# Galaxy for virologist training Exercise 4: Nanopore mapping and Assembly 101

Title Galaxy

Training dataset:

Nanopore Sequencing of a SARS-Cov-2

Questions:

How Nanopore reads are differently assembled from

Illumina<sup>4</sup>

Objectives:

· Understand the concept of assembly

· Learn how to interpret assembly quality control metrics

**Estimated** 

40 min

time:

# 1. Description

Nanopore techology is a third generation sequencing technique which allows to get longer sequences, but with reduced sequence quality. Different technologies have different formats, qualities, and specific known biases which make the analysis different among them. In this tutorial, we are going to see an example of how to assemble long reads from a Nanopore sequencing run.

## 2. Upload data to galaxy

#### **Training dataset**

- Experiment info: sequencing summary
- fastq:
  - fastq1
  - fastq2
  - o fastq3
- Reference genome MN908947.3 : fasta --- gff

#### Create new history

• Click the + icon at the top of the history panel and create a new history with the name nanopore assembly 101 tutorial as explained <a href="https://example.com/here">here</a>

#### **Upload data**

• Import and rename the read files fastq1, fastq2 and fastq3

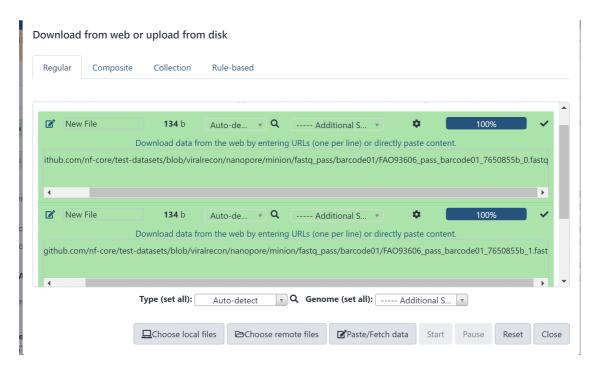
Note: Nanopore reads are commonly splitted in several files that we need to concatenate prior further analysis depending on the software we are going to use.

- 1. Click in upload data.
- 2. Click in paste/fetch data
- 3. Copy url for fastq R1 (select and Ctrl+C) and paste (Ctrl+V).
- 4. Click in Start.
- 5. Wait until the job finishes (green in history)
- 6. Do the same for the remaining files.

## **Table of Contents**

- 1. Description
- 2. Upload data to galaxy
  - Training dataset
  - Create new history
  - Upload data
  - Concatenate reads.
  - Mapping with Minimap2
  - Mapping stats with samtools
  - Assemble reads with Flye

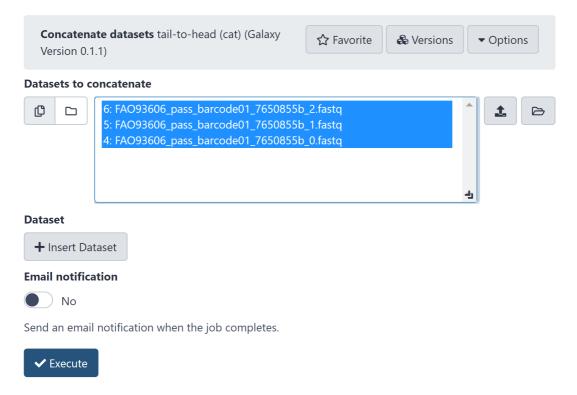




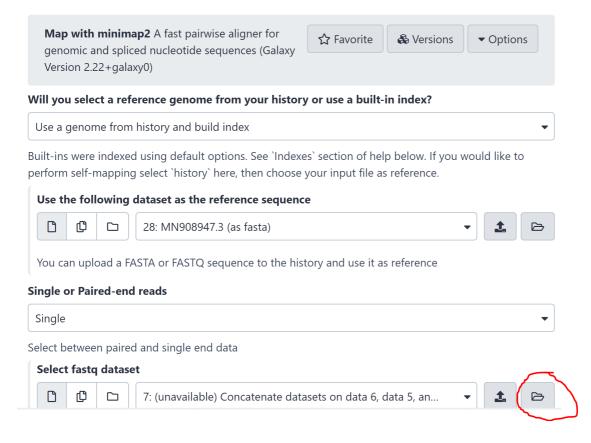
- · Rename files.
  - 1. Click in the ➡ in the history for all the files
  - 2. Change the name to fastq\_X
- Import the reference genome.
- Rename the reference genome.
  - 1. Click the for the reference file in the history.
  - 2. Change the name to MN908947.3

#### Concatenate reads.

- 1. Search Concatenate datasets using the search toolbox.
- 2. Select all three fastq files keeping Cntrl key clicked.
- 3. Click execute and wait.



- 1. Search minimap2 using the search toolbox.
- Will you select a reference genome from your history or use a built-in index?: Use a genome from history and index. Select MN908947.3
- Select fastq dataset: Concatenated fastqs. 
   <sup>≜</sup> The tool is not properly
  configured so you can't select directly the fastq, you need to use the folder
  icon and force the selection of the concatenated fastq dataset.
- 4. Click execute and wait.

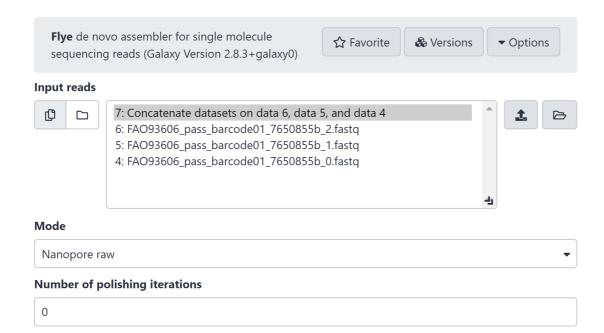


### Mapping stats with samtools

- 1. Search samtools flagstat using the search toolbox.
- 2. Bam file to convert: Minimap2 bam output.
- 3. Click execute and wait.
- 4. Click in the and see the bam stats.
- ▶ Which is the mapping rate?
- ▶ How many reads do we have in our dataset?

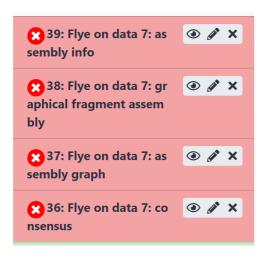
#### Assemble reads with Flye

- 1. Search Flye assembler using the search toolbox.
- 2. Input reads: Concatenate datasets.
- 3. Click execute and wait.



 $\triangle$ As mentioned in the illumina tutorial amplicon-based sequencing is not prepared for de novo assembly. Also this dataset is downsampled for time and performance issues so the depth of coverage is very limited.

 When the job ends, we see that it finished with an error, we can click in the icon and see the error description:



Tool generated the following standard error:

```
[2021-11-18 12:09:41] INFO: Starting Flye 2.8.3-b1695
[2021-11-18 12:09:41] INFO: >>>STAGE: configure
[2021-11-18 12:09:41] INFO: Configuring run
[2021-11-18 12:09:41] INFO: Total read length: 1612409
[2021-11-18 12:09:41] INFO: Reads N50/N90: 517 / 499
[2021-11-18 12:09:41] INFO: Minimum overlap set to 1000
[2021-11-18 12:09:41] INFO: >>>STAGE: assembly
[2021-11-18 12:09:41] INFO: Assembling disjointigs
[2021-11-18 12:09:41] INFO: Reading sequences
[2021-11-18 12:09:47] INFO: Counting k-mers:
0% 10% 20% 30% 40% 50% 60% 70% 80% 90% 100%
[2021-11-18 12:10:36] INFO: Filling index table (1/2)
0% 10% 20% 30% 40% 50% 60% 70% 80% 90% 100%
[2021-11-18 12:10:36] INFO: Filling index table (2/2)
0% 10% 20% 30% 40% 50% 60% 70% 80% 90% 100%
[2021-11-18 12:10:37] INFO: Extending reads
[2021-11-18 12:10:37] INFO: Overlap-based coverage: 58
[2021-11-18 12:10:37] INFO: Median overlap divergence: 0.18496
0% 100%
[2021-11-18 12:10:37] INFO: Assembled 0 disjointigs
[2021-11-18 12:10:37] INFO: Generating sequence
[2021-11-18 12:10:37] ERROR: No disjointigs were assembled - please check if the read type and
genome size parameters are correct
[2021-11-18 12:10:37] ERROR: Pipeline aborted
```

- If we search the error in google the flye developer suggests some
  possible fixes that we've tried and they don't work, but points to uneven
  depth of coverage as a probable source.
- 3. In conclusion, we can't do a de novo assembly in Galaxy using this data.

Note: Nanopore data is known to have more error than short sequencing reads. This is why assembly post-processing is strongly recommended, usually using combined sequencing aproximation with both Nanopore and Illumina reads.