**Pipeline of Relate software**

The original software information can be found here:

<https://myersgroup.github.io/relate/index.html>

This pipeline optimized the original software for more simple usage.

NOTE: CLUES has been updated to CLUES2.

\*\*IMPORTANT

If you want to use CLUES2, please download it from https://github.com/avaughn271/CLUES2 (this is not included in the Relate package). After downloading, please put the folder under the Relate folder and rename the folder name as "clues".

When choosing the type of data using -syn/-self/-hap, please only use one of them. If you don’t use any of these, it will assume that your data is normal diploid data.

When -syn is set, -pop cannot be set. Otherwise, please set -pop.

To do the analysis, first you need to prepare files:

1. VCF file(s). It is better to use multiple vcfs that every vcf contains only one contig/chromosome.
2. A poplabels file. Please see “K7\_07\_poplabels.txt” for format.
3. A genetic map file. Please see “rmap\_all.txt” for format example.
4. A reference fasta file. This is the genome sequence of the species that you made for the vcf file(s).
5. [Optional] A mask file. This is the same as the reference fasta file, but the mask regions are replaced by “N”.
6. For self-pollinating plant data, please use -self argument. Otherwise, you need to prepare synthetic F1 for running.
7. If you prefer to use synthetic F1 for self-pollinating plant data, please use -syn argument, and prepare a synthetic list file. (see “syn\_F1\_list.txt” for format.)

\*\*Make a masked fasta file by a bed file

If you want to make a masked reference fasta file, and you have a bed file indicating the repeated regions, you can follow the following steps to generate the masked fasta file:

1. Use “bedtools” to mask the fasta file. (If you don’t have bedtools, you can install it in the conda environment)

**bedtools maskfasta -fi INPUT\_FASTA\_FILE -bed BED\_FILE -fo OUTPUT\_FILE\_NAME**

1. The masked file still contains A,T,C,G at the unmasked regions, and Relate needs them to be “P”. Thus, you need to replace any A,T,G,G to P. The following command line will do this without replacing any A,T,C,G characters in the header line.

**sed '/^>/!s/[ATCG]/P/g' INPUT\_MASKED\_FASTA\_FILE > OUTPUT\_MASKED\_FASTA\_NAME**

\*\*IMPORTANT

1. When choosing the type of data using -syn/-self/-hap, please only use one of them. If you don’t use any of these, it will assume that your data is normal diploid data.
2. When -syn is set, -pop cannot be set. Otherwise, please set -pop.
3. Somehow, **h71** has some configuration error to run Relate. Be careful when using this server.

Steps:

1. Run Relate\_direct.pl script

**nohup perl Relate\_6\_direct\_5.pl -vcf VCF\_FILE -pop POPULATION\_LABEL\_FILE -map RECOMB\_MAP\_FILE -al ANCESTOR\_ID\_LIST [-am ANC/MUT\_FOLDER\_PATH] [-hap] [-o OUTPUT\_PATH] [-mask MASK\_FILE] [-bins LOWER,UPPER,STEPSIZE] [-rm REMOVE\_SAMPLE\_ID\_FILE] [-pre PREFIX] [-rr] [-coal COAL\_FILE] [-dps] [-clues] [-tvs] [-bp CHR:POS-POS] [-ns INT] [-tco INT] [-pf FLOAT] [-d FLOAT] [-cp COLOR\_PALETTE] [-m VALUE] [-n VALUE] [-spl VALUE] [-popi POP\_NAMES] [-year VALUE] [-rp all|clues] [-cb FILE] [-rc INT] [-epsrp INT] [-rpp] [-nat] [-mem INT] [--force] [-ow] [-sn SERIAL\_NUMBER] [-exc] [-h]**

**> LOGFILE 2>&1 &**

-r: Reference fasta file [split\_fasta.pl is required]

-vcf: vcf files, can point to a multi-contig vcf or a folder containing multiple vcfs with one chromosome/contig per vcf.

-pop: a poplabels file.

-o: output path, default: working folder

-am: If you already have \*.anc and \*.mut files, just indicate the folder path of these files by this argument.

-hap: If the data is haploid (or self-pollinated species), please use this argument.

-mask: a genome mask file [split\_gmap.pl is required]

-rm: A list file. Remove sample from the analysis, one sample ID per line

-pre: prefix of the chromosome/contig name. If you only want to process chromosomes/contigs beginning with a certain name, you can use this argument. (eg. “Chr”, then chloroplast and mitochondria might be excluded)

This is case sensitive.

-rr: re-run main Relate program to generate new \*.anc and \*.mut files without re-running vcf to \*.haps and \*.sample process.

-coal: if you have \*.coal file, you can use this argument

-bins: set boundary of the year for estimating effective population sizes

-dps: detect positive selection function (model)

-ci: file path and file name without extension for sample branch lengths used in CLUES

-bp: region of interests (chromosome\_name:start\_position-end\_position)

-ns: Number of times branch lengths are sampled for -ci. Integer >= 1, Default: 5

-m: mutation rate per base per generation. Default: 1e-8

-n: effective size. Default: 30000

-popi: population of interest. This argument can specify populations that you want to do analysis. multiple populations as a group could be indicated by [population\_1,population\_2]. Multiple independent runs can be indicated by comma as population\_1,population2.You can combine these two functions as [population\_1,population\_2],population\_3

The first run will be population\_1+population\_2, and the second run will be population\_3.

-clues: do sample branch lengths (CLUES, model)[need an additional CLUES software]

-tco: generation boundary of sample branch lengths (CLUES)

-pf: popFreq argument in CLUES plotting function.

-d: dom argument in CLUES plotting function.

-cp: color series for CLUES plotting.

-cb: set boundary of the year for CLUES’ plotting

-rp: redo CLUES2 calculation. possible values: all,clues

-rpp: re-plot CLUES2 results.

-nat: apply --noAlleleTraj in CLUES2, only for significance test.

-epsrp: define how many times to estimate the effective population size. It may take a lot of time. If you don’t need 95%CI for the effective population size, set it to 0. (default: 100)

-tvs: Calculate TreeViewSamples for SNP ages.

-rc: Repeat runs. (for CLUES and TreeViewSamples)

--force: when -rc > 1000, please use this option for force run.

-cf: output format for Sample branch lengths function. possible values: “a” or “b”. For CLUES, the default value is “b”. For TreeViewSamples, the default value is “a”.

-year: years per generation. Default: 1

-spl: smoothing genetic map by spline function. Please indicate the level of smoothing (integer value. Default = 25, if this argument is used.)

-ow: over-write all output files

-mem: memory used in the main analysis. (default: 12)

-sn: serial number. If you want to process the analysis using previous generated files, just set the –sn to the same serial number used for the previous run.

-exc: send the job to run

-h: help

\*\* I suggest you don’t use all model functions at the first time unless you are familiar with the scripts.

Note:

In Relate website, some function descriptions are wrong:

#### For coalescence rates:

#### 

#### No program named “RelateCoalescenceRate”, instead, please use “RelateCoalescentRate”

#### 

#### Also, please use “RelateCoalescentRate” to do this re-estimation. There is no argument –-anc and –-mut (caused core dump), please use “-i” instead. The file name for –i should be a prefix of the anc and mut files (“example” for example). If you have multiple chromosomes, you need to do looping, and the file name for –i should be “example\_chrX”. X is the number index.

For EstimatePopulationSize.sh with --pop\_of\_interest:

If you are using the files downloaded from the official website, please change “mv” to “cp” in line 432/452 (v.1.19/v.1.21) and line 970/982 (v.1.19/v.1.21) of EstimatePopulationSize.sh. Otherwise, an error will appear. (the bug has been reported to the author.)

If you see some information such as “killed” in the \*.out file of the job’s output folder, you need to increase the memory size using the “-mem” option and run the script again.