDATE] On November 24th at 17:45 pm CET and for the next 24 hours, we have scheduled a maintenance task on the Usegalaxy.eu infrastructure. All services will be shut down and running jobs at that time terminated. Please take it into account in your job schedule.

Tools search: artic

bcftools call SNP/indel variant calling from VCF/BCF

Naive Variant Caller - tabulate variable sites from BAM datasets

StringTie merge transcripts

FASTA header converter to append species information

Unipept retrieve taxonomy for peptides

ARTIC minion Build consensus sequence and call variants from amplicon-based nanopore sequence data

ARTIC guppyplex Filter Nanopore reads by read length and (optionally) quality

Analyze particles of binary image

Get the terms that are related by a concrete relationship type

Automated multiple sequence alignment with pipelign

Primer.design identify sequence fragments that are specific to particular

ARTIC guppyplex Filter Nanopore reads by read length and (optionally) quality (Galaxy Version 1.2.1+galaxy1)

Nanopore reads (FASTQ format
3: FAO93606_pass_barcode01_7650855b_2.fastq
2: FAO93606_pass_barcode01_7650855b_1.fastq
1: FAO93606_pass_barcode01_7650855b_0.fastq

Remove reads longer than
600
remove reads greater than this number of base pairs

Remove reads shorter than
300
remove reads less than this number of base pairs

Do not filter on quality score (speeds up processing)
Yes
(--skip-quality-check)

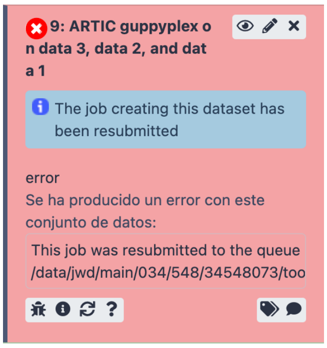
Email notification
No
Send an email notification when the job completes.

The ARTIC guppyplex tool filters reads by length and (optionally) quality. This filter is typically used as a pre-processing step in

Nanopore quality
8 shown
7.24 MB

7: NanoPlot on data 3, data 2, and data 1: Histogram Read Length
6: NanoPlot on data 3, data 2, and data 1: NanoStats post filtering
5: NanoPlot on data 3, data 2, and data 1: NanoStats
4: NanoPlot on data 3, data 2, and data 1: HTML report
3: FAO93606_pass_barcode01_7650855b_2.fastq
2: FAO93606_pass_barcode01_7650855b_1.fastq
1: FAO93606_pass_barcode01_7650855b_0.fast

We will come across one error in this job:



9: ARTIC guppyplex on data 3, data 2, and data 1

The job creating this dataset has been resubmitted

error
Se ha producido un error con este conjunto de datos:

This job was resubmitted to the queue
/data/jwd/main/034/548/34548073/too

This happens because Galaxy does not have the software to filter SARS-CoV-2 amplicon data properly installed in their server, which is something typical that we can find in Galaxy.

- This hands-on history URL: https://usegalaxy.eu/u/svarona/h/nanopore-quality