Galaxy for virologist training Exercise 6: Illumina Variant Calling 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 variant calling? • What is a vcf file? • How can I inspect a variant in a bam file to look for false positives? • How can I make a consensus genome based on a variant calling process?
Objectives:	• Understand variant calling concept • Learn how to interpret a vcf file • Learn how to make a reference consensus genome. • Learn how to visualize mapping and variant calling results
Estimated time:	2h

1. Description

After mapping, when we have a re-sequencing experiment, the next step usually comprises the variants calling step. Variant calling software tries to identify variants, positions that differ in our reads compared to a reference genome. We may want to have a consensus genome as well, which is obtained by including the variants we just identified in the published reference genome. We are going to address this type of analysis in this tutorial.

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 `variant calling 101 tutorial` as explained here

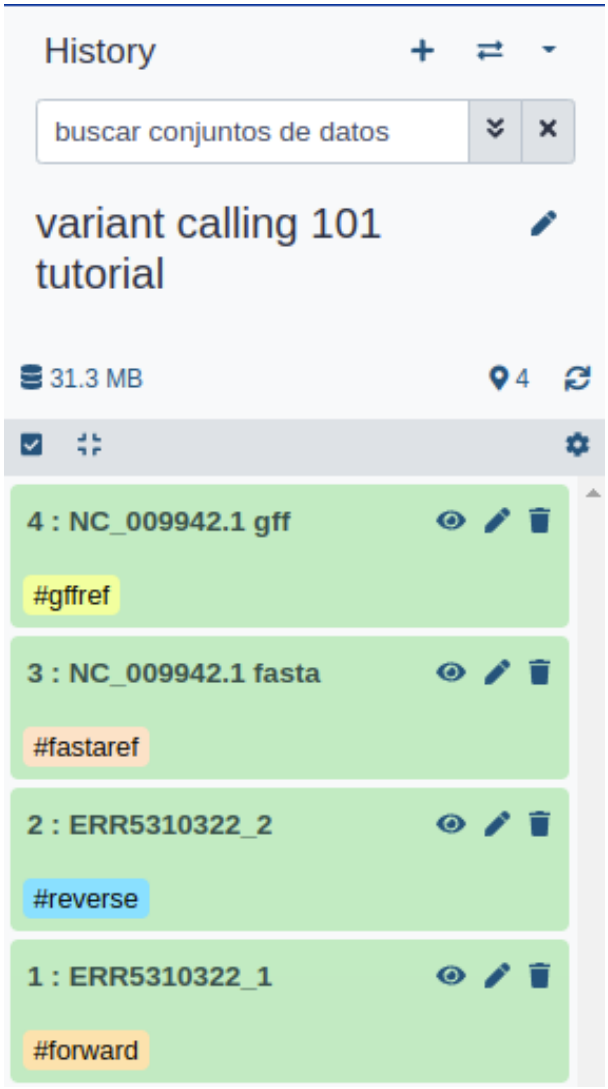
Upload data

Follow the same instructions here

```
ftp://ftp.sra.ebi.ac.uk/vol1/fastq/ERR531/002/ERR5310322/ERR5310322_1.fast
q.gz
ftp://ftp.sra.ebi.ac.uk/vol1/fastq/ERR531/002/ERR5310322/ERR5310322_2.fast
q.gz
https://ftp.ncbi.nlm.nih.gov/genomes/all/GCF/000/875/385/GCF_000875385.1_V
iralProj30293/GCF_000875385.1_ViralProj30293_genomic.fna.gz
https://ftp.ncbi.nlm.nih.gov/genomes/all/GCF/000/875/385/GCF_000875385.1_V
iralProj30293/GCF_000875385.1_ViralProj30293_genomic.gff.gz
```

Rename the data as follows:

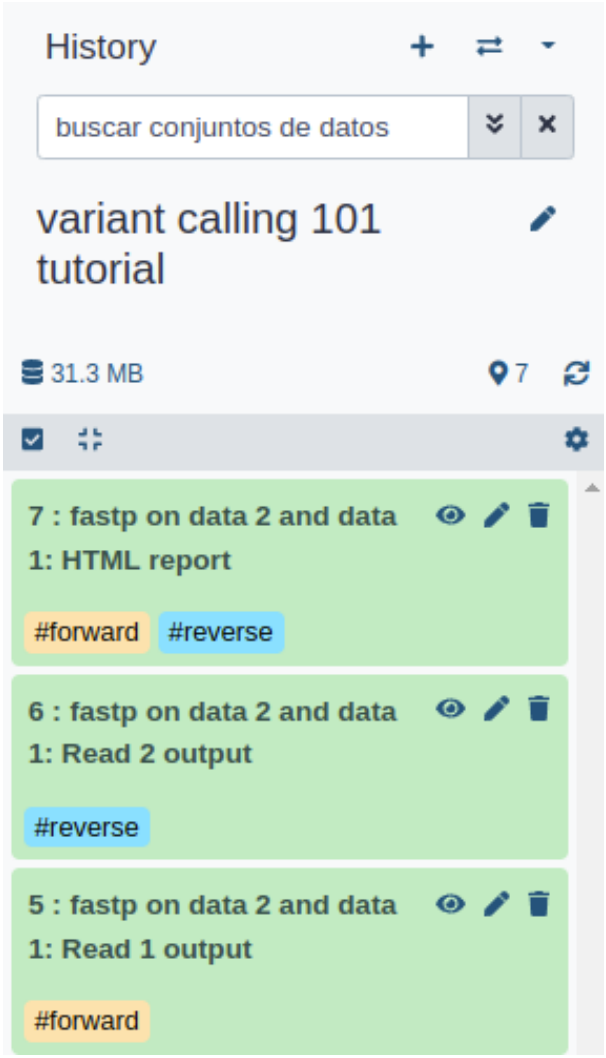
- `ERR5310322_1.fastq.gz` to `ERR5310322_1` with tag `#forward`
- `ERR5310322_2.fastq.gz` to `ERR5310322_2` with tag `#reverse`
- `GCF_000875385.1_ViralProj30293_genomic.fna.gz` to `NC_009942.1 fasta` with tag `#fastaref`
- `GCF_000875385.1_ViralProj30293_genomic.gff.gz` to `NC_009942.1 gff` with tag `#gffref`



3. Preprocess our reads.

Follow instructions here

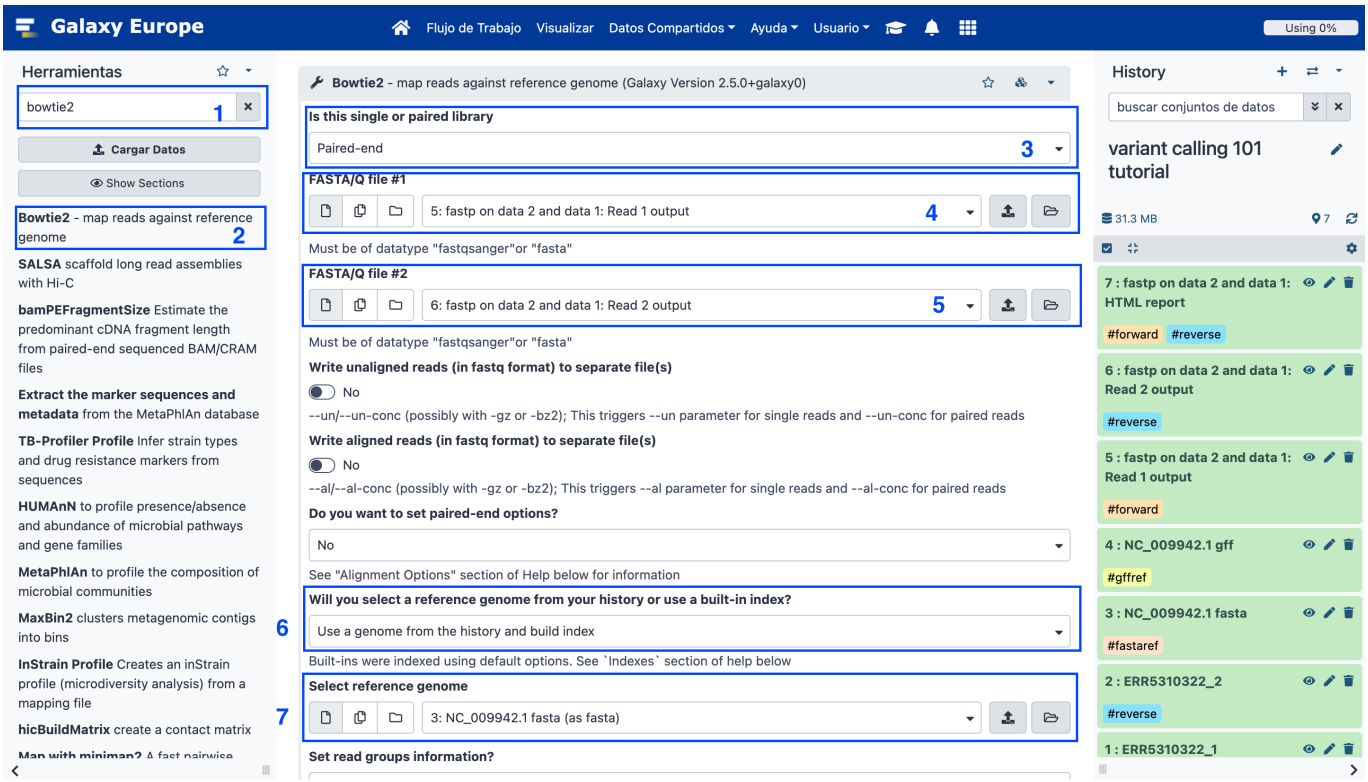
Then, fix fastp tags on the output data to be as follows:

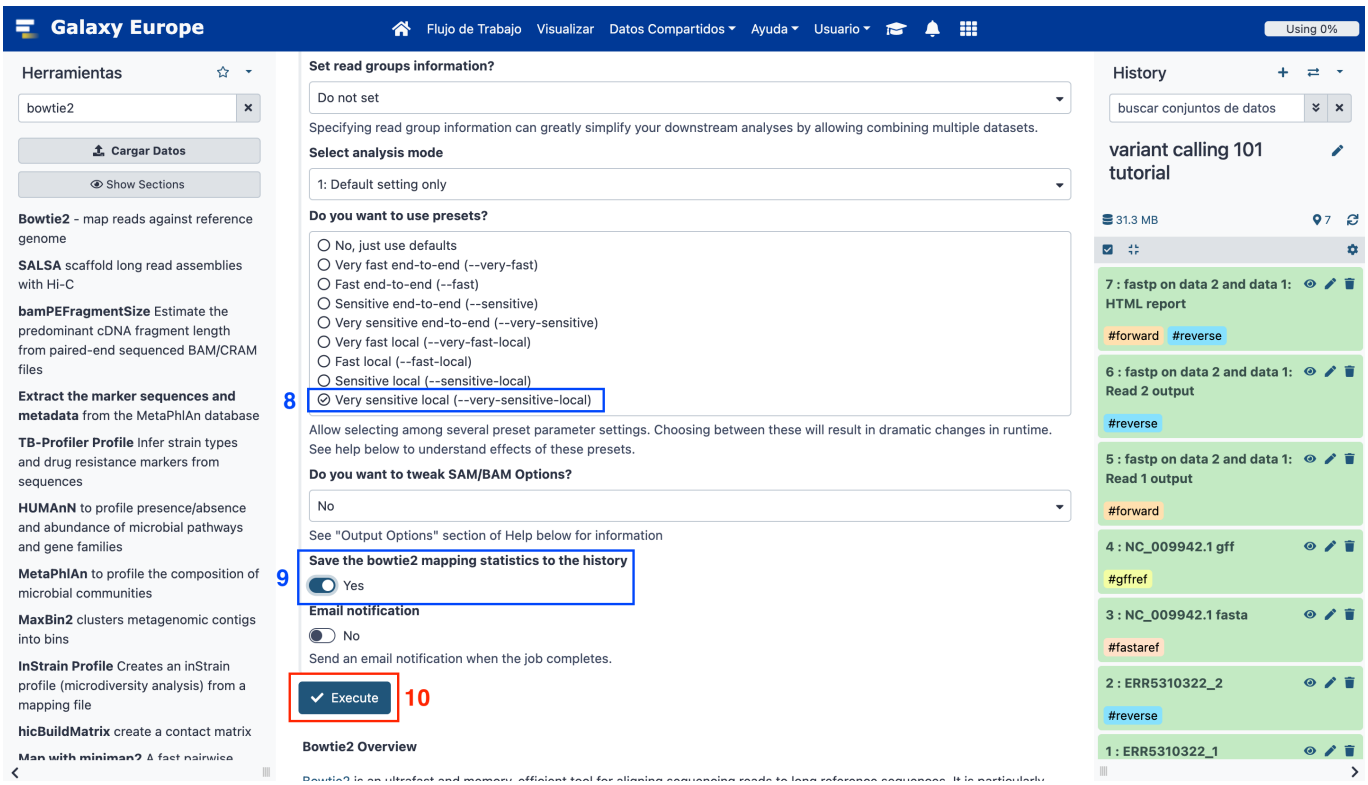


4. Map trimmed reads against the reference genome.

Follow:

1. Is this single or paired library: paired.
2. FASTA/Q file #1 : fastp Read 1 output #forward
3. FASTA/Q file #2 : fastp Read 2 output #reverse
4. Will you select a reference genome from your history or use a built-in index? : Use a genome from the history and build index.
5. Do you want to use presets? : Very sensitive local. This setting will hugly affect the mapping results, depending on the dataset/experiment must be tweaked (read bowtie2 manual)
6. Save the bowtie2 mapping statistics to the history



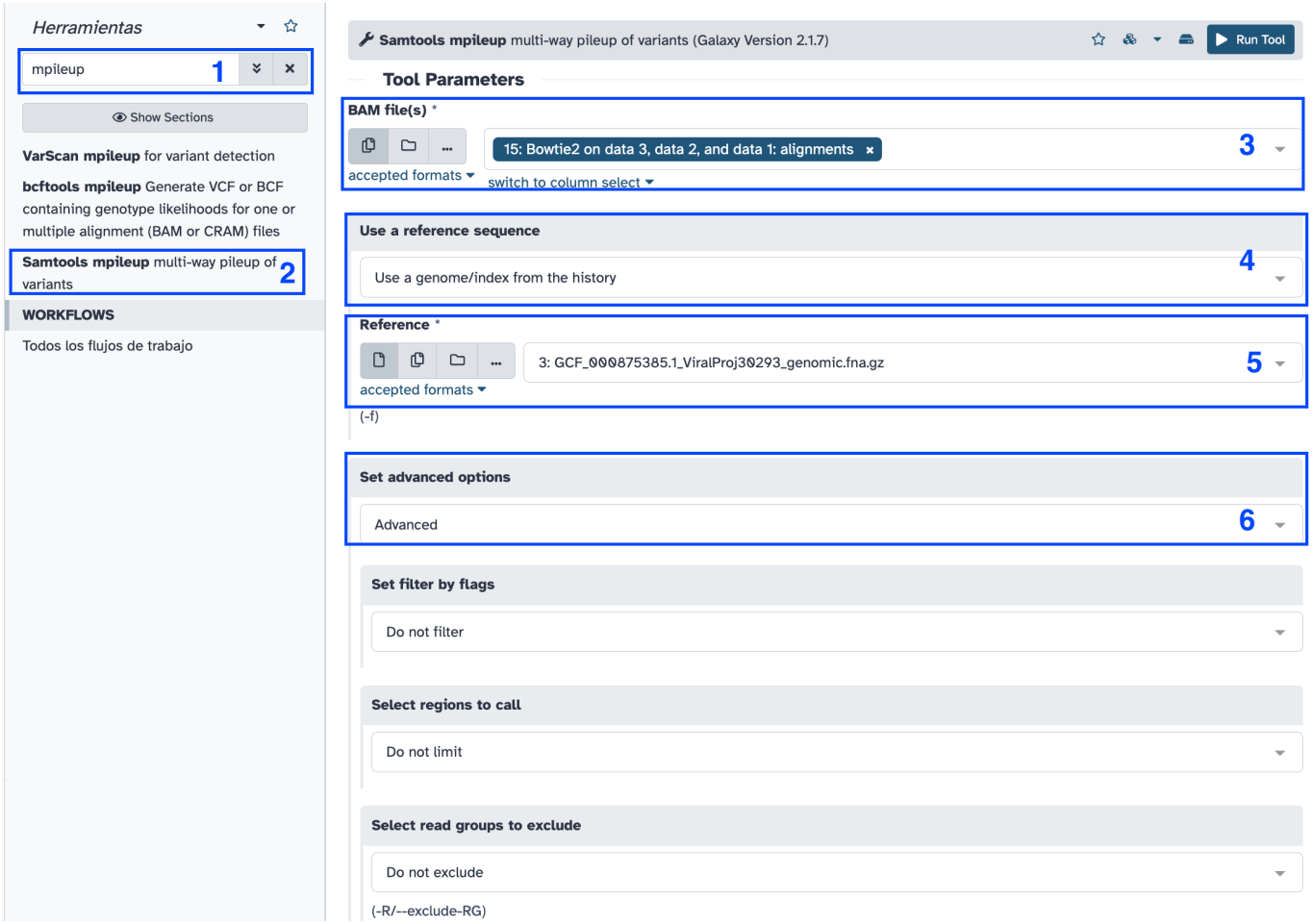


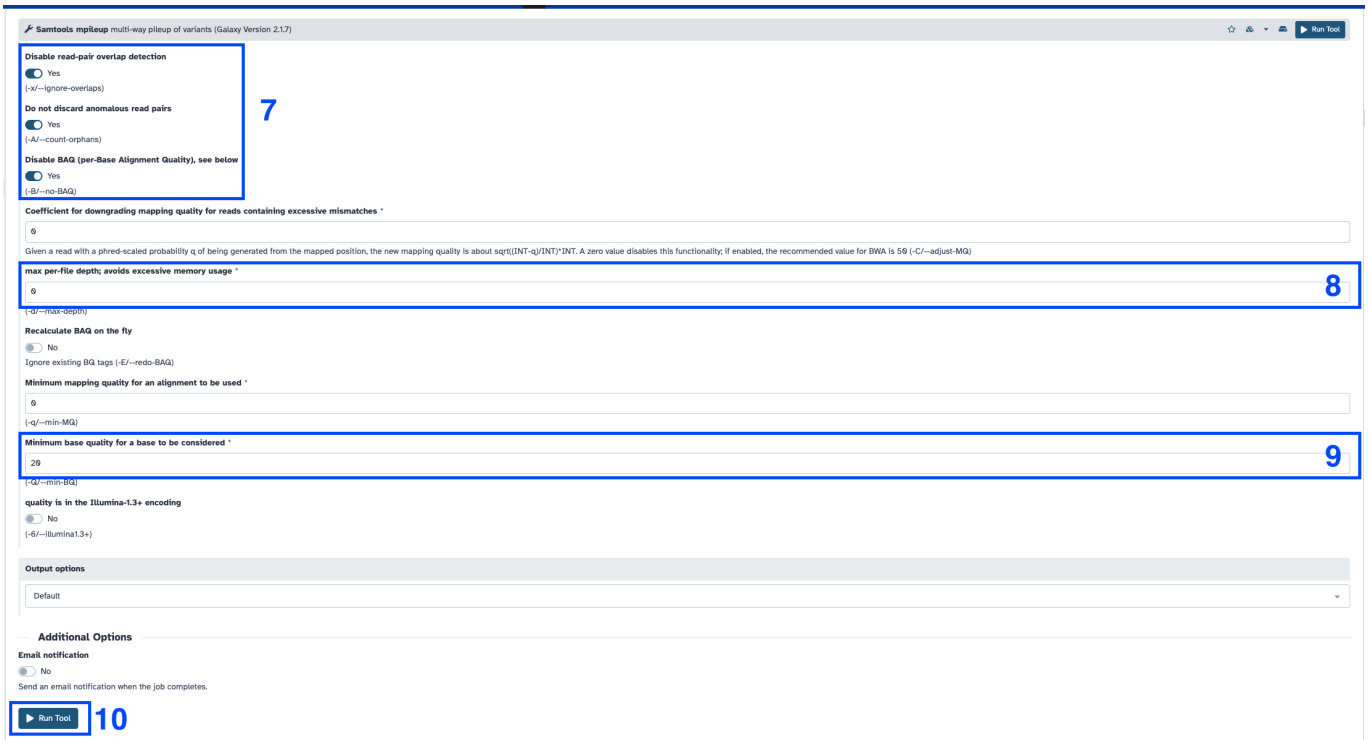
5. Variant Calling

Samtools mpileup

1. Search samtools mpileup in the search toolbox, scroll down and select `Samtools mpileup multi-way pileup of variants`
2. Bam files: Bowtie2 bam file
3. Use reference: Use reference/genome from history. NC_009942.1.
4. Set advanced options: Advanced
5. Disable read-pair overlap detection: Yes

6. Disable BAQ (per-Base Alignment Quality), see below: Yes
7. Do not discard anomalous read pairs: Yes
8. max per-file depth; avoids excessive memory usage: 0
9. Minimum base quality for a base to be considered: 20
10. Click execute and wait.





11. Click the :eye: icon on the history and inspect the mpileup output.

VarScan

1. Search `VarScan Mpileup` in the search toolbox.
2. Samtools pileup dataset: samtools mpileup output
3. Minimum read depth: 10
4. Minimum supporting reads: 5
5. Minimum base quality at a position to count a read: 20
6. Minimum variant allele frequency threshold: 0,75
7. Default p-value threshold for calling variants: 0,05
8. Click execute and wait

1. Click the :eye: icon and inspect the vcf file.

VCF stats

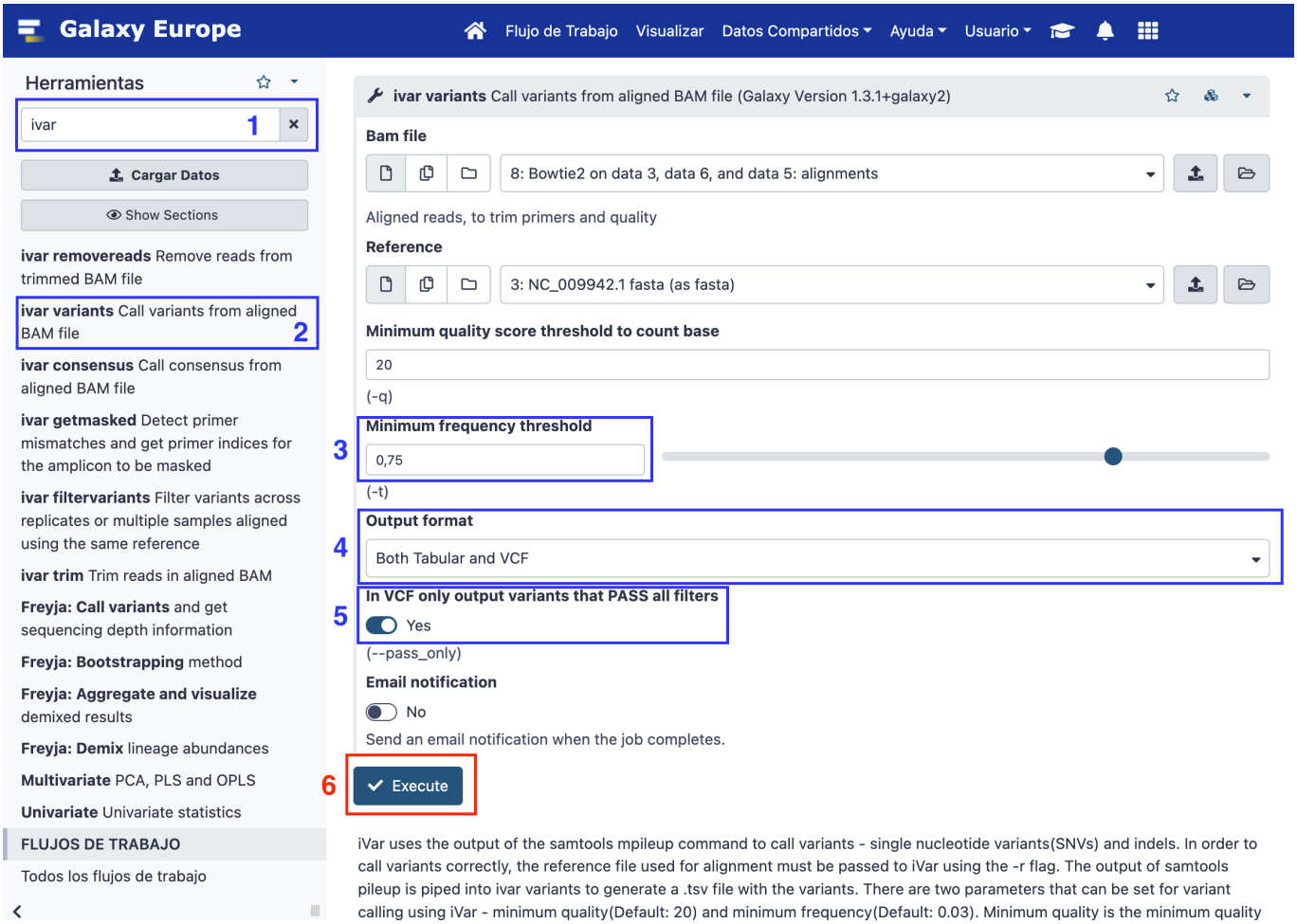
1. Search `bcftools stats` in the search toolbox.
2. VCF/BCF Data: varscan vcf output.
3. Click execute and wait.
4. Click the :eye: icon and inspect the stats.

▶ How may variants do we have in our vcf file?

Ivar variants

1. Search `ivar variants` in the search toolbox.
2. Select `ivar variants Call variants from aligned BAM file`
3. Bam file: bowtie bam output
4. Reference: NC_009942.1
5. Minimum quality score threshold to count base: 20
6. Minimum frequency threshold: 0.75
7. Output format: Both tabular and vcf

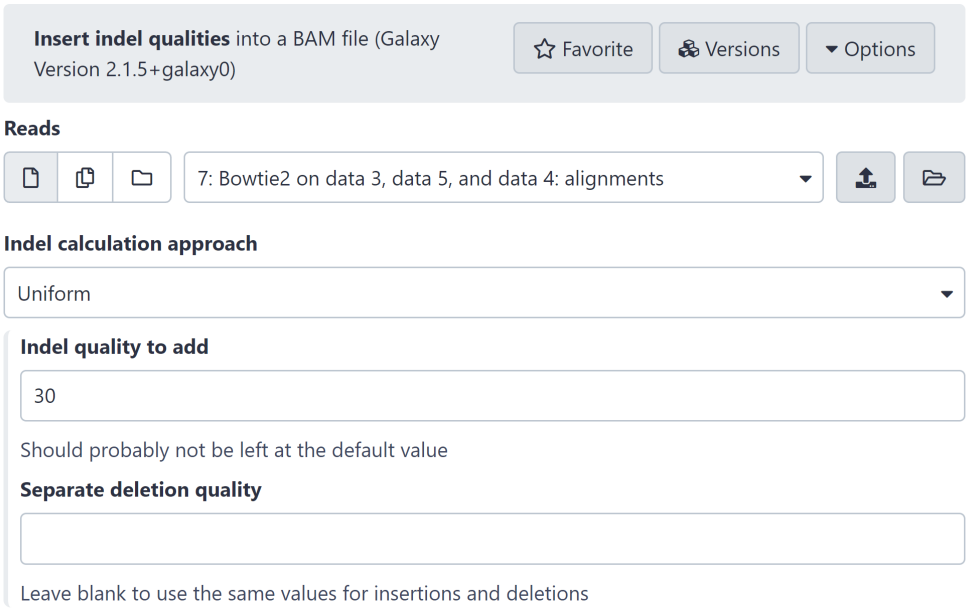
8. In VCF only output variants that PASS all filters > Yes
9. Click execute and wait.



Lofreq

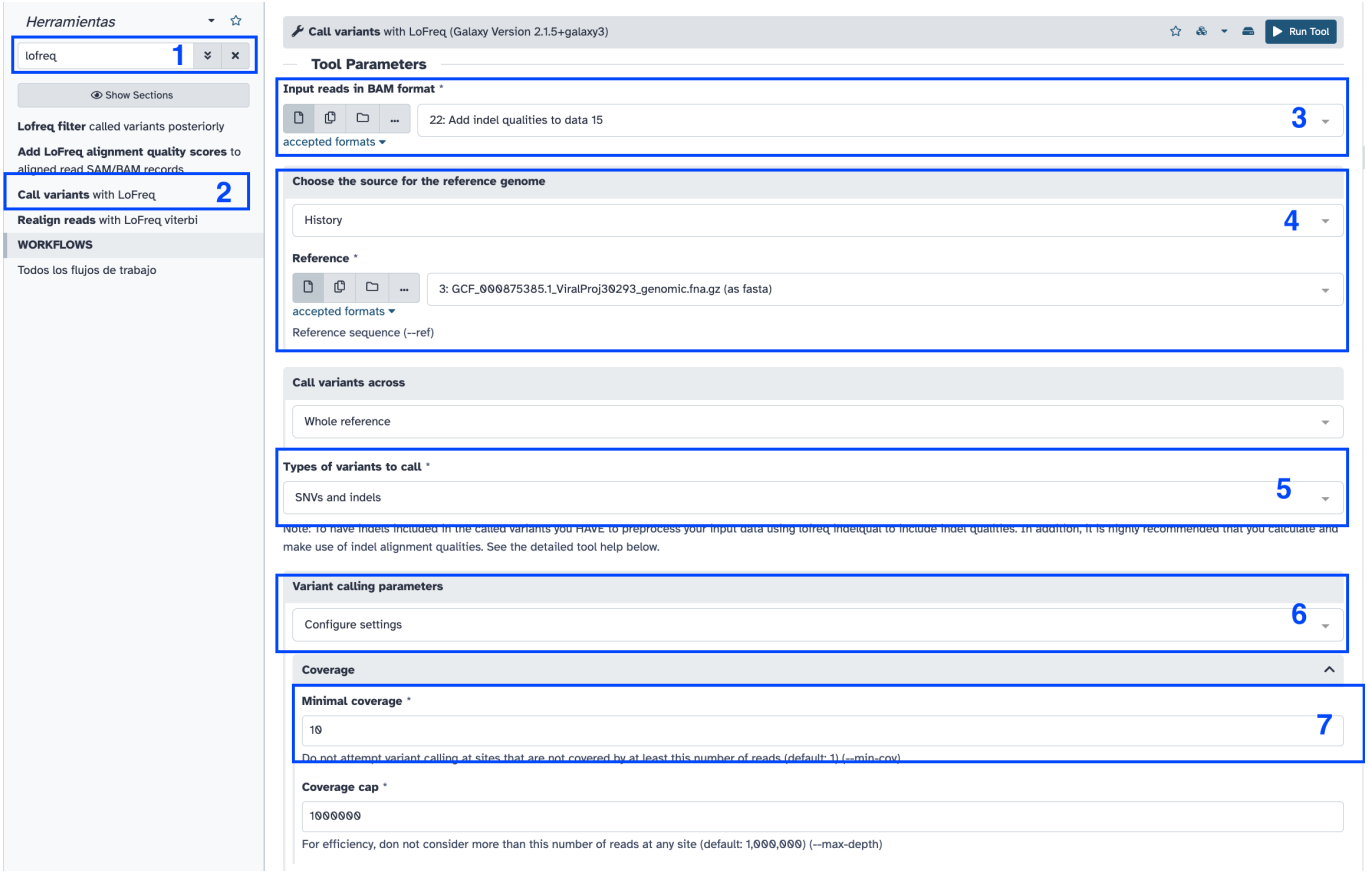
Insert indel qualities

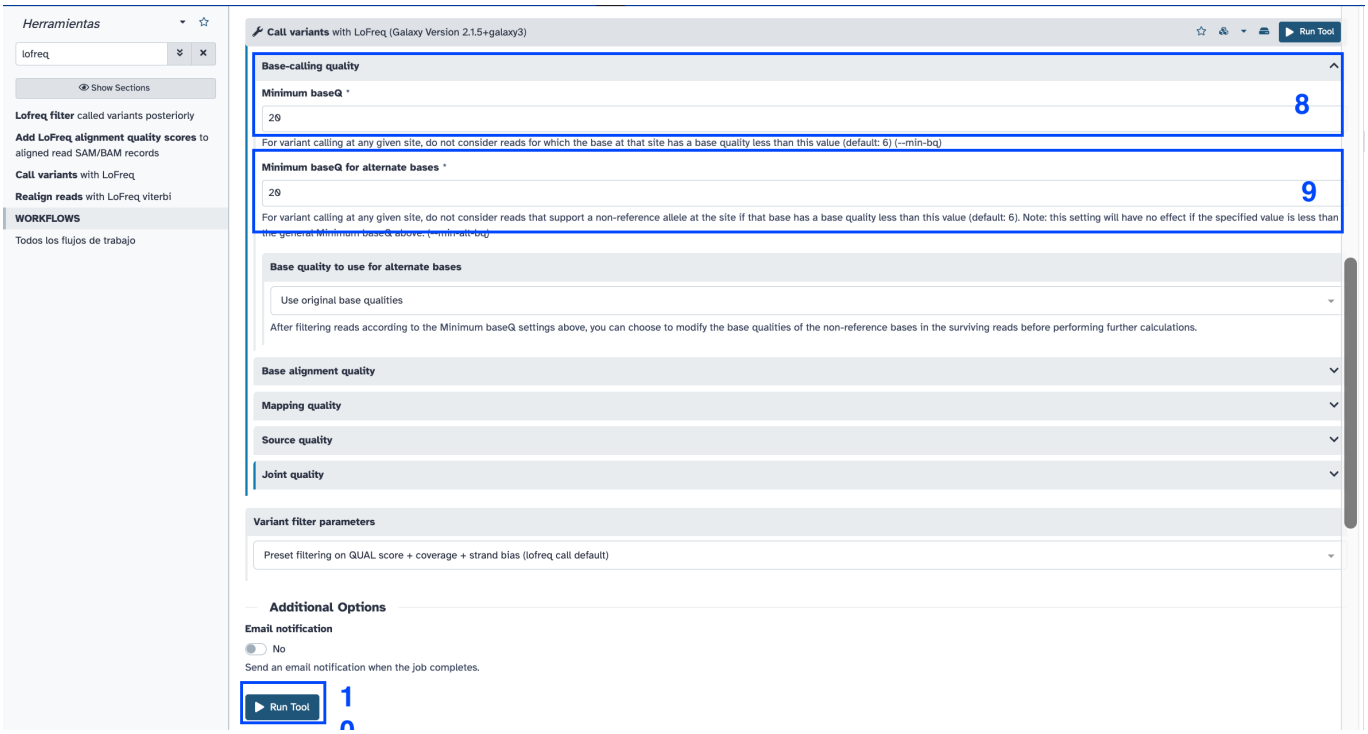
1. Search `Insert indel qualities` in the search toolbox. Select `Insert indel qualities into a BAM file`
2. Reads: bowtie2 bam output.
3. Click execute and wait.



Call variants

1. Search `lofreq` in the search toolbox. Select Call variants with lofreq.
2. Input reads in BAM format: indel qualities bam output.
3. Choose the source for the reference genome: History. NC_009942.1
4. Types of variants to call: SNVs and INDELs
5. Variant calling parameters: Configure settings
6. Minimal coverage: 10
7. Minimum baseQ: 20
8. Minimum baseQ for alternate bases: 20
9. Click execute and wait.





Compare vcfs among callers

Visualize datasets

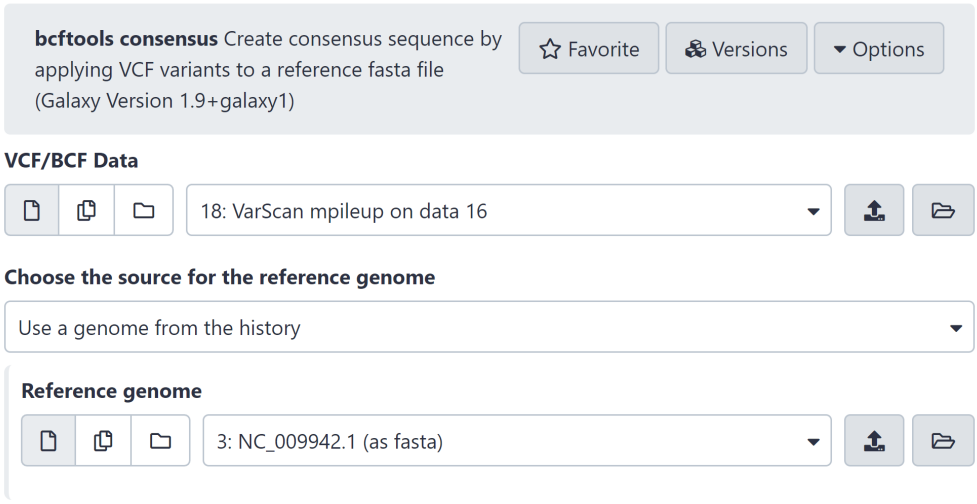
1. Search `upSet diagram` in the search toolbox.
2. Select input files for which to produce intersections: select vcf from varscan, vcf from lofreq filter and vcf from ivar variants.
3. Click execute and wait.
4. Click the :eye: icon and check the diagram.

▶ How many variants differ among the vcfs?

7. Consensus genome

Bcftools consensus

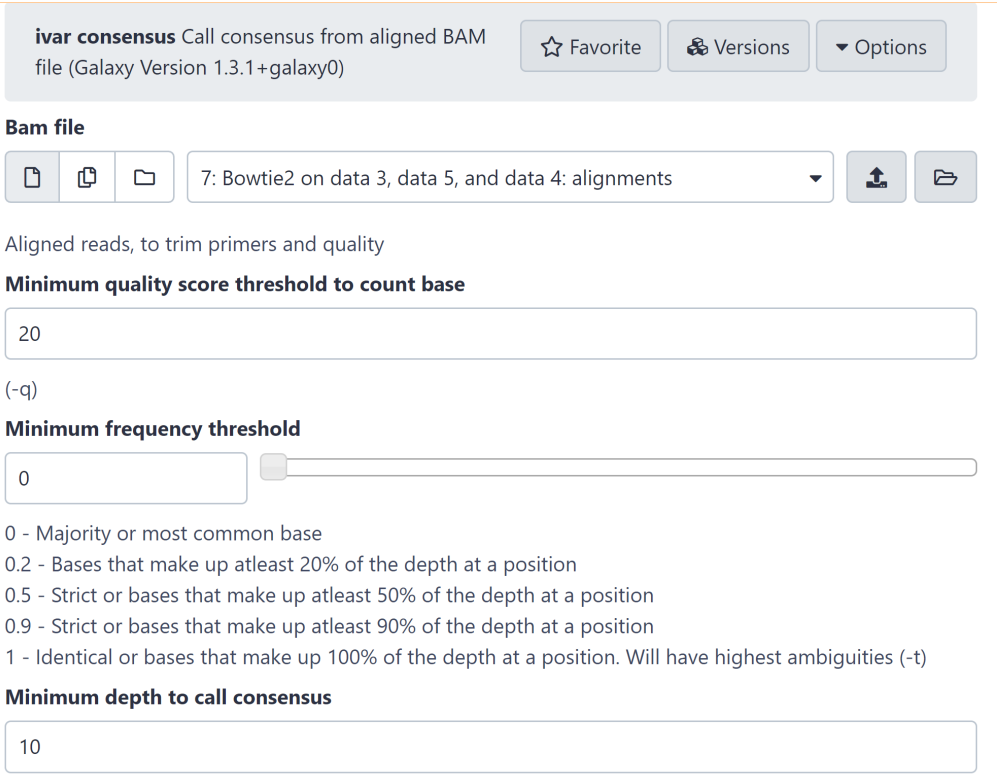
1. Search `bcftools consensus` in the search toolbox.
2. VCF/BCF Data: varscan vcf output.
3. Choose a reference genome: use genome/reference from history. Select NC_009942.1.
4. Click execute and wait.



> Note: for this example we are not going to mask any position with low coverage, this will be addressed in the exercise 8, with a real example.

Ivar Consensus

1. Search `ivar consensus` in the search toolbox.
2. Bam file: bowtie bam output.
3. Use N instead of - for regions with less than minimum coverage: Yes



> Here is the galaxy history for this exercise: https://usegalaxy.eu/u/smonzon/h/variant-calling-101-tutorial-1