Ancestry HMM Manual

Russ Corbett-Detig

Download and Compile:

git clone https://github.com/russcd/Ancestry_HMM.git
cd Ancestry_HMM/
g++ -std=c++0x -ltcmalloc -O3 -o ancestry_hmm ancestry_hmm_V0.4.cpp

Basic usage:

./ancestry_hmm (options) –i (int, ploidy) (string, input_file) (string, posterior_file)

Options and Usage:

Parameter	Default	Meaning
-a <long double>	N/A	Fixed ancestry proportion. It is recommended that this option be set.
--ai <long double>	N/A	Initial ancestry proportion that will be updated during optimization. If neither –a nor --ai is set, the program will attempt to estimate the proportion. Using this program without setting one of these options has not been thoroughly tested, and it is not recommended.
--t <int>	N/A	Fix number of generations since admixture. HMM will evaluate the supplied time and exit without optimization.
--tmax <int>	10,000	Maximum time for golden section search optimization
--tmin <int>	1	Minimum time for golden section search optimization
--tolerance <int>	10	Minimum distance between successive test points in golden section search. i.e. the search will terminate when within this distance.
--ne <int>	1e4	Diploid effective population size of admixed population
-d <long double>	0	Minimum distance (in morgans) between adjacent markers. If LD pruning has not been performed on the ancestral genotype data, this option may be useful for ensuring markers are not in strong LD.
-g	False	If set, sample counts are assumed to be genotypes, not read counts. Genotype counts must sum to the sample ploidy.
-e <long double>	1e-2	Error rate per read per site.
--ge <long double>	1e-5	Error rate of sample genotypes (i.e. used with –g). This should only be necessary when used with --fix.
--fix	False	Ancestral allele frequencies are assumed to be fixed. This might apply to *e.g.* qtl mapping or experimental evolution applications where the ancestral allele frequencies are known with high confidence.

-i <int> <string> <string>	N/A	Required. This option must be specified for all input files for a given run of the program, and therefore will be specified multiple times if multiple samples are to be included in LAI. In brief, this specifies sample ploidy with the first argument, the input file name with the second argument, and the output file name for outputting posterior distributions at all sites in that sample.

Input File Format

In the present version, all samples must be provided in a separate file. The format of this file is tab delimited with columns:

1. Position in basepairs
2. Allele counts of allele A in reference panel 0
3. Allele counts of allele a in reference panel 0
4. Allele counts of allele A in reference panel 1
5. Allele counts of allele a in reference panel 1
6. Read counts of allele A in the sample
7. Read counts of allele a in the sample
8. Recombinational distance in morgans between the previous marker to the left and this position. For the first position, this may take any value.

Output File Format

The output file is also tab delimited. Each sample will have a separate out file, specified via the command line as described above. The format of this file is

1. The position in basepairs of the observation
2. The posterior probability of state 0.
3. –n posterior probability for states 1-n.