Galaxy for virologist training

Exercise 4: Nanopore mapping and Assembly 101

Table of Contents

Title	Galaxy
Training dataset:	Nanopore MinION Sequencing of a Monkey Pox Virus (MPXV) from Spain 2022 oubreak. Data is publicly available at SRA with ID ERR10297654. Paper
Questions:	• How Nanopore reads are differently assembled from Illumina?
Objectives:	• Understand the concept of assembly • Learn how to interpret assembly quality control metrics
Estimated time:	40 min

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

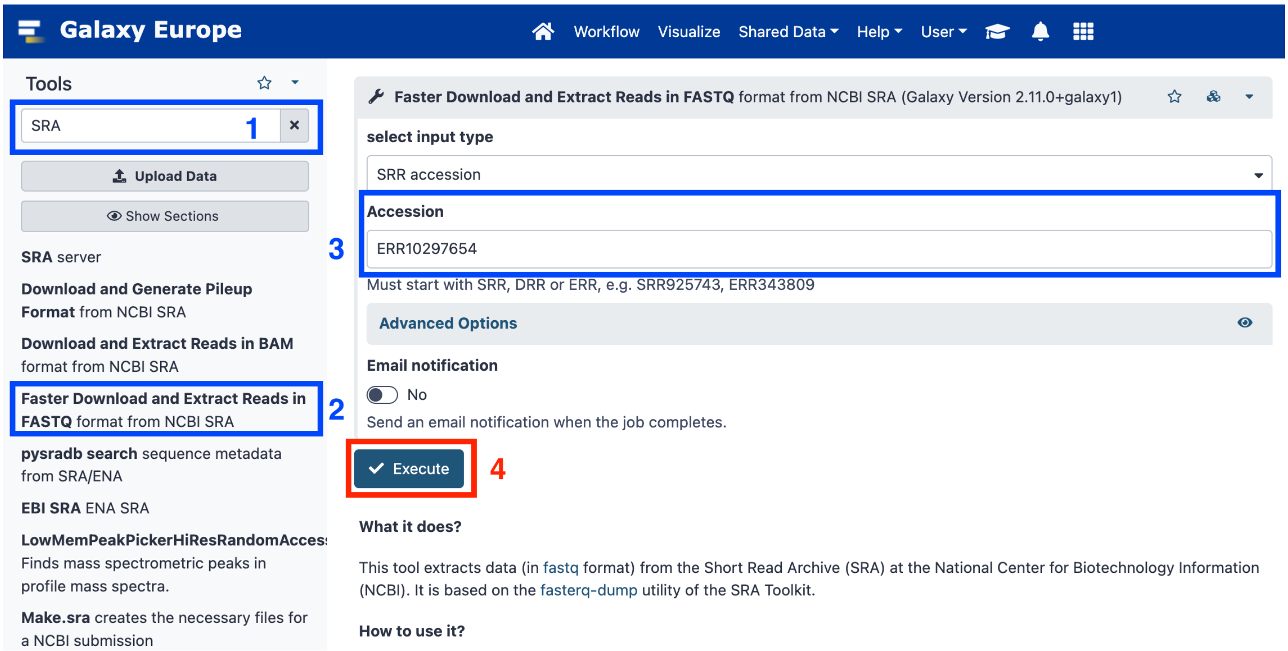
- [SRA ID: ERR10297654](https://trace.ncbi.nlm.nih.gov/Traces/?view=run_browser&acc=ERR10297654&display=metadata

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



SRA

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

pysradb search sequence metadata from SRA/ENA

EBI SRA ENA SRA

LowMemPeakPickerHiResRandomAcces
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

Email notification

No

Send an email notification when the job completes.

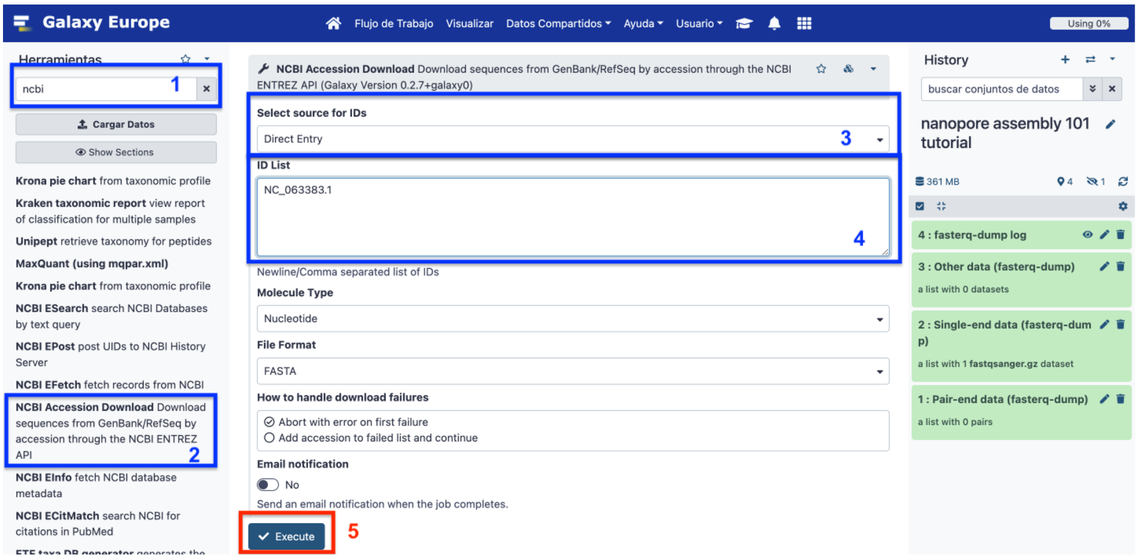
What it does?

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



ncbi

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

NCBI EInfo fetch NCBI database metadata

NCBI ECitMatch search NCBI for citations in PubMed

ETE taxa DB generator generates 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

ID List

NC_063383.1

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.

buscar conjuntos de datos

nanopore assembly 101 tutorial

361 MB 4 1

4 : fasterq-dump log

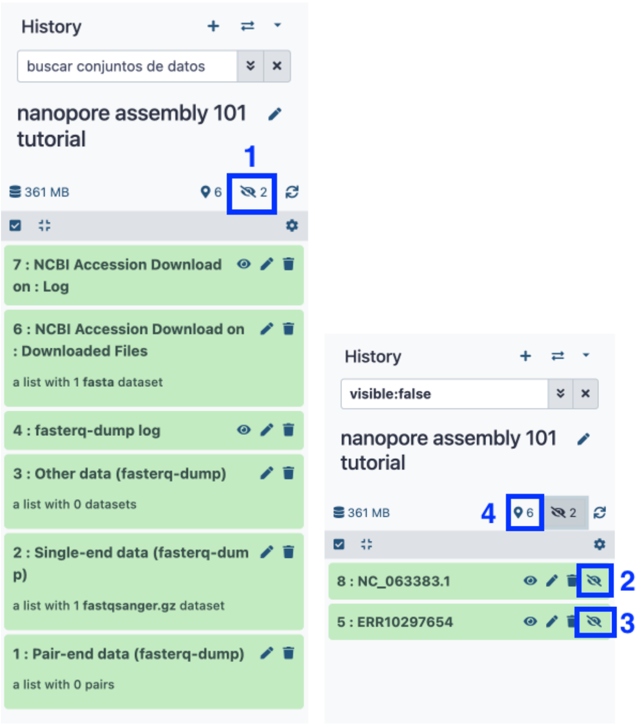
3 : Other data (fasterq-dump)
a list with 0 datasets

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

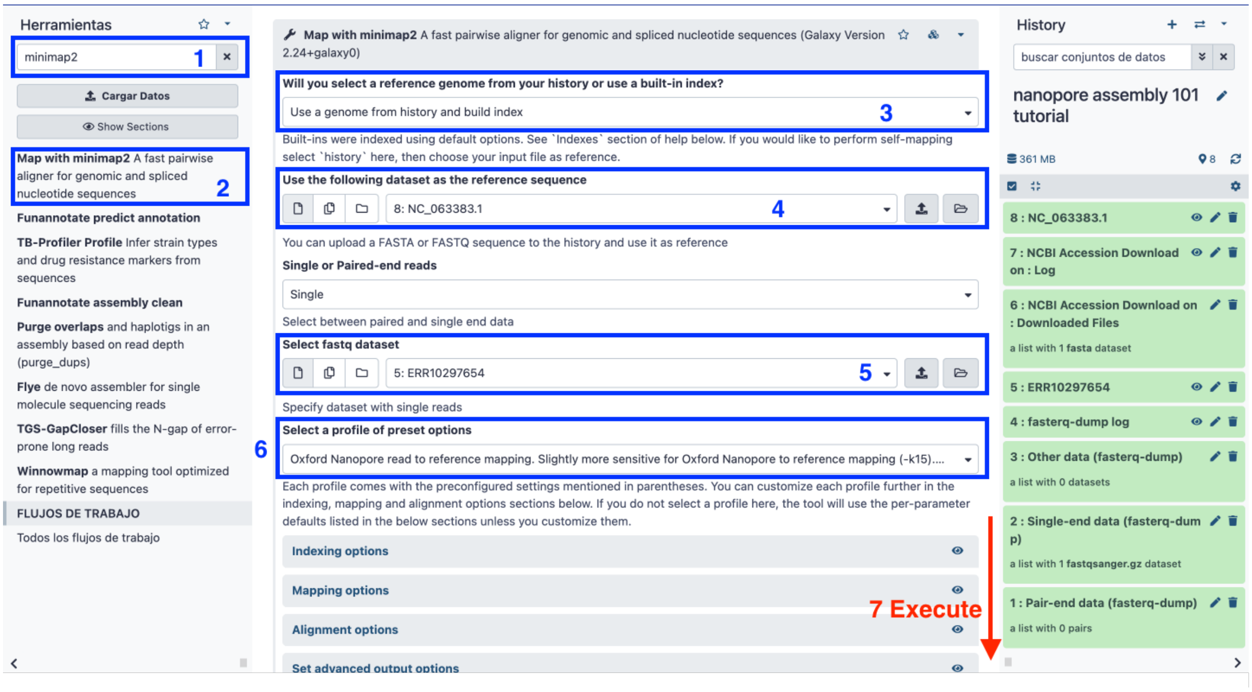
Using SRA and NCBI API downloads data as hidden so we are going to unhidde this data as follows: 1. Click on the strikethrough eye (Show hidden) 2. Select the strikethrough for ERR10297654 and NC_063383.1 datas. 3. Then select the location icon (show active)


History
buscar conjuntos de datos
nanopore assembly 101 tutorial
1
361 MB 6 2
7 : NCBI Accession Download on : Log
6 : NCBI Accession Download on : Downloaded Files
a list with 1 fasta dataset
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

History
visible:false
nanopore assembly 101 tutorial
361 MB 4 6 2
8 : NC_063383.1 2
5 : ERR10297654 3

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?
Use a genome from history and build index 3
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
8: NC_063383.1 4
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: ERR10297654 5
Specify dataset with single reads

6
Select a profile of preset options
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

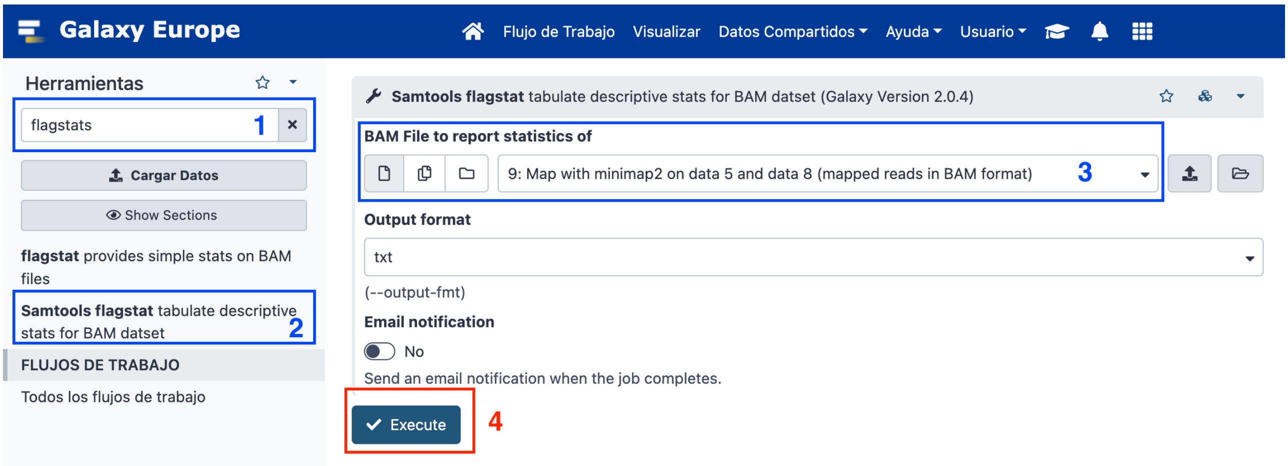
7 Execute

History
buscar conjuntos de datos
nanopore assembly 101 tutorial
361 MB 8
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 `flagstatst` using the search toolbox and select `Samtools flagstat tabulate descriptive stats for BAM datset`
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.



flagstats

flagstat provides simple stats on BAM files

Samtools flagstat tabulate descriptive stats for BAM datset

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

BAM File to report statistics of

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

Output format

txt

(--output-fmt)

Email notification

No

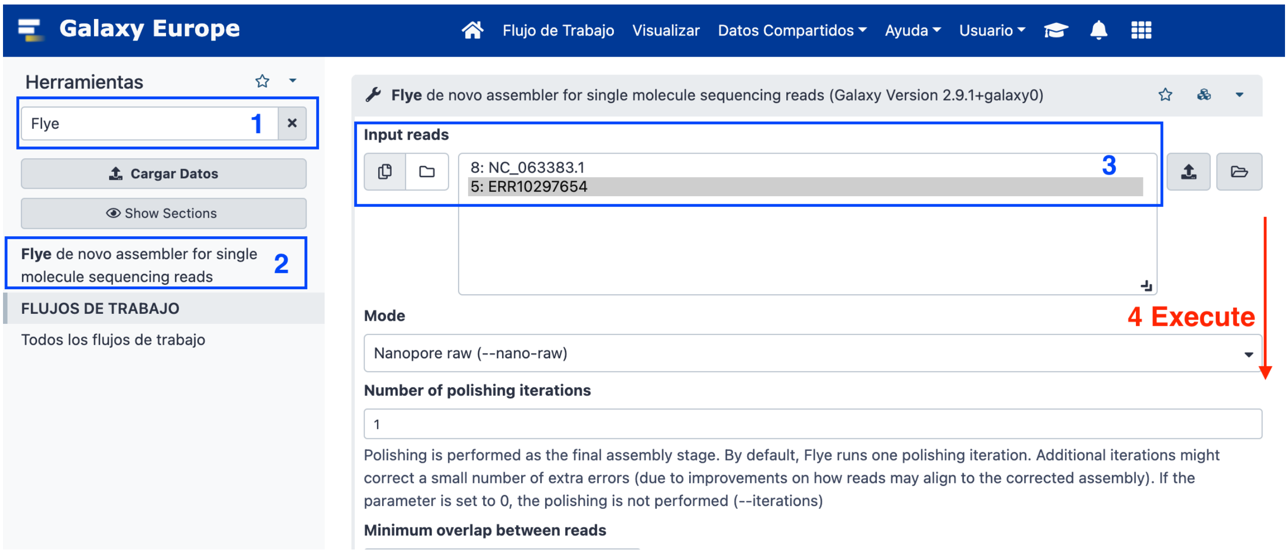
Send an email notification when the job completes.

▶ 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 and select `Flye de novo assembler for single molecule sequencing reads`
2. Input reads: ERR10297654
3. Click execute and wait.



Flye

Flye de novo assembler for single molecule sequencing reads

Flye de novo assembler for single molecule sequencing reads (Galaxy Version 2.9.1+galaxy0)

Input reads

8: NC_063383.1
5: ERR10297654

Mode

Nanopore raw (--nano-raw)

Number of polishing iterations

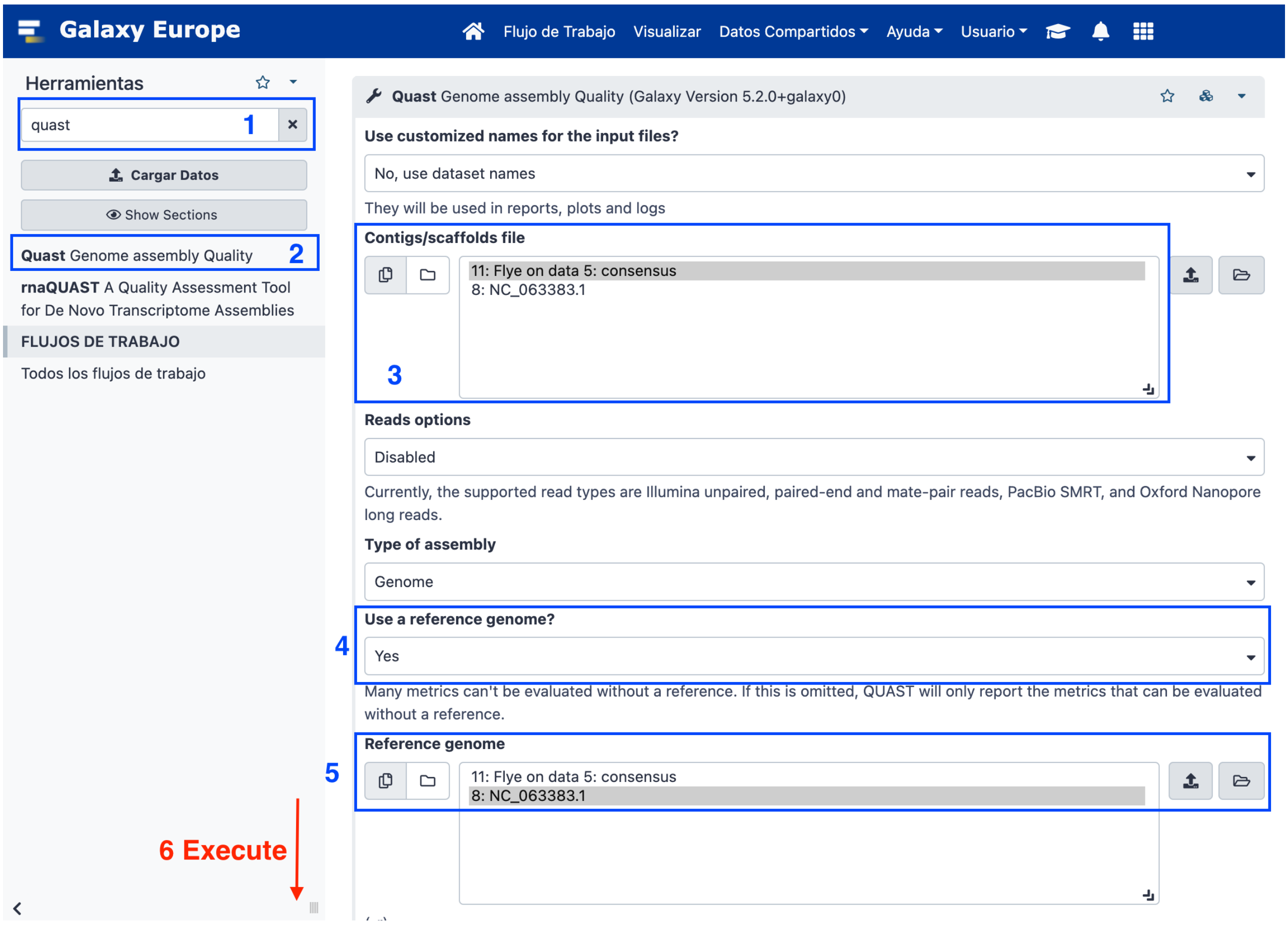
1

Polishing is performed as the final assembly stage. By default, Flye runs one polishing iteration. Additional iterations might correct a small number of extra errors (due to improvements on how reads may align to the corrected assembly). If the parameter is set to 0, the polishing is not performed (--iterations)

Minimum overlap between reads

Assembly quality control with Quast

1. Search Quast in the search tool box.
2. Contigs/scaffolds file > Flye results
3. Use a reference genome: Yes. Select the NC_063383.1 fasta file previously loaded.



Quast Genome assembly Quality (Galaxy Version 5.2.0+galaxy0)

Use customized names for the input files?
No, use dataset names
They will be used in reports, plots and logs

Contigs/scaffolds file
11: Flye on data 5: consensus
8: NC_063383.1

Reads options
Disabled
Currently, the supported read types are Illumina unpaired, paired-end and mate-pair reads, PacBio SMRT, and Oxford Nanopore long reads.

Type of assembly
Genome

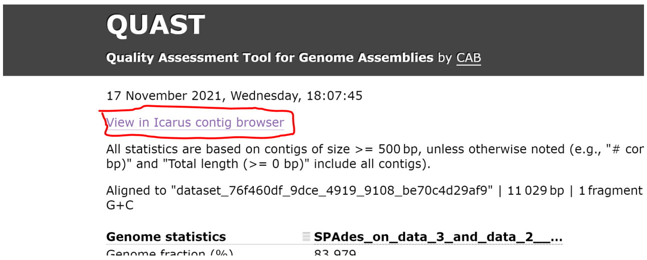
Use a reference genome?
Yes
Many metrics can't be evaluated without a reference. If this is omitted, QUAST will only report the metrics that can be evaluated without a reference.

Reference genome
11: Flye on data 5: consensus
8: NC_063383.1

1. Click the :eye: icon Quast HTML report.

   - ▶ How much of or reference genome have we reconstructed?
   - ▶ How many contigs do we have greater than 1000 pb?
   - ▶ How long is the largest contig in the assembly?
   - ▶ Which is the N50?

2. Open the Icarus viewer in the quast report.



QUAST
Quality Assessment Tool for Genome Assemblies by CAB

17 November 2021, Wednesday, 18:07:45

View in Icarus contig browser

All statistics are based on contigs of size >= 500 bp, unless otherwise noted (e.g., "# cor bp)" and "Total length (>= 0 bp)" include all contigs).

Aligned to "dataset_76f460df_9dce_4919_9108_be70c4d29af9" | 11 029 bp | 1 fragment G+C

Genome statistics	**SPAdes_on_data_3_and_data_2__...**
Genome fraction (%)	83.979

▶ Which contig align against our reference genome?

> 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 aproximation with both Nanopore and Illumina reads.