MALDsoft

Release 1.0, October 2004
Software for Mapping by Admixture Linkage Disequilibrium
By Giovanni Montana and Jonathan Pritchard

The program MALDsoft computes statistical tests for admixture mapping based on both case-only and case-control data, as described in Montana and Pritchard (2004); see also Falush et al. (2003).

Installation

In order to install this software, simply unzip all the files contained in the current distribution in a folder of your choice. A makefile script is also included for those who wish to compile the program from the source.

General Usage

Please make sure that the STRUCTURE2 program is also available for your analysis, as the MALDsoft computations are based on the results produced by STRUCTURE2. This program can also be obtained from:

http://pritch.bsd.uchicago.edu/software/

A typical admixture mapping project involving STRUCTURE2 and MALDsoft is performed in three steps:

1. The input data file and the configuration files (usually named *mainparams* and *extraparams)* are set up. *The configuration files must be accessible to both STRUCTURE2 and MALDsoft, and therefore should be located in the appropriate folders.* See the section “Setting Up Data and Configuration Files” for guidance on how to perform this step.

2. STRUCTURE2 is run, and an output file (usually with extension *_f)* is created. *This output file is needed by MALDsoft, and therefore must be located in the same folder where MALDsoft is run*.

3. MALDsoft is run. The program computes the required statistical tests, and generates several output files. See the section “MALDsoft Output Files” for more information on what the output files contain.

Setting Up Data and Configuration Files

The file containing the multilocus genotype data can be set up in any format accepted by STRUCTURE2. Two configuration files, *mainparams* and *extraparms*, need to be created and will contain several parameters indicating all the elements of the input data file. These two configuration files are required in order to run both STRUTURE2 and MALDsoft, and should therefore be accessible to both programs. The user can create them once, and then copy them in the appropriate folders.

A few rules must be observed when setting up an input data set:

- **Map distances**. The data set must contain a row with map distances between loci, and the parameter `MAPDISTANCES=1` must be specified in the *mainparams file*. The first value should be coded as -1. The map distances should be specified in the top row (or in the second row, if the first row has the marker names; see Table A below).

- **Learning samples.** If samples of non-admixted individuals are available, they can be used as learning samples for the estimation of the population-specific allele frequencies. In this case, the configuration file *mainparams* should contain the parameter `POPDATA=1`, and the input data file should contain a `POPFLAG` column; individuals with `POPFLAG=1` will be used as learning samples, while the remaining individuals will have `POPGLAG=0`. Non-admixted individuals can also be used to better assist the clustering algorithm implemented in STRUCTURE2; this is done by setting `USEPOPINFO=1` in the *extraparams* file; in this case, individuals with `POPFLAG=1` will be used for the clustering. We suggest setting `USEPOPINFO=1`. See example below.

- **Phenotype information.** A `PHENOTYPE` column is also necessary in order for the program to distinguish between cases and controls. The current version of MALDsoft only deals with a dichotomous phenotype variable: individuals that present the phenotype of interest (cases) should be labeled as `PHENOTYPE=1` while the controls should have `PHENOTYPE=0`. A `PHENOTYPE` column should always be included in the data set, even when a case-only test is required. See example below.

Tables A, B and C at the end of this document provide a complete description of all the parameters that need to be specified within each configuration file. Some of these parameters allow to customize the format of the input data file. In particular, Table B specifies the admixture mapping parameters.

Please consult the STRUCTURE2 documentation for a comprehensive description of some extra options available.

The example configuration files included with this distribution can be used as templates for your own project

Example

As an example, the following diagram shows on how to set up the first columns of a data file relating to learning samples and phenotype information as described above. Here it is assumed that there are two ancestral populations (`MAXPOPS=2`), and samples of individuals representing each one of the two population have been collected and coded as `POPDATA=1` and `POPDATA=2`. Individuals sampled from the admixted population are coded as `POPDATA=3`. From this population, it is assumed that a sample of controls (coded as `PHENOTYPE=0`) is available.

LABEL	POPFLAG	POPDATA	PHENOTYPE	
1	1	1	0	Learning samples, Population 1
2	1	1	0	
3	1	2	0	Learning samples, Population 2
4	1	2	0	
5	0	3	0	Admixed samples, Controls
6	0	3	0	
7	0	3	1	Admixed samples, Cases
8	0	3	1	

The `LABEL` parameter used in this example uniquely identifies each individual. The format used here also assumes `ONEROWPERIND=1` (the data for each individual are arranged in a single row).

MALDsoft Usage

All the parameters for admixture mapping are specified in the *mainparams* file as described in Table B. Once the MCMC estimation performed by STRUCTURE2 is completed, the program can be run. This is usually done from a shell; for instance, Windows users can open an MS-DOS shell and type the command “maldsoft“.

The program can also be lunched as “maldsoft [options]”, where [options] can be any combination from the following parameters:

-i	`INFILE`	-H	`PHASE`
-o	`OUTFILE`	-T	`WHICHTEST`
-N	`NUMINDS`	-R	`MCREPS`
-L	`NUMLOCI`	-X	`EXTREMES`

Be advised that the parameter values specified via command line will override those in the *mainparams* configuration file.

MALDsoft Output Files

MALDsoft generates several output files, depending on which parameters have been specified. The name of each output files is given by the value of the `INFILE` parameter followed by an extension. The format of each output file name is "`INFILE_EXTENSION`". Table D describes all the MALDsoft output files.

References

Montana, G. and J.K. Pritchard, *Statistical tests for admixture mapping with case-control and cases-only data.* Am J Hum Genet, 2004. **75**(5): p. 771-89.

Falush, D., M. Stephens, and J.K. Pritchard, *Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies.* Genetics, 2003. **164**(4): p. 1567-87.

TABLE A. Mainparams File, general parameters.

Parameter	Type	Description
`INFILE`	String	The name of the input data file
`OUTFILE`	String	The name of the output data file. The output file created by STRUCTURE2 will have this name plus an extension "_f", and it is read when running MALDsoft.
`NUMINDS`	Integer	The number of individuals in the data file.
`NUMLOCI`	Integer	The number of loci in the data file.
`LABEL`	Boolean	It says whether or not (1=yes, 0=not) the input file contains labels for each individual.
`POPDATA`	Boolean	It says whether or not (1=yes, 0=not) the input file contains an indicator variable that identifies the population of origin for each individual. If learning samples are available, then set `POPDATA=1`.
`POPFLAG`	Boolean	It says whether or not (1=yes, 0=not) the input file contains an indicator variable used for the identification of the learning samples. See example above.
`PHENOTYPE=1`	Boolean	It says whether or not (1=yes, 0=not) the input file contains an indicator variable used for the identification of individuals having the phenotype of interest (cases versus controls). This parameter must be set to 1.
`EXTRACOLS`	Integer	Number of additional columns of data after the phenotype before the genotype data start. These columns are ignored by the program.
`PHASEINFO`	Boolean	It says whether or not (1=yes, 0=not) haplotype phase is known.
`MARKOVPHASE=0`	Boolean	This option must be set to 0.
`MISSING`	Integer	The value given to missing genotype data
`PLOIDY`	Integer	Ploidy of the organism. Default is 2 (diploid).
`ONEROWPERIND`	Boolean	It says whether or not (1=yes, 0=not) the data for each individual are arranged in a single row.
`GENENAMES`	Boolean	It says whether or not (1=yes, 0=not) the top row of the data file contains a list of names corresponding to the markers used.
`MAPDISTANCES=1`	Boolean	It says whether or not (1=yes, 0=not) there is a top row containing a list of map distances between neighboring loci. If `GENENAMES=0`, this is the first row, otherwise is the second. This parameter must be set to 1.
`MAXPOPS=2`	Integer	The number of populations assumed. Set to 2 for this release of MALDsoft.
`BURNIN`	Integer	The length of burn-in period before the start of data collection. Used by STRUCTURE2 only.
`NUMREPS`	Integer	Number of MCMC reps after burn-in. Used by STRUCTURE2 only.

TABLE B. Mainparams File, MALDsoft-specific parameters.

Parameter	Type	Description
WHICHTEST	Integer	It says which admixture mapping test has to be computed: 1=cases-only and 2=case-control. If 2 is chosen, case-only tests will also be computed by default.
MCREPS	Integer	The number of required replicates within the parametric bootstrap estimation. We suggest setting this parameter to any integer greater than 100.
EXTREMES	Integer	The number of simulated data sets required to approximate the distribution of extreme tests at each locus. We suggest setting this parameter to any integer greater than 100. It can be set to zero if the distribution of extreme values is not required.

Please notice: the fact that the *mainparams* file contains parameters that are MALDsoft specific will cause STRUCTURE2 to print some warnings and notify the user that a few parameters are not being recognized; these warnings are to be expected and can simply be ignored.

TABLE C. Extraparams File, essential parameters.

Parameter	Type	Description
LINKAGE=1	Boolean	It says whether or not (1=yes, 0=not) the Linkage Model has to be used. This parameter must be set to 1.
NOADMIX=0	Boolean	It says whether or not (1=yes, 0=not) to assume the model without admixture. This parameter must be set to 0.
USEPOPINFO=1	Boolean	It says whether or not (1=yes, 0=not) to use prior population information to assist clustering. If learning samples are available, this parameter should be set to 1. See also POPDATA and POPFLAG.
PFROMPOPGLAGONLY=1	Boolean	It says whether or not (1=yes, 0=not) to use the learning individuals (if available) to update the allele frequencies. See also POPDATA and POPFLAG.

TABLE D. MALDsoft output files by extension.

Extension	File contents
`ests`	It is used internally by MALDsoft and contains no useful information for the end user. It is generated by default and contains the model parameters estimated by STRUCTURE2.
`test1`	Genome-wide case-only tests.
`test2`	Genome-wide case-control tests. Generated if `WHICHTEST=2`
`extrs1`	Extreme values for the case-only test. Generated if `EXTREMES>0`
`extrs2`	Extreme values for the case-control test. Generated if `WHICHTEST=2 and EXTREMES>0`
`acas`	Genome-wide average ancestry proportion in cases.
`actr`	Genome-wide average ancestry proportion in controls. Generated if `WHICHTEST=2`