

## ROBIN manual

Welcome to the manual of ROBIN (ROBust Inference of admixture time). The goal of ROBIN is to infer the time since admixture from patterns in local ancestry. ROBIN is a utility written in python, with usability and portability in mind. However, issues might still remain due to unforeseen user-cases, and if you run into any trouble installing, using or applying ROBIN to your own data, please feel free to contact me to iron out issues at [thijs.janzen@uni-oldenburg.de](mailto:thijs.janzen@uni-oldenburg.de).

## INSTALLATION

ROBIN works both on Python 2.7.x and on Python 3.5. However, it does require a number of packages to run:

- numpy                    - matrix operations
- pandas                  - read/write files
- scikit-allel            - VCF loading
- h5py                    - storing VCF as a local hdf5 file to reduce memory load
- configparser           - reading config file
- scipy.optimize         - Nelder-Mead optimization routine to find best fitting age

All these packages can be installed using 'pip install packagename' on the command line. ROBIN checks upon execution that these packages are indeed installed, and throws an error if they are not.

ROBIN requires two command line parameters:

- c path to the config file
- s random seed

The first option specifies where the user has placed the config file (relative or absolute paths are both accepted), and the second option specifies the random seed used (only for non-scaffolded analyses).

The config file has the following options:

```
[Analysis Type]
# Choice of: 'scaffolds', 'contigs' or 'assembly_free'
analysis = scaffolds

[File Names]
vcf_path = source.vcf
sample_list = sample_list.txt
hdf5_file = anc_panel.h5
panel_file = scaffold_panel.txt
genome_size_file = genome_size.txt
ancestry_hmm = /Ancestry_HMM/src/ancestry_hmm
contig_assignment_file = contig_list.txt

[Parameters]
max_dp = 200
min_gq = 40
required_alleles = 14
freq_ancestor_1 = 0.5
number_of_chromosomes = 24
```

### *[Analysis type]*

Here, the user can specify the type of analysis, and can choose from 'scaffolds', 'contigs' and 'assembly\_free'. Please note that these entries are case and letter specific, mistyping will result in the absence of analysis.

### *[File Names]*

In this section, the user can provide absolute or relative paths to different files used by ROBIN. These files are:

- `vcf_path` : path to the source vcf used as input for the analysis
- `sample_list` : list of all sample names, see below for the format
- `hdf5_file` : the vcf file is not loaded in memory, but rather locally stored in a hdf5 container, which reduces computational strain. The user can specify the name of the hdf5 file, which in turn is read from disk if already present (this saves time)
- `panel_file` : this is the input file for ANCESTRY\_HMM, and is generated by ROBIN. If already present from a previous ROBIN analysis, the existing file is used. See below for the format
- `genome_size_file` : text file with a single entry: the size of the genome (or scaffold) in base pairs. Is calculated from the VCF if not supplied.
- `ancestry_hmm` : path to the executable of ANCESTRY\_HMM, see below for more details
- `contig_assignment_file` : text file indicating the linkage group a contig is assigned to, see below for the format.

### *[Parameters]*

- `max_dp` : maximum read depth for a variant call, to avoid including repetitive regions. Default is 200.
- `min_gq` : minimum phred 10 score to include a variant call.
- `required_alleles` : minimum number of alleles required from both ancestors in order to include a SNP
- `freq_ancestor_1` : frequency of the first ancestor in the hybrid swarm.
- `number_of_chromosomes` : total number of autosomes in the dataset. Only used for assembly-free analysis.

### *VCF layout*

ROBIN has been tested on VCFs generated with VCFv4.2. For further information about the VCF format, we redirect the user to <http://samtools.github.io/hts-specs/VCFv4.3.pdf>. Currently, ROBIN can not yet analyze a VCF file containing multiple chromosomes (for scaffold based analysis), so rather the user needs to provide a separate VCF file for each chromosome. Ideally it requires entries for DP and GQ, but both can be inferred from RR/VR and PL as well if necessary.

### *Sample list*

To extract the correct information regarding the ancestry panel and the hybrid genotypes, ROBIN requires information about which samples in the VCF belong

to which ancestor, and which are hybrid. Information about this should be supplied in the sample list file, which looks the following:

Cperi_LarseBeek1PGA_March_2018.ngm.mapped.sort	Ancestor_1
Cperi_Molenbeek1PGA_March_2018.ngm.mapped.sort	Ancestor_1
Cperi_RiverTwin1PGA_March_2018.ngm.mapped.sort	Ancestor_1
Crhen_BroelWPGA_March_2018.ngm.mapped.sort	Ancestor_2
Crhen_FokkenbachWPGA_March_2018.ngm.mapped.sort	Ancestor_2
PeriInv3_2015PGA_March_2018.ngm.mapped.sort	Hybrid

Samples belonging to the panel of the first ancestor (usually the most frequent ancestor) are indicated with 'Ancestor\_1'. Samples belonging to the panel of the second ancestor are indicated with 'Ancestor\_2', and lastly, samples belonging to the focal hybrids are indicated with 'Hybrid'.

### *Panel file*

This is the file used as input for ANCESTRY\_HMM, but which can also be used for other analyses. The file contains 7 fixed columns, plus 2 columns per hybrid sample. A typical line from such a panel file (for non-scaffolded analyses) might look like this:

```
tig000000014 18704 5 9 14 0 18704 1 1 0 0 2 0
tig000000014 18725 6 8 14 0 18725 1 1 1 1 2 0
tig000000020 9874 6 8 14 0 9874 1 1 1 1 0 2
tig000000020 21634 5 9 14 0 21634 1 1 2 0 0 0
tig000000024 10381 0 16 14 2 10381 0 2 0 2 0 0
tig000000024 22882 6 8 14 0 22882 1 1 0 0 0 0
tig000000024 22944 2 12 14 0 22944 0 2 0 0 0 0
```

The following 7 fixed columns are used:

- contig name, or chromosome name (in the case of scaffold based analysis)
- location within the contig (in base pairs)
- number of reference alleles in 'Ancestor\_1'
- number of alternative alleles in 'Ancestor\_1'
- number of reference alleles in 'Ancestor\_2'
- number of alternative alleles in 'Ancestor\_2'
- location within the contig (will later be re-calculated to recombination distance)

Then, per hybrid, two columns are added which denote:

- the number of reference alleles
- the number of alternative alleles

The numbers of reference and alternative alleles are extracted by ROBIN from the VCF, and the sum of alleles between the two ancestral panels is kept artificially equal by down sampling the ancestor with a higher number of alleles (see the manuscript for details).

To speed up repeated assembly-free analysis, the total panel is saved to a text file (path and name of the text file are provided by the user in the config file), and will be reloaded if it is available.

### *Ancestry\_HMM*

To infer local ancestry, ROBIN makes use of the external program 'ANCESTRY\_HMM', which can be found here:

[https://github.com/russcd/Ancestry\\_HMM](https://github.com/russcd/Ancestry_HMM). In order for ROBIN to work, it

requires the path to the executable of ANCESTRY\_HMM (usually in the /src/ folder after installation), and it requires the executable to have the corresponding rights (typically, `chmod +x`). For full installation details we redirect the user to the manual of ANCESTRY\_HMM ([https://github.com/russcd/Ancestry\\_HMM/blob/master/Manual\\_V0.94.pdf](https://github.com/russcd/Ancestry_HMM/blob/master/Manual_V0.94.pdf)), but in short:

```
> git clone https://github.com/russcd/Ancestry_HMM.git
> cd Ancestry_HMM/src/
> make
```

This should compile the executable. The executable requires the C++ linear algebra library Armadillo, which can be obtained from:

<http://arma.sourceforge.net> or by using homebrew on OSX:

```
> brew install homebrew/science/armadillo
```

Further details and testing of the installation of Ancestry\_HMM can be found in the manual of Ancestry\_HMM, which can be found here:

[https://github.com/russcd/Ancestry\\_HMM/blob/master/Manual\\_V0.94.pdf](https://github.com/russcd/Ancestry_HMM/blob/master/Manual_V0.94.pdf).

### *Contig\_assignment file*

In the case of a contig based analysis, ROBIN needs to know on which linkage group each contig is found. The contig file has two columns, with the first indicating the linkage group number, and the second indicating the contig name (as used in the VCF):

```
3      tig00010465
3      tig00010467
3      tig00010462
3      tig00010698
3      tig00010768
3      tig00010860
3      tig00010862
```

### *Output*

ROBIN leaves behind a number of output files:

- hybrid\_input\_CHROM\_HYBRIDINDEX.txt'
- SAMPLENAME\_CHROM.posterior
- sample\_HYBRIDINDEX.txt
- output.txt

For every chromosome (true chromosomes as in the scaffolded analysis, completely artificial chromosomes as in the assembly\_free analysis, or semi-artificial chromosomes from the contigs analysis), a separate ANCESTRY\_HMM analysis is performed to infer local ancestry, and a separate local input and output file are created. The input file is called 'hybrid\_input\_CHROM\_HYBRIDINDEX.txt', where CHROM indicates the chromosome, and HYBRIDINDEX indicates the numbered indicator of the number of hybrids, e.g. if the user provided three hybrid samples in 'sample\_list.txt', HYBRIDINDEX of 0 indicates the first hybrid sample (counting is 0,1,2.. etc). The output file of ANCESTRY\_HMM is named 'SAMPLENAME\_CHROM.posterior', where 'SAMPLENAME' refers to the name of

the sample of that hybrid (as indicated in 'sample\_list.txt'), and CHROM again refers to the analyzed chromosome (real or artificial). Lastly, the file 'sample\_HYBRIDINDEX.txt' is an intermediate file and is only used to inform ANCESTRY\_HMM of sample names and ploidy numbers. Lastly, the file 'output.txt' provides the overall output that the user is after, and is organized in a number of columns:

- Sample name
- Chromosome number
- Ancestry uncertainty
- Population Size
- Number of detected junctions
- Inferred age

A snapshot of such an output file:

PeriInv3_2015contigs.ngm.mapped.sort	2	1e-05	10000	7	16
PeriInv3_2015contigs.ngm.mapped.sort	2	1e-05	100000	7	16
PeriInv3_2015contigs.ngm.mapped.sort	2	0.0001	1000	23	77
PeriInv3_2015contigs.ngm.mapped.sort	2	0.0001	10000	23	57
PeriInv3_2015contigs.ngm.mapped.sort	2	0.0001	100000	23	53
PeriInv3_2015contigs.ngm.mapped.sort	2	0.001	1000	65	310
PeriInv3_2015contigs.ngm.mapped.sort	2	0.001	10000	65	180