Galaxy for virologist training Exercise 8: Viralrecon

Title	Galaxy
Training dataset:	SARS-CoV-2 downsampled sequencing data used to report variants and lineages to national Spanish epidemiologist.
Questions:	• How many variants does the samples have • Whoch lineage do the samples belong to?
Objectives:	• Learn how to run viralrecon in Galaxy's interface • Understand the results generated
Estimated time:	1h 15 min

In this report you will find all the information necessary to follow the steps to analyze SARS-CoV-2 data with Galaxy.

Training overview

During this training we will following these steps:

1. Create a new history and name it **Viralrecon**
2. Upload data: Upload data for the analysis.
3. Quality: Analysis of the quality of the raw reads.
4. Trimming: Quality trimming using fastp
5. Mapping: Mapping reads to reference genome with Bowtie2
6. Stats: Mapping statistics with samtools and picard.
7. Amplicons: Preprocessing steps mandatory for amplicon sequencing data.
8. Variants: Variant calling and filtering.
9. Consensus: Consensus genome generation

From now on, each job we run in Galaxy will have a unique number for identifying each process. This numbers can differ depending on the number of samples and the times you run or delete any process. This training's snapshots were taken using other samples and some process were deleted for any reason, so numbers and names MAY DIFFER. However, the steps you have to run are THE SAME

History

1. Create a new history in the + and name it `Viralrecon`

Data

We are going to upload files using these URLS as seen in the Galaxy tutorial first day

```
https://zenodo.org/record/5724464/files/SARSCOV2-1_R1.fastq.gz?download=1
https://zenodo.org/record/5724464/files/SARSCOV2-1_R2.fastq.gz?download=1
https://zenodo.org/record/5724464/files/SARSCOV2-2_R1.fastq.gz?download=1
https://zenodo.org/record/5724464/files/SARSCOV2-2_R2.fastq.gz?download=1
```

Prior to any analysis, we have to download the fasta reference genome using the following URL:

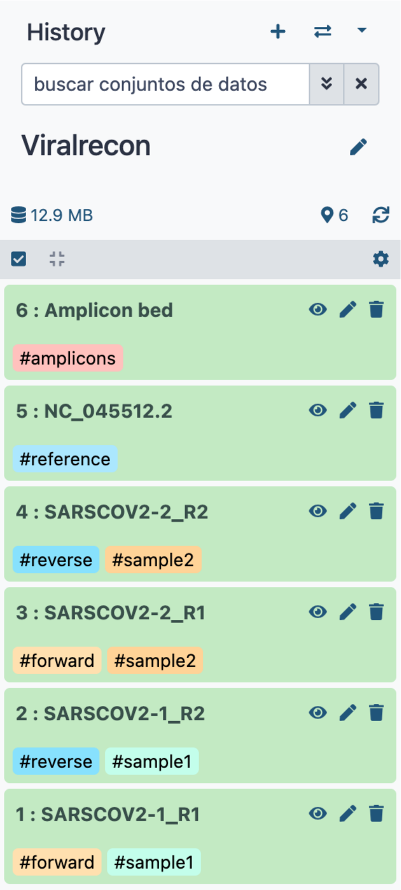
```
https://zenodo.org/record/5724970/files/GCF_009858895.2_ASM985889v3_genomic.200409.fna.gz?download=1
```

Also, you will download the bed file of the amplicon primers, which contains the positions in the reference genome of each of the amplicon primers. Use this URL in the window:

```
https://zenodo.org/record/5724970/files/nCoV-2019.artic.V3.scheme.bed.txt?download=1
```

Finally, rename and tag the data as follows:

- `SARSCOV2-1_R1.fastq.gz`to `SARSCOV2-1_R1`with tagS `#sample1` and `#forward`
- `SARSCOV2-1_R2.fastq.gz`to `SARSCOV2-1_R2`with tagS `#sample1` and `#reverse`
- `SARSCOV2-2_R1.fastq.gz`to `SARSCOV2-2_R1`with tagS `#sample2` and `#forward`
- `SARSCOV2-2_R2.fastq.gz`to `SARSCOV2-2_R2`with tagS `#sample2` and `#reverse`
- `GCF_009858895.2_ASM985889v3_genomic.200409.fna.gz?download=1` to `NC_045512.2` with tag `#reference`
- `nCoV-2019.artic.V3.scheme.bed.txt?download=1` to `Amplicon bed` with tag `#amplicons`



History

buscar conjuntos de datos

Viralrecon

12.9 MB 6

6 : Amplicon bed
#amplicons

5 : NC_045512.2
#reference

4 : SARSCOV2-2_R2
#reverse #sample2

3 : SARSCOV2-2_R1
#forward #sample2

2 : SARSCOV2-1_R2
#reverse #sample1

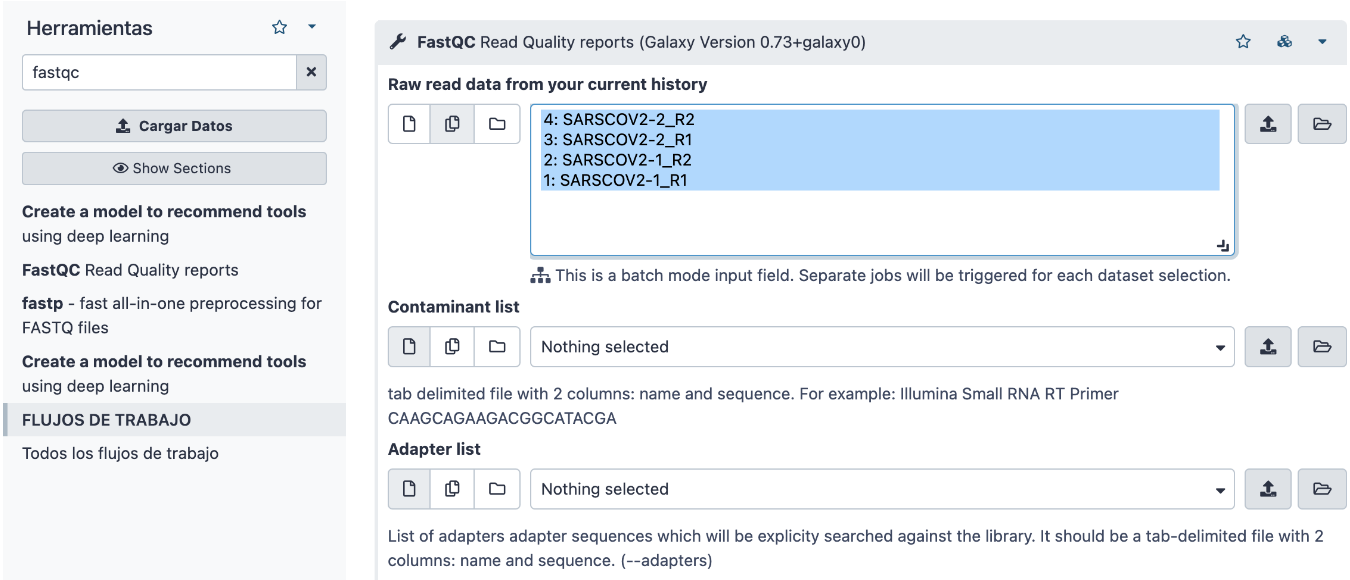
1 : SARSCOV2-1_R1
#forward #sample1

Quality

Quality Analysis (FastQC)

Once we have the raw data, an important step is to analyze the quality of the reads, to know if the reads are worth it. To do this, we have to look for the program "*FastQC*" in the search bar, then select **FastQC Read Quality reports** and set the following parameters, same as here:

- Select multiple file data set and select the fastq files R1 and R2 for both samples
- With *Ctrl* select the two datasets
- Then go down and select **Execute**



Herramientas

fastqc

Create a model to recommend tools using deep learning

FastQC Read Quality reports

fastp - fast all-in-one preprocessing for FASTQ files

Create a model to recommend tools using deep learning

FLUJOS DE TRABAJO

Todos los flujos de trabajo

FastQC Read Quality reports (Galaxy Version 0.73+galaxy0)

Raw read data from your current history

4: SARSCOV2-2_R2
3: SARSCOV2-2_R1
2: SARSCOV2-1_R2
1: SARSCOV2-1_R1

This is a batch mode input field. Separate jobs will be triggered for each dataset selection.

Contaminant list

Nothing selected

tab delimited file with 2 columns: name and sequence. For example: Illumina Small RNA RT Primer CAAGCAGAAGACGGCATACGA

Adapter list

Nothing selected

List of adapters adapter sequences which will be explicity searched against the library. It should be a tab-delimited file with 2 columns: name and sequence. (--adapters)

FastQC results visualization and interprepation questions

To visualize the information coming from FastQC we just have to select the job of interest. In this case we are interested in the "*Web page results*" so for the sample we want to see the results we have to click in the *eye* to visualize galaxy results:

For more information about FastQC output visit FasxstQC website

Question

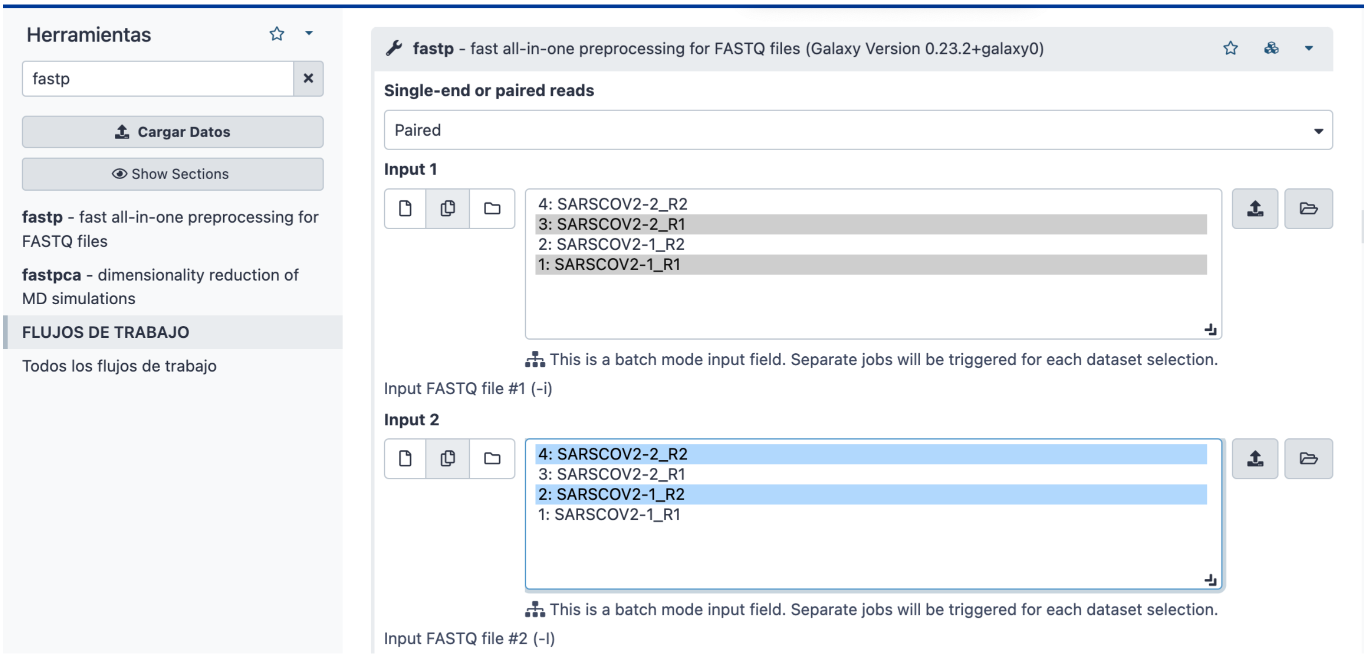
▶ Which is the read length? What type of sequencing are we doing?
▶ How many reads has samlpe1 before trimming?
▶ How many reads has samlpe2 before trimming?

Trimming

Quality trimming (Fastp)

Once we have check the quality of our reads, it's important to trim low quality nucleotides from those reads, for which we will use *Fastp*. So, in the search bar you look for fastp and then select "*fastp - fast all-in-one preprocessing for FASTQ files*". There, we will have to change some parameters ensure the trimming accuracy for this amplicon data. First of all we are going to do the analysis for the sample we gave to you (201569). These are the field we will have to change:

1. Search for **fastp** in the tools and select **fastp - fast all-in-one preprocessing for FASTQ files**
2. Select custom parameters:
    - Single-end or paired reads > Paired
        - Input 1 > Browse datasets (right folder icon) > Select both forward reads for both samples
        - Input 2 > Browse datasets > Select reverse reads for both samples
    - Display Filter Options
        - Quality Filtering options
            - Qualified Quality Phred = 30
            - Unqualified percent limit = 10
        - Length Filtering Options
            - Length required = 50
    - Read modification options
        - PoliX tail trimming > Enable polyX tail trimming
        - Per read cutting by quality options
            - Cut by quality in front (5') > Yes
            - Cut by quality in tail (3') > Yes
            - Cutting mean quality = 30
3. Finally, click on **Execute**



fastp - fast all-in-one preprocessing for FASTQ files

fastpca - dimensionality reduction of MD simulations

fastp - fast all-in-one preprocessing for FASTQ files (Galaxy Version 0.23.2+galaxy0)

Single-end or paired reads

Paired

Input 1

4: SARSCOV2-2_R2
3: SARSCOV2-2_R1
2: SARSCOV2-1_R2
1: SARSCOV2-1_R1

This is a batch mode input field. Separate jobs will be triggered for each dataset selection.

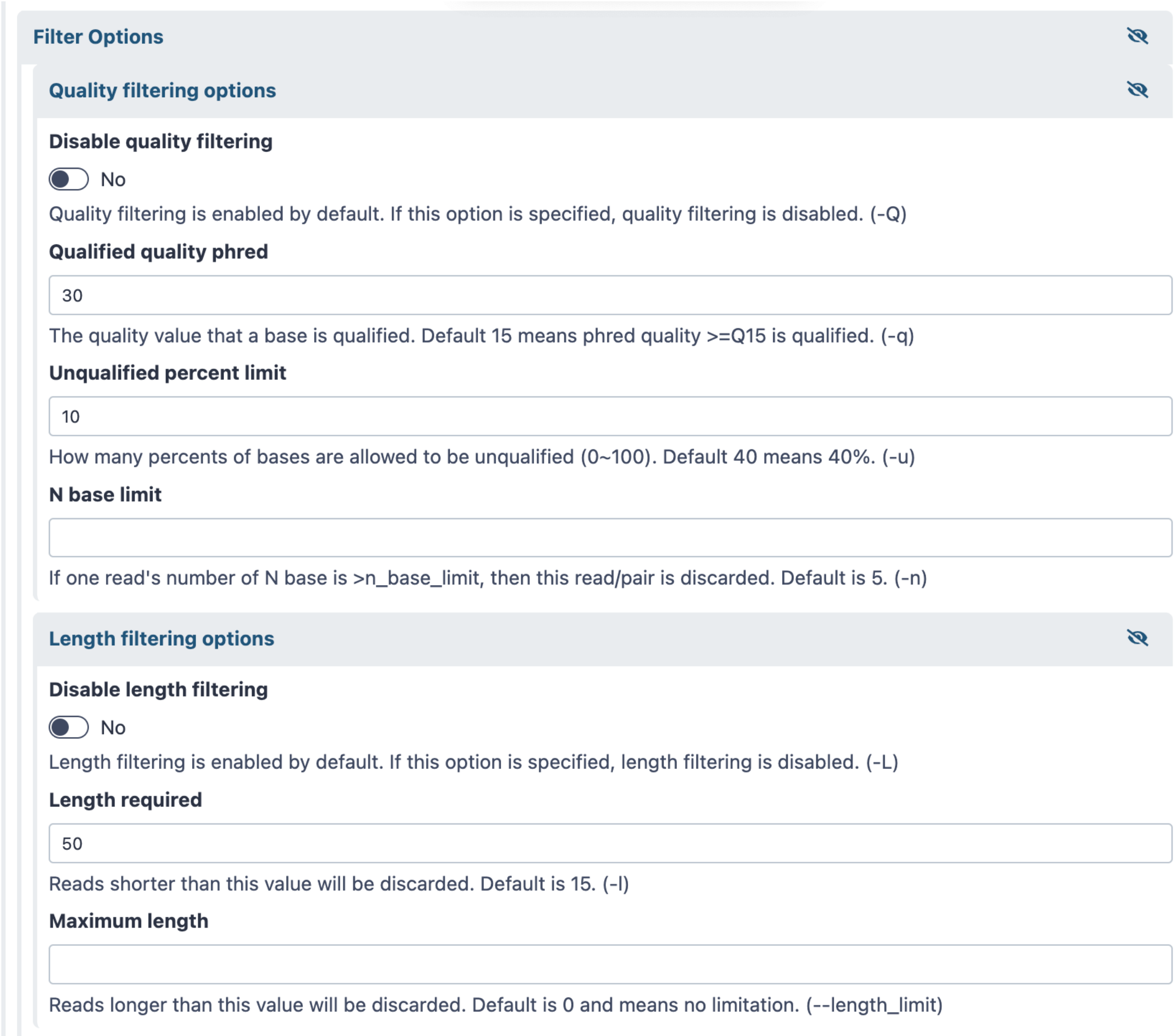
Input FASTQ file #1 (-i)

Input 2

4: SARSCOV2-2_R2
3: SARSCOV2-2_R1
2: SARSCOV2-1_R2
1: SARSCOV2-1_R1

This is a batch mode input field. Separate jobs will be triggered for each dataset selection.

Input FASTQ file #2 (-I)



Filter Options

Quality filtering options

Disable quality filtering

No

Quality filtering is enabled by default. If this option is specified, quality filtering is disabled. (-Q)

Qualified quality phred

30

The quality value that a base is qualified. Default 15 means phred quality >=Q15 is qualified. (-q)

Unqualified percent limit

10

How many percents of bases are allowed to be unqualified (0~100). Default 40 means 40%. (-u)

N base limit

If one read's number of N base is >n_base_limit, then this read/pair is discarded. Default is 5. (-n)

Length filtering options

Disable length filtering

No

Length filtering is enabled by default. If this option is specified, length filtering is disabled. (-L)

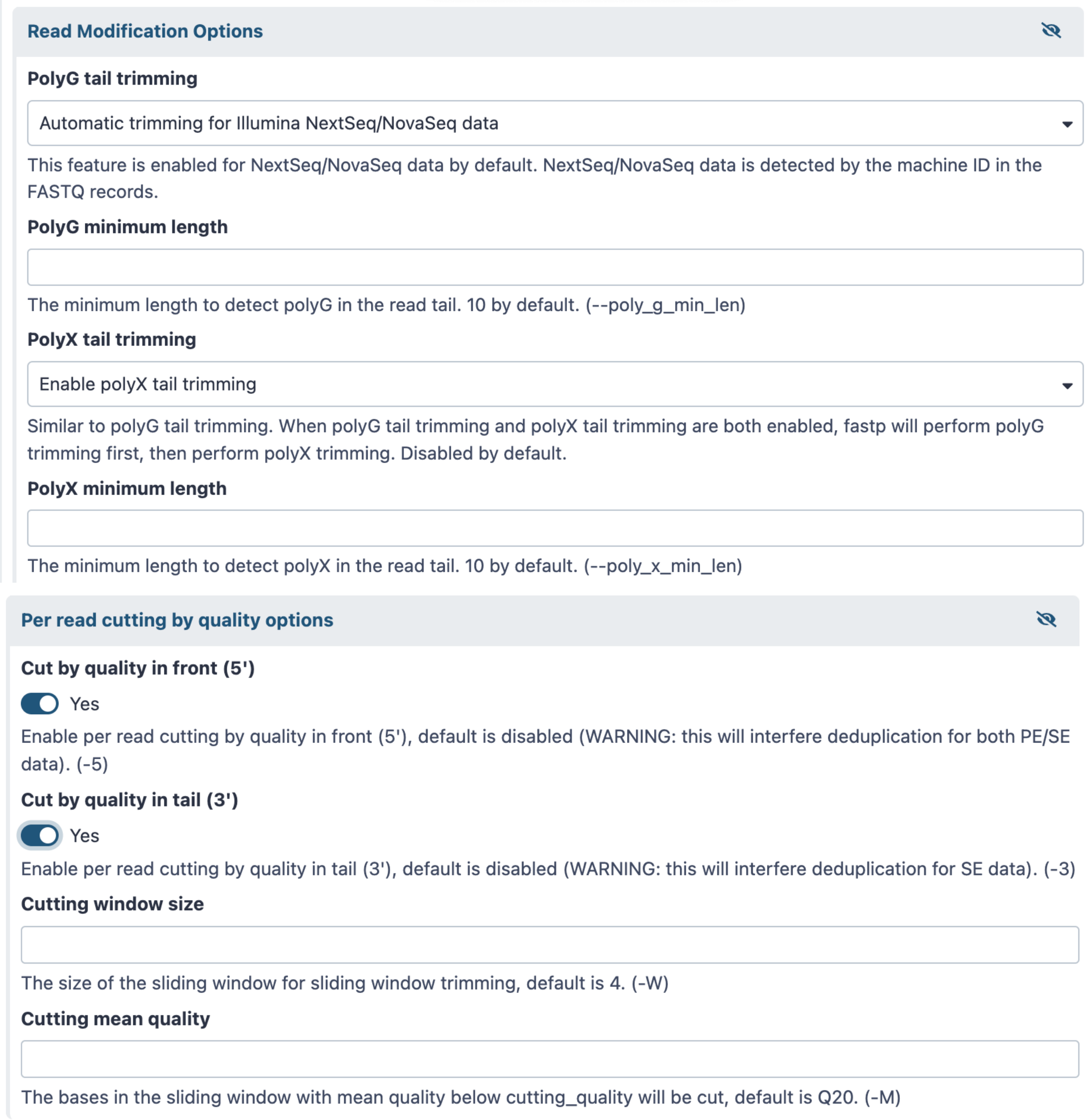
Length required

50

Reads shorter than this value will be discarded. Default is 15. (-l)

Maximum length

Reads longer than this value will be discarded. Default is 0 and means no limitation. (--length_limit)



Read Modification Options

PolyG tail trimming

Automatic trimming for Illumina NextSeq/NovaSeq data

This feature is enabled for NextSeq/NovaSeq data by default. NextSeq/NovaSeq data is detected by the machine ID in the FASTQ records.

PolyG minimum length

The minimum length to detect polyG in the read tail. 10 by default. (--poly_g_min_len)

PolyX tail trimming

Enable polyX tail trimming

Similar to polyG tail trimming. When polyG tail trimming and polyX tail trimming are both enabled, fastp will perform polyG trimming first, then perform polyX trimming. Disabled by default.

PolyX minimum length

The minimum length to detect polyX in the read tail. 10 by default. (--poly_x_min_len)

Per read cutting by quality options

Cut by quality in front (5')

Yes

Enable per read cutting by quality in front (5'), default is disabled (WARNING: this will interfere deduplication for both PE/SE data). (-5)

Cut by quality in tail (3')

Yes

Enable per read cutting by quality in tail (3'), default is disabled (WARNING: this will interfere deduplication for SE data). (-3)

Cutting window size

The size of the sliding window for sliding window trimming, default is 4. (-W)

Cutting mean quality

The bases in the sliding window with mean quality below cutting_quality will be cut, default is Q20. (-M)

A message will appear, which means that 6 results will be generated:

1. Two, one with the R1 trimmed reads, for each sample
2. Another two, one with the R2 trimmed reads, for each sample
3. Two, one with the HTML results, for each sample

Fastp results

Once fastp analysis is done, you can see the results by clicking in the eye ("*View Data*") in the fatp HTML results.

Among the most relevant results, you have the:

- Summary: Stats summary
  - After filtering: Statistics of the reads after quality filtering
    - reads passed filters: Reads remaining after quality filter trimming

        - reads with low quality: Reads that were remove due to low quality
        - reads too short: Reads that didn't pass the minimum length filter.
    - After filtering: Plots after filtering
        - After filtering: read1: quality: Plot with the evolution of R1 quality over read position. Usually it decays in the last nucleotides.
        - After filtering: read2: quality: Same plot for R2.

For more information about FastQC output visit Fastp github

Question

▶ How many reads are we keeping from sample1?

▶ How many reads did we lost for sample1 and why?

▶ How many reads are we keeping from sample2?

▶ How many reads did we lost for sample2 and why?

Mapping

In order to call for variants between the samples and the reference, it's mandatory to map the sample reads to the reference genome. To do this we need the fasta file of the reference and the Bowtie2 index of that fasta file.

Mapping reads with reference genome (Bowtie2)

Now we can start with the main mapping process. The first thing we have to do is look for the program "*Bowtie2*" in the search bar and then select "*Bowtie2 - map reads against reference genome*". Here we will have to set the following parameters, for the first sample, same as here

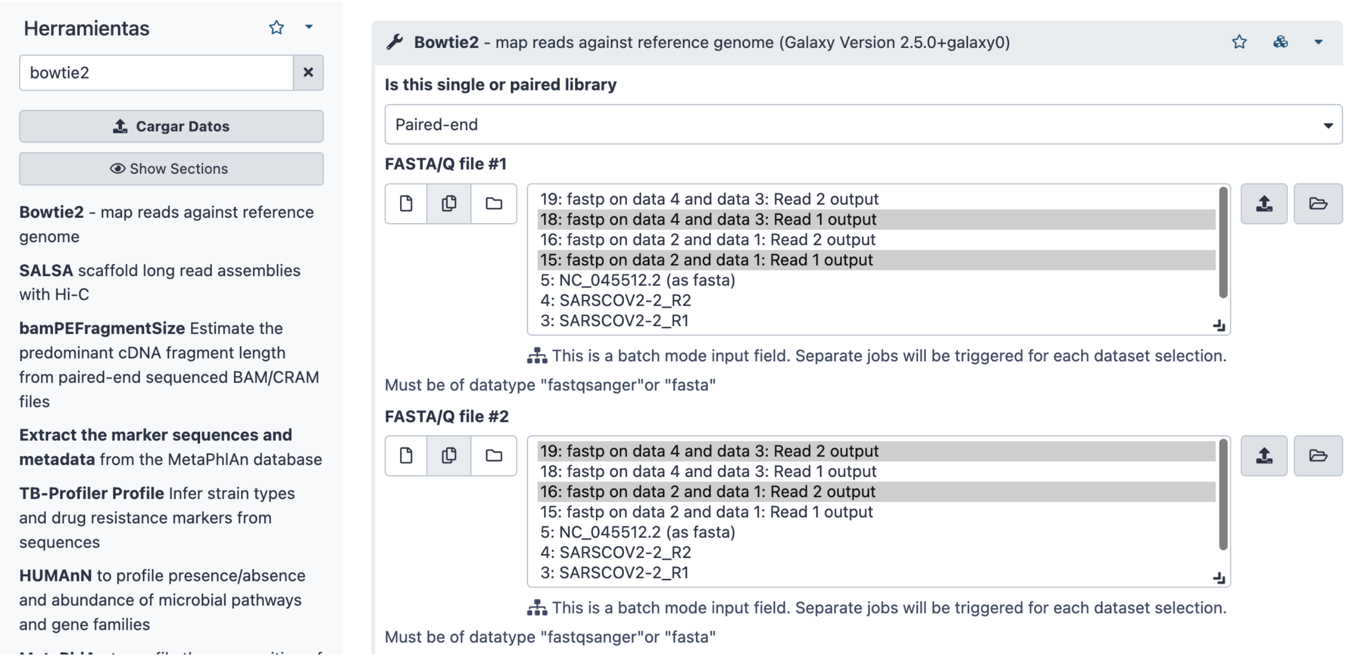
3. Is this single or paired library > Paired-end
4. Fasta/Q file #1: **fastp Read 1 output** for both samples
5. Fasta/Q file #2: **fastp Read 2 output** for both samples
6. Will you select a reference genome from your history or use a built-in index? > Use a genome from the history and create index

- **This is very important because we haven't previously created the SARS-Cov2 genome index, si bowtie 2 will generate it automatically.**

7. Select reference genome > GCF_009858895.2_ASM985889v3_genomic.200409.fna.gz

- It's important to select the file we downloaded from URL.

8. Do you want to use presets? > Very sensitive local
9. Save the bowtie2 mapping statistics to the history > Yes
10. Execute



Herramientas

bowtie2

Cargar Datos

Show Sections

Bowtie2 - map reads against reference genome

SALSA scaffold long read assemblies with Hi-C

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

Extract the marker sequences and metadata from the MetaPhlAn database

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

HUMAnN to profile presence/absence and abundance of microbial pathways and gene families

Bowtie2 - map reads against reference genome (Galaxy Version 2.5.0+galaxy0)

Is this single or paired library

Paired-end

FASTA/Q file #1

19: fastp on data 4 and data 3: Read 2 output
18: fastp on data 4 and data 3: Read 1 output
16: fastp on data 2 and data 1: Read 2 output
15: fastp on data 2 and data 1: Read 1 output
5: NC_045512.2 (as fasta)
4: SARSCOV2-2_R2
3: SARSCOV2-2_R1

This is a batch mode input field. Separate jobs will be triggered for each dataset selection.

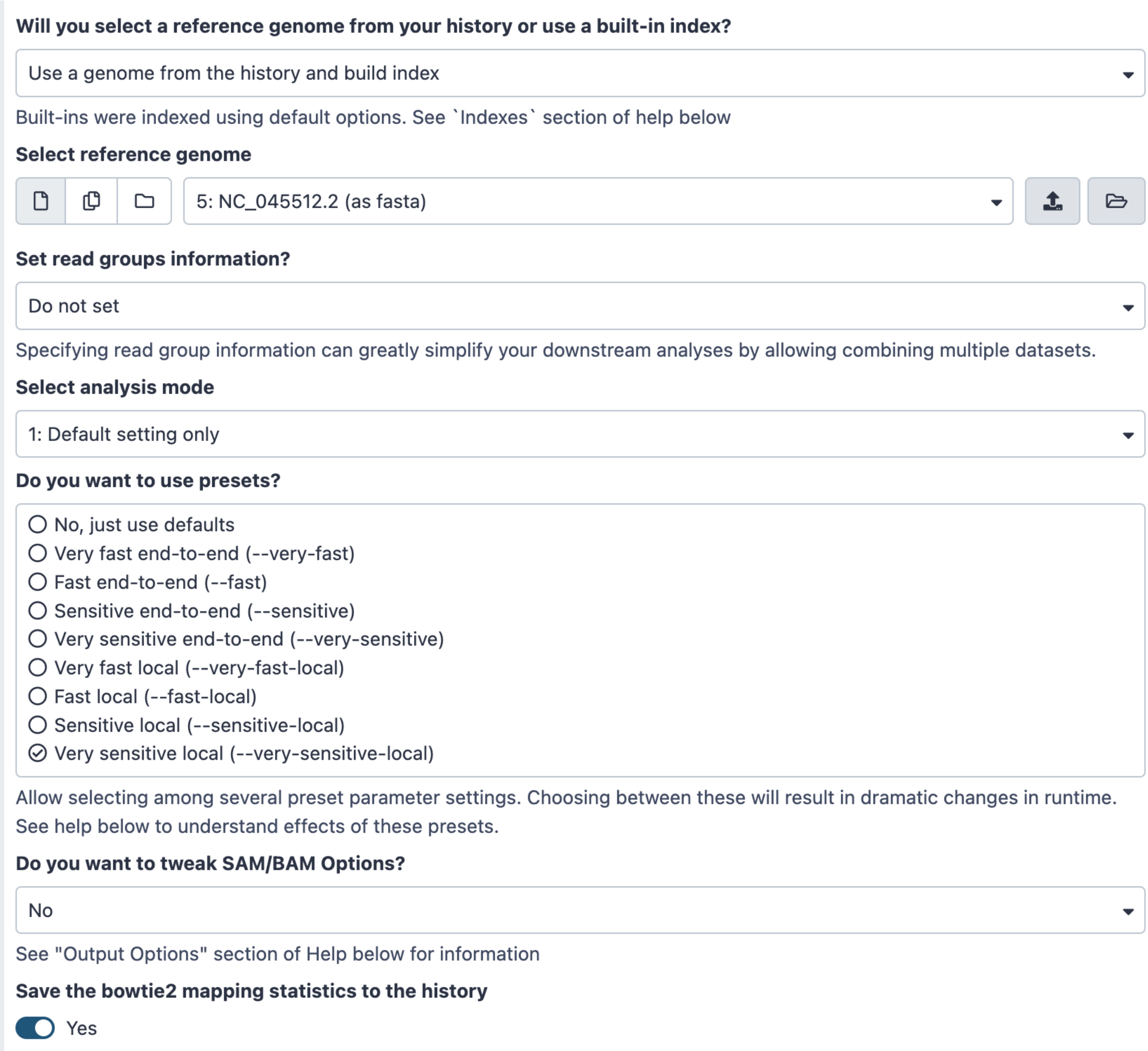
Must be of datatype "fastqsanger"or "fasta"

FASTA/Q file #2

19: fastp on data 4 and data 3: Read 2 output
18: fastp on data 4 and data 3: Read 1 output
16: fastp on data 2 and data 1: Read 2 output
15: fastp on data 2 and data 1: Read 1 output
5: NC_045512.2 (as fasta)
4: SARSCOV2-2_R2
3: SARSCOV2-2_R1

This is a batch mode input field. Separate jobs will be triggered for each dataset selection.

Must be of datatype "fastqsanger"or "fasta"



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

5: NC_045512.2 (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

Mapping results

Now we can see the mapping results for the samples. The bowtie2 resulting file is a .bam file, which is not easy to read by humans. This .bam file can be downloaded by clicking in the alignment file and then into download. Then, the resulting .gz file will contain the alignment .bam file that can be introduced in a software such as IGV with the reference genome fasta file.

In our case, the file that can be visualize is the `mapping stats` file, which contains information such as the percentage of reads that aligned.

Question

▶ Which is the overall alignment rate for sample1?

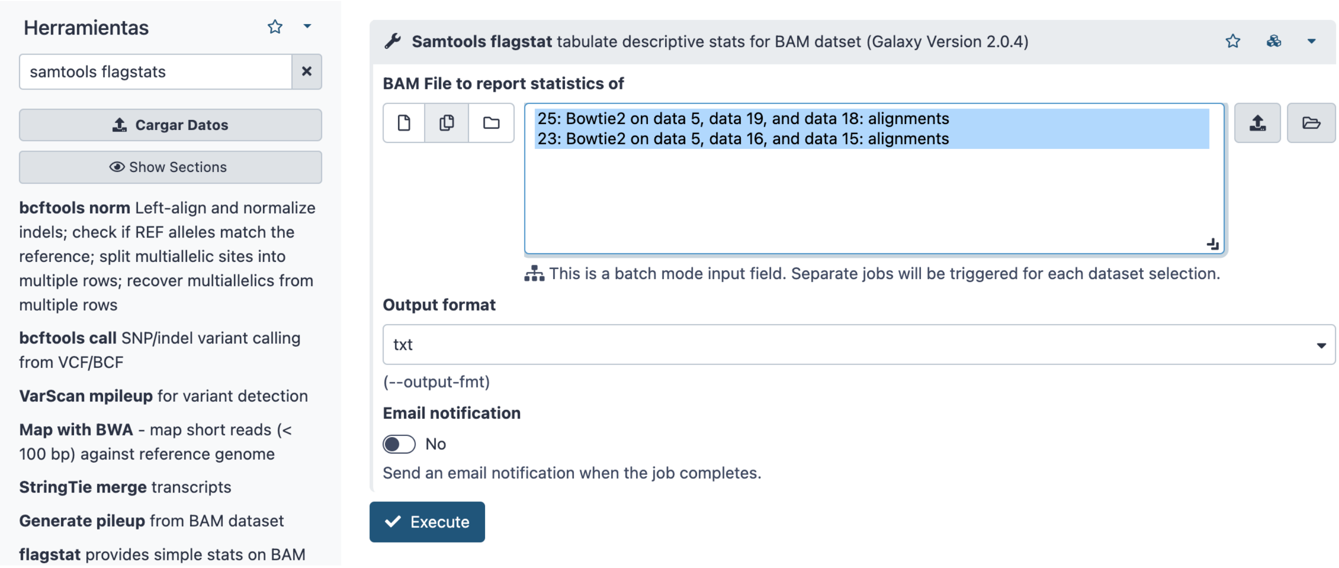
▶ And the overall alignment rate for sample2?

Stats

The previously shown files give few human readable information, because mapping files are supposed to be used by other programs. In this sense, we can use some programs to extract relevant statistical information about the mapping process.

Samtools flagstat

The first program is Samtools, from which we will use the module samtools flagstat. To do this, we have to look in the search bar for "*samtools flagstat*" and then select "*Samtools flagstat tabulate descriptive stats for BAM datset*". There, we just have to select the samples we want to perform the mapping stats (in the example there are two samples, you just have to use one): *Bowtie2 on data X, data X and data X: alingment*. You can select the samples from the list in *Multiple datasets* or select the folder icon (*Browse datasets*) to select the file from the history. Finally, select *Execute*



samtools flagstats

bcftools norm Left-align and normalize indels; check if REF alleles match the reference; split multiallelic sites into multiple rows; recover multiallelics from multiple rows

bcftools call SNP/indel variant calling from VCF/BCF

VarScan mpileup for variant detection

Map with BWA - map short reads (< 100 bp) against reference genome

StringTie merge transcripts

Generate pileup from BAM dataset

flagstat provides simple stats on BAM

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

BAM File to report statistics of

25: Bowtie2 on data 5, data 19, and data 18: alignments
23: Bowtie2 on data 5, data 16, and data 15: alignments

This is a batch mode input field. Separate jobs will be triggered for each dataset selection.

Output format

txt

(--output-fmt)

Email notification

No

Send an email notification when the job completes.

Samtools results

The results of the samtools program gives information about the number and percentage of reads that mapped with the reference genome.

Question

▶ How many reads mapped against the refernce genome for sample1?

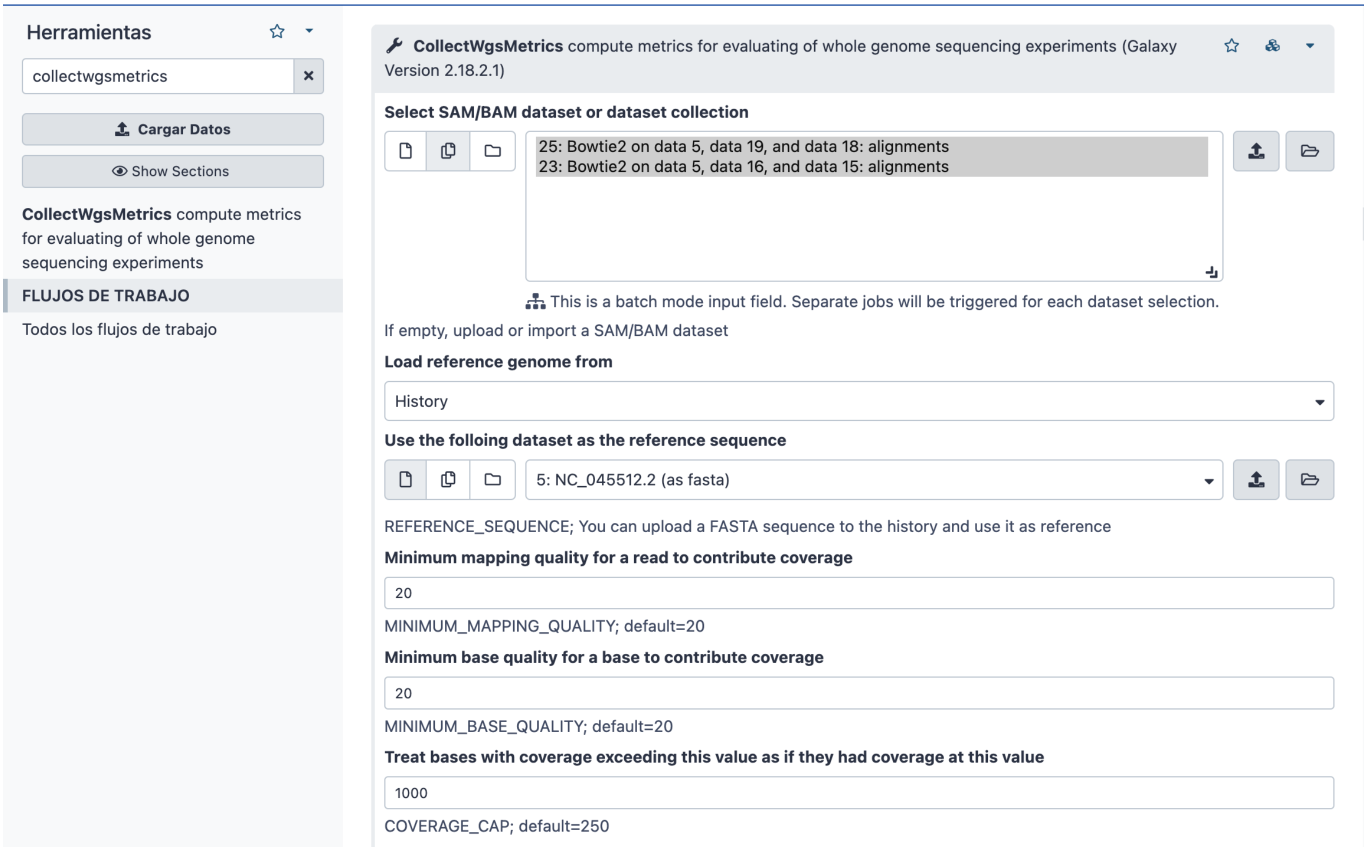
▶ And how many for sample2?

Picard CollectWgsMetrics

Another program that gives statistical information about the mapping process is Picard. To run this program you just have to search "*Collect Wgs Metrics*" and then select "*CollectWgsMetrics compute metrics for evaluating of whole genome sequencing experiments*".

You have to change the following parameters:

1. Select SAM/BAM dataset or dataset collection > Dataset collection > Select both bam files at once
2. Load reference genome from > History
3. Select the fasta file we uploaded with the reference genome (NC_045512.2).
4. Treat bases with coverage exceeding this value as if they had coverage at this value = 1000
5. Select validation stringency > Lenient
6. Execute.


Herramientas
collectwgsmetrics
Cargar Datos
Show Sections
CollectWgsMetrics compute metrics for evaluating of whole genome sequencing experiments
FLUJOS DE TRABAJO
Todos los flujos de trabajo

CollectWgsMetrics compute metrics for evaluating of whole genome sequencing experiments (Galaxy Version 2.18.2.1)

Select SAM/BAM dataset or dataset collection
25: Bowtie2 on data 5, data 19, and data 18: alignments
23: Bowtie2 on data 5, data 16, and data 15: alignments
This is a batch mode input field. Separate jobs will be triggered for each dataset selection.
If empty, upload or import a SAM/BAM dataset

Load reference genome from
History

Use the folloing dataset as the reference sequence
5: NC_045512.2 (as fasta)
REFERENCE_SEQUENCE; You can upload a FASTA sequence to the history and use it as reference

Minimum mapping quality for a read to contribute coverage
20
MINIMUM_MAPPING_QUALITY; default=20

Minimum base quality for a base to contribute coverage
20
MINIMUM_BASE_QUALITY; default=20

Treat bases with coverage exceeding this value as if they had coverage at this value
1000
COVERAGE_CAP; default=250

This process will generate one output file per .bam alignment file selected as input.

Picard results

Picard results consist in quite long files, so the best is to download those results and visualize them in your computer. Yo you have to click in the CollectWgsMetrics job you want to download, and then click in the save button:

Then you just have to open the file with Excell in your computer, and you will see a file with different columns with information about the percentage of the reference genome that is covered by the reads at a specific depth or the mean depth of coverage of the reference genome.

Question

- ▶ Which is the mean coverage for sample1?
- ▶ Which percetage of the reference genome is covered to more than 10X by sample1 reads?
- ▶ Which is the mean coverage for sample2?
- ▶ Which percetage of the reference genome is covered to more than 10X by sample2 reads?

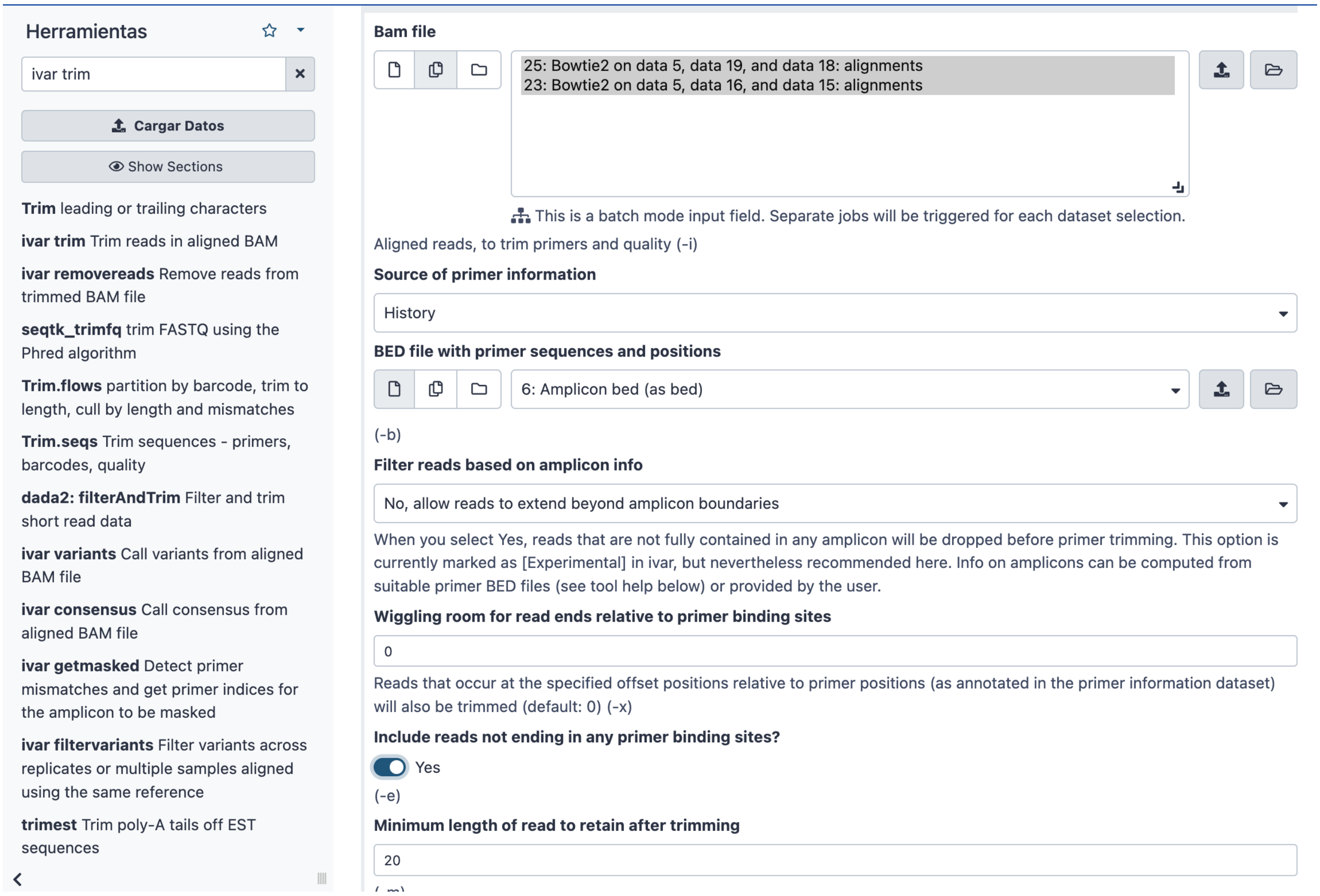
Amplicons

After mapping the reads to the reference genome, we are interested in removing the sequences of the amplicon primers. To do that you will use a program called iVar, and you will need a bed file with the positions of those amplicon primers.

Trim amplicon sequences

Once you have the bed file, you just have to search for "*ivar trim*" in the search bar and select "*ivar trim Trim reads in aligned BAM*". Then follow these steps:

1. Bam file > Select the aligment bam file generated with Bowtie2 for both samples.
2. BED file with primer sequences and positions > Select the Amplicon bed file.
3. Include reads with no primers > Yes.
4. Minimum length of read to retain after trimming = 20



Trim leading or trailing characters

ivar trim Trim reads in aligned BAM

ivar removereads Remove reads from trimmed BAM file

seqtk_trimfq trim FASTQ using the Phred algorithm

Trim.flows partition by barcode, trim to length, cull by length and mismatches

Trim.seqs Trim sequences - primers, barcodes, quality

dada2: filterAndTrim Filter and trim short read data

ivar variants Call variants from aligned BAM file

ivar consensus Call consensus from aligned BAM file

ivar getmasked Detect primer mismatches and get primer indices for the amplicon to be masked

ivar filtervariants Filter variants across replicates or multiple samples aligned using the same reference

trimest Trim poly-A tails off EST sequences

Bam file

25: Bowtie2 on data 5, data 19, and data 18: alignments
23: Bowtie2 on data 5, data 16, and data 15: alignments

This is a batch mode input field. Separate jobs will be triggered for each dataset selection.

Aligned reads, to trim primers and quality (-i)

Source of primer information

History

BED file with primer sequences and positions

6: Amplicon bed (as bed)

(-b)

Filter reads based on amplicon info

No, allow reads to extend beyond amplicon boundaries

When you select Yes, reads that are not fully contained in any amplicon will be dropped before primer trimming. This option is currently marked as [Experimental] in ivar, but nevertheless recommended here. Info on amplicons can be computed from suitable primer BED files (see tool help below) or provided by the user.

Wiggling room for read ends relative to primer binding sites

0

Reads that occur at the specified offset positions relative to primer positions (as annotated in the primer information dataset) will also be trimmed (default: 0) (-x)

Include reads not ending in any primer binding sites?

● Yes

(-e)

Minimum length of read to retain after trimming

20

(-m)

iVar trim results

The resulting file from iVar will be a new BAM file where amplicon primer positions will be removed, so there's no result to visualize.

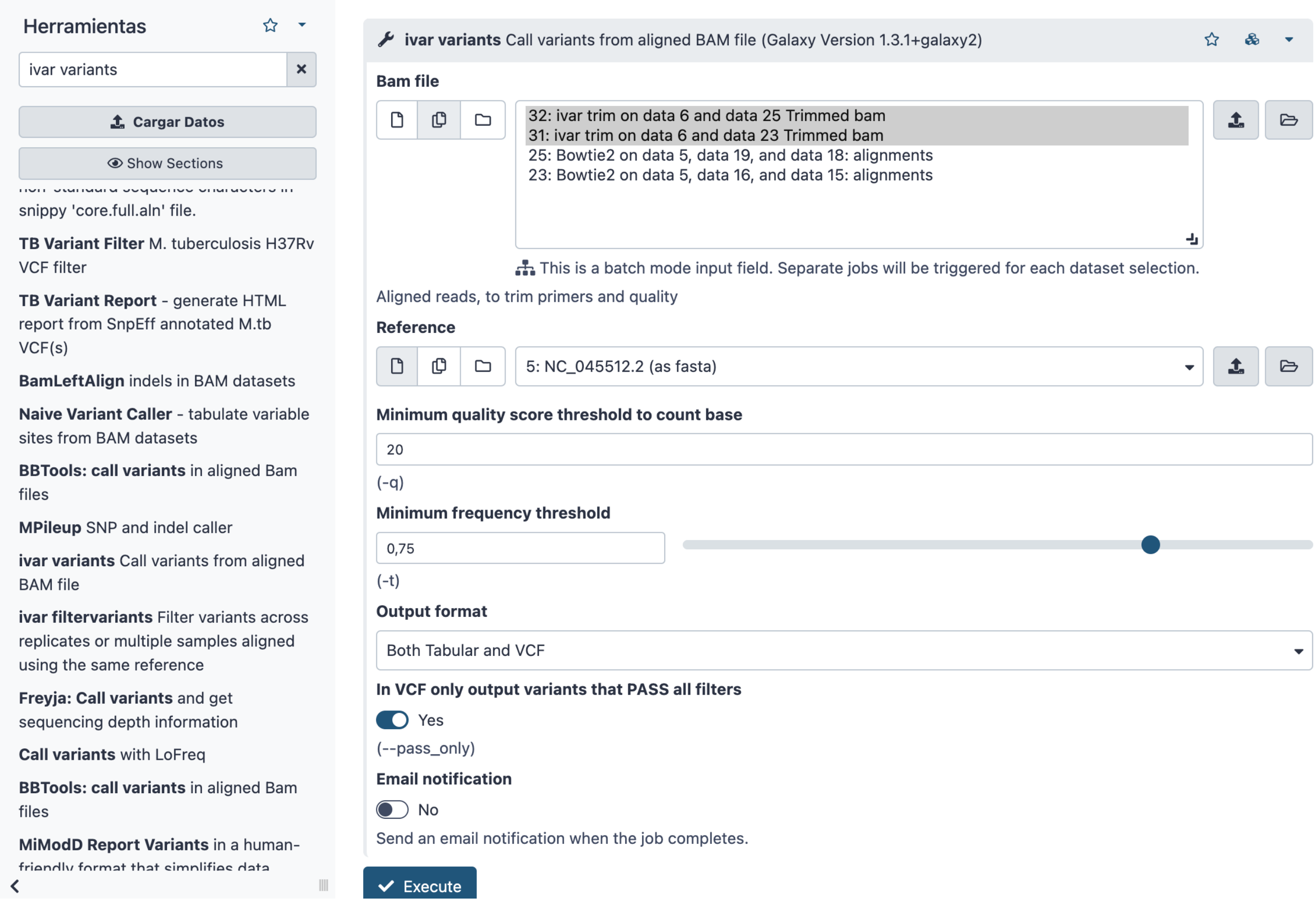
Variants

Once we have the alignment statistics and files with amplicon primers trimmed, we can start with the variant calling process.

iVar variants

iVar uses primer positions supplied in a BED file to soft clip primer sequences from an aligned and sorted BAM file. Following this, the reads are trimmed based on a quality threshold(Default: 20). To do the quality trimming, iVar uses a sliding window approach(Default: 4). The windows slides from the 5' end to the 3' end and if at any point the average base quality in the window falls below the threshold, the remaining read is soft clipped. If after trimming, the length of the read is greater than the minimum length specified(Default: 30), the read is written to the new trimmed BAM file.

1. Search for `ivar variants` and select `ivar variants Call variants from aligned BAM file`
2. Bam file > Select ivar trimmed bam files for both samples
3. Minimum frequency threshold > 0,75
4. Output format > Both tabular and VCF
5. In VCF only output variants that PASS all filters > Yes



snippy 'core.full.aln' file.

TB Variant Filter M. tuberculosis H37Rv VCF filter

TB Variant Report - generate HTML report from SnpEff annotated M.tb VCF(s)

BamLeftAlign indels in BAM datasets

Naive Variant Caller - tabulate variable sites from BAM datasets

BBTools: call variants in aligned Bam files

MPileup SNP and indel caller

ivar variants Call variants from aligned BAM file

ivar filtervariants Filter variants across replicates or multiple samples aligned using the same reference

Freyja: Call variants and get sequencing depth information

Call variants with LoFreq

BBTools: call variants in aligned Bam files

MiModD Report Variants in a human-

ivar variants Call variants from aligned BAM file (Galaxy Version 1.3.1+galaxy2)

Bam file

32: ivar trim on data 6 and data 25 Trimmed bam
31: ivar trim on data 6 and data 23 Trimmed bam
25: Bowtie2 on data 5, data 19, and data 18: alignments
23: Bowtie2 on data 5, data 16, and data 15: alignments

This is a batch mode input field. Separate jobs will be triggered for each dataset selection.

Aligned reads, to trim primers and quality

Reference

5: NC_045512.2 (as fasta)

Minimum quality score threshold to count base

20

(-q)

Minimum frequency threshold

0,75

(-t)

Output format

Both Tabular and VCF

In VCF only output variants that PASS all filters

Yes

(--pass_only)

Email notification

No

Send an email notification when the job completes.

iVar results

iVar results consist in a VCF file containing all the variants found between the reference and the samples. Each line represents a variant the columns give information about that variant, such as the position in the reference genome, the reference allele, the alternate allele, if that variant passed the filters, and so on.

This variants have passed the minimum quality filter, which we set as 20, and the minimum allele frequency of 75%.

Question

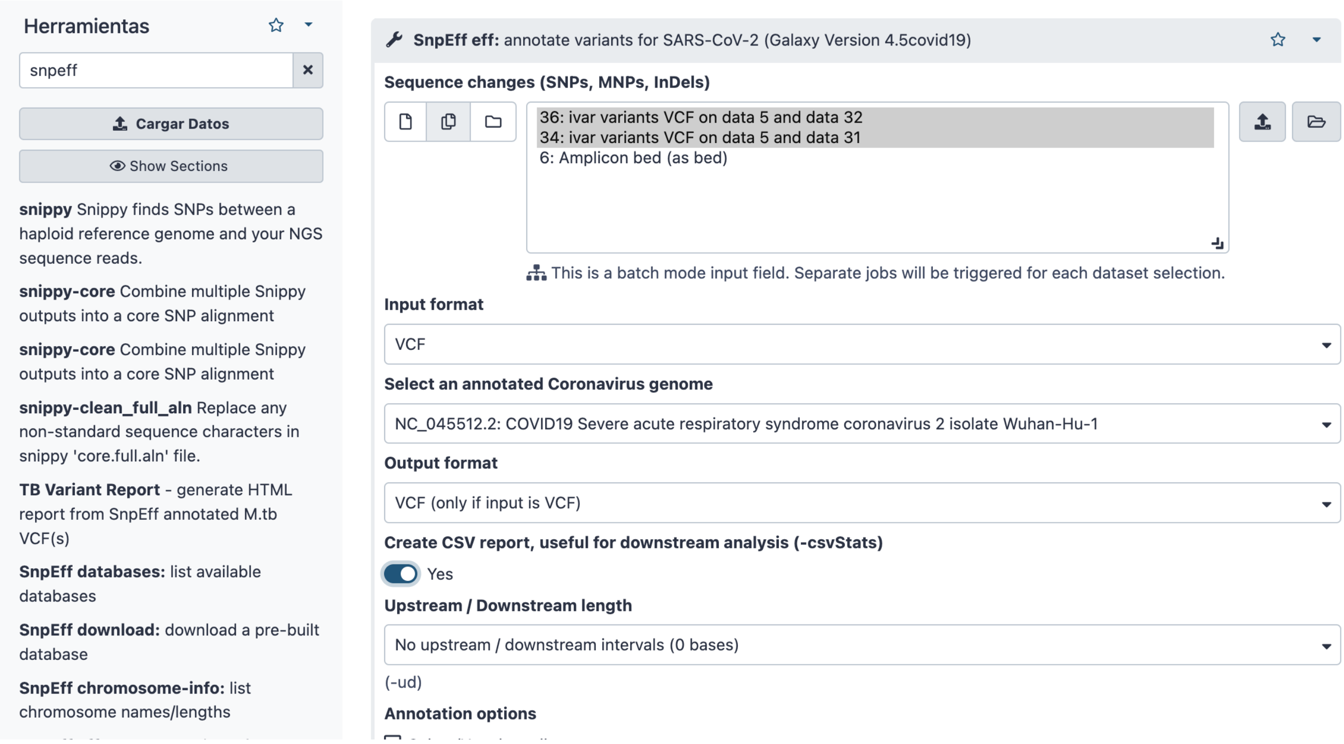
▶ How many possitions are diferent (variant) between reference and sample1 that pass all filters?

▶ And how many between reference and sample2 that pass all filters?

Annotation with SnpEff

Once we have the variants called, it's interesting to annotate those variants, for which you will use SnpEff. Search for "*snpeff*" in the searh bar and select "*SnpEff eff: annotate variants for SARS-CoV-2*", then change the following parameters:

3. Sequence changes (SNPs, MNPs, InDels) > Select ivar output VCF for both samples
4. Create CSV report, useful for downstream analysis (-csvStats) > Yes



Herramientas

snpeff

Cargar Datos

Show Sections

snippy Snippy finds SNPs between a haploid reference genome and your NGS sequence reads.

snippy-core Combine multiple Snippy outputs into a core SNP alignment

snippy-core Combine multiple Snippy outputs into a core SNP alignment

snippy-clean_full_aln Replace any non-standard sequence characters in snippy 'core.full.aln' file.

TB Variant Report - generate HTML report from SnpEff annotated M.tb VCF(s)

SnpEff databases: list available databases

SnpEff download: download a pre-built database

SnpEff chromosome-info: list chromosome names/lengths

SnpEff eff: annotate variants for SARS-CoV-2 (Galaxy Version 4.5covid19)

Sequence changes (SNPs, MNPs, InDels)

36: ivar variants VCF on data 5 and data 32
34: ivar variants VCF on data 5 and data 31
6: Amplicon bed (as bed)

This is a batch mode input field. Separate jobs will be triggered for each dataset selection.

Input format

VCF

Select an annotated Coronavirus genome

NC_045512.2: COVID19 Severe acute respiratory syndrome coronavirus 2 isolate Wuhan-Hu-1

Output format

VCF (only if input is VCF)

Create CSV report, useful for downstream analysis (-csvStats)

Yes

Upstream / Downstream length

No upstream / downstream intervals (0 bases)

(-ud)

Annotation options

SnpEff results

The SnpEff gives three different results, from which the most interesting ones are:

1. Snpeff eff: Which is a VCF file with the annotation results. It is a very similar file to the ones we saw before for VarScan and Bcftools but with the last column different, containing relevant information about that variant.

2. Snpeff eff CSV stats: This file is a CSV file that contains statistics about the variant annotation process, such as the percentage of variants annotated, the percentage of variants that are MISSENSE or SILENT, the percentage that have HIGH, MODERATE or LOW effect, and so on.

Question

▶ How many missense variants has sample1?

- ▶ How many INDELs has sample1?
- ▶ How many missense variants has sample2?
- ▶ How many INDELs has sample2?

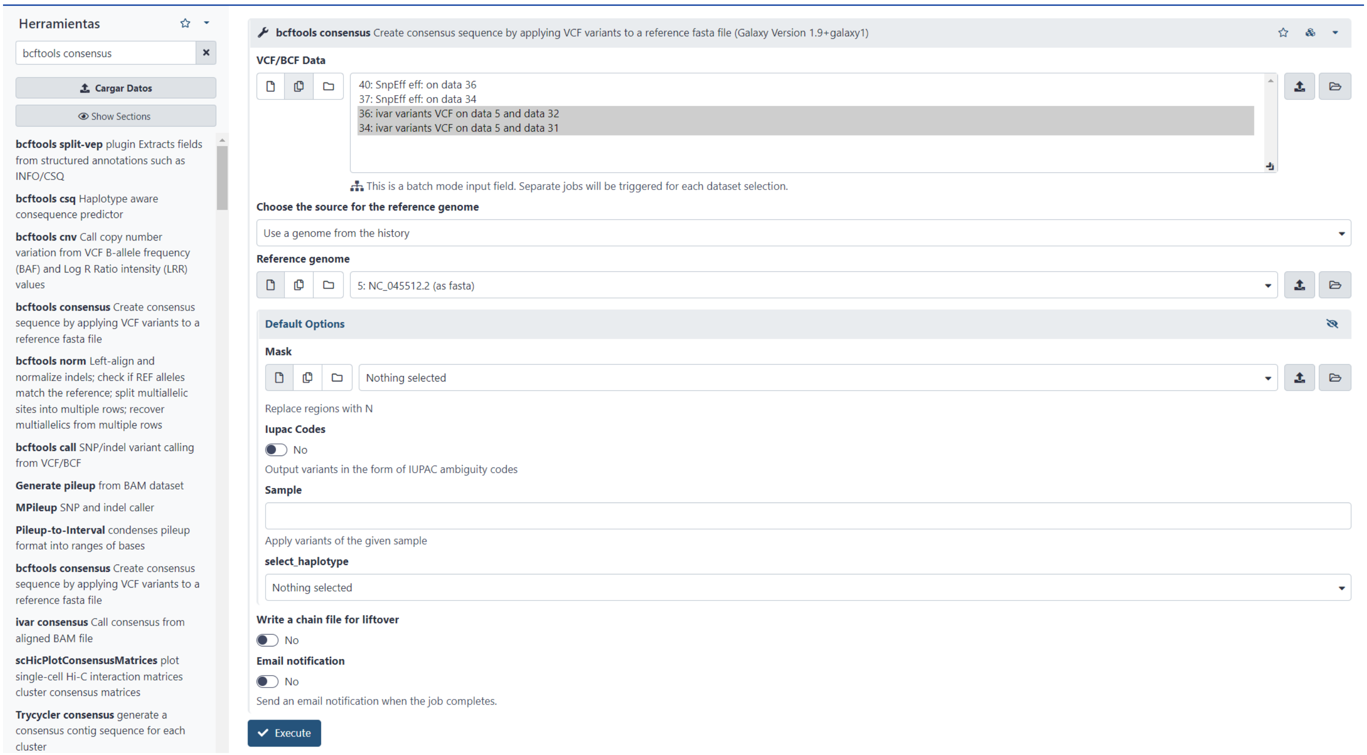
Consensus

Once we have the most relevant variants that can be considered to include in the consensus genome, you can start with the consensus genome generation.

Bcftools consensus

The first step consist in including the called variants into the reference genome, for which you will search for "*bcftools consensus*" in the search bar and then select "*bcftools consensus Create consensus sequence by applying VCF variants to a reference fasta file*". In this module you have to select:

1. VCF/BCF Data > VCF resulting from iVar variants for both samples
2. Choose the source for the reference genome > Use a genome from the history
3. Reference genome > Fasta file uploaded at the begining.
4. Execute



Herramientas

bcftools consensus ✕

Cargar Datos

Show Sections

bcftools split-vep plugin Extracts fields from structured annotations such as INFO/CSQ

bcftools csq Haplotype aware consequence predictor

bcftools cnv Call copy number variation from VCF B-allele frequency (BAF) and Log R Ratio intensity (LRR) values

bcftools consensus Create consensus sequence by applying VCF variants to a reference fasta file

bcftools norm Left-align and normalize indels; check if REF alleles match the reference; split multiallelic sites into multiple rows; recover multiallelics from multiple rows

bcftools call SNP/indel variant calling from VCF/BCF

Generate pileup from BAM dataset

MPileup SNP and indel caller

Pileup-to-Interval condenses pileup format into ranges of bases

bcftools consensus Create consensus sequence by applying VCF variants to a reference fasta file

ivar consensus Call consensus from aligned BAM file

scHicPlotConsensusMatrices plot single-cell Hi-C interaction matrices cluster consensus matrices

Trycycler consensus generate a consensus contig sequence for each cluster

bcftools consensus Create consensus sequence by applying VCF variants to a reference fasta file (Galaxy Version 1.9+galaxy1)

VCF/BCF Data

40: SnpEff eff: on data 36
37: SnpEff eff: on data 34
36: ivar variants VCF on data 5 and data 32
34: ivar variants VCF on data 5 and data 31

This is a batch mode input field. Separate jobs will be triggered for each dataset selection.

Choose the source for the reference genome

Use a genome from the history

Reference genome

5: NC_045512.2 (as fasta)

Default Options

Mask

Nothing selected

Replace regions with N

Iupac Codes

No

Output variants in the form of IUPAC ambiguity codes

Sample

Apply variants of the given sample

select_haplotype

Nothing selected

Write a chain file for liftover

No

Email notification

No

Send an email notification when the job completes.

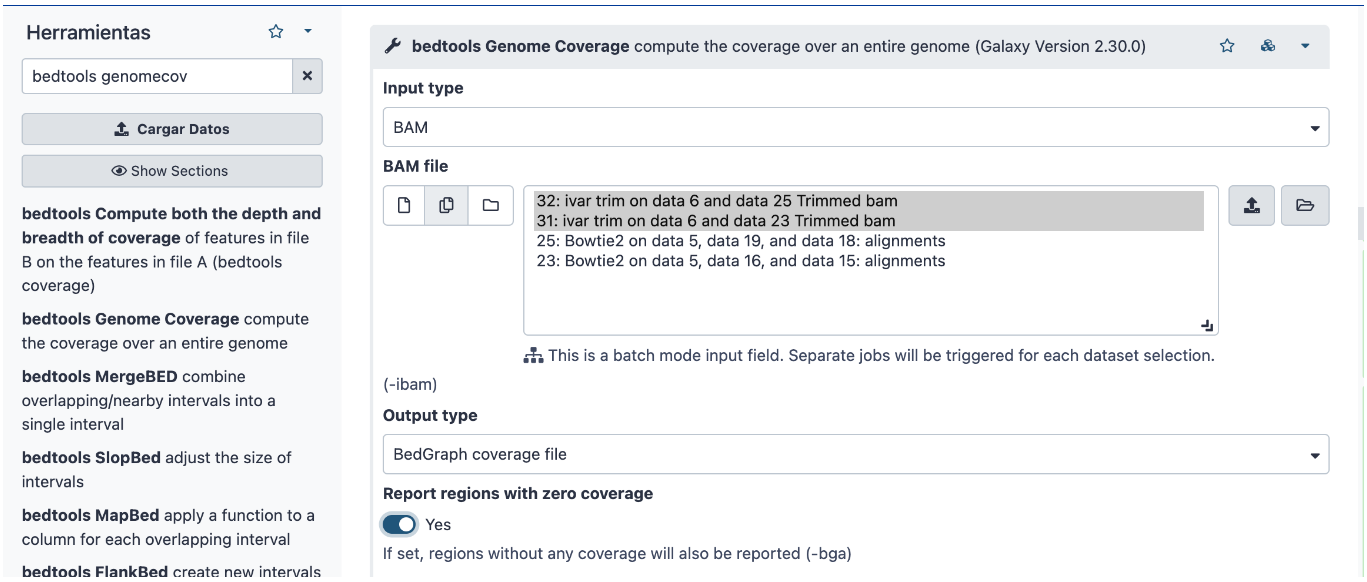
This will just generate a fasta file identical to the reference one, except for those nucleotides that are variants from the VCF file.

Genome coverage calculation

At this point, we have the consensus viral genome, but we know that we have filtered the variants based on the coverage, selecting only those that had a coverage depth higher than 10X. So we cannot ensure that the consensus genome doesn't have any variant that we have filter in those regions with a coverage lower than 10X. So the next step is to determine which regions of the reference genome have a coverage lower than 10X.

To do that you will search for "*bedtools genomecov*" in the search bar and select "*bedtools Genome Coverage compute the coverage over an entire genome*", the you will have to select the following files:

1. Input type > BAM
2. BAM file > iVar trim output bam files for both samples
3. Output type > BedGraph coverage file
4. Report regions with zero coverage > Yes
5. Execute



bedtools Compute both the depth and breadth of coverage of features in file B on the features in file A (bedtools coverage)

bedtools Genome Coverage compute the coverage over an entire genome

bedtools MergeBED combine overlapping/nearby intervals into a single interval

bedtools SlopBed adjust the size of intervals

bedtools MapBed apply a function to a column for each overlapping interval

bedtools FlankBed create new intervals

bedtools Genome Coverage compute the coverage over an entire genome (Galaxy Version 2.30.0)

Input type

BAM

BAM file

32: ivar trim on data 6 and data 25 Trimmed bam
31: ivar trim on data 6 and data 23 Trimmed bam
25: Bowtie2 on data 5, data 19, and data 18: alignments
23: Bowtie2 on data 5, data 16, and data 15: alignments

This is a batch mode input field. Separate jobs will be triggered for each dataset selection.

(-ibam)

Output type

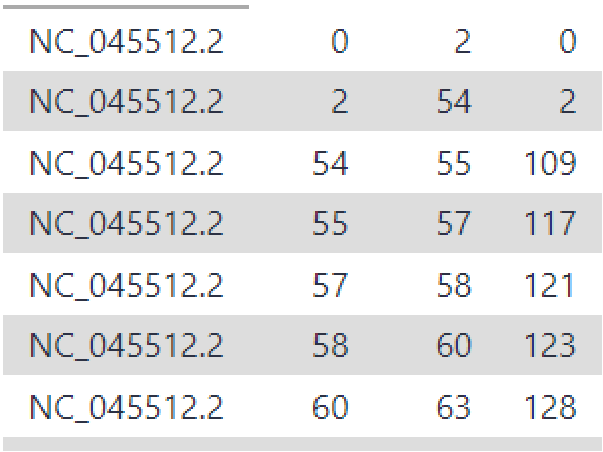
BedGraph coverage file

Report regions with zero coverage

Yes

If set, regions without any coverage will also be reported (-bga)

This process will generate a BED file where each genomic position range of the reference genome has the coverage calculated. In this example you can see that for the positions of the reference genome from the nucleotide 2 to 54 they have a coverage of 2X and then will be masked.



NC_045512.2	0	2	0
NC_045512.2	2	54	2
NC_045512.2	54	55	109
NC_045512.2	55	57	117
NC_045512.2	57	58	121
NC_045512.2	58	60	123
NC_045512.2	60	63	128

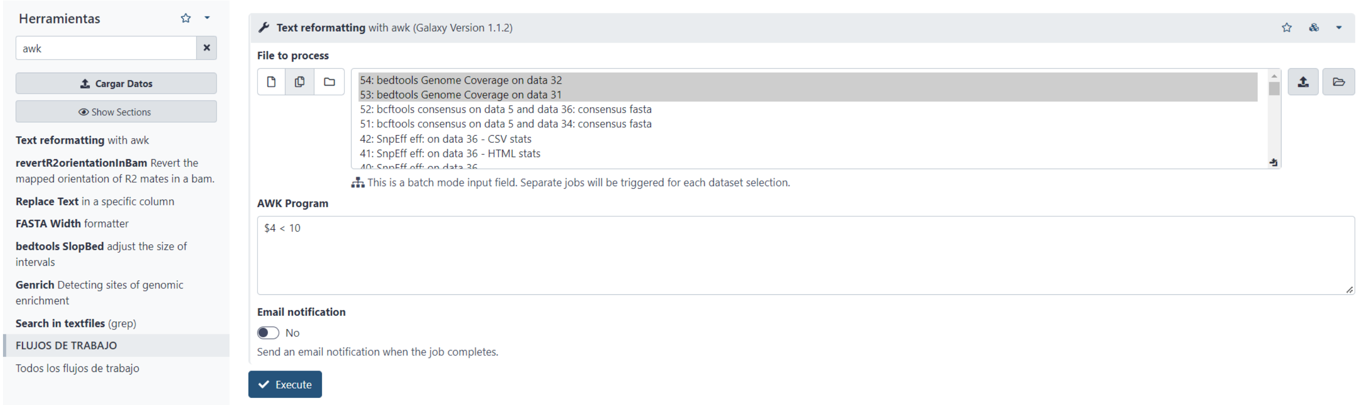
Regions filtering

From this resulting file from betdools genomecoverage you are going to select those regions with a coverage lower than 10X. Writing in the search bar "*awk*" and selecting "*Text reformatting with awk*", you are going to change:

3. File to process > Bedtools genome coverage file with the coverage regions for both samples
4. AWK Program = `$4 < 10`

- **This will filter all the lines (genomic regions) that have a value lower than 10 in the 4th column (coverage)**

5. Execute


Herramientas
awk
Cargar Datos
Show Sections
Text reformatting with awk
revertR2orientationInBam Revert the mapped orientation of R2 mates in a bam.
Replace Text in a specific column
FASTA Width formatter
bedtools SlopBed adjust the size of intervals
Genrich Detecting sites of genomic enrichment
Search in textfiles (grep)
FLUJOS DE TRABAJO
Todos los flujos de trabajo

Text reformatting with awk (Galaxy Version 1.1.2)

File to process

54: bedtools Genome Coverage on data 32
53: bedtools Genome Coverage on data 31
52: bcftools consensus on data 5 and data 36: consensus fasta
51: bcftools consensus on data 5 and data 34: consensus fasta
42: SnpEff eff: on data 36 - CSV stats
41: SnpEff eff: on data 36 - HTML stats

This is a batch mode input field. Separate jobs will be triggered for each dataset selection.

AWK Program

```
$4 < 10
```

Email notification

No

Send an email notification when the job completes.

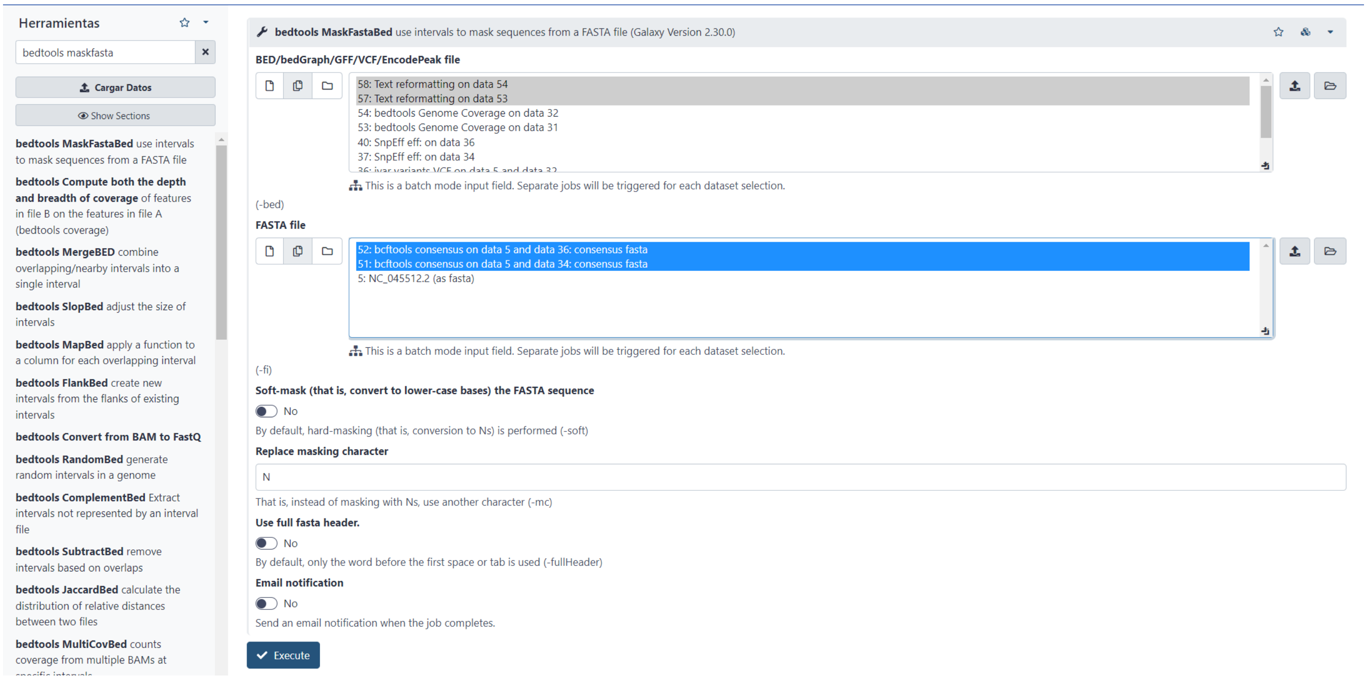
Execute

The resulting file is exactly the same as the one in Bedtools genomecoverage but only containing those lines with the genomic region coverage lower than 10X.

Masking the consensus genome

Now that you have the consensus genome and the regions with a sequencing depth lower than 10X, you are going to "mask" those regions in the consensus genome replacing the nucleotides in those regions with "N"s. You have to search for "*bedtools maskfasta*", select "*bedtools MaskFastaBed use intervals to mask sequences from a FASTA file*" and then select the following parameters:

3. BED/bedGraph/GFF/VCF/EncodePeak file > Select the BED files resulting from AWK text filter for both samples
4. FASTA file > Select the consensus genome fasta file generated with Bcftools consensus, for both samples
5. Execute


Herramientas
bedtools maskfasta
Cargar Datos
Show Sections
bedtools MaskFastaBed use intervals to mask sequences from a FASTA file
bedtools Compute both the depth and breadth of coverage of features in file B on the features in file A (bedtools coverage)
bedtools MergeBED combine overlapping/nearby intervals into a single interval
bedtools SlopBed adjust the size of intervals
bedtools MapBed apply a function to a column for each overlapping interval
bedtools FlankBed create new intervals from the flanks of existing intervals
bedtools Convert from BAM to FastQ
bedtools RandomBed generate random intervals in a genome
bedtools ComplementBed Extract intervals not represented by an interval file
bedtools SubtractBed remove intervals based on overlaps
bedtools JaccardBed calculate the distribution of relative distances between two files
bedtools MultiCovBed counts coverage from multiple BAMs at

bedtools MaskFastaBed use intervals to mask sequences from a FASTA file (Galaxy Version 2.30.0)

BED/bedGraph/GFF/VCF/EncodePeak file

58: Text reformatting on data 54
57: Text reformatting on data 53
54: bedtools Genome Coverage on data 32
53: bedtools Genome Coverage on data 31
40: SnpEff eff: on data 36
37: SnpEff eff: on data 34

This is a batch mode input field. Separate jobs will be triggered for each dataset selection.

(-bed)

FASTA file

52: bcftools consensus on data 5 and data 36: consensus fasta
51: bcftools consensus on data 5 and data 34: consensus fasta
5: NC_045512.2 (as fasta)

This is a batch mode input field. Separate jobs will be triggered for each dataset selection.

(-fi)

Soft-mask (that is, convert to lower-case bases) the FASTA sequence

No

By default, hard-masking (that is, conversion to Ns) is performed (-soft)

Replace masking character

N

That is, instead of masking with Ns, use another character (-mc)

Use full fasta header.

No

By default, only the word before the first space or tab is used (-fullHeader)

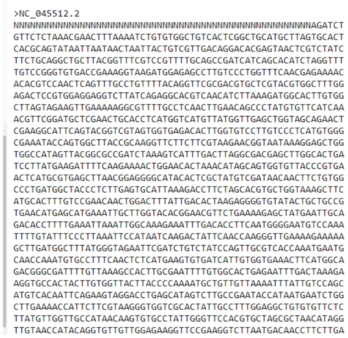
Email notification

No

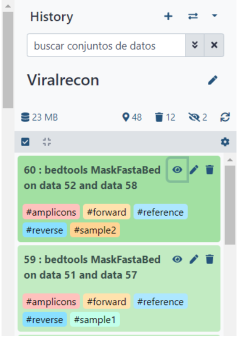
Send an email notification when the job completes.

Execute

The resulting file is the consensus genome generated previously but now only contains Ns instead of A, T, G or C in the regions with less than 10X depth of coverage



```
>NC_045512.2
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNAGATCT
GTTCTCTAAACGAACTTTAAAATCTGTGTGGCTGTCACTCGGCTGCATGCTTAGTGCACT
CACGCAGTATAATTAATAACTAATTACTGTCGTTGACAGGACACGAGTAACTCGTCTATC
TTCTGCAGGCTGCTTACGGTTTCGTCCGTTTTGCAGCCGATCATCAGCACATCTAGGTTT
TGTCCGGGTGTGACCGAAAGGTAAGATGGAGAGCCTTGTCCCTGGTTTCAACGAGAAAAC
ACACGTCCAACTCAGTTTGCCTGTTTTACAGGTTCGCGACGTGCTCGTACGTGGCTTTGG
AGACTCCGTGGAGGAGGTCTTATCAGAGGCACGTCAACATCTTAAAGATGGCACTTGTGG
CTTAGTAGAAGTTGAAAAAGGCGTTTTGCCTCAACTTGAACAGCCCTATGTGTTCATCAA
ACGTTCGGATGCTCGAACTGCACCTCATGGTCATGTTATGGTTGAGCTGGTAGCAGAACT
CGAAGGCATTCAGTACGGTCGTAGTGGTGAGACACTTGGTGTCCTTGTCCCTCATGTGGG
CGAAATACCAGTGGCTTACCGCAAGGTTCTTCTTCGTAAGAACGGTAATAAAGGAGCTGG
TGGCCATAGTTACGGCGCCGATCTAAAGTCATTTGACTTAGGCGACGAGCTTGGCACTGA
TCCTTATGAAGATTTTCAAGAAAACTGGAACACTAAACATAGCAGTGGTGTTACCCGTGA
ACTCATGCGTGAGCTTAACGGAGGGGCATACACTCGCTATGTCGATAACAACTTCTGTGG
CCCTGATGGCTACCCTCTTGAGTGCATTAAAGACCTTCTAGCACGTGCTGGTAAAGCTTC
ATGCACTTTGTCCGAACAACTGGACTTTATTGACACTAAGAGGGGTGTATACTGCTGCCG
TGAACATGAGCATGAAATTGCTTGGTACACGGAACGTTCTGAAAAGAGCTATGAATTGCA
GACACCTTTTGAAATTAAATTGGCAAAGAAATTTGACACCTTCAATGGGGAATGTCCAAA
TTTTGTATTTCCCTTAAATTCCATAATCAAGACTATTCAACCAAGGGTTGAAAAGAAAAA
GCTTGATGGCTTTATGGGTAGAATTCGATCTGTCTATCCAGTTGCGTCACCAAATGAATG
CAACCAAATGTGCCTTTCAACTCTCATGAAGTGTGATCATTGTGGTGAAACTTCATGGCA
GACGGGCGATTTTGTTAAAGCCACTTGCGAATTTTGTGGCACTGAGAATTTGACTAAAGA
AGGTGCCACTACTTGTGGTTACTTACCCCAAAATGCTGTTGTTAAAATTTATTGTCCAGC
ATGTCACAATTCAGAAGTAGGACCTGAGCATAGTCTTGCCGAATACCATAATGAATCTGG
CTTGAAAACCATTCTTCGTAAGGGTGGTCGCACTATTGCCTTTGGAGGCTGTGTGTTCTC
TTATGTTGGTTGCCATAACAAGTGTGCCTATTGGGTTCCACGTGCTAGCGCTAACATAGG
TTGTAACCATACAGGTGTTGTTGGAGAAGGTTCCGAAGGTCTTAATGACAACCTTCTTGA
```



History

buscar conjuntos de datos

Viralrecon

23 MB 48 12 2

60 : bedtools MaskFastaBed on data 52 and data 58

#amplicons #forward #reference #reverse #sample2

59 : bedtools MaskFastaBed on data 51 and data 57

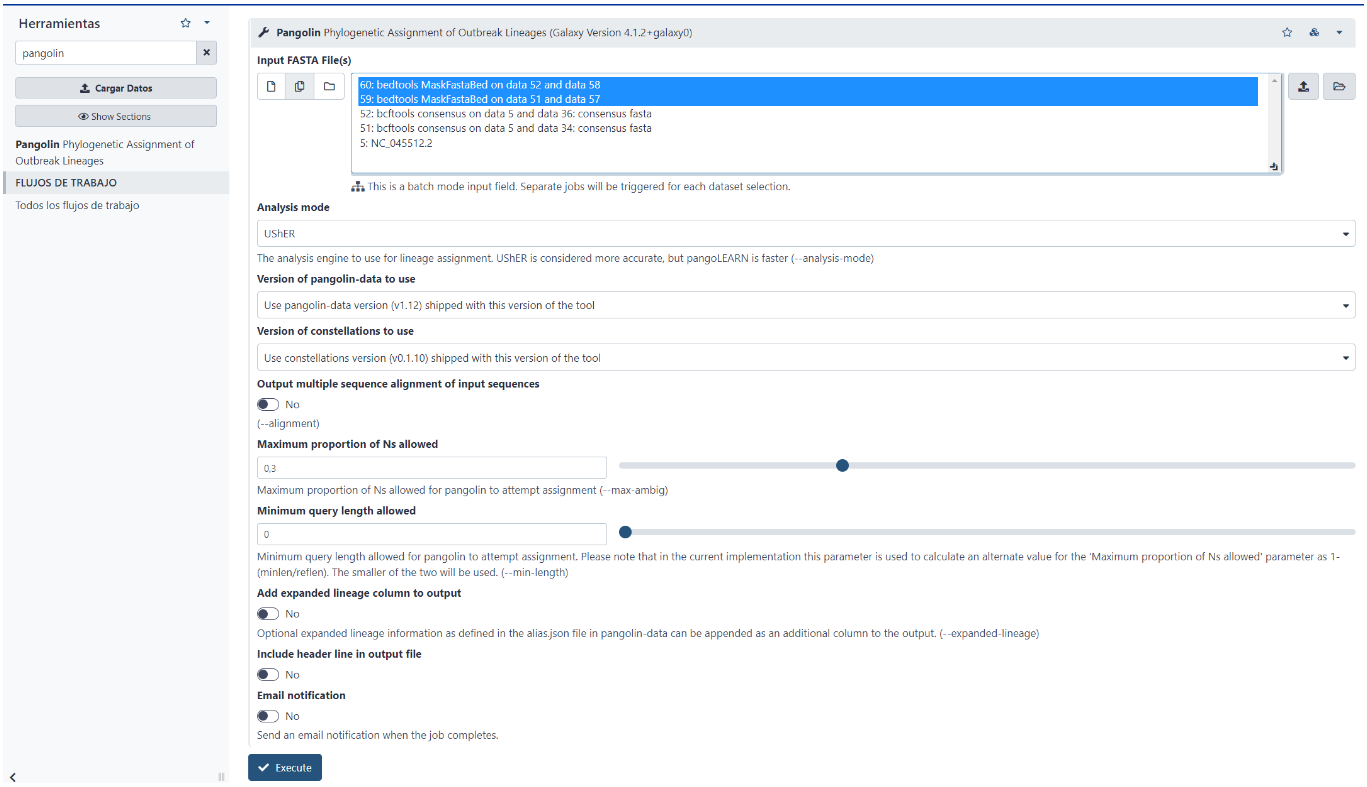
#amplicons #forward #reference #reverse #sample1

You can download this fasta file and use it to upload it to any public repository such as ENA or GiSaid. Also you can use it to perform phylogenetic trees or whatever else you want to do with the SARS-CoV-2 consensus fasta file.

Lineage

Now we are going to determine the lineage of the samples. We will use a software called pangolin. We are going to use the masked consensus genomes generated in the previous steps as follows:

1. Search for the **pangolin** tool
2. Select **Pangolin Phylogenetic Assignment of Outbreak Lineages** and set the following parameters:
3. Select the *bedtools MaskFastaBed* generated in the previous step as input fasta file for both samples
4. **Execute**



Herramientas
pangolin
Cargar Datos
Show Sections
Pangolin Phylogenetic Assignment of Outbreak Lineages
FLUJOS DE TRABAJO
Todos los flujos de trabajo

Pangolin Phylogenetic Assignment of Outbreak Lineages (Galaxy Version 4.1.2+galaxy0)

Input FASTA File(s)

60: bedtools MaskFastaBed on data 52 and data 58
59: bedtools MaskFastaBed on data 51 and data 57
52: bcftools consensus on data 5 and data 36: consensus fasta
51: bcftools consensus on data 5 and data 34: consensus fasta
5: NC_045512.2

This is a batch mode input field. Separate jobs will be triggered for each dataset selection.

Analysis mode

UShER

The analysis engine to use for lineage assignment. UShER is considered more accurate, but pangoLEARN is faster (--analysis-mode)

Version of pangolin-data to use

Use pangolin-data version (v1.12) shipped with this version of the tool

Version of constellations to use

Use constellations version (v0.1.10) shipped with this version of the tool

Output multiple sequence alignment of input sequences

No

(--alignment)

Maximum proportion of Ns allowed

0,3

Maximum proportion of Ns allowed for pangolin to attempt assignment (--max-ambig)

Minimum query length allowed

0

Minimum query length allowed for pangolin to attempt assignment. Please note that in the current implementation this parameter is used to calculate an alternate value for the 'Maximum proportion of Ns allowed' parameter as 1-(minlen/reflen). The smaller of the two will be used. (--min-length)

Add expanded lineage column to output

No

Optional expanded lineage information as defined in the alias.json file in pangolin-data can be appended as an additional column to the output. (--expanded-lineage)

Include header line in output file

No

Email notification

No

Send an email notification when the job completes.

Now we are going to have a look to the results from pangolin. As you can see, results are in table format, where you have in first place the reference genome and then de lineage assingned.

Question

▶ Which is the lineage of sample1?

▶ And the lineage of sample2?

All results

If you have any problem following this training and you want to visualize the resulting file you can access them through this URL:

https://usegalaxy.eu/u/s.varona/h/viralrecon

And viralrecon workflfow in:

https://usegalaxy.eu/u/s.varona/w/viralrecon2022