Polymorphism analysis (SNP) from whole genome resequencing

This tutorial will describe how the workflow Gandalf analyse raw data from sequencing to do SNP calling and convert VCF files to do genetic mapping analysis.

The workflow Gandalf allows to do : URGI Gandalf tutorial

1) Trimming....1
2) Mapping....3
3) Filtering....4
a) SAMfilters....4
b) Markduplicate....5
4) Statistics on the reads....5
5) SNP calling....6
VCFFiltering....9
VCFStorage....11
VCFCarto....12

If you have any trouble to import your raw data into galaxy, or for your reads quality control, please read the NGS: reads quality control document.

Two workflows are share for the use of gandalf : « workflow gandalf 1 set » for one strain, and « workflow gandalf 2 sets » on 2 datasets applied on genetic mapping. First you need to import them as described on Galaxy Tutorial. You can try them with the files available on download repository. (note the data are already trimmed, so you can go directly to step 2) :

1) Trimming

Trimmomatic is a tool dedicated to reads cleaning in Single or Paired End mode. It performs adaptor removal, trimming by sliding window or cut off, minimum length, …

To perform trimming, please follows the trimming section on document NGS: reads quality control.

Trimmomatic flexible read trimming tool for Illumina NGS data (Galaxy Tool Version 0.36.0)

Paired end data?

Yes | No

Input Type

Pair of datasets

Input FASTQ file (R1/first of pair)

2: 2-Strain1_gal1.fq

Input FASTQ file (R2/second of pair)

3: 3-Strain1_gal2.fq

Perform initial ILLUMINACLIP step?

Yes | No

Cut adapter and other illumina-specific sequences from the read

Adapter sequences to use

TruSeq2 (single-ended, for Illumina GAII)

Maximum mismatch count which will still allow a full match to be performed

2

How accurate the match between the two 'adapter ligated' reads must be for PE palindrome read alignment

30

How accurate the match between any adapter etc. sequence must be against a read

12

Trimmomatic Operation

1: Trimmomatic Operation

Select Trimmomatic operation to perform

Sliding window trimming (SLIDINGWINDOW)

Number of bases to average across

4

Average quality required

20

2: Trimmomatic Operation

Select Trimmomatic operation to perform

Drop reads below a specified length (MINLEN)

Minimum length of reads to be kept

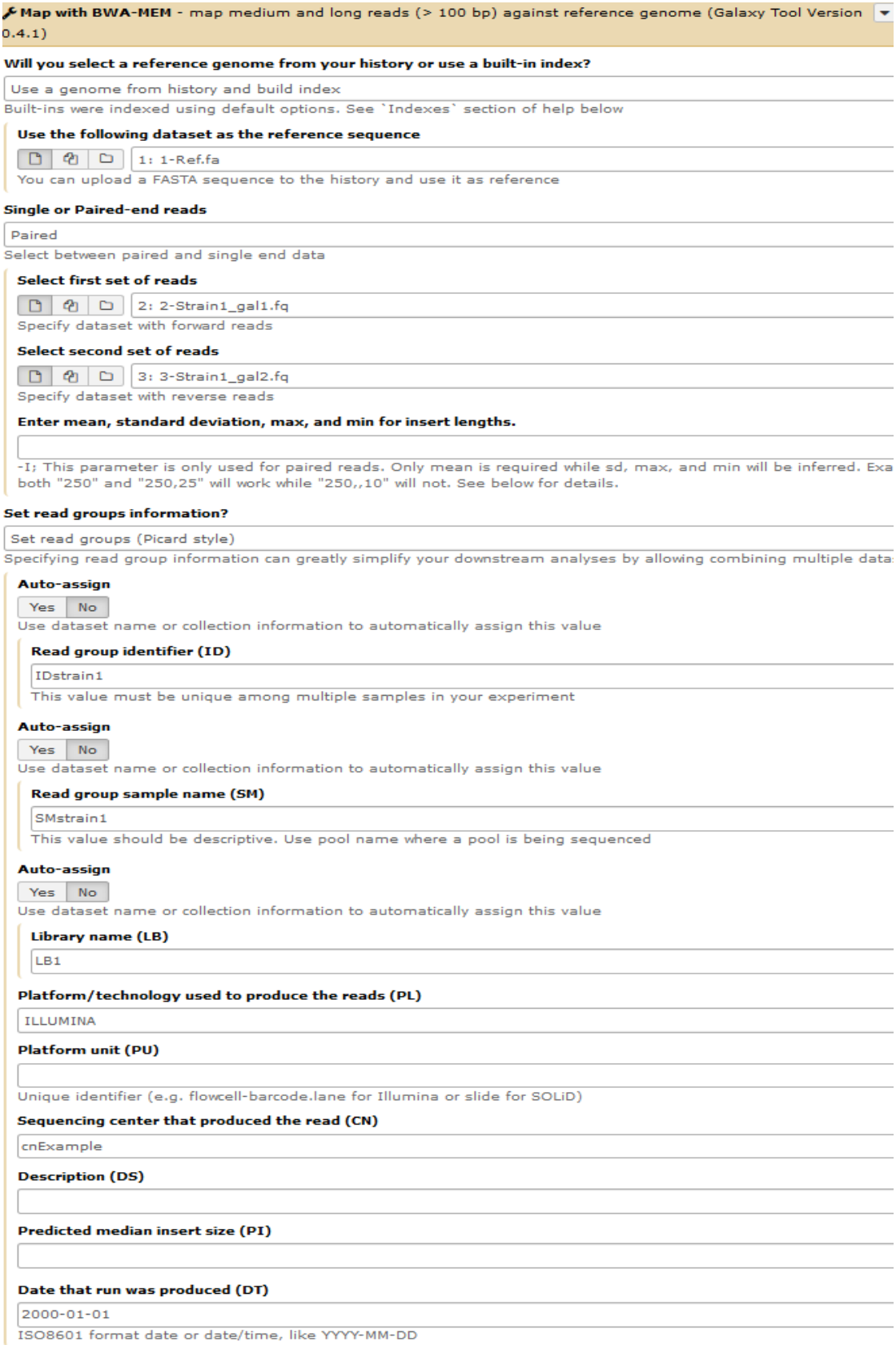
20

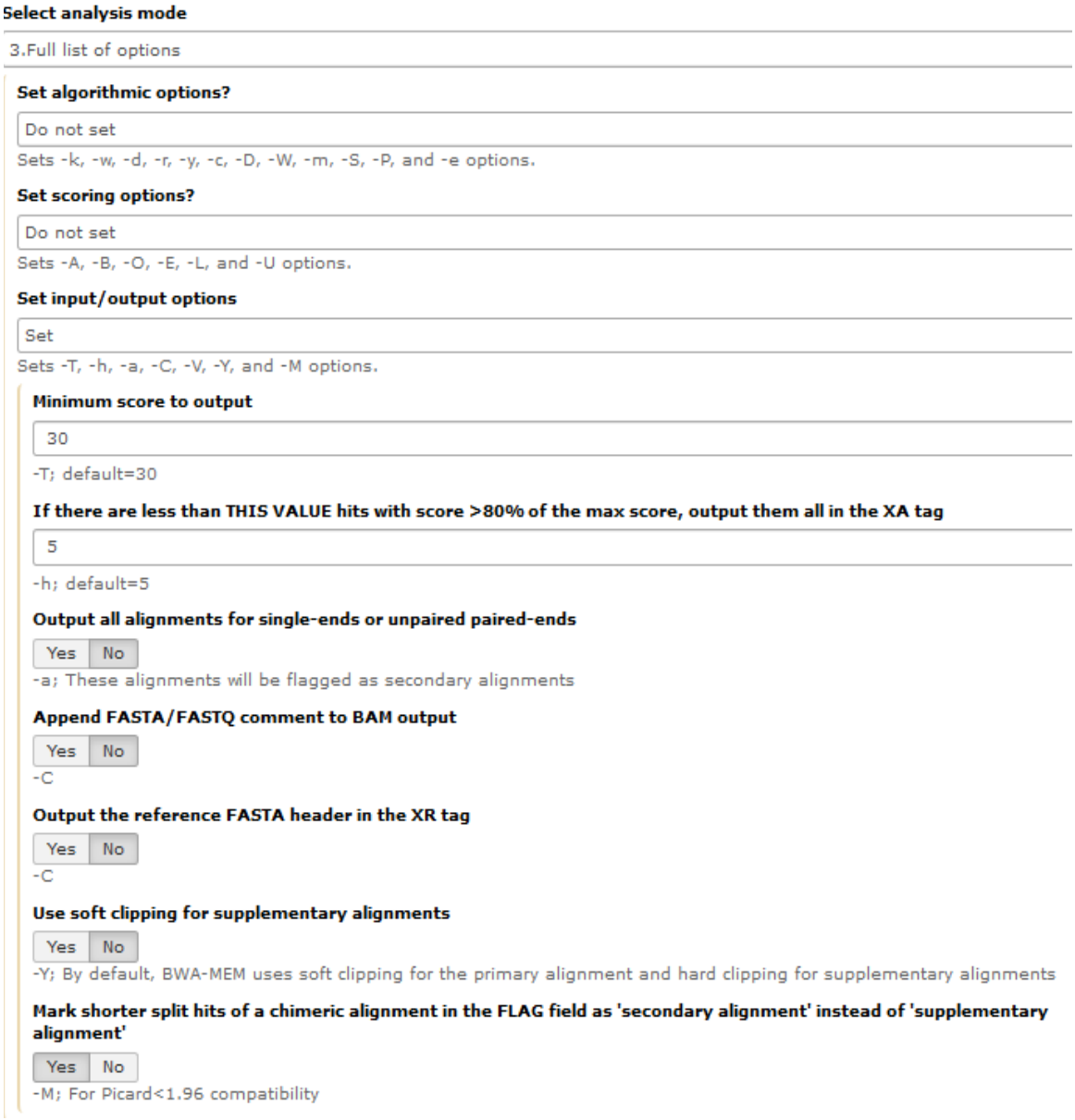
+ Insert Trimmomatic Operation

2) Mapping

BWA is a software package for mapping low-divergent sequences against a large reference genome.

Select Map with BWA-MEM in (**Gandalf**). A lot of options are possible. In the example below, optional read groups information are set. On the full list of option, check "Yes" to flag correctly secondary alignments for the picard tools compatibility (option M).



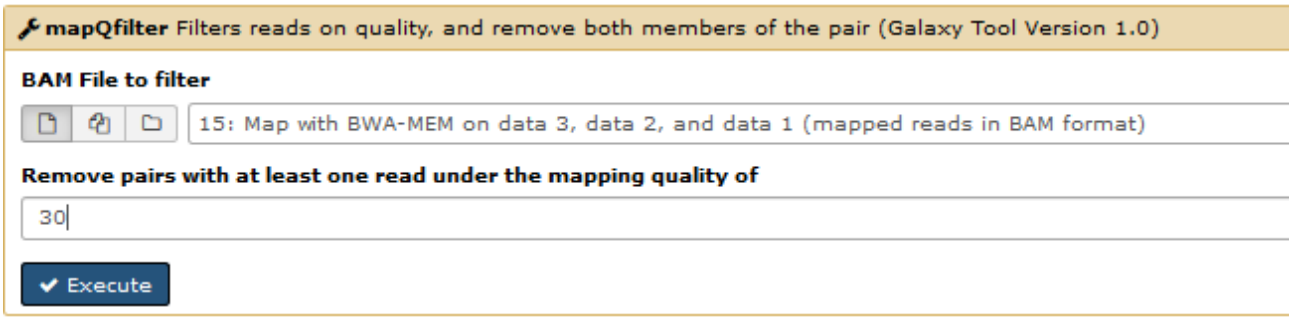


3) Filtering

a) SAMfilters

MapQfilter is a tool using the SAMtools suite. It allows to filter mapped reads depending on the quality of the mappingn and only keep properly paired reads

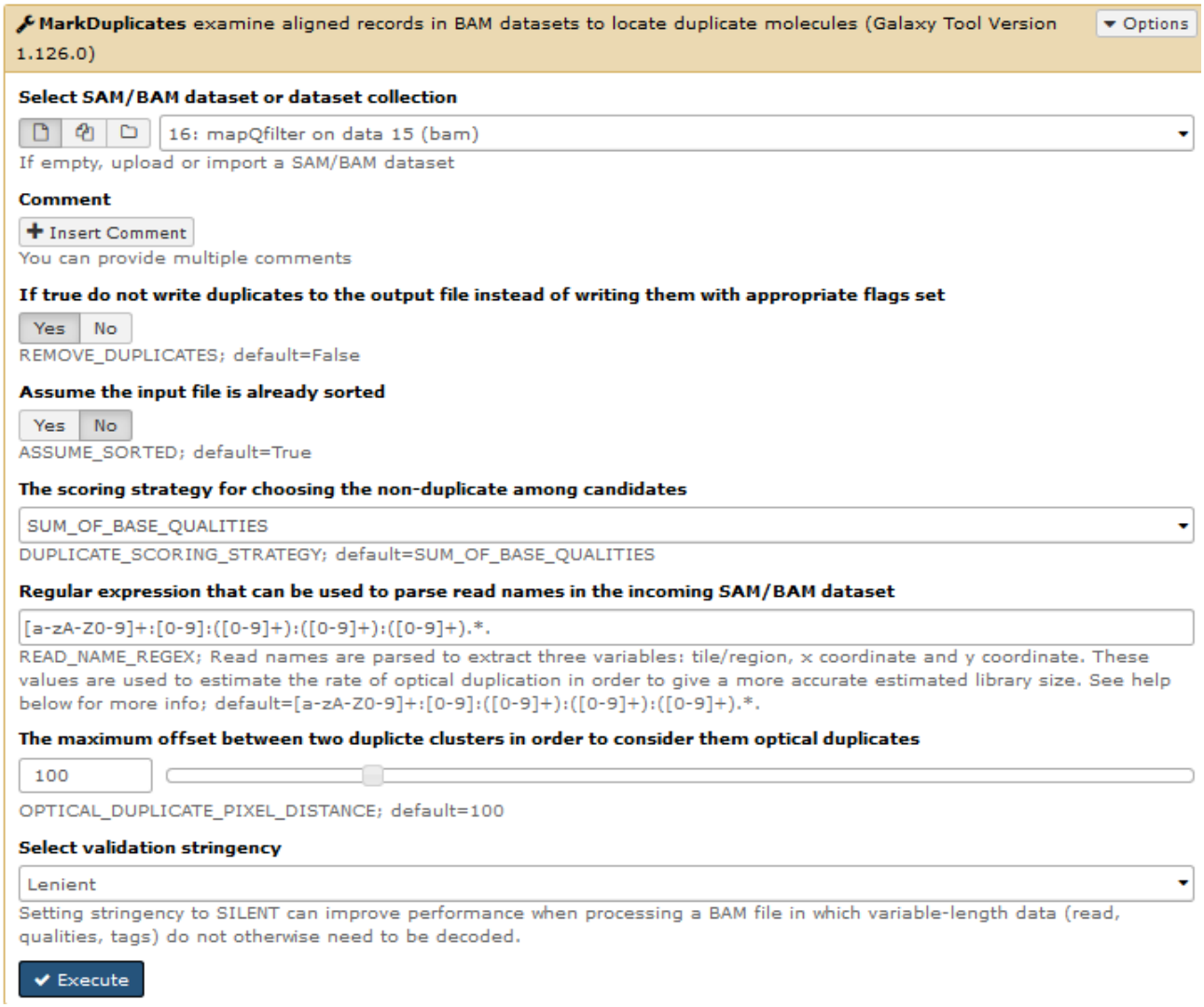
Select mapQfilter in (**Gandalf**). Keep the default option for the mapping quality threshold (30)



b) Markduplicate

Markduplicate is a picard tool that examine aligned records in BAM or datasets to locate duplicate molecules.

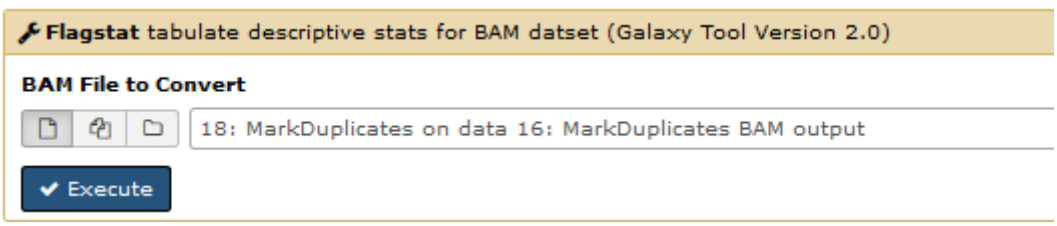
Select MarkDuplicates in (**Gandalf**). Choose to remove duplicates, and to not assume the output is sorted.



4) Statistics on the reads

Flagstat is a SAM tool that gives stats on the final set of reads after filtering. While being optional, it's a fast step that allows to be sure the previous steps went well.

Select flagstat in (**NGS: SAM Tools**). No options.



On the history panel, explore the output file content (click on the eye icone).

```
283254 + 0 in total (QC-passed reads + QC-failed reads)
0 + 0 duplicates
283254 + 0 mapped (100.00%:nan%)
283254 + 0 paired in sequencing
141627 + 0 read1
141627 + 0 read2
283254 + 0 properly paired (100.00%:nan%)
283254 + 0 with itself and mate mapped
0 + 0 singletons (0.00%:nan%)
0 + 0 with mate mapped to a different chr
0 + 0 with mate mapped to a different chr (mapQ>=5)
```

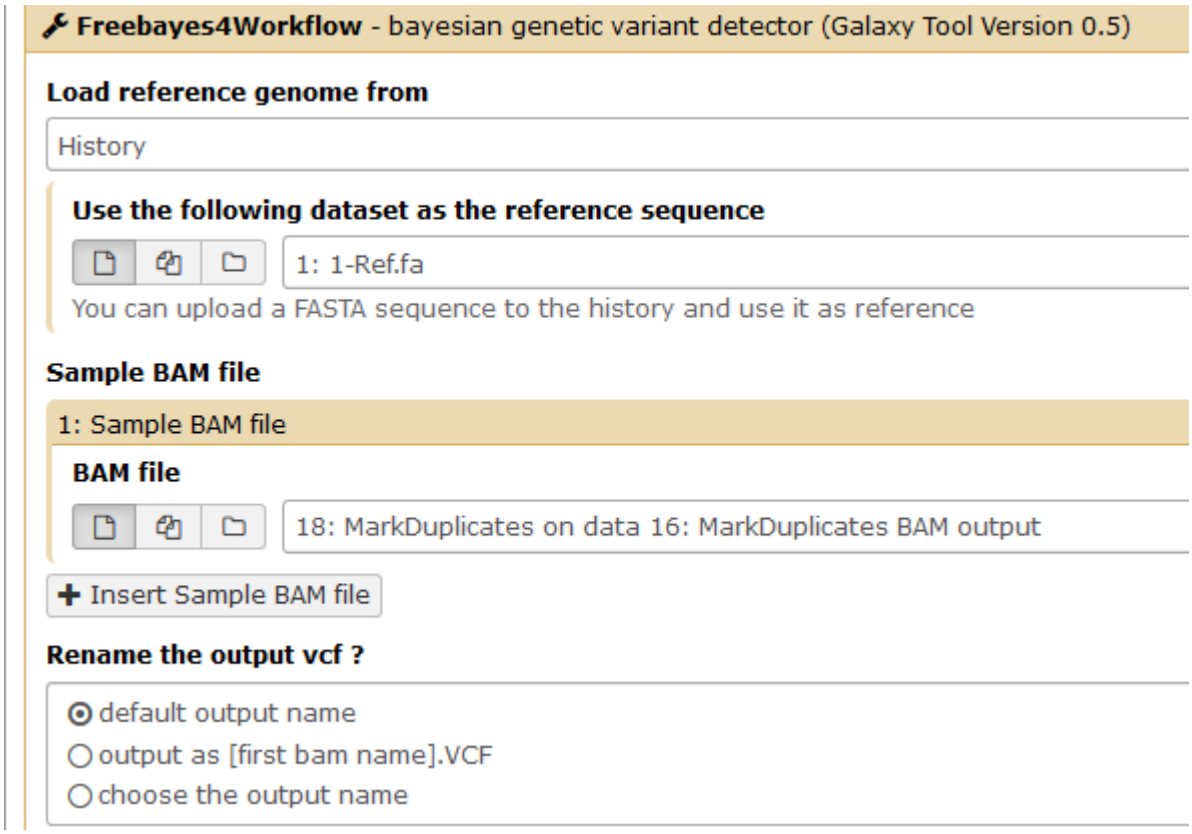
There should not be any duplicates, nor any unpaired or not properly paired reads.

5) SNP calling

Freebayes is a bayesian haplotype-based polymorphism discovery and genotyping tool (SNP caller)

Select Freebayes4Workflow in (**Gandalf**). Choose to load reference genome from history, then choose the corresponding BAM input and reference sequence. Choose the complete list of all options.

- Ask for additional inputs, and in particular choose to report even loci which appear to be monomorphic, and to report all considered alleles (option --report-monomorphic)
- Don't set reporting options
- don't set population model
- don't use reference alleles
- choose to set allelic scope ; In that section choose to ignore MNPs (option -X) and complexe events (option -u)
- check no for the next options : Turn off left alignment of Indels, Set input filters, set population and mappability priors, tweak genotyping likelihood, tweak algorithmic features.



Limit variant calling to a set of regions?

Do not limit

Sets --targets or --region options

Choose parameter selection level

5:Complete list of all options

Select how much control over the freebayes run you need

Do you want to provide additional inputs?

Yes | No

Sets --samples, --populations, --cnv-map, --trace, --failed-alleles, --varinat-input, --only-use-input-alleles, --haplotype-basis-alleles, --report-all-haplotype-alleles, --report-monomorphic options, --observation-bias, and --contamination-estimates

Write out failed alleles file

Yes | No

--failed-alleles

Write out algorithm trace file

Yes | No

--trace

Limit analysis to samples listed (one per line) in the FILE

Nothing selected

-s --samples; default=By default FreeBayes will analyze all samples in its input BAM files

Populations File

Nothing selected

--populations; default=False. Each line of FILE should list a sample and a population which it is part of. The population-based bayesian inference model will then be partitioned on the basis of the populations

Read a copy number map from the BED file FILE

Nothing selected

-A --cnv-map; default=copy number is set to as specified by --ploidy. Read a copy number map from the BED file FILE, which has the format: reference sequence, start, end, sample name, copy number ... for each region in each sample which does not have the default copy number as set by --ploidy.

Provide variants file

Do not provide

Only use variant alleles provided in this input VCF for the construction of complex or haplotype alleles

Nothing selected

--haplotype-basis-alleles

Report even loci which appear to be monomorphic, and report all considered alleles, even those which are not in called genotypes.

Yes | No

--report-monomorphic

Load read length-dependent allele observation biases from

Nothing selected

--observation-bias; The format is [length] [alignment efficiency relative to reference] where the efficiency is 1 if there is no relative observation bias

Upload per-sample estimates of contamination from

Nothing selected

--contamination-estimates; The format should be: sample p(read=R|genotype=AR) p(read=A|genotype=AA) Sample '*' can be used to set default contamination estimates.

Set reporting option?

Yes | No

Sets -P --pvar option

Set population model?

Yes | No

Sets --theta, --ploidy, --pooled-discrete, and --pooled-continuous options

Use reference allele?

Yes | No

Sets --use-reference-allele and --reference-quality options

Set allelic scope?

Yes | No

Sets -I, i, -X, -u, -n, --haplotype-length, --min-repeat-size, --min-repeat-entropy, and --no-partial-observations options

Ignore SNP alleles

Yes | No

-I --no-snps; default=False

Ignore indels alleles

Yes | No

-i --no-indels; default=False

Ignore multi-nucleotide polymorphisms, MNPs

Yes | No

-X --no-mnps; default=False

Ignore complex events (composites of other classes).

Yes | No

-u --no-complex; default=False

How many best SNP alleles to evaluate

0

-n --use-best-n-alleles; default=0 (all). Alleles are ranked by the sum of supporting quality scores. Set to 0 to evaluate all

Allow haplotype calls with contiguous embedded matches of up to (nucleotides)

3

-E --max-complex-gap --haplotype-length; default=3.

When assembling observations across repeats, require the total repeat length at least this many bp

5

--min-repeat-size; default=5.

To detect interrupted repeats, build across sequence until it has entropy > (bits per bp)

0

--min-repeat-entropy; default=0 (off).

Exclude observations which do not fully span the dynamically-determined detection window

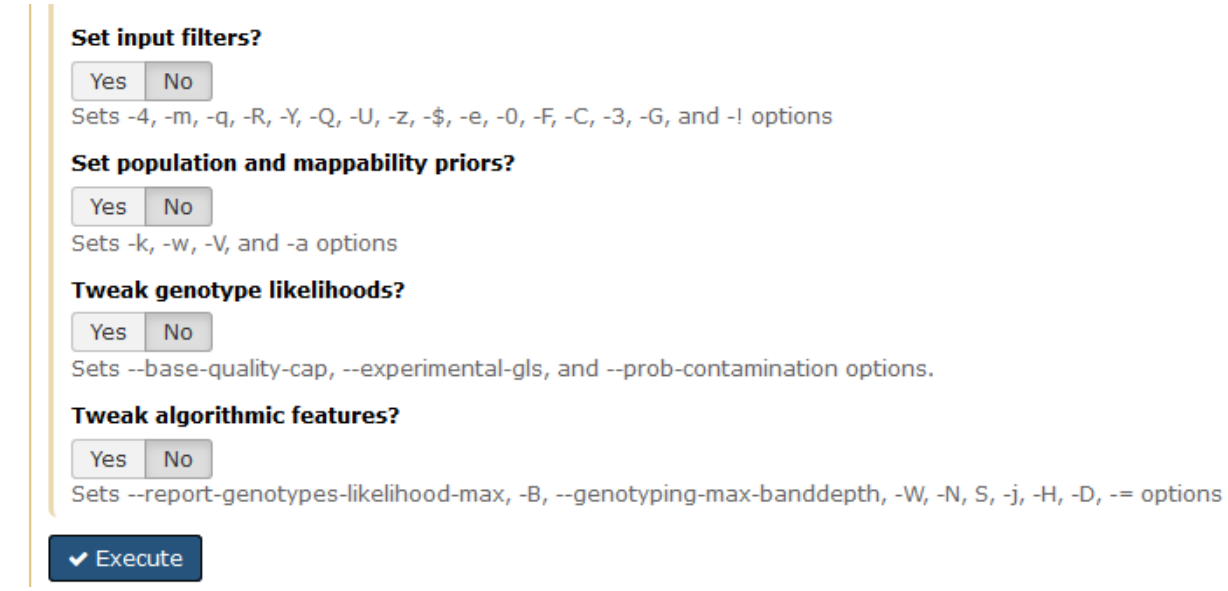
Yes | No

--no-partial-observations; default=use all observations, dividing partial support across matching haplotypes when generating haplotypes.

Turn off left-alignment of indels?

Yes | No

-O --dont-left-align-indels; default=False (do left align).



The set of options purposed here is adapted for the post analysis described below. If you desire a more standard SNP calling, do not hesitate to choose preset options .

6) SNP calling post analysis

The next steps are done with python scripts developped during the Gandalf project.

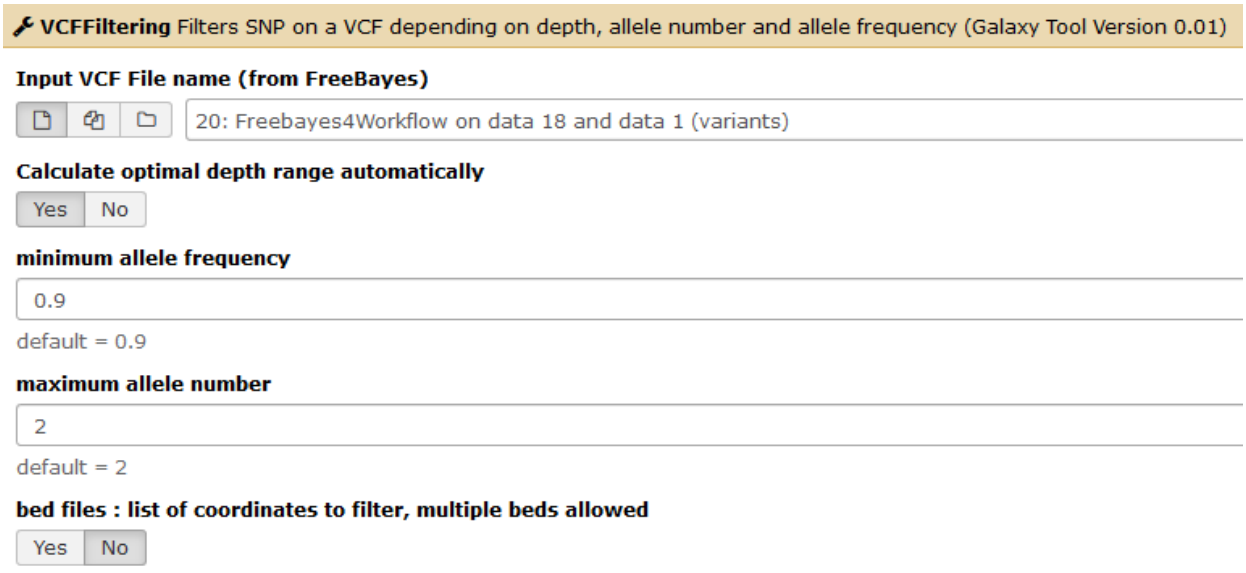
VCFFiltering

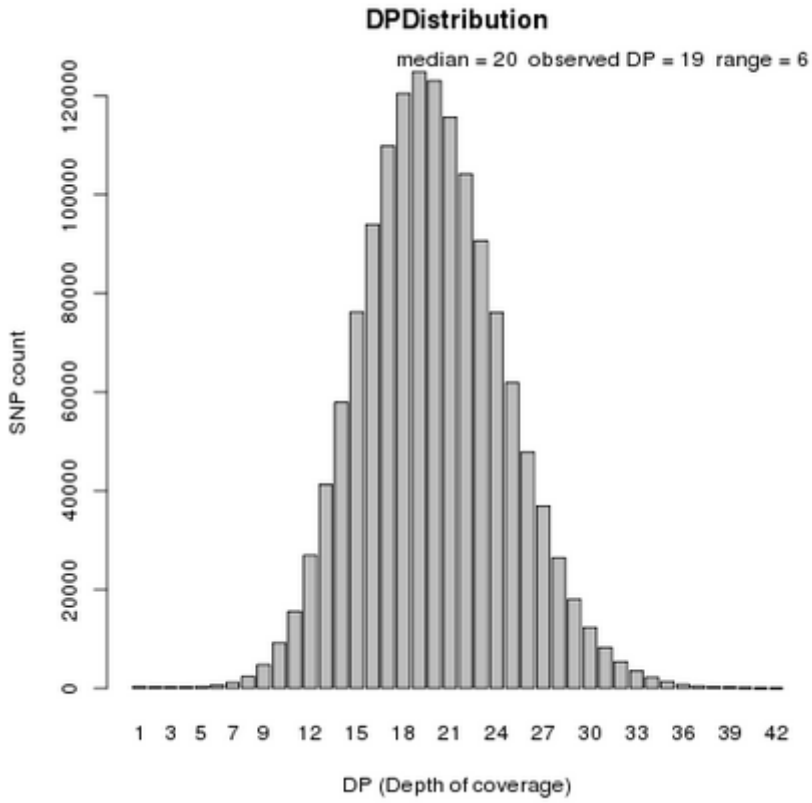
VCFFiltering is a python script that allows to filter SNP results from freebayes on multiple criterias as once. The filters are :

- Allele number : number of possible allele at the genomic position
- Allele frequency : frenquency of the most represented allele ; note that if the most represented allele is the reference (a "." in the 4th column of the VCF, the allele frequency will still work but allele frequency should be under 1-x)
- Depth : Higher and lower bound of the depth ; the depth is the number of reads mapped on the genomic positions.

select VCFFiltering in (**Gandalf**).

If necessary, add a bed file to filter out SNP positions on unwanted regions (e.g. low complexity).



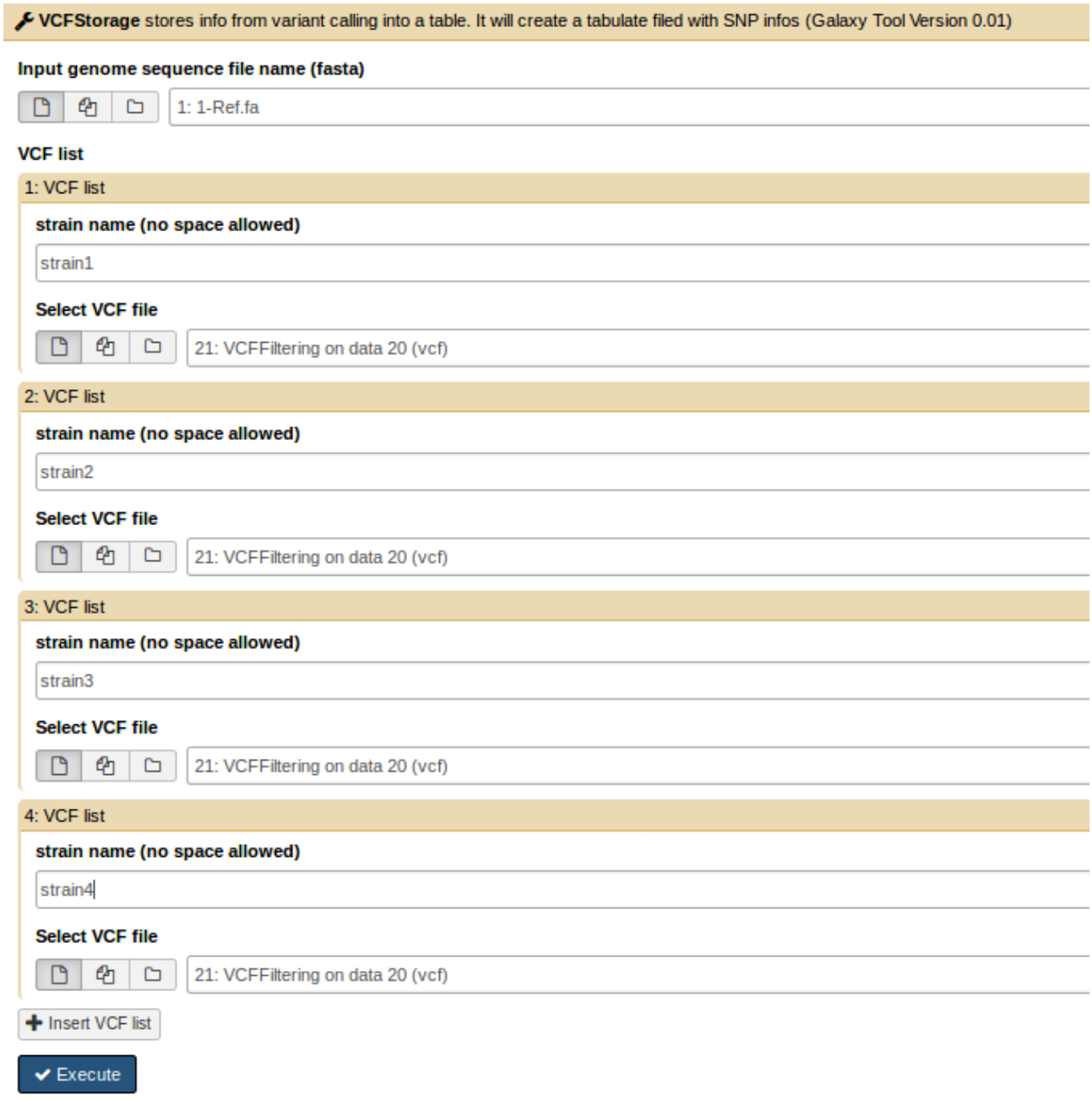

DPDistribution
median = 20 observed DP = 19 range = 6
SNP count
0
20000
40000
60000
80000
100000
120000
1 3 5 7 9 12 15 18 21 24 27 30 33 36 39 42
DP (Depth of coverage)

The next step can be done with multiple results from VCFFiltering.

VCFStorage

VCFStorage allows to store data from multiple VCF files into a single tabular marker file.

select VCFStorage in (**Gandalf**). Put as much VCF as you want. Each new VCF will create a new column on the final output.



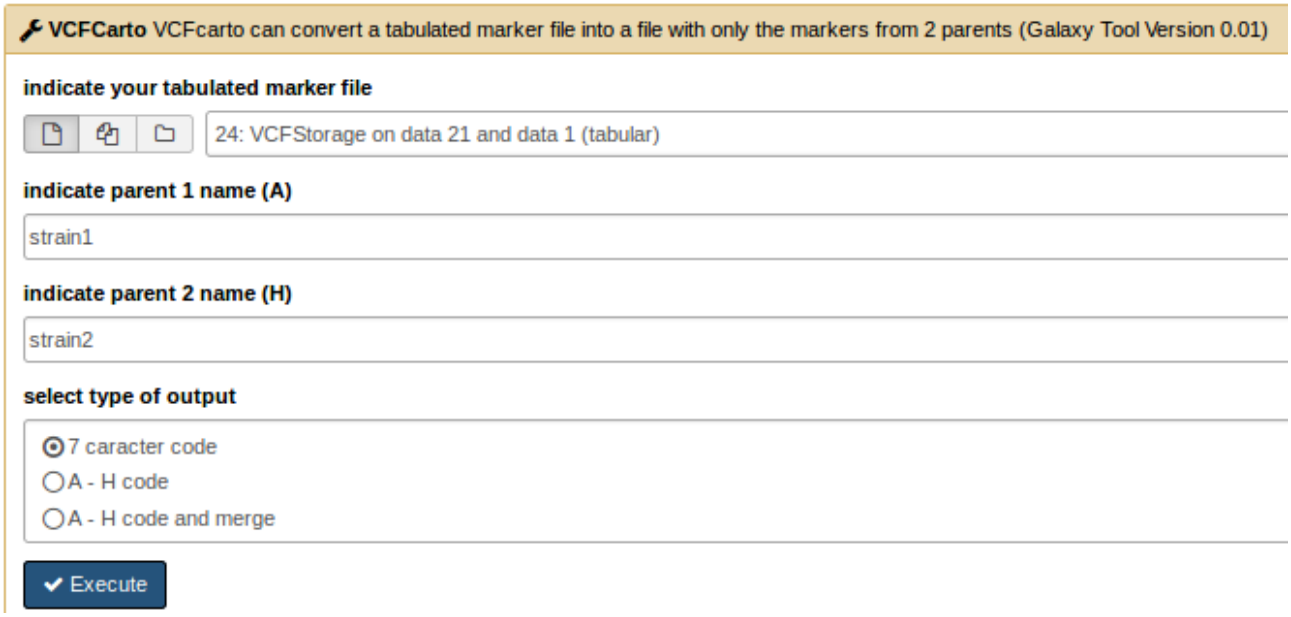
VCFCarto

VCFcarto can convert an output from VCFStorage into a matrix for genetic mapping or GWAS for genetic mapping, use A - H code, otherwise use caracter code.

select VCFCarto in (**Gandalf**).

- for genetic mapping :

write your 2 parents names, then select the "A - H" output code. If you want to merge similar markers, select "A - H code and merge"



3 outputs are generated. Explore the tabulated output content:

The tabulated output content can be used by a cartographic tool (e.g. MSTMap)