

Galaxy for virologist training

Exercise 4: Nanopore mapping and Assembly 101

Title	Galaxy
Training dataset:	Nanopore Sequencing of a SARS-Cov-2
Questions:	<ul style="list-style-type: none">How Nanopore reads are differently assembled from Illumina?
Objectives:	<ul style="list-style-type: none">Understand the concept of assemblyLearn how to interpret assembly quality control metrics
Estimated time:	40 min

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](#)

1. Description


Nanopore technology 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](#)
 - [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 [here](#)

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.

Download from web or upload from disk

Regular Composite Collection Rule-based

New File
134 b
Auto-de...
----- Additional S...
100%
✓

Download data from the web by entering URLs (one per line) or directly paste content.

ithub.com/nf-core/test-datasets/blob/viralrecon/nanopore/minion/fastq_pass/barcode01/FAO93606_pass_barcode01_7650855b_0.fastq

New File
134 b
Auto-de...
----- Additional S...
100%
✓

Download data from the web by entering URLs (one per line) or directly paste content.

github.com/nf-core/test-datasets/blob/viralrecon/nanopore/minion/fastq_pass/barcode01/FAO93606_pass_barcode01_7650855b_1.fastq

Type (set all): Auto-detect Genome (set all): ----- Additional S...

Choose local files Choose remote files Paste/Fetch data Start Pause Reset Close

- 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.

Concatenate datasets tail-to-head (cat) (Galaxy Version 0.1.1)
Favorite Versions Options

Datasets to concatenate

Copy Paste

6: FAO93606_pass_barcode01_7650855b_2.fastq
5: FAO93606_pass_barcode01_7650855b_1.fastq
4: FAO93606_pass_barcode01_7650855b_0.fastq

Upload Folder

Dataset

+ Insert Dataset

Email notification

☐ No

Send an email notification when the job completes.

✓ Execute

Mapping with Minimap2

1. Search `minimap2` using the search toolbox.
2. 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
3. Select fastq dataset: Concatenated fastqs. ⚠ 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.

Map with minimap2 A fast pairwise aligner for genomic and spliced nucleotide sequences (Galaxy Version 2.22+galaxy0)

☆ Favorite

🔄 Versions

▼ Options

Will you select a reference genome from your history or use a built-in index?

Use a genome from history and build index ▼

Built-ins were indexed using default options. See `Indexes` section of help below. If you would like to perform self-mapping select `history` here, then choose your input file as reference.

Use the following dataset as the reference sequence



28: MN908947.3 (as fasta) ▼



You can upload a FASTA or FASTQ sequence to the history and use it as reference

Single or Paired-end reads

Single ▼

Select between paired and single end data

Select fastq dataset



7: (unavailable) Concatenate datasets on data 6, data 5, an... ▼



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.



Flye de novo assembler for single molecule sequencing reads (Galaxy Version 2.8.3+galaxy0)

☆ Favorite

🔄 Versions

▼ Options

Input reads





7: Concatenate datasets on data 6, data 5, and data 4

6: FAO93606_pass_barcode01_7650855b_2.fastq

5: FAO93606_pass_barcode01_7650855b_1.fastq

4: FAO93606_pass_barcode01_7650855b_0.fastq




Mode

Nanopore raw

Number of polishing iterations

0

△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.

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

✖ 39: Flye on data 7: as
sembly info



✖ 38: Flye on data 7: gr
aphical fragment assem
bly



✖ 37: Flye on data 7: as
sembly graph



✖ 36: Flye on data 7: co
nsensus



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
```

2. 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 approximation with both Nanopore and Illumina reads.