Galaxy for virologist training
Exercise 4: Illumina Assembly 101

Table of Contents

Title	Galaxy
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:	• What is assembly? • How can I evaluate my assembly?
Objectives:	• Understand assembly concept • Learn how to interpret assembly quality control metrics
Estimated time:	40 min

1. Description

Sometimes, we don't have a reference genome to map against, or we want to reconstruct a genome without any bias caused by a reference. In such cases, we need to do a *de novo assembly*. This type of analysis tries to reconstruct the original genome without any template, using only the reads. Some considerations:

- When we assemble, the longer the reads are and the longer the size of the library fragments the easier it gets for the assembler. That's why pacbio or nanopore are recommended for assembly. Think of it like a puzzle, the bigger the pieces, the easier it is to form the image.
- It's almost imposible to reconstruct the entire genome of a large-genome microorganism with only one sequencing, although it can be done for smaller ones, like viruses.
- Assembly is not recommended for amplicon based libraries due to the depth of coverage uneveness and the amplicons intrinsic bias.

2. Upload data to galaxy

Training dataset

- Experiment info: PRJEB43037, WGS, Illumina MiSeq, paired-end
- Fastq R1: ERR5310322_1 - url : `ftp://ftp.sra.ebi.ac.uk/vol1/fastq/ERR531/002/ERR5310322/ERR5310322_1.fastq.gz`
- Fastq R2: ERR5310322_2 url : `ftp://ftp.sra.ebi.ac.uk/vol1/fastq/ERR531/002/ERR5310322/ERR5310322_2.fastq.gz`
- Reference genome NC_009942.1: fasta -- gff

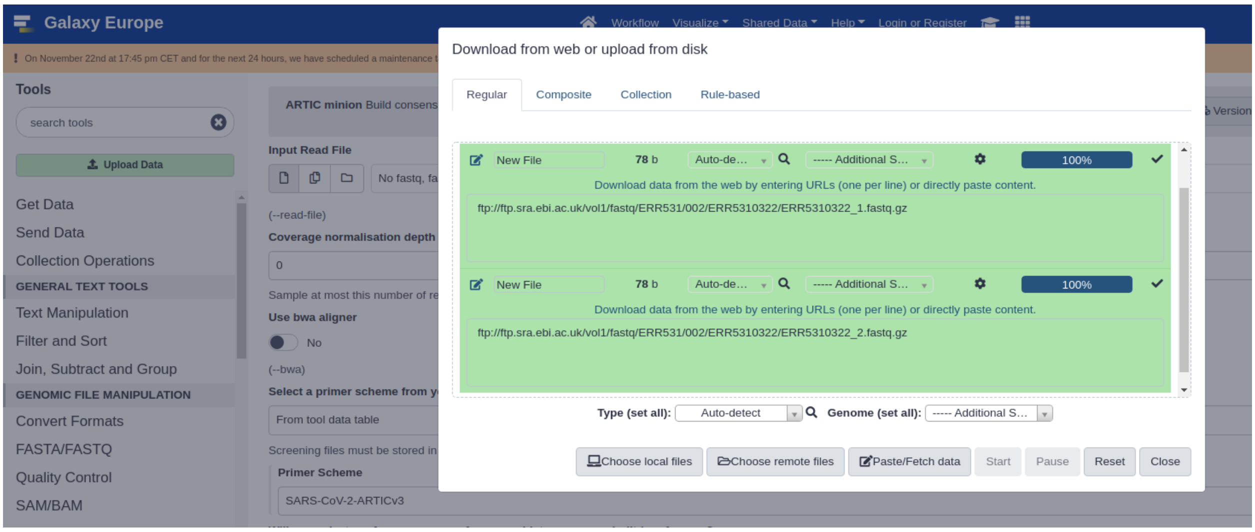
Create new history

- Click the `+` icon at the top of the history panel and create a new history with the name `illumina assembly 101 tutorial` as explained here

Upload data

- Import and rename the read files `ERR5310322_1` and `ERR5310322_2`

  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 fastq R2.



On November 22nd at 17:45 pm CET and for the next 24 hours, we have scheduled a maintenance t

ARTIC minion Build consens

Input Read File

No fastq, fa

(--read-file)

Coverage normalisation depth

0

Sample at most this number of re

Use bwa aligner

No

(--bwa)

Select a primer scheme from y

From tool data table

Screening files must be stored in

Primer Scheme

SARS-CoV-2-ARTICv3

Download from web or upload from disk

New File 78 b Auto-de... ----- Additional S... 100%

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

ftp://ftp.sra.ebi.ac.uk/vol1/fastq/ERR531/002/ERR5310322/ERR5310322_1.fastq.gz

New File 78 b Auto-de... ----- Additional S... 100%

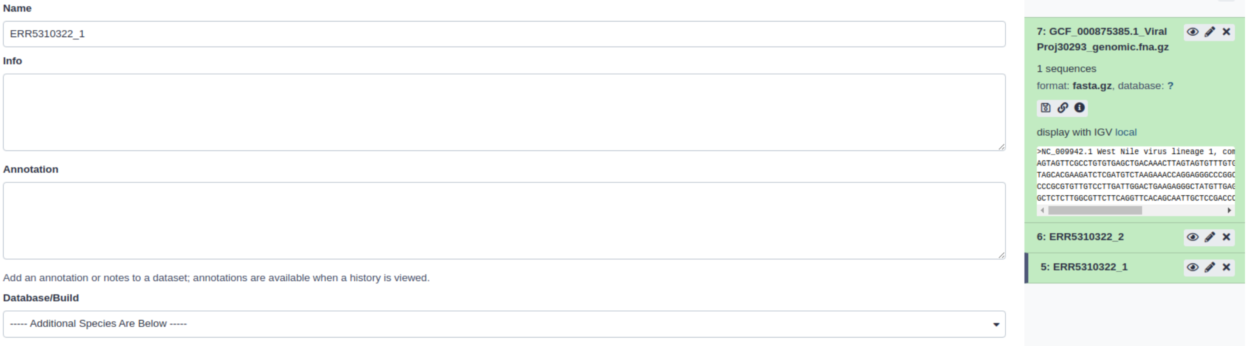
Download data from the web by entering URLs (one per line) or directly paste content.

ftp://ftp.sra.ebi.ac.uk/vol1/fastq/ERR531/002/ERR5310322/ERR5310322_2.fastq.gz

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

- Rename R1 and R2 files.

  1. Click in the ✏ in the history for `ERR5310322_1.fastq.gz`
  2. Change the name to `ERR5310322_1`
  3. Do the same for R2.



Name

ERR5310322_1

Info

Annotation

Add an annotation or notes to a dataset; annotations are available when a history is viewed.

Database/Build

----- Additional Species Are Below -----

7: GCF_000875385.1_ViralProj30293_genomic.fna.gz

1 sequences

format: **fasta.gz**, database: ?

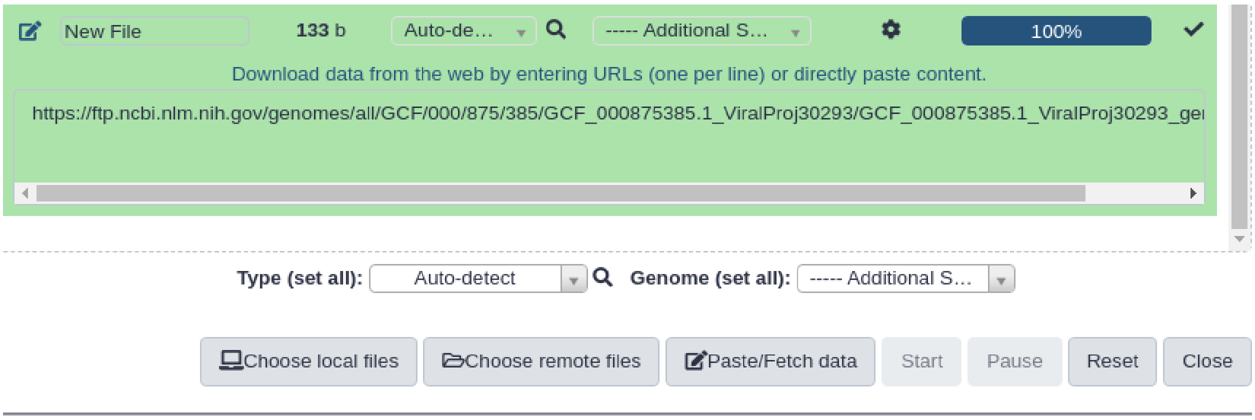
display with IGV local

```
>NC_009942.1 West Nile virus lineage 1, com
AGTAGTTCGCCTGTGTGAGCTGACAAACTTAGTAGTGTTTGTG
TAGCACGAAGATCTCGATGTCTAAGAAACCAGGAGGGCCCGGC
CCCGCGTGTTGTCCTTGATTGGACTGAAGAGGGCTATGTTGAG
GCTCTCTTGGCGTTCTTCAGGTTCACAGCAATTGCTCCGACCC
```

6: ERR5310322_2

5: ERR5310322_1

- Import the reference genome.



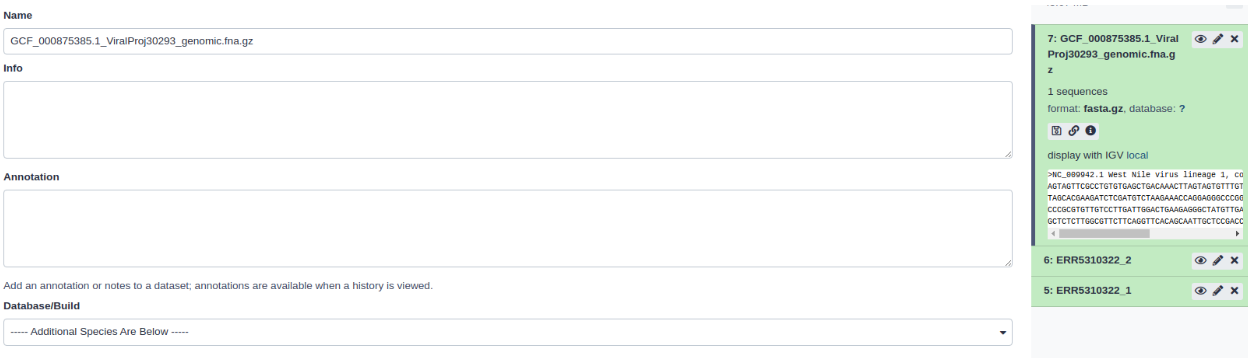
New File 133 b Auto-de... ----- Additional S... 100%

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

https://ftp.ncbi.nlm.nih.gov/genomes/all/GCF/000/875/385/GCF_000875385.1_ViralProj30293/GCF_000875385.1_ViralProj30293_ge

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

- Rename the reference genome.
  1. Click the ✏ for the reference file in the history.
  2. Change the name to `NC_009942.1`



Name

GCF_000875385.1_ViralProj30293_genomic.fna.gz

Info

Annotation

Add an annotation or notes to a dataset; annotations are available when a history is viewed.

Database/Build

----- Additional Species Are Below -----

7: GCF_000875385.1_ViralProj30293_genomic.fna.gz

1 sequences

format: **fasta.gz**, database: ?

display with IGV local

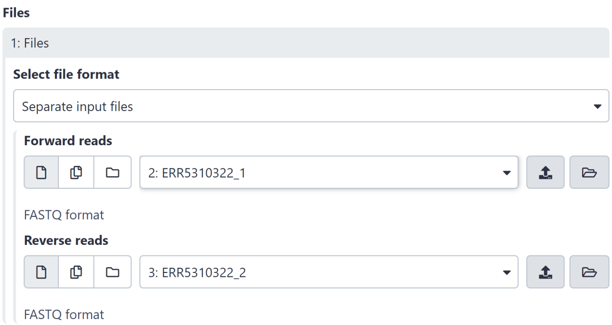
```
>NC_009942.1 West Nile virus lineage 1, co
AGTAGTTCGCCTGTGTGAGCTGACAAACTTAGTAGTGTTTGT
TAGCACGAAGATCTCGATGTCTAAGAAACCAGGAGGGCCCGG
CCCGCGTGTTGTCCTTGATTGGACTGAAGAGGGCTATGTTGA
GCTCTCTTGGCGTTCTTCAGGTTCACAGCAATTGCTCCGACC
```

6: ERR5310322_2

5: ERR5310322_1

Assemble reads with Spades

1. Search Spades in the search tool box.
2. Automatically choose k-mer values: Yes
3. Select File format: Separate input reads. Forward reads: ERR5310322_1, Reverse Reads: ERR5310322_2
4. Click execute and wait.



Files

1: Files

Select file format

Separate input files

Forward reads

2: ERR5310322_1

FASTQ format

Reverse reads

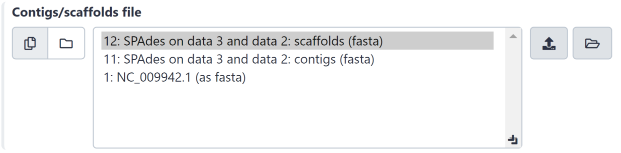
3: ERR5310322_2

FASTQ format

5. Click the :eye: icon in the history: Spades Contigs stats.
   ▶ How many contigs has been assembled?
6. Click the :eye: icon in the history: Spades scaffolds.

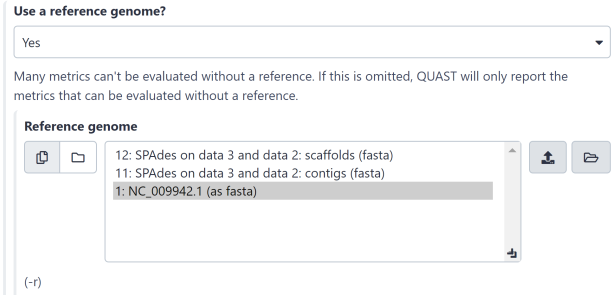
Assembly quality control with Quast

1. Search Quast in the search tool box.
2. Contigs/Scaffolds file: Spades scaffolds
3. Use a reference genome: Yes. Select the NC_009942.1 fasta file previously loaded.



Contigs/scaffolds file

12: SPAdes on data 3 and data 2: scaffolds (fasta)
11: SPAdes on data 3 and data 2: contigs (fasta)
1: NC_009942.1 (as fasta)



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

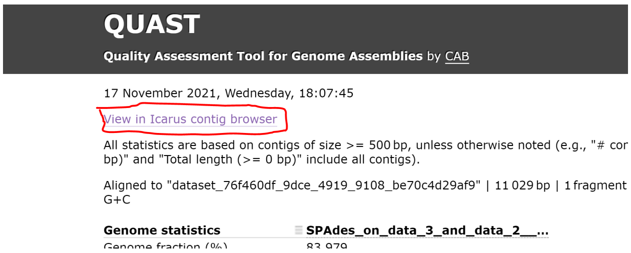
12: SPAdes on data 3 and data 2: scaffolds (fasta)
11: SPAdes on data 3 and data 2: contigs (fasta)
1: NC_009942.1 (as fasta)

(-r)

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

▶ How many contigs align against our reference genome?