

# Galaxy for virologist training Exercise 4: Nanopore mapping 101

---

Title	Galaxy
<b>Training dataset:</b>	Nanopore MinION Sequencing of a Monkey Pox Virus (MPXV) from Spain 2022 outbreak. Data is publicly available at SRA with <a href="#">ID ERR10297654</a> . <a href="#">Paper</a>
<b>Questions:</b>	<ul style="list-style-type: none"><li>• How Nanopore reads are differently assembled from Illumina?</li></ul>
<b>Objectives:</b>	<ul style="list-style-type: none"><li>• Understand the concept of assembly</li><li>• Learn how to interpret assembly quality control metrics</li></ul>
<b>Estimated time:</b>	40 min

## 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

- [SRA ID: ERR10297654]([https://trace.ncbi.nlm.nih.gov/Traces/?view=run\\_browser&acc=ERR10297654&display=metadata](https://trace.ncbi.nlm.nih.gov/Traces/?view=run_browser&acc=ERR10297654&display=metadata))

### 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

1. Look for **SRA** in the tool search bar and select **Faster Download and Extract Reads in FASTQ format from NCBI SRA**
2. Accession = **ERR10297654**
3. Execute

**Galaxy Europe**

Tools

SRA 1

Upload Data

Show Sections

SRA server

Download and Generate Pileup Format from NCBI SRA

Download and Extract Reads in BAM format from NCBI SRA

**Faster Download and Extract Reads in FASTQ format from NCBI SRA** 2

pysradb search sequence metadata from SRA/ENA

EBI SRA ENA SRA

LowMemPeakPickerHiResRandomAccess Finds mass spectrometric peaks in profile mass spectra.

Make.sra creates the necessary files for a NCBI submission

**Faster Download and Extract Reads in FASTQ format from NCBI SRA** (Galaxy Version 2.11.0+galaxy1)

select input type

SRR accession

**Accession**

ERR10297654

Must start with SRR, DRR or ERR, e.g. SRR925743, ERR343809

Advanced Options

Email notification

☐ No

Send an email notification when the job completes.

**Execute** 4

What it does?

This tool extracts data (in fastq format) from the Short Read Archive (SRA) at the National Center for Biotechnology Information (NCBI). It is based on the [fasterq-dump](#) utility of the SRA Toolkit.

How to use it?

## Load reference file from NCBI

1. Search **NCBI** using the search toolbox and select **NCBI Accession Download Download sequences from GenBank/RefSeq by accession through the NCBI ENTREZ API**
2. Select source for IDs > Direct entry
3. ID List = NC\_063383.1
4. Execute

**Galaxy Europe**

Herramientas

ncbi 1

Cargar Datos

Show Sections

Krona pie chart from taxonomic profile

Kraken taxonomic report view report of classification for multiple samples

Unipept retrieve taxonomy for peptides

MaxQuant (using mqpar.xml)

Krona pie chart from taxonomic profile

NCBI ESearch search NCBI Databases by text query

NCBI EPost post UIDs to NCBI History Server

NCBI EFetch fetch records from NCBI

**NCBI Accession Download Download sequences from GenBank/RefSeq by accession through the NCBI ENTREZ API** 2

NCBI EInfo fetch NCBI database metadata

NCBI ECIIMatch search NCBI for citations in PubMed

ETE tree DB generator generate the

**NCBI Accession Download Download sequences from GenBank/RefSeq by accession through the NCBI ENTREZ API** (Galaxy Version 0.2.7+galaxy0)

Select source for IDs

Direct Entry 3

ID List

NC\_063383.1 4

Newline/Comma separated list of IDs

Molecule Type

Nucleotide

File Format

FASTA

How to handle download failures

☒ Abort with error on first failure

☐ Add accession to failed list and continue

Email notification

☐ No

Send an email notification when the job completes.

**Execute** 5

History

buscar conjuntos de datos

nanopore assembly 101 tutorial

361 MB

4 : fasterq-dump log

3 : Other data (fasterq-dump)

a list with 0 datasets

2 : Single-end data (fasterq-dump)

a list with 1 fastqsanger.gz dataset

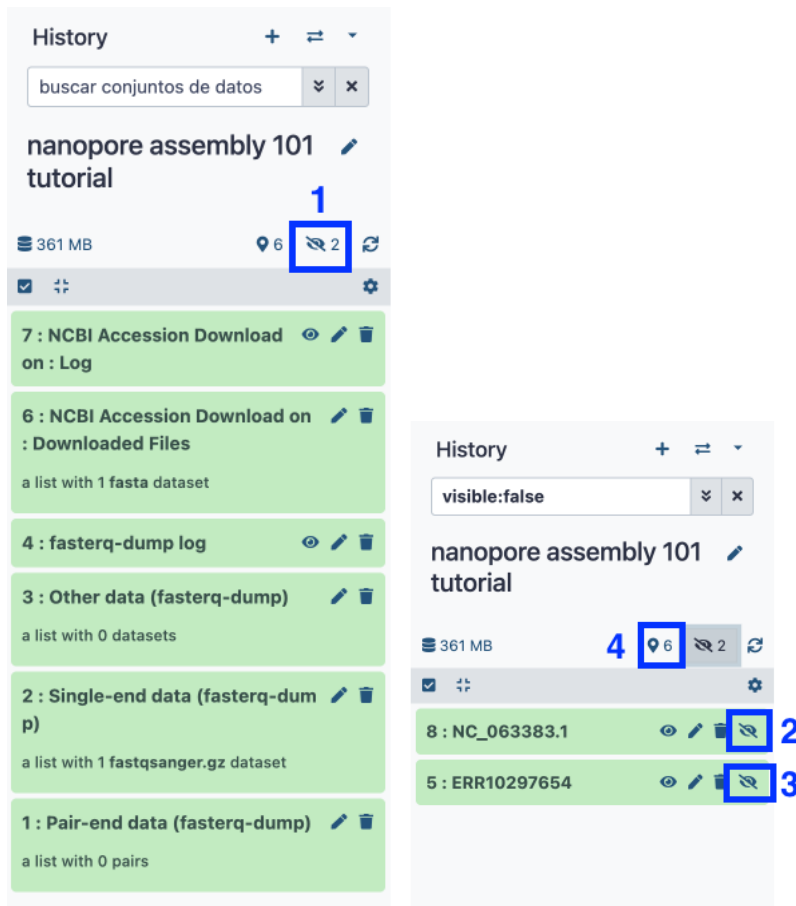
1 : Pair-end data (fasterq-dump)

a list with 0 pairs

## Unhide data

Using SRA and NCBI API downloads data as hidden so we are going to unhide this data as follows:

1. Click on the strikethrough eye (Show hidden)
2. Select the strikethrough for ERR10297654 and NC\_063383.1 datas.
3. Then select the location icon (show active)



## Mapping with Minimap2

1. Search **minimap2** using the search toolbox and select **Map with minimap2 A fast pairwise aligner for genomic and spliced nucleotide sequences**
2. Will you select a reference genome from your history or use a built-in index?: Use a genome from history and built-in index
  - Select NC\_063383.1
3. Select fastq dataset: ERR10297654
4. Select a profile of preset options > Oxford Nanopore Read to reference mapping (map-ont)
5. Click execute and wait.

**Herramientas**

minimap2 1

Cargar Datos

Show Sections

**Map with minimap2** A fast pairwise aligner for genomic and spliced nucleotide sequences 2

Funannotate predict annotation

TB-Profiler Profile Infer strain types and drug resistance markers from sequences

Funannotate assembly clean

Purge overlaps and haplotigs in an assembly based on read depth (purge\_dups)

Flye de novo assembler for single molecule sequencing reads

TGS-GapCloser fills the N-gap of error-prone long reads

Winnowmap a mapping tool optimized for repetitive sequences

**FLUJOS DE TRABAJO**

Todos los flujos de trabajo

**Map with minimap2** A fast pairwise aligner for genomic and spliced nucleotide sequences (Galaxy Version 2.24+galaxy0)

Will you select a reference genome from your history or use a built-in index? 3

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 4

8 : NC\_063383.1

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 5

5 : ERR10297654

Specify dataset with single reads

Select a profile of preset options 6

Oxford Nanopore read to reference mapping. Slightly more sensitive for Oxford Nanopore to reference mapping (-k15)....

Each profile comes with the preconfigured settings mentioned in parentheses. You can customize each profile further in the indexing, mapping and alignment options sections below. If you do not select a profile here, the tool will use the per-parameter defaults listed in the below sections unless you customize them.

Indexing options

Mapping options

Alignment options

Set advanced output options

7 Execute

**History**

buscar conjuntos de datos

nanopore assembly 101 tutorial

361 MB

8 : NC\_063383.1

7 : NCBI Accession Download on : Log

6 : NCBI Accession Download on : Downloaded Files

a list with 1 fasta dataset

5 : ERR10297654

4 : fasterq-dump log

3 : Other data (fasterq-dump)

a list with 0 datasets

2 : Single-end data (fasterq-dump)

a list with 1 fastqsanger.gz dataset

1 : Pair-end data (fasterq-dump)

a list with 0 pairs

## Mapping stats with samtools

1. Search **flagstat** using the search toolbox and select **Samtools flagstat tabulate descriptive stats for BAM dataset**
2. BAM File to report statistics of > Select Minimap2 bam output
3. Click execute and wait.
4. Click in the and see the bam stats.

**Galaxy Europe**

Flujo de Trabajo Visualizar Datos Compartidos Ayuda Usuario

**Herramientas**

flagstats 1

Cargar Datos

Show Sections

**flagstat** provides simple stats on BAM files

**Samtools flagstat** tabulate descriptive stats for BAM dataset 2

**FLUJOS DE TRABAJO**

Todos los flujos de trabajo

**Samtools flagstat** tabulate descriptive stats for BAM dataset (Galaxy Version 2.0.4)

BAM File to report statistics of 3

9 : Map with minimap2 on data 5 and data 8 (mapped reads in BAM format)

Output format

txt

(--output-fmt)

Email notification

No

Send an email notification when the job completes.

4 Execute

- Which is the mapping rate?
- How many reads do we have in our dataset?

This training history is available at: <https://usegalaxy.eu/u/s.varona/h/nanopore-assembly-101-tutorial>

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.