Galaxy for virologist training Exercise 2: Quality control and trimming

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

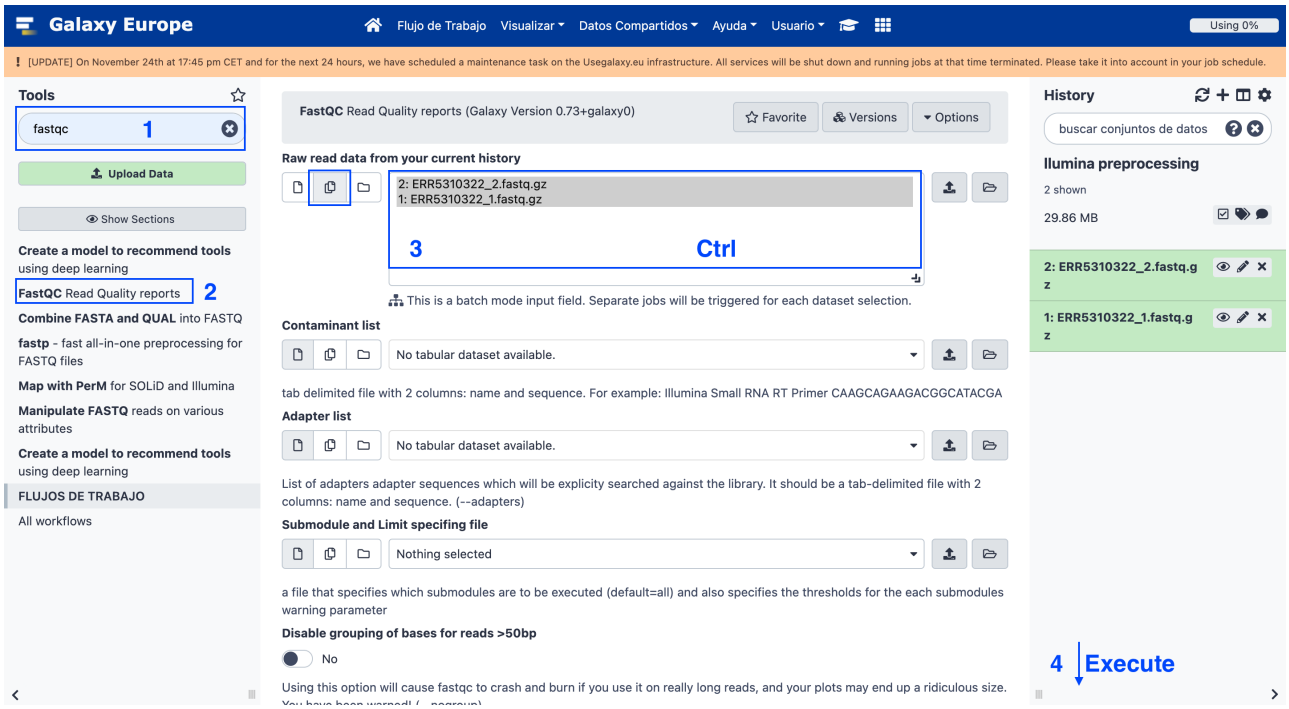
1. Quality control

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:

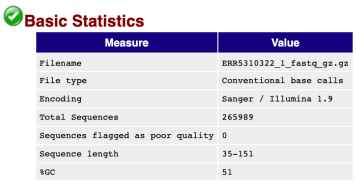
```
ftp://ftp.sra.ebi.ac.uk/vol1/fastq/ERR531/002/ERR5310322/ERR5310
ftp://ftp.sra.ebi.ac.uk/vol1/fastq/ERR531/002/ERR5310322/ERR5310
```

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**

Table of Contents

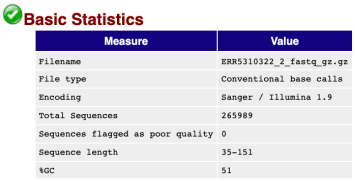


To see the results we are going to open the jobs with **Web page** in their name for both data 1 and data 2.



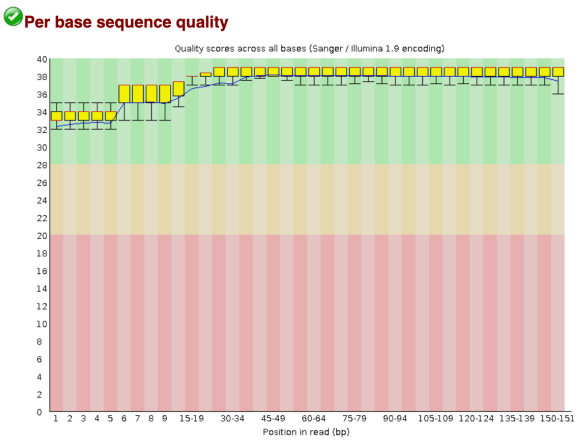
Basic Statistics

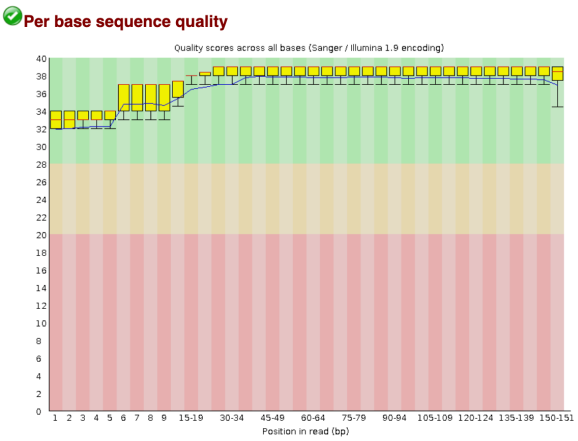
Measure	Value
Filename	ERR5310322_1_fastq_gz.gz
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	265989
Sequences flagged as poor quality	0
Sequence length	35-151
%GC	51



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

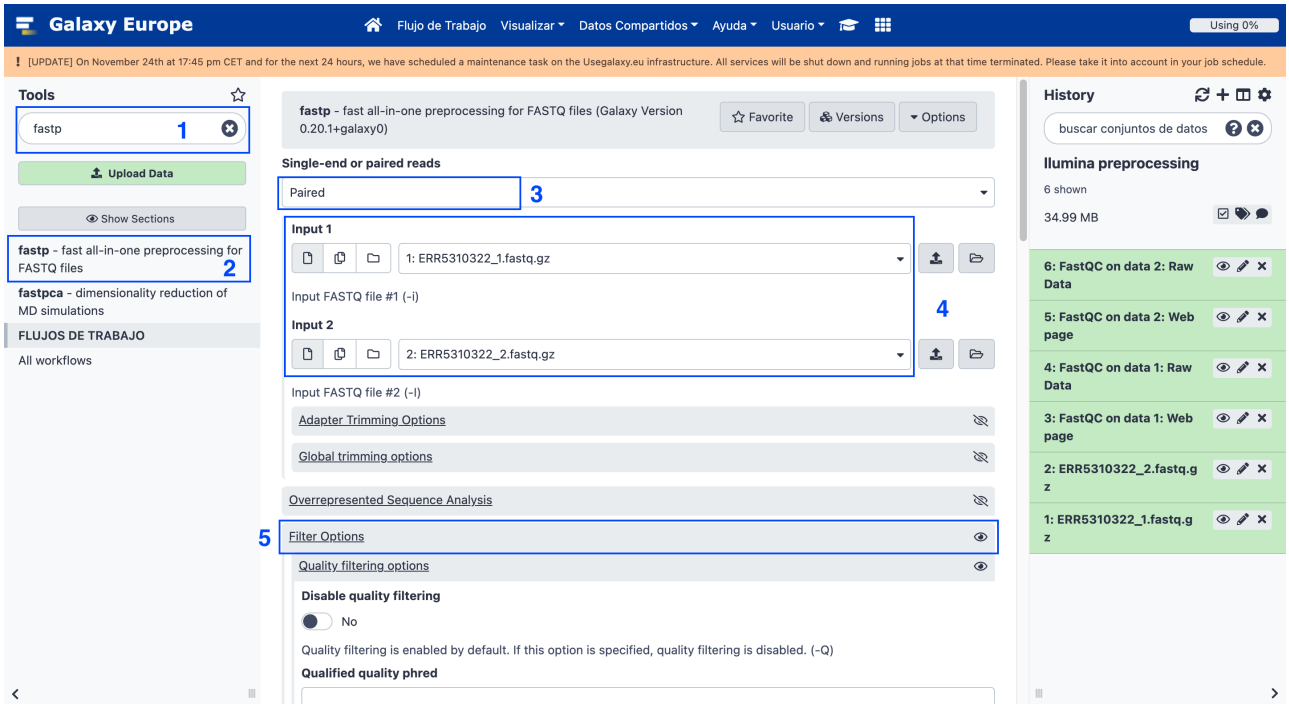
▶ How do I check whether my Illumina data was correctly sequenced?

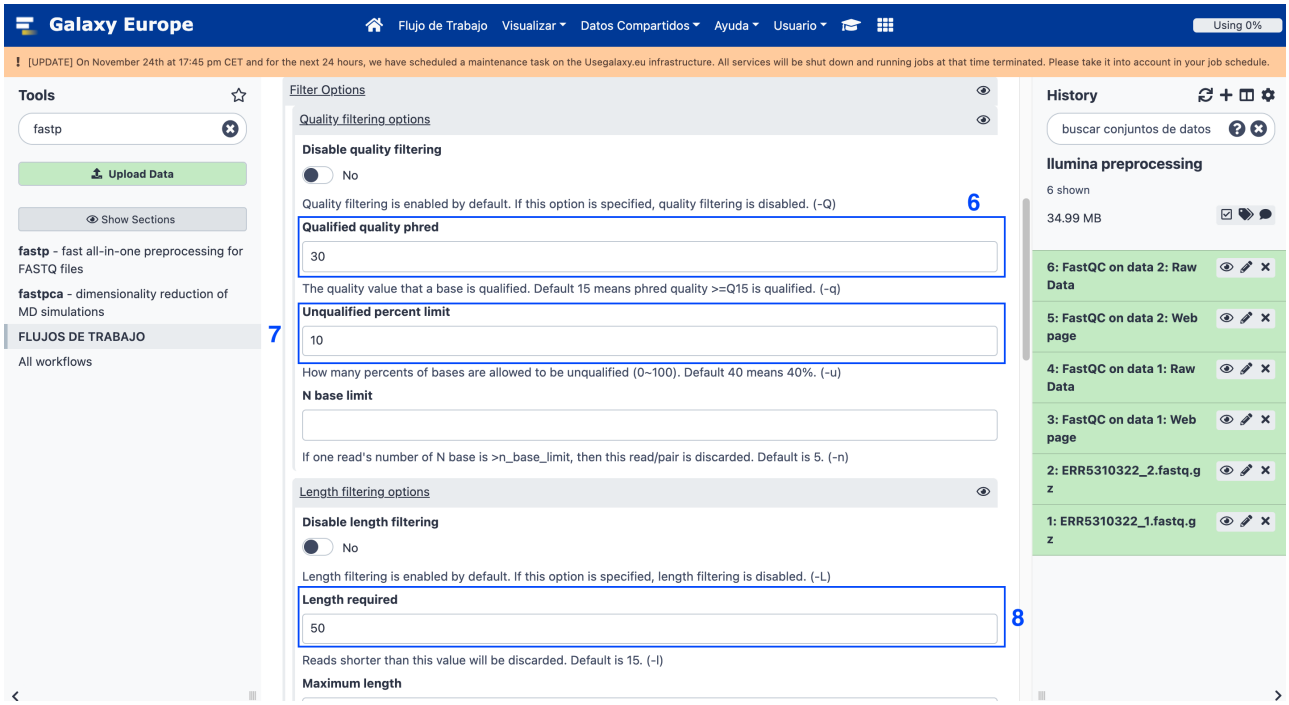
2. Trimming

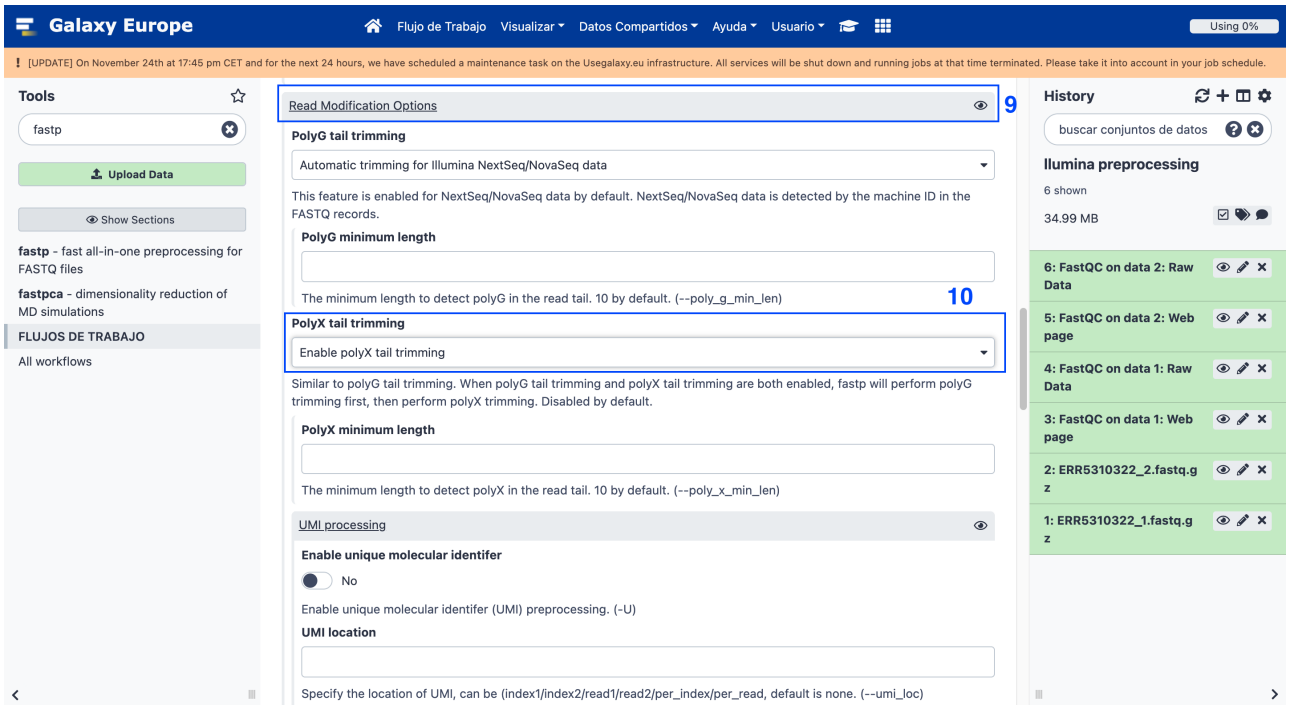
Once we have performed the quality control, we have to perform the quality and read length trimming:

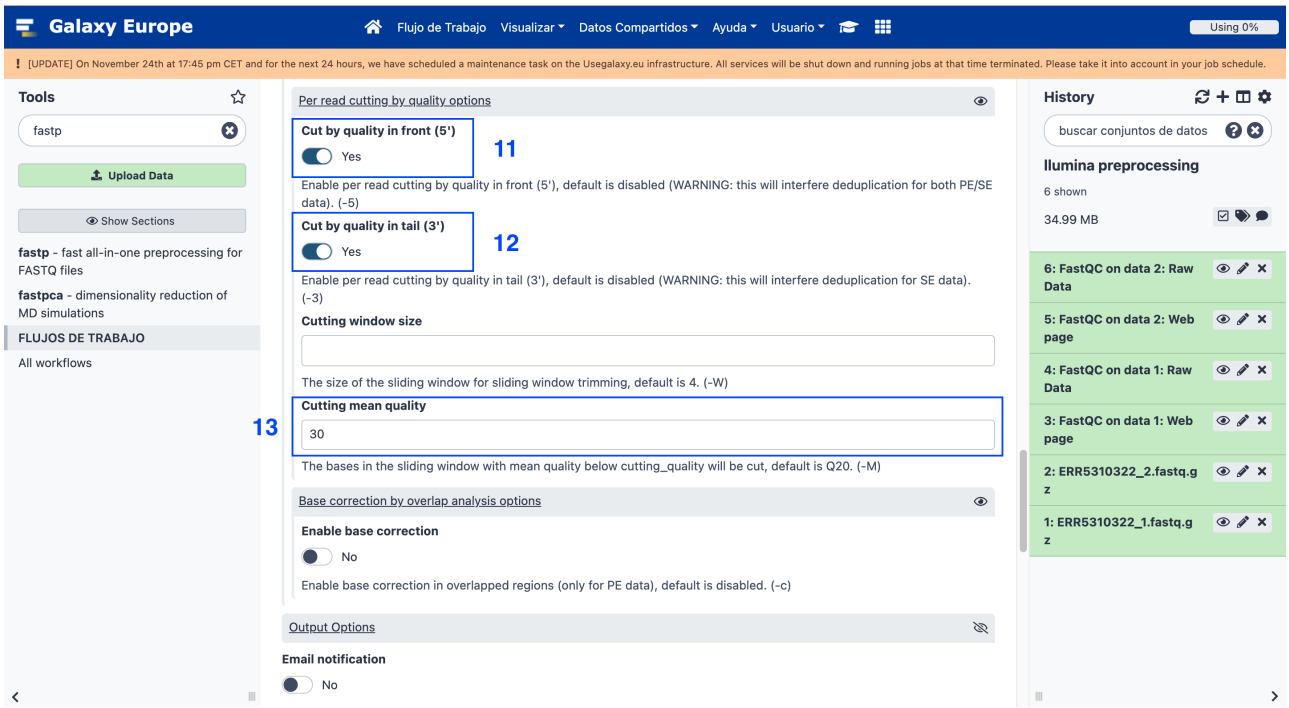
1. Search for **fastp** in the tools and select **fastp - fast all-in-one preprocessing for FASTQ files**

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**









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

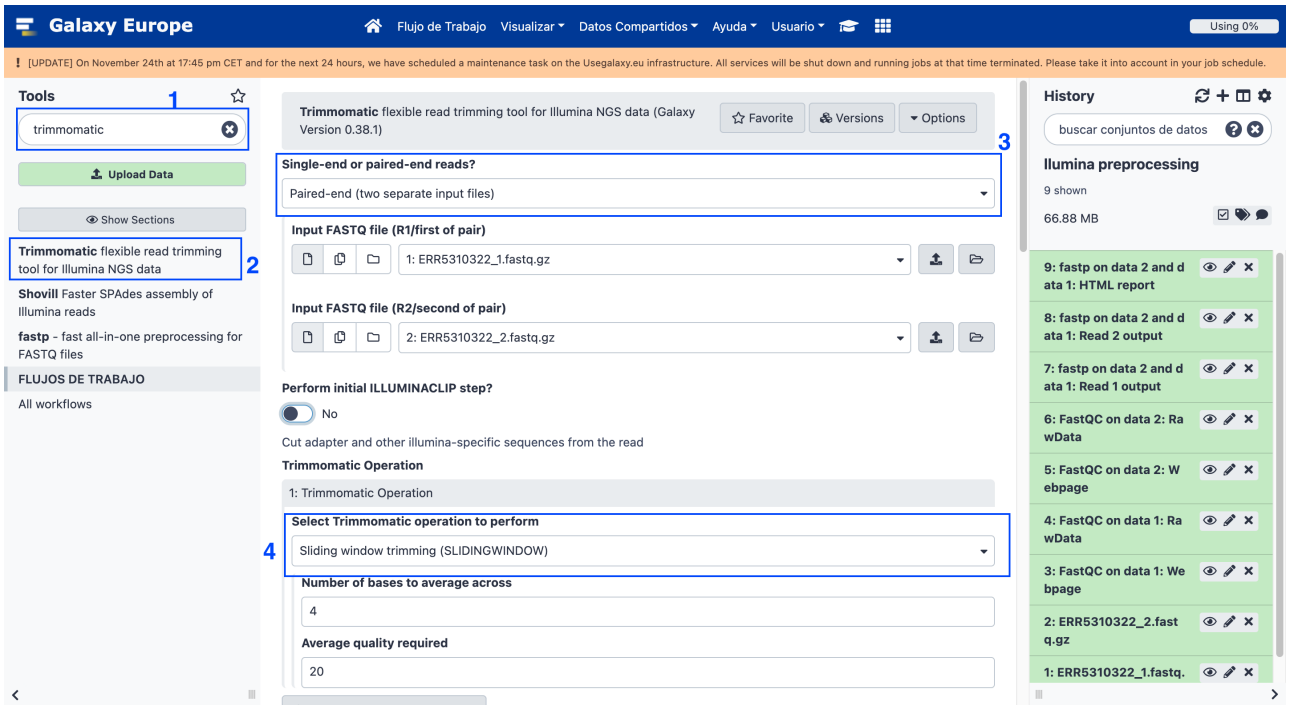
total reads:	433.314000 K
total bases:	49.003611 M
Q20 bases:	48.876432 M (99.740470%)
Q30 bases:	48.825481 M (99.636496%)
GC content:	51.087943%

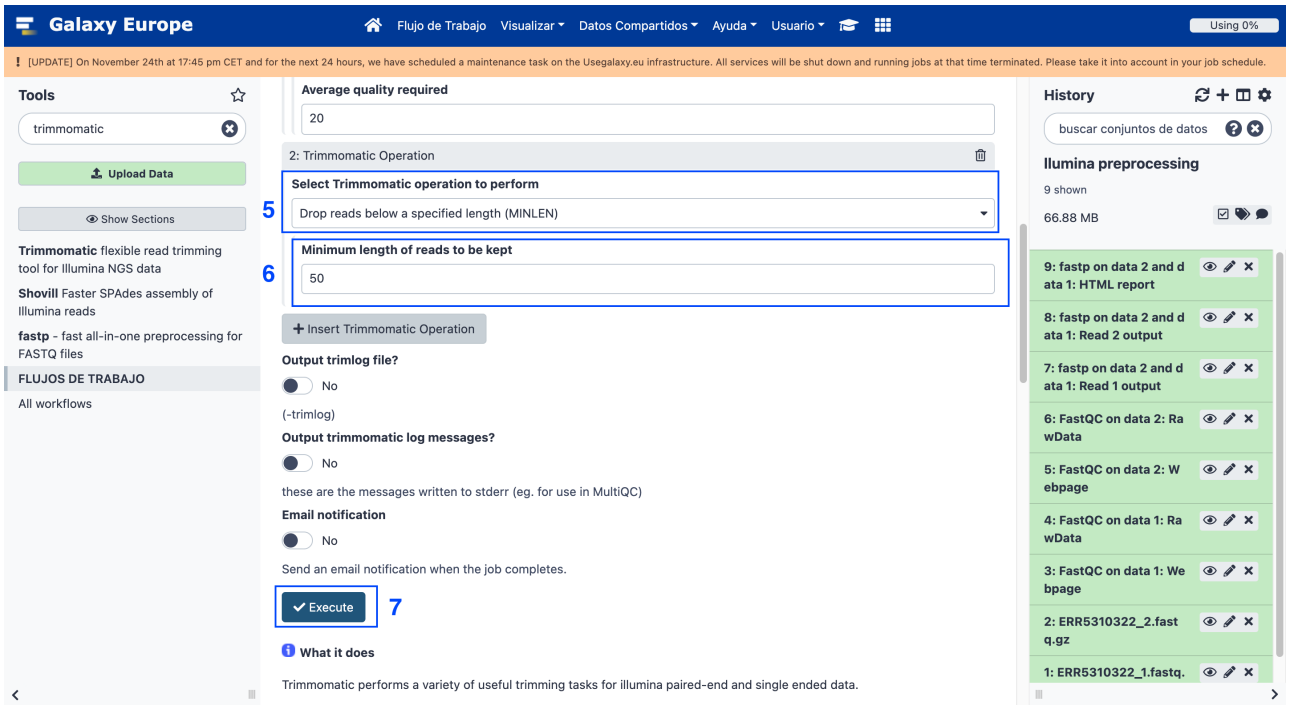
Filtering result

▶ 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**





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/s.varona/h/illumina-preprocessing

2. Nanopore Quality control and trimming

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.

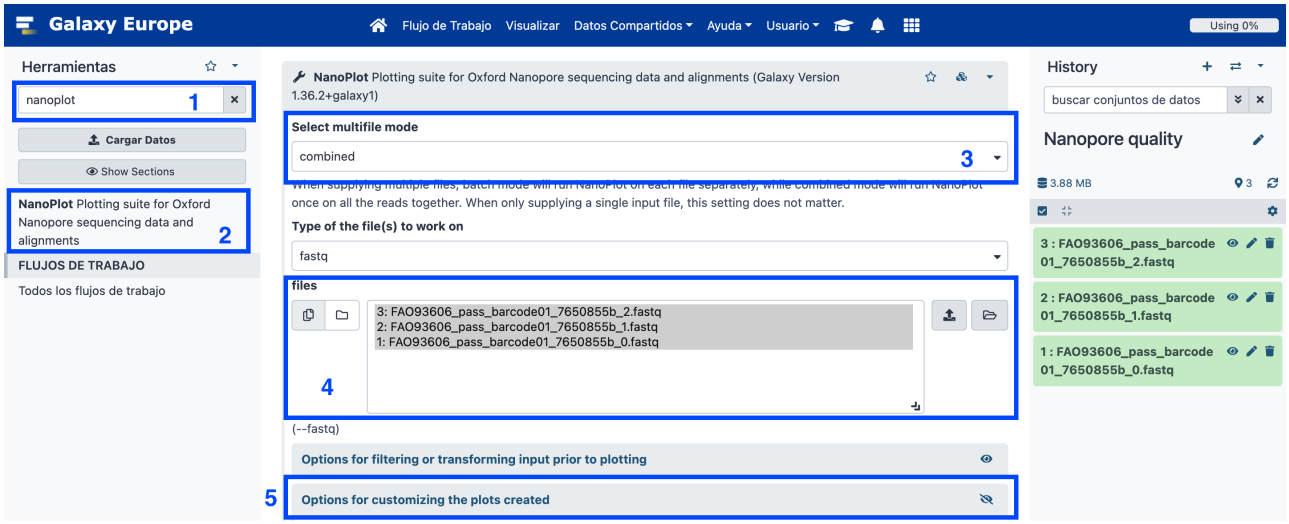
Title	Galaxy
Questions:	• How do I know if my Nanopore data was correctly sequenced?
Objectives:	• Perform a quality control in raw Illumina reads • Perform a quality trimming in raw Nanopore reads • Perform a quality control in trimmed Nanopore reads
Estimated time:	15 min

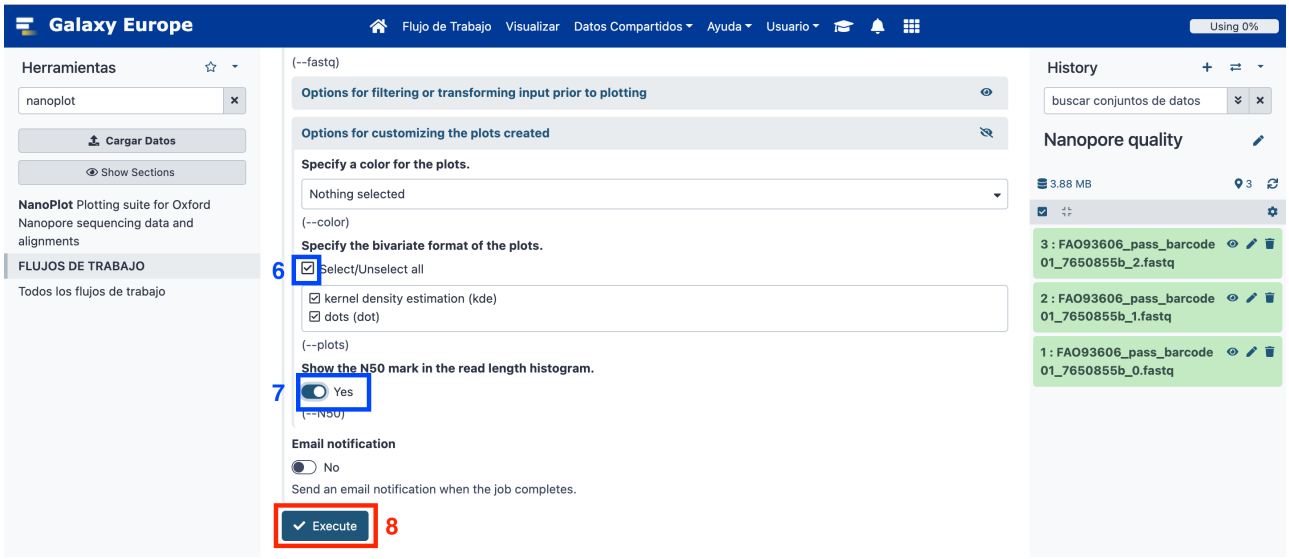
1. Quality control

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
```

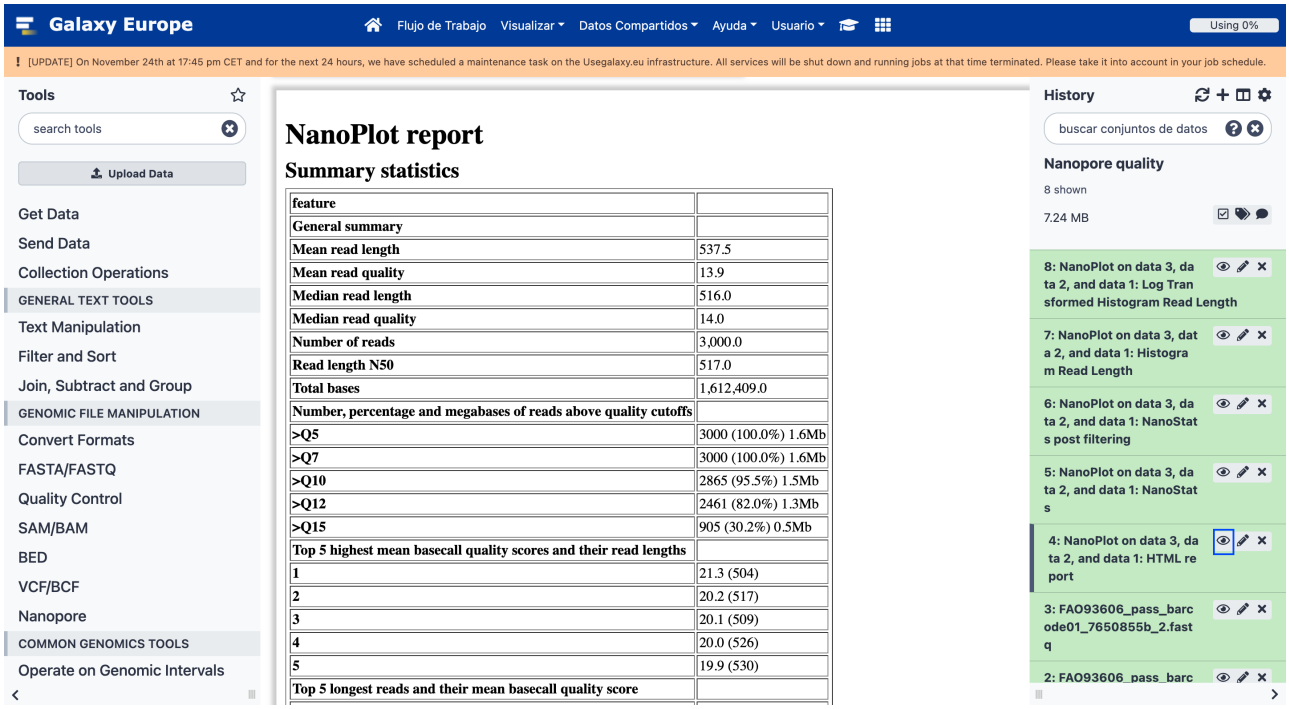
1. Search for the **Nanoplot** tool and select **NanoPlot Plotting suite for Oxford Nanopore sequencing data and alignments**
2. Run the tool as follows:
    - In *Select multifile mode*: **Combined** (as we are working with 3 different fastq files for the same sample, we can analyze them in batch)
    - In the *files* part, use *Ctrl* to select the three fastq files.
    - Display **Options for customizing the plots created**:
        - **Specify the bivariate format of the plots** > *Select all*
        - **Show the N50 mark in the read length histogram** > *Yes*
    - Select **Execute**





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.



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?

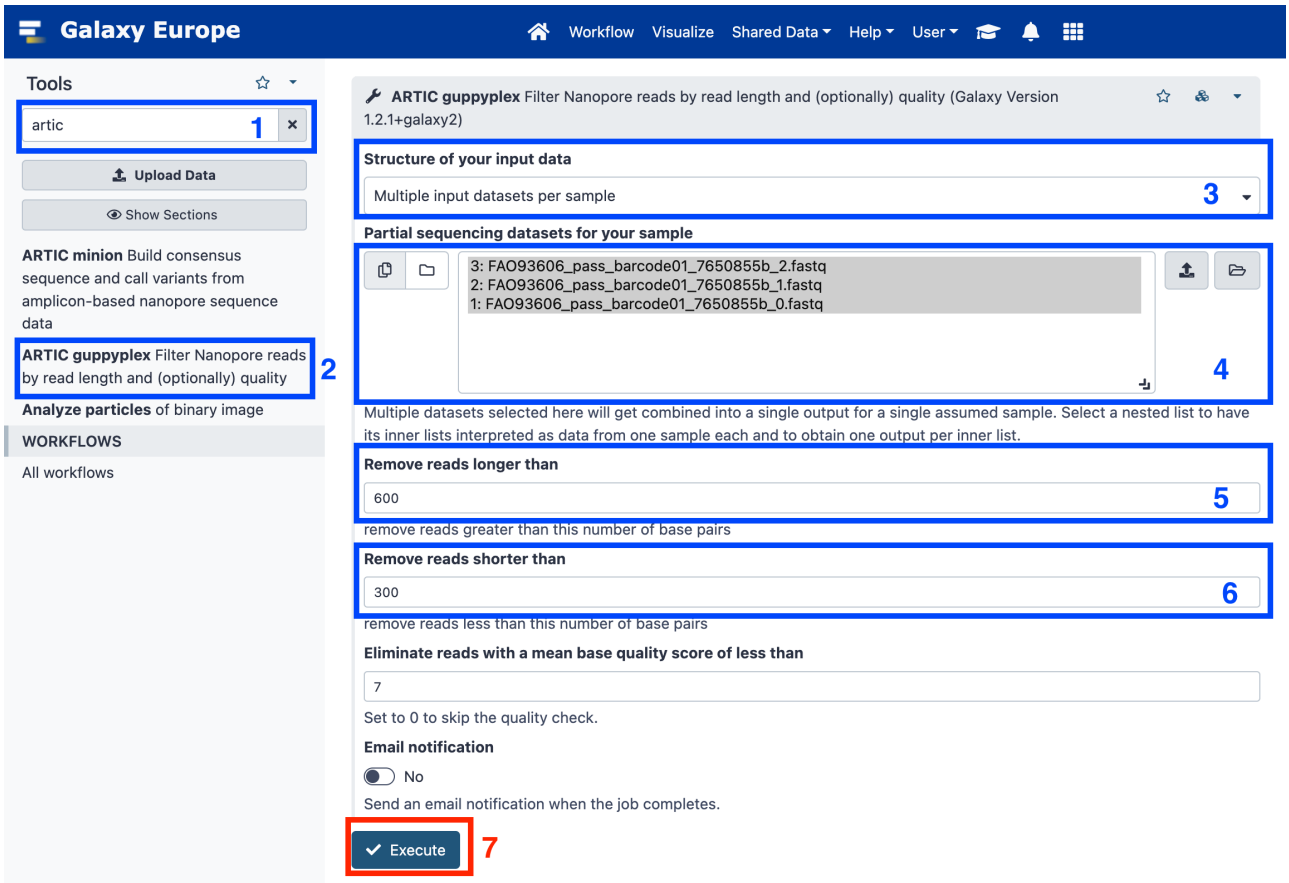
2. Trimming

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

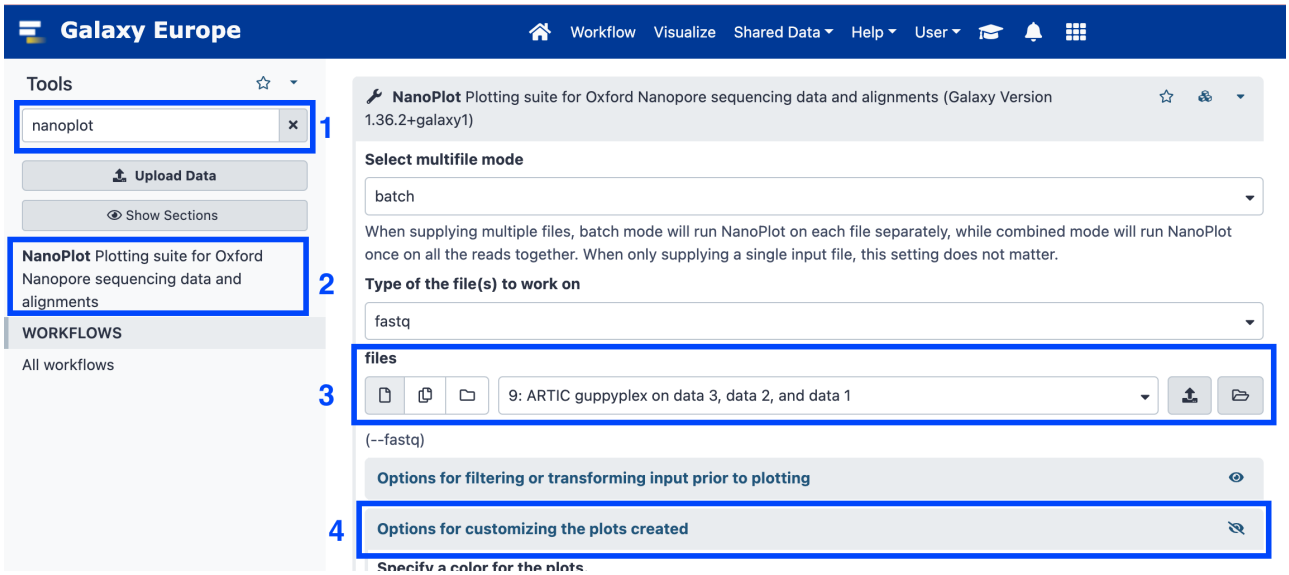
reads.

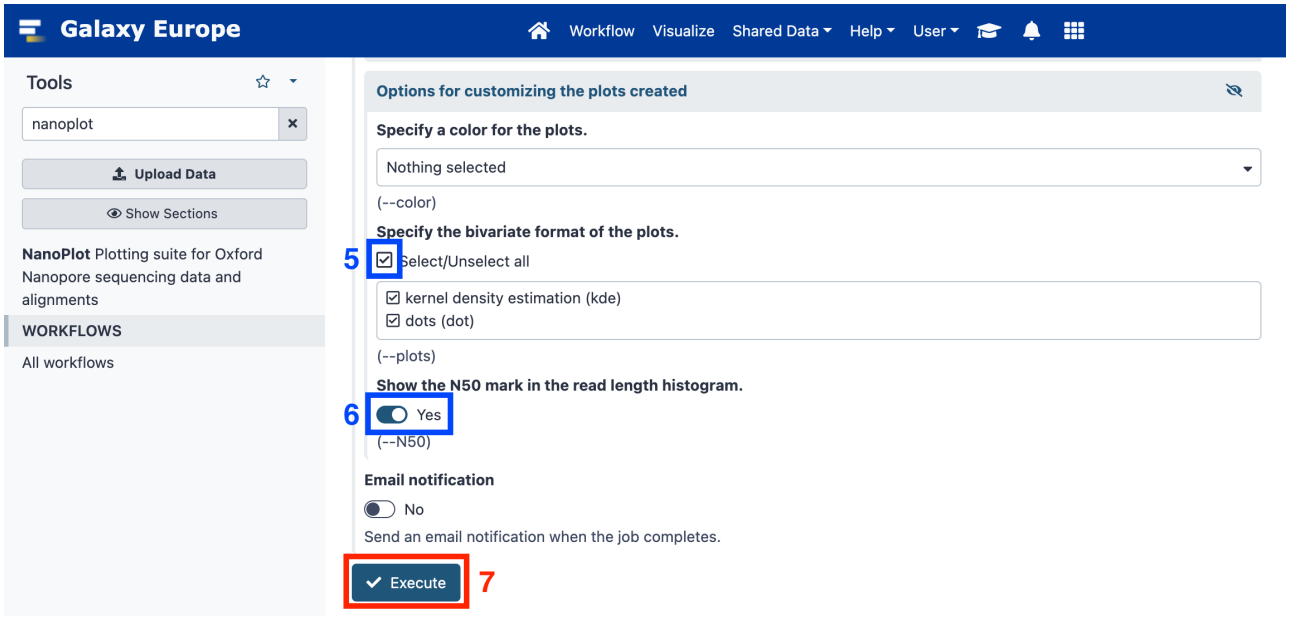
1. Search for **artic** tool
2. Select **ARTIC guppyplex Filter Nanopore reads by read length and (optionally) quality**
3. While pressing the *Ctrl* key, select the three samples
4. Remove reads longer than = 600
5. Remove reads shorter than = 300
6. Do not filter on quality score (speeds up processing) = Yes (we had already select pass reads)



Now we are going to repeat NanoPlot on filtered data:

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 the *files* part, select ARTIC output file.
    - Display **Options for customizing the plots created**:
        - **Specify the bivariate format of the plots** > *Select all*
        - **Show the N50 mark in the read length histogram** > *Yes*
    - Select **Execute**





Questions

▸ Did our data length and quality improve?

▸ How many reads did we lost during trimming step?

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

> ***NOTE:*** We can't use PycoQC because it needs MinION `sequencing_summary.txt` file which we don't have.
>
> ***NOTE:*** We can't use nanofilt because it is not installed in Galaxy