

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?

Input Type

Input FASTQ file (R1/first of pair)

Input FASTQ file (R2/second of pair)

Perform initial ILLUMINACLIP step?

 Cut adapter and other illumina-specific sequences from the read

Adapter sequences to use

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

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

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

Trimmomatic Operation

1: Trimmomatic Operation

Select Trimmomatic operation to perform

Number of bases to average across

Average quality required

2: Trimmomatic Operation

Select Trimmomatic operation to perform

Minimum length of reads to be kept

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).

Map with BWA-MEM - map medium and long reads (> 100 bp) against reference genome (Galaxy Tool Version 0.4.1)

Will you select a reference genome from your history or use a built-in index?
Use a genome from history and build index
Built-ins were indexed using default options. See `Indexes` section of help below

Use the following dataset as the reference sequence
 1: 1-Ref.fa
You can upload a FASTA sequence to the history and use it as reference

Single or Paired-end reads
Paired
Select between paired and single end data

Select first set of reads
 2: 2-Strain1_gal1.fq
Specify dataset with forward reads

Select second set of reads
 3: 3-Strain1_gal2.fq
Specify dataset with reverse reads

Enter mean, standard deviation, max, and min for insert lengths.

-I; This parameter is only used for paired reads. Only mean is required while sd, max, and min will be inferred. Exa both "250" and "250,25" will work while "250,,10" will not. See below for details.

Set read groups information?
Set read groups (Picard style)
Specifying read group information can greatly simplify your downstream analyses by allowing combining multiple data:

Auto-assign
 Yes No
Use dataset name or collection information to automatically assign this value

Read group identifier (ID)

This value must be unique among multiple samples in your experiment

Auto-assign
 Yes No
Use dataset name or collection information to automatically assign this value

Read group sample name (SM)

This value should be descriptive. Use pool name where a pool is being sequenced

Auto-assign
 Yes No
Use dataset name or collection information to automatically assign this value

Library name (LB)

Platform/technology used to produce the reads (PL)

Platform unit (PU)

Unique identifier (e.g. flowcell-barcode.lane for Illumina or slide for SOLiD)

Sequencing center that produced the read (CN)

Description (DS)

Predicted median insert size (PI)

Date that run was produced (DT)

ISO8601 format date or date/time, like YYYY-MM-DD

Select analysis mode

3.Full list of options

Set algorithmic options?

Do not set

Sets -k, -w, -d, -r, -y, -c, -D, -W, -m, -S, -P, and -e options.

Set scoring options?

Do not set

Sets -A, -B, -O, -E, -L, and -U options.

Set input/output options

Set

Sets -T, -h, -a, -C, -V, -Y, and -M options.

Minimum score to output

30

-T; default=30

If there are less than THIS VALUE hits with score >80% of the max score, output them all in the XA tag

5

-h; default=5

Output all alignments for single-ends or unpaired paired-ends

Yes No

-a; These alignments will be flagged as secondary alignments

Append FASTA/FASTQ comment to BAM output

Yes No

-C

Output the reference FASTA header in the XR tag

Yes No

-C

Use soft clipping for supplementary alignments

Yes No

-Y; By default, BWA-MEM uses soft clipping for the primary alignment and hard clipping for supplementary alignments

Mark shorter split hits of a chimeric alignment in the FLAG field as 'secondary alignment' instead of 'supplementary alignment'

Yes No


-M; For Picard<1.96 compatibility

3) Filtering




a) SAMfilters

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

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

 **mapQfilter** Filters reads on quality, and remove both members of the pair (Galaxy Tool Version 1.0)

BAM File to filter

   15: Map with BWA-MEM on data 3, data 2, and data 1 (mapped reads in BAM format)

Remove pairs with at least one read under the mapping quality of

30

Execute

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.

MarkDuplicates examine aligned records in BAM datasets to locate duplicate molecules (Galaxy Tool Version 1.126.0) Options

Select SAM/BAM dataset or dataset collection

If empty, upload or import a SAM/BAM dataset

Comment

You can provide multiple comments

If true do not write duplicates to the output file instead of writing them with appropriate flags set

REMOVE_DUPLICATES; default=False

Assume the input file is already sorted

ASSUME_SORTED; default=True

The scoring strategy for choosing the non-duplicate among candidates

DUPLICATE_SCORING_STRATEGY; default=SUM_OF_BASE_QUALITIES

Regular expression that can be used to parse read names in the incoming SAM/BAM dataset

READ_NAME_REGEX; Read names are parsed to extract three variables: tile/region, x coordinate and y coordinate. These values are used to estimate the rate of optical duplication in order to give a more accurate estimated library size. See help below for more info; default=[a-zA-Z0-9]+:[0-9]:([0-9]+):([0-9]+):([0-9]+).*

The maximum offset between two duplicate clusters in order to consider them optical duplicates

OPTICAL_DUPLICATE_PIXEL_DISTANCE; default=100

Select validation stringency

Setting stringency to SILENT can improve performance when processing a BAM file in which variable-length data (read, qualities, tags) do not otherwise need to be decoded.

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.

Flagstat tabulate descriptive stats for BAM dataset (Galaxy Tool Version 2.0)

BAM File to Convert

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

```
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 complex 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.

Freebayes4Workflow - bayesian genetic variant detector (Galaxy Tool Version 0.5)

Load reference genome from

History

Use the following dataset as the reference sequence

1: 1-Ref.fa

You can upload a FASTA sequence to the history and use it as reference

Sample BAM file

1: Sample BAM file

BAM file

18: MarkDuplicates on data 16: MarkDuplicates BAM output

Rename the output vcf ?

default output name

output as [first bam name].VCF

choose the output name

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 NoSets `--P` `--pvar` option**Set population model?** Yes NoSets `--theta`, `--ploidy`, `--pooled-discrete`, and `--pooled-continuous` options**Use reference allele?** Yes NoSets `--use-reference-allele` and `--reference-quality` options**Set allelic scope?** Yes NoSets `--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**`--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)**`--E` `--max-complex-gap` `--haplotype-length`; default=3.**When assembling observations across repeats, require the total repeat length at least this many bp**`--min-repeat-size`; default=5.**To detect interrupted repeats, build across sequence until it has entropy > (bits per bp)**`--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).

Set input filters?

Yes No

Sets -4, -m, -q, -R, -Y, -Q, -U, -z, -\$, -e, -0, -F, -C, -3, -G, and -! options

Set population and mappability priors?

Yes No

Sets -k, -w, -V, and -a options

Tweak genotype likelihoods?

Yes No

Sets --base-quality-cap, --experimental-gls, and --prob-contamination options.

Tweak algorithmic features?

Yes No

Sets --report-genotypes-likelihood-max, -B, --genotyping-max-banddepth, -W, -N, S, -j, -H, -D, -= options

Execute



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 developed 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 : frequency 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).

VCFFiltering Filters SNP on a VCF depending on depth, allele number and allele frequency (Galaxy Tool Version 0.01)

Input VCF File name (from FreeBayes)

Calculate optimal depth range automatically

Yes No

minimum allele frequency

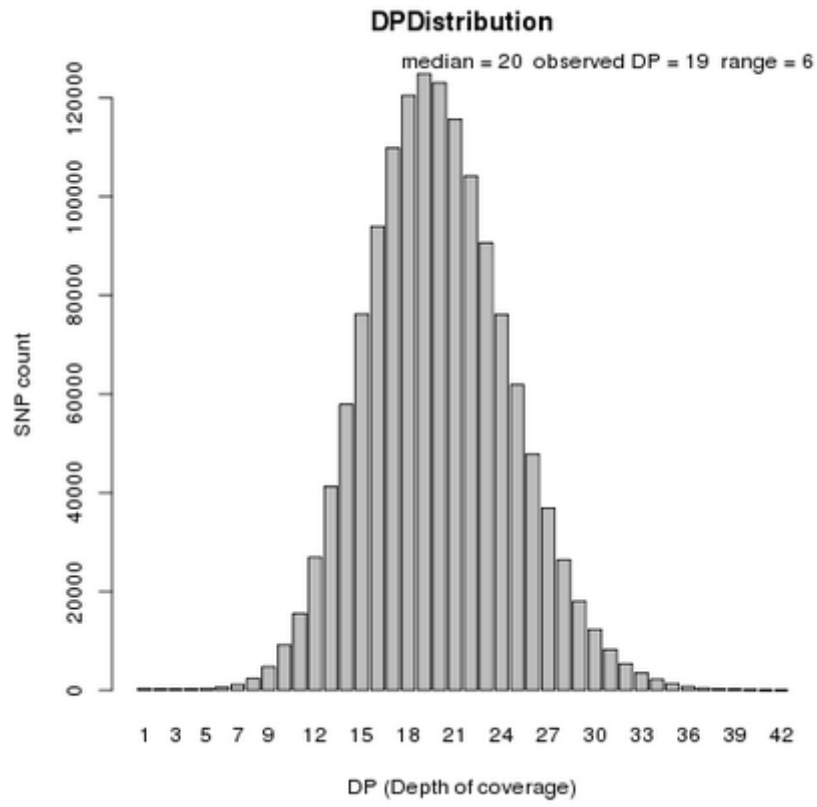
default = 0.9

maximum allele number

default = 2

bed files : list of coordinates to filter, multiple beds allowed

Yes No




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.

 **VCFStorage** stores info from variant calling into a table. It will create a tabulate filed with SNP infos (Galaxy Tool Version 0.01)

Input genome sequence file name (fasta)

   1: 1-Ref.fa




VCF list

1: VCF list

strain name (no space allowed)

strain1

Select VCF file




   21: VCFFiltering on data 20 (vcf)

2: VCF list

strain name (no space allowed)

strain2

Select VCF file




   21: VCFFiltering on data 20 (vcf)

3: VCF list

strain name (no space allowed)

strain3

Select VCF file




   21: VCFFiltering on data 20 (vcf)


4: VCF list


strain name (no space allowed)

strain4

Select VCF file

   21: VCFFiltering on data 20 (vcf)

 Insert VCF list

 Execute


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 character 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"

 **VCFCarto** VCFCarto can convert a tabulated marker file into a file with only the markers from 2 parents (Galaxy Tool Version 0.01)

indicate your tabulated marker file

24: VCFStorage on data 21 and data 1 (tabular)

indicate parent 1 name (A)

strain1

indicate parent 2 name (H)

strain2

select type of output

7 character code
 A - H code
 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)