

Galaxy for virologist training Exercise 3: Illumina Assembly 101

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:	<ul style="list-style-type: none">• What is assembly?• How can I evaluate my assembly?
Objectives:	<ul style="list-style-type: none">• 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 unevenness and the amplicons intrinsic bias.

2. Upload data to galaxy

Training dataset

- Experiment info: PRJEB43037, WGS, Illumina MiSeq, paired-end
- Fastq R1: [ERR5310322_1](ftp://ftp.sra.ebi.ac.uk/vol1/fastq/ERR531/002/ERR5310322/ERR5310322_1.fastq.gz) - url :
ftp://ftp.sra.ebi.ac.uk/vol1/fastq/ERR531/002/ERR5310322/ERR5310322_1.fastq.gz
- Fastq R2: [ERR5310322_2](ftp://ftp.sra.ebi.ac.uk/vol1/fastq/ERR531/002/ERR5310322/ERR5310322_2.fastq) url :
ftp://ftp.sra.ebi.ac.uk/vol1/fastq/ERR531/002/ERR5310322/ERR5310322_2.fastq

.gz

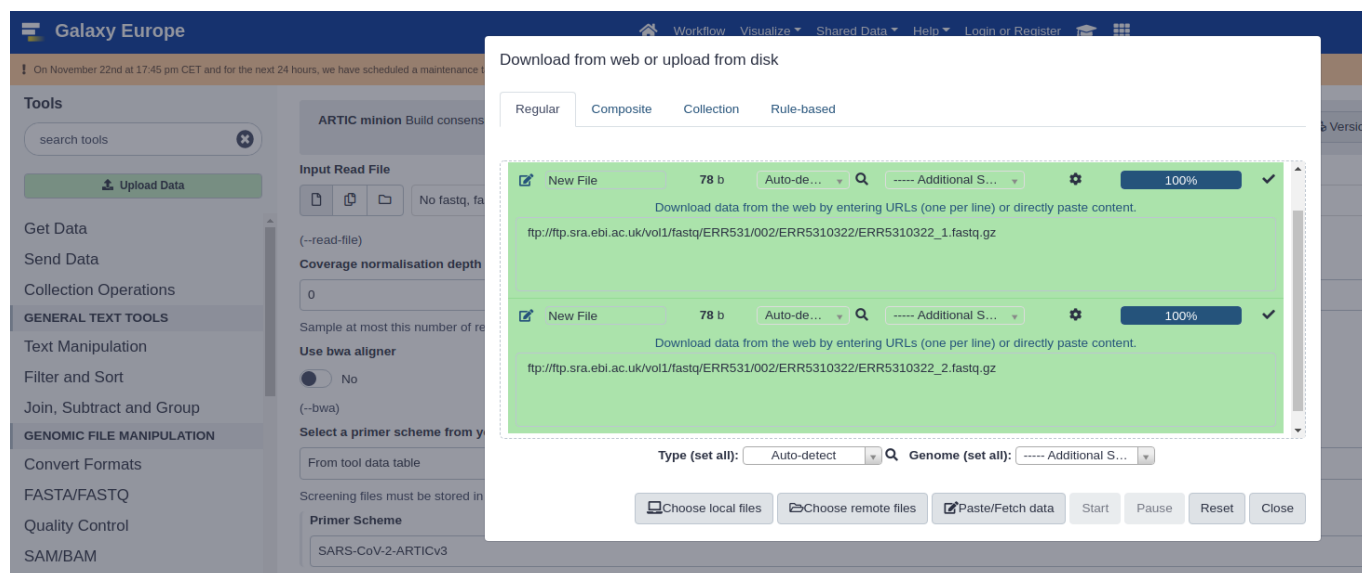
- Reference genome NC_009942.1: **fasta** -- **gff**


Create new history

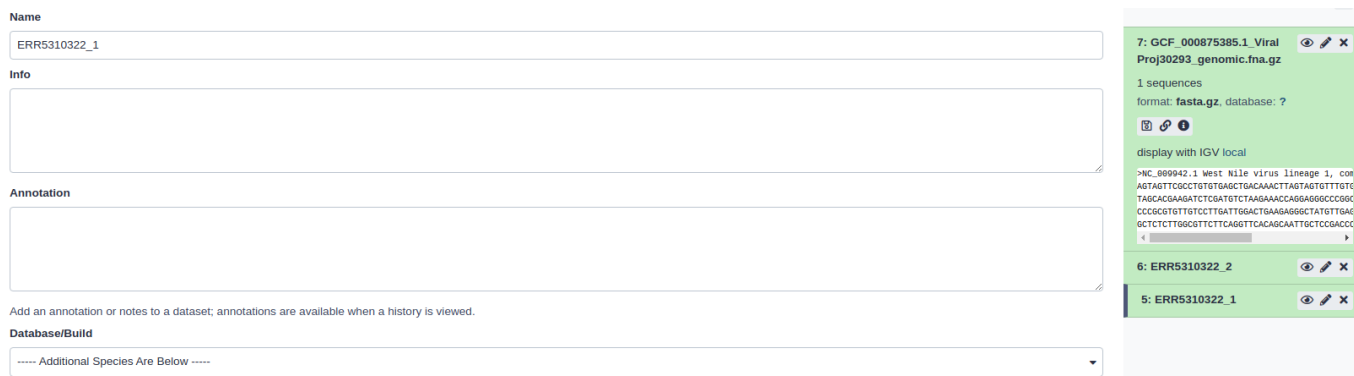
- Click the **+** icon at the top of the history panel and create a new history with the name **Illumina Assembly** as explained [here](#)

Upload data

- Import and rename the read files **ERR5310322_1** and **ERR5310322_2**
 - Click in upload data.
 - Click in paste/fetch data
 - Copy url for fastq R1 (select and Ctrl+C) and paste (Ctrl+V).
 - Click in Start.
 - Wait until the job finishes (green in history)
 - Do the same for fastq R2.



- Rename R1 and R2 files.
 - Click in the  in the history for **ERR5310322_1.fastq.gz**
 - Change the name to **ERR5310322_1**
 - Do the same for R2.



- Import the reference genome and GFF file.

```
https://ftp.ncbi.nlm.nih.gov/genomes/all/GCF/000/875/385/GCF_000875385.1_ViralProj30293/GCF_000875385.1_ViralProj30293_genomic.fna.gz
https://ftp.ncbi.nlm.nih.gov/genomes/all/GCF/000/875/385/GCF_000875385.1_ViralProj30293/GCF_000875385.1_ViralProj30293_genomic.gff.gz
```

Descargar de la red o cargar desde disco

RegularCompositeCollectionRule-based

You added 1 file(s) to the queue. Add more files or click 'Start' to proceed.

Name	Size	Type	Genome	Settings	Status
New File	267 b	Auto-de...	unspecified (?)		0%

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

pi.nlm.nih.gov/genomes/all/GCF/000/875/385/GCF_000875385.1_ViralProj30293/GCF_000875385.1_ViralProj30293_genomic.fna.gz
pi.nlm.nih.gov/genomes/all/GCF/000/875/385/GCF_000875385.1_ViralProj30293/GCF_000875385.1_ViralProj30293_genomic.gff.gz

Type (set all): Auto-detect Genome (set all): unspecified (?)

Elegir archivos locales

Choose remote files

Paste/Fetch data

Start

Pause

Reset

Close

- Rename the reference genome and gff file.
 - Click the for the reference file in the history.
 - 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
AGTAGTTGGCTGTGTGAGCTGACAACTTAGTAGTTTGT
TAGCAGAGATCTCGATGTTCTAAGAACCAAGAGGCCCG
CCGCGGTGTGTCTCTGATGAGCTGAAGAGGCTATGTTGA
GCTCTCTGGCGTCTCTCAAGTTCCAGCAATGCTCCGACC

6: ERR5310322_2

5: ERR5310322_1

- Finally, add some usefull tags

Change name 2

Assemble reads with Spades

- Search **Spades** in the search tool box and select *rnaviralSPAdes de novo assembler for transcriptomes, metatranscriptomes and metaviromes*

3 / 6

- Single-end or paired-end short-reads > Paired-end: individual datasets
- FASTQ RNA-seq file(s): forward reads: ERR5310322_1; FASTQ RNA-seq file(s): reverse reads: ERR5310322_2
- Select optional output file(s) > Scaffolds stats
- Click execute and wait.

Herramientas

spades 1

SPAdes genome assembler for genomes of regular and single-cell projects

metaviralSPAdes extract and assembly viral genomes from metagenomic data

metaplasmidSPAdes extract and assembly plasmids from metagenomic data

plasmidSPAdes extract and assembly plasmids from WGS data

metaSPAdes metagenome assembler

rnaviralSPAdes de novo assembler for transcriptomes, metatranscriptomes and metaviromes 2

biosyntheticSPAdes biosynthetic gene cluster assembly

coronaSPAdes SARS-CoV-2 de novo genome assembler

rnaSPAdes de novo transcriptome assembler

Shovill Faster SPAdes assembly of Illumina reads

Operation mode

Assembly and error correction

To run read error correction, reads should be in FASTQ format.

Single-end or paired-end short-reads

Paired-end: individual datasets

It assumes that all samples belong to the same library. If you want to use samples from two different libraries, include the second library as additional set of short-reads.

FASTQ RNA-seq file(s): forward reads

3: NC_009942.1
2: ERR5310322_2
1: ERR5310322_1

FASTQ RNA-seq file(s): reverse reads

3: NC_009942.1
2: ERR5310322_2
1: ERR5310322_1

Type of paired-reads

Default (--pe)

In paired-end sequencing, the library preparation yields a set of fragments, and the machine sequences each fragment from

History

buscar conjuntos de datos

illumina assembly 101 tutorial

31.4 MB

4: NC_009942.1 GFF
#gff
3: NC_009942.1
#reference
2: ERR5310322_2
#reverse
1: ERR5310322_1
#forward

Galaxy Europe

Flujo de Trabajo Visualizar Datos Compartidos Ayuda Usuario

Herramientas

spades

SPAdes genome assembler for genomes of regular and single-cell projects

metaviralSPAdes extract and assembly viral genomes from metagenomic data

metaplasmidSPAdes extract and assembly plasmids from metagenomic data

plasmidSPAdes extract and assembly plasmids from WGS data

metaSPAdes metagenome assembler

rnaviralSPAdes de novo assembler for transcriptomes, metatranscriptomes and metaviromes

biosyntheticSPAdes biosynthetic gene cluster assembly

coronaSPAdes SARS-CoV-2 de novo genome assembler

Set Phred quality offset

Auto

Phred quality offset in the input reads. Default: auto-detect (--phred-offset)

Select optional output file(s)

Select/Unselect all

Assembly graph Assembly graph with scaffolds Contigs Scaffolds 6

Contigs paths

Corrected reads

Contigs stats

Log

Scaffolds paths

Scaffolds stats

What it does

SPAdes - St. Petersburg genome assembler - is an assembly toolkit containing various assembly pipelines.

rnaviralSPAdes is a pipeline specially designed for de novo assembler tailored for RNA viral datasets (transcriptome, metatranscriptome and metavirome).

Input

SPAdes takes as input paired-end reads, mate-pairs and single (unpaired) reads in FASTA and FASTQ. For IonTorrent data SPAdes also supports unpaired reads in unmapped BAM format (like the one produced by Torrent Server). However, in order to run read error correction, reads should be in FASTQ or BAM format. Sanger, Oxford Nanopore and PacBio CLR reads can be

History

buscar conjuntos de datos

illumina assembly 101 tutorial

31.4 MB

4: NC_009942.1 GFF
#gff
3: NC_009942.1
#reference
2: ERR5310322_2
#reverse
1: ERR5310322_1
#forward

Warning ☕️🔪⌚ **Assembly takes time!** There is no such thing as Assembly in real time. It can take anywhere between 90 minutes and two hours.

Questions:

Click the :eye: icon in the history: Spades Contigs stats.

► How many contigs has been assembled?

Click the :eye: icon in the history: Spades scaffolds.

Assembly quality control with Quast

1. Search Quast in the search tool box.
2. rnaviralSpades Scaffolds
3. Use a reference genome: Yes. Select the NC_009942.1 fasta file previously loaded.
4. Genomic feature positions in the reference genome > NC_009942.1 gff file previously loaded.

Galaxy Europe Flujo de Trabajo Visualizar Datos Compartidos Ayuda Usuario

Herramientas ☆

quast 1

Cargar Datos

Show Sections

Quast Genome assembly Quality 2

rnaQUAST A Quality Assessment Tool for De Novo Transcriptome Assemblies

FLUJOS DE TRABAJO

Todos los flujos de trabajo

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 3

12: rnaviralSPAdes on data 2 and data 1: Scaffolds
11: rnaviralSPAdes on data 2 and data 1: Contigs
3: NC_009942.1 (as fasta)

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? 4

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 5

12: rnaviralSPAdes on data 2 and data 1: Scaffolds
11: rnaviralSPAdes on data 2 and data 1: Contigs
3: NC_009942.1 (as fasta)

Galaxy Europe Flujo de Trabajo Visualizar Datos Compartidos Ayuda Usuario

Herramientas ☆

quast

Cargar Datos

Show Sections

Quast Genome assembly Quality

rnaQUAST A Quality Assessment Tool for De Novo Transcriptome Assemblies

FLUJOS DE TRABAJO

Todos los flujos de trabajo

Genomic feature positions in the reference genome 6

4: NC_009942.1 GFF (as bed)

Gene coordinates for the reference genome (--features)

Operon positions in the reference genome

Nothing selected

Operon coordinates for the reference genome (--operons)

Compute k-mer-based quality metrics? 7

No

It is recommended for large genomes. This may significantly increase memory and time consumption on large genomes (--k-mer-stats)

Generate Circos plot

No

Plot Circos version of Icarus contig alignment viewer (--circos)

7 Execute

5. 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?

6. 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 (%) 83.070

► How did the contig align against our reference genome?

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