Galaxy for virologist training Exercise 5: Illumina Mapping 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:	- What is mapping? - What is a BAM file? - Which metrics are important to check after mapping?
Objectives:	- Understand the concept of mapping - Learn how to interpret mapping metrics - Learn how to visualize mapping results
Estimated time:	40 min

1. Description

One of the most common experiments using massive sequencing are re-sequencing experiments. This type of experiments sequence already known microorganisms, with the goal to discover variation between an already assembled and known reference, and our reads. Mapping is a mandatory step for this kind of experiments, where we need to sort all the short sequences (reads) we have in our fastq file, lacking any genomic context. After the mapping step, we will transform our fastq file into a bam file that contains information about where a read came from, meaning we are going to have the coordinates where each read is placed inside our reference genome.

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 `mapping 101 tutorial` as explained here
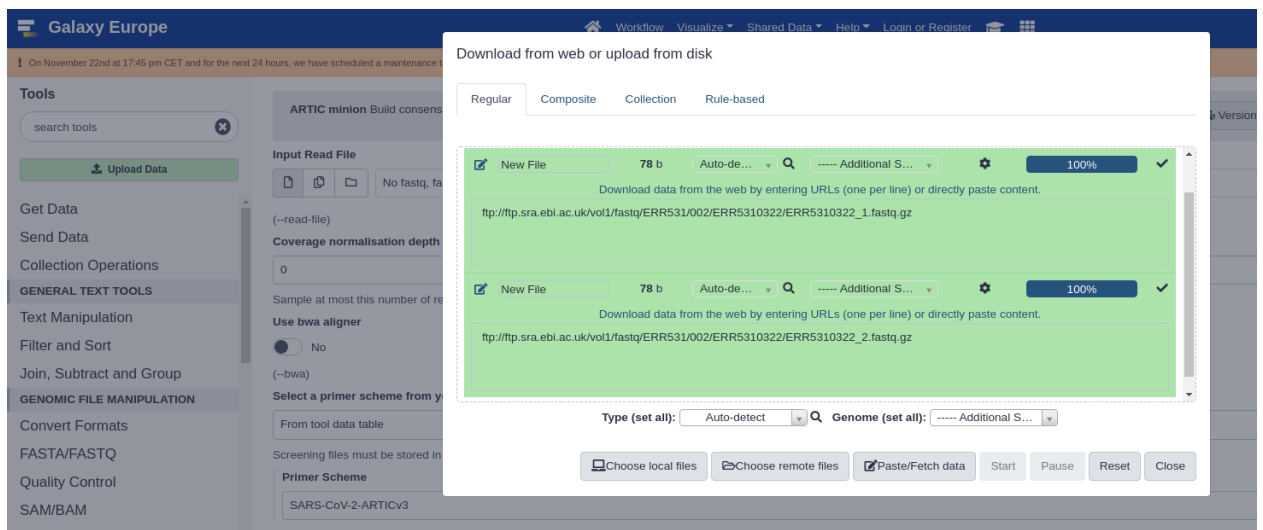
Upload data

- 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.

- Rename R1 and R2 files.

  1. Click in ✎ 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 -----

- Import the reference genome and GFF file.

https://ftp.ncbi.nlm.nih.gov/genomes/all/GCA/009/858/895/GCA_009858
https://ftp.ncbi.nlm.nih.gov/genomes/all/GCA/009/858/895/GCA_009858

Descargar de la red o cargar desde disco

| | Regular | Composite | Collection | Rule-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.

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

1. Click the ✏ for the reference file in the history.
2. Change the name to `MN908947.3`

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 -----

Map reads using Bowtie2

1. Search bowtie2 software in the search tools box on the left.

Tools

bowtie2 ✕

⬆ Upload Data

👁 Show Sections

Bowtie2 - map reads against reference genome

Extract the marker sequences and metadata from the MetaPhlAn database

bamPEFragmentSize Estimate the predominant cDNA fragment length from paired-end sequenced BAM/CRAM files

MetaPhlAn to profile the composition of microbial communities

TB-Profiler Profile Infer strain types

Edit dataset at

≡ Attributes

Edit attributes

Name

GCF_0008753851

Info

Annotation

1. Set bowtie2 parameters:

```
- Is this single or paired library: paired.
- FASTA/Q file #1 : ERR5310322_1
- FASTA/Q file #2 : ERR5310322_2
- Will you select a reference genome from your history or use a buil
- Do you want to use presets? : Very sensitive local. This setting
- Save the bowtie2 mapping statistics to the history
```

Bowtie2 - map reads against reference genome (Galaxy Version 2.4.2+galaxy0)

⚙ Versions **▾ Options**

Is this single or paired library

Paired-end ▾

FASTA/Q file #1

5: ERR5310322_1 ▾

Must be of datatype "fastqsanger"or "fasta"

FASTA/Q file #2

6: ERR5310322_2 ▾

Must be of datatype "fastqsanger"or "fasta"

Write unaligned reads (in fastq format) to separate file(s)

⚫ No

--un/--un-conc (possibly with -gz or -bz2); This triggers --un parameter for single reads and --un-conc for paired reads

Write aligned reads (in fastq format) to separate file(s)

⚫ No

--al/--al-conc (possibly with -gz or -bz2); This triggers --al parameter for single reads and --al-conc for paired reads

Do you want to set paired-end options?

No ▾

See "Alignment Options" section of Help below for information

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

Use a genome from the history and build index ▾

Built-ins were indexed using default options. See `Indexes` section of help below

Select reference genome

7: GCF_000875385.1_ViralProj30293_genomic.fna.gz (as fasta) ▾

Set read groups information?

Do not set ▾

Specifying read group information can greatly simplify your downstream analyses by allowing combining multiple datasets.

Select analysis mode

1: Default setting only ▾

Do you want to use presets?

○ No, just use defaults
○ Very fast end-to-end (--very-fast)
○ Fast end-to-end (--fast)
○ Sensitive end-to-end (--sensitive)
○ Very sensitive end-to-end (--very-sensitive)
○ Very fast local (--very-fast-local)
○ Fast local (--fast-local)
○ Sensitive local (--sensitive-local)
⊙ Very sensitive local (--very-sensitive-local)

Allow selecting among several preset parameter settings. Choosing between these will result in dramatic changes in runtime. See help below to understand effects of these presets.

Do you want to tweak SAM/BAM Options?

No ▾

See "Output Options" section of Help below for information

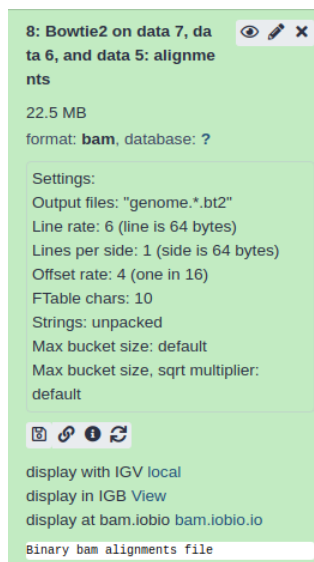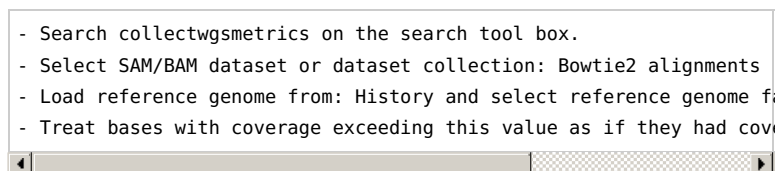**Save the bowtie2 mapping statistics to the history**

⚫ Yes

✔ Execute

1. Click execute and wait.

Visualize bam file and calculate metrics

1. Click the   icon in the Bowtie2 aligments in history.

```
8: Bowtie2 on data 7, da    👁 ✎ ✕
ta 6, and data 5: alignme
nts

22.5 MB
format: bam, database: ?

   Settings:
   Output files: "genome.*.bt2"
   Line rate: 6 (line is 64 bytes)
   Lines per side: 1 (side is 64 bytes)
   Offset rate: 4 (one in 16)
   FTable chars: 10
   Strings: unpacked
   Max bucket size: default
   Max bucket size, sqrt multiplier:
   default

🖹 🔗 ❶ ⟳

display with IGV local
display in IGB View
display at bam.iobio bam.iobio.io
Binary bam alignments file
```

1. Interpret the columns in the bam format according to the theory from class.

2. Visualize mapping metrics

    - Click on the eye icon on Bowtie2 mapping stats history.
        - ▶ Which is the mapping rate?

3. Calculate depth of coverage metrics using picard collectWGSMetrics.

```
- Search collectwgsmetrics on the search tool box.
- Select SAM/BAM dataset or dataset collection: Bowtie2 alignments
- Load reference genome from: History and select reference genome fa
- Treat bases with coverage exceeding this value as if they had cove
◄                                                                  ►
```

1. Click execute and wait.

Visualize bam file using IGV

In order to visualize our mapping we will use IGV (Integrative Genomics Viewer). This is an open source, freely available and lightweight visualization tool that enables intuitive real-time exploration of diverse, large-scale genomic data sets on standard desktop computers. It supports flexible integration of a wide range of genomic data types including aligned sequence reads, mutations, copy number, RNA interference screens, gene expression, methylation and genomic annotations.

Navigation through a data set is similar to that of Google Maps, allowing the user to zoom and pan seamlessly across the genome at any level of detail, from whole genome to base pair. Data sets can be loaded from either local or remote sources, including cloud-based resources, enabling investigators to view their own genomic data sets alongside publicly available data.

1. Install IGV
2. Launch IGV on your computer
3. Expand the param-file output of Bowtie2 tool
4. Click on the local in display with IGV to load the reads into the IGV browser
5. Here you have a galaxy training document for IGV usage.