MASTOR Software Documentation

Version 0.7
Oct 16, 2015

Johanna Jakobsdottir[1] and Mary Sara McPeek[2]

[1]The Icelandic Heart Association, Kopavogur, Iceland
johannaj@hjarta.is, jjakobsdottir@gmail.com

[2]Departments of Statistics and Human Genetics
The University of Chicago, Chicago, IL, USA
mcpeek@galton.uchicago.edu

Support

1. Theoretical and software development: The National Institutes of Health grant R01HG001645 to MSM.

2. Software development: From Nov 1 2011, The Icelandic Heart Association.

3. Software development: From Jan 1, 2013, The Icelandic Research Fund grant 130726-051 to JJ.

MASTOR (Mixed-model Association Score Test On Related individuals)
A C program for association testing of quantitiative traits in related individuals.

Homepage: http://galton.uchicago.edu/∼mcpeek/software/MASTOR
Release 0.3 March 26, 2013
Release 0.7 Oct 16, 2015

==========

==========

This program includes code provided by others under licenses compatible with GNU GPL as free software:

CLAPACK, which is under the 3-clause BSD license.
Numerical recipes utility functions, which are public domain.
GNU Scientific Library, which is under GNU GPL version 3 (or later).
Hash table library by Attractive Cahos, which is under an Open Source MIT License.

==========

We request that use of this software be cited in publications as follows:
Jakobsdottir, J. and McPeek, M.S. 2013. MASTOR: Mixed-Model Association Mapping of Quantitative Traits in Samples with Related Individuals. American Journal of Human Genetics. 92(5):625-666.

==========

Contents

1 Overview of MASTOR

MASTOR is a C program that performs a single-SNP association testing in sample with related individuals with known pedigree structures. It is applicable to completely general combinations of related and unrelated individuals, provided the family structures are known. Analysis can be performed genome-wide, though currently only for autosomes and only for binary markers (i.e. SNPs). Extensions to multi-allelic markers are described in the paper [1].

The MASTOR testing method is an extension of the M_{QLS} testing method of Thornton and McPeek from binary to quantitative traits [2]. In addition to individuals who have complete data on phenotype, covariates and genotype at the variant of interest, MASTOR will also include in the analysis (1) individuals who are genotyped at the variant but are missing phenotype and covariate information, and (2) individuals who have phenotype and covariate information, but have missing genotype at the variant of interest, provided that they have at least one relative who is genotyped at the variant of interest.

For each SNP, the MASTOR program can compute test statistics and p-values for 3 different tests of association: 1) MASTOR which is an extension of the M_{QLS} test, 2) ASTOR, which is a simpler and computationally more tractable version of MASTOR, and 3) GTAM, which is a method closely related to the MASTOR method, but where the analysis are prospective instead of retrospective and only individuals with complete information can contribute to the analysis. For MASTOR and ASTOR, the p-value is calculated based on an asymptotic χ^2_1 null distribution, while GTAM is a t-test and the p-value is calculated based on an asymptotic t_{n-k-2} null distribution (where $n = \#$ of individuals, $k = \#$ of covariates). For more information on the 3 test statistics and recommendations for their use, see Chapter 8.

The MASTOR program also allows the option of fitting a linear mixed model to the data without performing a genome scan. This serves two purposes: First, it is useful for preliminary analyses of the phenotype and covariate data in order to formulate the null model. Second, after genetic variants of interest have been identified by an initial analysis with MASTOR, the mixedMLE option can be applied, with one or more variants included as covariates, to estimate the parameters of the alternative model including effect size(s) of variants or even interactions among them.

2 Installing MASTOR

Installation instructions:

1. Download the MASTOR package. This package contains the documentation, source code, pre-compiled binaries, example files, and the GNU GPL license.

2. Read the entire documentation (this document) carefully to understand the purpose of this program and how it works. It will also be helpful to read the paper on the MASTOR statistic [1].

3. Decompress the archive: `tar xvfz MASTOR_v0.7.tar.gz`

- The compression program `gzip` is GNU software and can be obtained from: http://www.gnu.org/software/gzip/, in case it is not available on your system.

4. Switch to the newly created directory: `cd mastor/`
5. This directory contains the GNU GPL license in file `gpl.txt` and four sub-directories:
   - `src` contains the source code.
   - `doc` contains this document `MASTOR_v0.7_doc.pdf`.
   - `examples` contains example input and output files.
   - `executables` contains pre-compiled binaries for Mac and Unix, see the `README.txt` file.
6. If none of the pre-compiled binaries are suitable for your use, switch to the `src` directory:
   - `cd src/`
7. Type `make`. This will build an executable program called `mastor`
   - It is assume that a GNU gcc compiler is available on your system, in case it is not it can be obtained from: http://gcc.gnu.org/.

3 Running MASTOR

1. To run the executable program (see Chapter 2), first, prepare the input files (see Chapter 4). Then, to run MASTOR, using the default filenames simply type on the command line type:

   ```
   ./mastor
   ```

   Alternatively, to change the input filenames or use other options, use flags in the command line and type:

   ```
   ./mastor -p ped.txt -g geno.txt -k kin.txt
   ```

   or:

   ```
   ./mastor --pedfile ped.txt --genofile geno.txt --kinfile kin.txt
   ```

where:

-p ped.txt or **--pedfile ped.txt** Allows the user to specify the name for the pedigree and phenotype data file. The filename defaults to `ped.txt`. To specify another filename, replace `ped.txt` with appropriate filename.

-g geno.txt or **--genofile geno.txt** Allows the user to specify the name for the genotype data file with the individual genotypes. The filename defaults to `geno.txt`. To specify another filename, replace `geno.txt` with appropriate filename.

-k kin.txt or **--kinfile kin.txt** Allows the user to specify the name for the kinship information file. The filename defaults to `kin.txt`. To specify another filename, replace `kin.txt` with appropriate filename.

Additional flags

-a gtam or **--analysis gtam** This flag is optional and tells the program to only perform GTAM analysis instead of MASTOR analysis. If omitting this options the program will by default perform the MASTOR analysis only.

-m or **--mixedMLE** This flag instructs the program to only estimate the MLEs under the null hypothesis and not perform any association analysis. Depending on the `-a / --analysis` flag either null MLEs are estimated from the MASTOR or GTAM sample.

--like This flag instructs the program to calculate, and write to a file, range of values of the $-\log(\mathcal{L}_p)$. The flags below are required if this flag is turned on.

--from 0 This tells the program to calculate the $-\log(\mathcal{L}_p)$ for α values starting at $\alpha = 0$. To specify another value to start at simply replace `0` by the desired value.

--to 1 This tells the program to calculate the $-\log(\mathcal{L}_p)$ for α values ending at $\alpha = 1$. To specify another value to end at simply replace `1` by the desired value.

--by 0.01 This tells the program to calculate the negative log profile likelihood for α at every 0.01 between the values given by the `--from` and `--to` flags. To specify another increment value, simply replace `0.01` by the desired value.

--prefix PREFIX is the prefix string to be added to the default output filenames. If skipped the default file names for the MASTOR analysis are `MASTOR.txt` and `MASTOR_MLEs.txt` o.w. if the prefix is specified as PREFIX the filenames are `PREFIX_MASTOR.txt` and `PREFIX_MASTOR_MLEs.txt` Similar rule applies for GTAM.

--subset S Allows the user to specify which subset of individuals to use in the calculation of variance components: S=0 uses everyone phenotyped, S=1 uses those phenotyped and genotyped (default in GTAM), S=2 uses those phenotyped and genotyped as well as those phenotype with missing genotype but with a genotyped relative, S=3 uses everyone with S=1 and S=2 as well as those phenotyped with missing genotype but with relative who is phenotyped and genotyped or phenotyped with a genotyped relative (default in MASTOR).

See the **Example** chapter 6 for further ways to run the program.

4 Input

1. **-p pedigree and phenotype file** The pedigree file contains the pedigree structure (in linkage format), value of the phenotype and any covariates. The requirements are following:

   - Tab or space delimited.
   - The individual ID is assumed to be *unique across* all pedigrees. Numeric and alphanumeric IDs are allowed.
   - Individuals in the same family should be *grouped together* in the file (order within pedigrees does not matter). The current version of the program will give an error if this is not the case.
   - Single individuals who are unrelated to everyone else in the data need to be in the pedigree file with unique pedigree ID and zero for their father and mother IDs.
   - All father and mother IDs must be present in the file. If they are founders, they have zero for their father and mother IDs. The current version of the program will give error if mother or father IDs are missing.
   - The missing value code for phenotype and covariates is -9.0. This can be changed in the file mastor.h by changing the value in the line `#define MISSVAL -9.0` and recompile the program (first type `make clean` and then `make`)
   - The columns in the file are:

     ```
     pedigree ID
     individual ID
     father ID
     ```

```
mother ID
sex
phenotype
covariate1
covariate2
...
```

- A file with 2 pedigrees with 4 individuals each, one phenotype, and 2 covariates might look like:

```
3 1 0 0 1 1.2 1.3 1.4
3 2 0 0 2 2.2 2.3 2.4
3 3 1 2 2 3.3 3.4 -9.0
3 IND4 1 2 2 4.4 4.4 5.5
2 5 0 0 1 1.2 1.3 1.4
2 6 0 0 2 2.2 2.3 2.4
2 7 5 6 2 3.3 3.4 -9.0
2 IND2 5 6 2 4.4 4.4 5.5
```

2. **-g genotype file** The genotype data file contains the genotype data in a row-format where the columns refer to the individuals and rows refer to the marker. The requirements are following:

   - Space-delimated
   - The first row is a header line with the unique individual IDs. The first four columns are annotation information about each marker: Chr, SNP name, centiMorgan, base pair
     - First column has the chromosome number of each marker. Will be printed out with the results
     - Second column has the marker names. Alpha-numeric names, such as rs numbers, are allowed. Will be printed out with the results.
     - Third column is location such as centiMorgan location. Will not be printed out with the results
     - Fourth column is location such as base pair location. Will be printed out with the results
     - Columns 5 to 5+'number of individuals' have the unique individual IDs
     - The order of individuals does not matter.
   - Genotypes are stored with a space, e.g `1 1 1 2` for genotypes of 2 individuals. This means that effectively there are more columns in the rows with the genotype data than in the header row. The rows with the genotype data have exactly the format that can be generated from `plink` with `--transpose --recode12` commands. Thus the MASTOR format can be easily generated from `plink` by adding a header line.

- Every person with genotype data (that is every person in this file) has to be present in the pedigree file. The current version of the program will result an error if this is not the case.
- Alleles are labeled with '1' and '2' and missing alleles are '0'. Currently, the program does not support other allele labels. Currently the program will NOT check if other labels are present. If there are non-integer values the program will likely result in 'segmentation fault' but if the values are integers the program will likely run but will not produce correct results.
- It is assumed that no one is half-typed. Currently the program does NOT check for half-typing.
- Currently the program does NOT handle markers with more than 2 alleles.
- First 3 rows for a study with 3 genotyped individuals and two markers might look like:

```
Chr Marker cm bp IND41 3 IND25
1 1a 1 1 1 1 1
2 5e 5 5 2 1 1
```

- When no association test is performed, that is when the `--mixedMLE` or `--like` options, then the program only uses the header line of this file and thus it can be used even when no genotype data are available. In that case this file is simply a list file of individuals to include in the analysis. In the MASTOR analysis the individuals included will be everyone who is phenotyped and either present in this file or related to an individual present in the file. In the GTAM analysis the individuals included will be everyone who is phenotyped and present in this file. Thus to include all phenotyped individuals in the analysis simply include everyone in this file (in that case the MASTOR and GTAM analysis should give identical results).

3. **-k kinship coefficient file** The kinship coefficent file contains the pair-wise kinship coefficients between individuals. The requirements are following:

   - The columns in the file should be

     `ped ind1 ind2 coef`

     If ind1= ind 2, the coef is the inbreeding coefficient for the individual and if ind1 $\neq$ ind2 then the coef is the kinship coefficient between the individuals.
   - The program KinInbCoef.c, which can be found at:
     `http://galton.uchicago.edu/~mcpeek/software/KinInbcoef/index.html`
     will output the format required for MASTOR, however note that it does not handle alpha-numeric IDs.
   - It is assumed that everyone in the pedigree file is also in the kinship file and everyone in the kinship file is also in the pedigree file. The program will only check whether everyone in the kinship file is also in the pedigree file and not vice versa.

5 Output

Program will output up to six files, two association results files, two parameter estimate files, and two files with value of negative log-profile likelihood: $-\log(\mathcal{L}_p)$. One association results file has 6 columns for the MASTOR results (called `MASTOR.txt`) and the other has 4 columns for the GTAM results (called `GTAM.txt`). Both association results files contain a header line and then one line with the results for each marker.

- For MASTOR the file with analysis of two markers might look like:

```
Chr Marker bp MASTOR MASTOR-P ASTOR ASTOR-P Freq
1 rs2 2 0.578427 0.44693 1.572955 0.209778 0.215385
1 rs3 3 4.275925 0.0386559 4.052727 0.0441 0.301923
```

  where 1) `MASTOR` is the MASTOR statistic with the variance components estimated from the phenotyped individuals who contribute to the analysis and `MASTOR-P` is the associated p-value from a χ^2_1 distribution, 2) `ASTOR` is the ASTOR statistic where the identity matrix used in place of $\widehat{\Omega}$ and `ASTOR-P` is the corresponding p-value, and 3) `Freq` is the best linear unbiased estimator of the allele frequency.

- For GTAM the file with analysis of two markers might look like:

```
Chr Marker bp GTAM GTAM-P n
1 rs2 2 0.675750 0.4994 780
1 rs3 3 2.177901 0.0297131 780
```

  where `GTAM` is the GTAM statistic with the variance components estimated from the phenotyped and genotyped individuals, `GTAM-P` is the associated p-value from a t_{n-k-2} distribution (n=# individuals, k=# covariates), and `n` is the number of individuals used in the GTAM analysis.

- Output files with variance components and parameter estimates are produced (named `MASTOR_MLEs.txt` and `GTAM_MLEs.txt`). These files have 4 columns and might look like:

```
Parameter       nullMLE     SE_nullMLE   nullMLE_OLS
Heritability    0.512986    0.003957     NA
Additive_Var    6.626302    1.076178     NA
Error_Var       6.290817    0.526706     NA
Intercept       6.442246    0.169722     6.395470
Covariate_1     0.692167    0.056009     0.725662
```

where `nullMLE` is the MLE from the numerical maximization of the log profile likelihood from the generalized regression of the null model, `SE_nullMLE` is the standard error corresponding to the `nullMLE` estimator and `nullMLE_OLS` is the MLE from ordinary least square analysis of the null model.

Note, that:

- if the additive variance is estimated at the boundary of the parameter space (i.e. ≤ 0) it is set equal to zero and the error variance and fixed effects of covariates will be estimated using OLS.

- the OLS values of the fixed effects of the covariates are used in the calculations of the M_I statistic.

- `Covariate_1` is the first covariate in the user's input file (column 7 of the pedigree and phenotype data file).

- Output files with the negative log-profile likelihood values, $-\log(\mathcal{L}_p)$, are produced if the `-like` flag is turned on and called `like_plot_MASTOR.txt` and `like_plot_GTAM.txt`. Both files have two columns: `Alpha` and `negLogLike` and might look like:

```
Alpha negLogLike
0.000000 2437.331358
0.010000 2432.655073
0.020000 2428.419064
0.030000 2424.571835
0.040000 2421.069538
0.050000 2417.874606
0.060000 2414.954667
0.070000 2412.281671
0.080000 2409.831198
0.090000 2407.581885
0.100000 2405.514964
```

6 Examples

Example input and output files are provided in the directory `mastor/examples`. The input files are: `ped_ex.txt`, `geno_ex.txt`, and `kin_ex.txt`. To run the program on these files type on the command line one of several commands:

1. `./mastor -p ped_ex.txt -g geno_ex.txt -k kin_ex.txt`

This command generates the files `MASTOR.txt` and `MASTOR_MLEs.txt` which can be compared to the supplied output files `EX_*.txt`.

Equivalent command is:

```
./mastor --pedfile ped_ex.txt -g geno_ex.txt -k kin_ex.txt
```

2. `./mastor -p ped_ex.txt -g geno_ex.txt -k kin_ex.txt -a gtam`

   This command generates the files `GTAM.txt` and `GTAM_MLEs.txt` which can be compared to the supplied output files `EX_*.txt`.

3. `./mastor -p ped_ex.txt -g geno_ex.txt -k kin_ex.txt --mixedMLE`

   This command generates the file `MASTOR_MLEs.txt`

4. `./mastor -p ped_ex.txt -g geno_ex.txt -k kin_ex.txt -mixedMLE -a gtam`

   This command generates the file `GTAM_MLEs.txt`

5. ```
   ./mastor -p ped_ex.txt -g geno_ex.txt -k kin_ex.txt --like
            --from -0.5 --to 5 --by 0.05
   ```

   This command generates files `MASTOR_MLEs.txt` and `like_plot_MASTOR.txt`

6. ```
   ./mastor -p ped_ex.txt -g geno_ex.txt -k kin_ex.txt
            --like --from -0.5 --to 5 --by 0.05 -analysis gtam
   ```

   This command generates files `GTAM_MLEs.txt` and `like_plot_GTAM.txt`

7 Tips

- Before running the MASTOR software it is recommended to use the `--mixedMLE` option to evaluate the null model and use the results to help choose covariates as well as possible transformations of the phenotypes and/or covariates. Note, however, that this option does not replace a thorough quality and integrity checks of the phenotype data necessary prior to any statistical analysis, for an example the user should check the distribution of the phenotype with respect to outliers and normality.

- If you want to output values of the log-likelihood for plotting purposes (i.e. use the `--like` flag) then supplying the program with the range (i.e. flags `--from` and `--to`) and increment (i.e. flag `--by`) is required because the program will not use any default values.

  To choose these values we recommend first running the program with the `--mixedMLE` but without the `--like` flag then check the `MASTOR_MLEs.txt` and `GTAM_MLEs.txt` files for the MLE of α (which equals `Additive_Vari/Error_Var`) and then choose a range of values that includes both the MASTOR and GTAM α values.

- If there are no relative pairs with both members of the pairs phenotyped available then program will either ignore the `--like` flag even if set or in the case of `MASTOR` analysis try to include more phenotype members of the family to estimate the null MLEs (if that does not work then the flag is completely ignored).

  The reason for this is that when there are no phenotype relative paris available the additive genetic variance cannot be estimated from the data. However, the program will still run and simply output ordinary least square estimates for the error variance and other parameters.

- If the input files are prepared on Windows, they may have dos format line endings and may not work with MASTOR (this has not been tested yet, they may work or they may not). To be safe the user should be sure to change the format of the input files to unix. This may be done on the command line, for example using vi. Type the following four comments on the command line:

  `vi filename` (then you are in your file)

  `:set ff=unix` (hit enter)
  `:w` (hit enter to save your file in the unix format)
  `:q` (hit enter to exit)

- We use a C routine adapted from GNU GSL to calculate the P-values. It is also the basis for the P-value calculations in R. However, as a sanity check you may also double check the P-values using R directly:

  For example, if for MASTOR the test statistic value is 44.56 then the command to get the p-value in R is `pchisq(44.56,1,lowerF)`= which yields the p-value `[1] 2.466805e-11`.

  For GTAM either t-distribution (with the correct degrees of freedom, which can be calculated from the number of individuals as $n-k-2$, where n is the number of individuals, k is the number of covariates) or χ^2_1 distribution can be used. If the GTAM test statistic is -2.01 (with 1026 df) then in R:

  `pt(abs(-2.01),1026,lower=F)*2` yields p-value `[1] 0.04469257`

  `pchisq((-2.01)\^2,1,lower=F)` yields p-value `[1] 0.04443119`

- The easiest way to prepare the input files for KinInbcoef is to use `awk` to pull out the right colums from the pedigree file (assuming the families are numbered consecutively and the pedigree file is in the format required by MASTOR). Two input files are needed for KinInbcoef, the pedfile and the listfile, which are possible to get by typing in the command line:

```
awk '{print($1,$2,$3,$4)}' pedigreefile > pedfile
awk '{print($1,$2)}' pedigreefile > listfile
```

- If the genotype data file is in either linkage format or contains other allele labels than '1' and '2' it may be best to use `plink` (which can be found at `http://pngu.mgh.harvard.edu/~purcell/plink/`) to reformat the data, use `--transpose` and `--recode12` options, respectively.

- When running the program, it will print out few counts before starting analysis. The user should make sure those counts are correct. If they are not that might indicate problems with the format of the input files, which the program does not yet recognize (please let us know if you discover such bugs so that we may add that to our format error checking routines). The counts that are printed out are:

  1. `n_fam`, which is the number of families
  2. `n_marker`, which is the total number of markers in the genotype data file (should be the number of lines -1 in the genotype data file)
  3. `n_total`, which is the total number of individuals in the pedigree file (should be the number of lines in the pedigree file)
  4. `n_typed`, which is the total number of individuals in the genotype data file (should be the number of columns - 1 in the genotype data file)
  5. `n_cov`, which is the total number of covariates (including intercept, should be the number of columns - 5 in the pedigree file).

- Four global fixed parameters are defined in the `mastor.h` file. The defaults are (but all can be changed):

  ```
  #define MAXLEN 1024
  #define MISSVAL -9.0
  #define NUMTOL 1e-6
  #define ANALYSIS_STATUS 50000
  ```

  where:

  - `MAXLEN` is the maximum length for character vectors when reading in the data (this includes maximum length of the file names and family and individual IDs).
  - `MISSVAL` is the missing value code for trait and covariates.
  - `NUMTOL` is the tolerance set when comparing float and double numbers as well as for the converging of the maximization of the log-likelihood when the variance components are estimated.

- `ANALYSIS_STATUS` is control variable to tell the program how many status updates it should print, by default it prints out a message when it has analyzed every 50,000th marker.

If you do need to change the values. Please do so and then recompile the program. It is advisable to recompile it by first cleaning up all `.o` files, do that by typing `make clean` first and then type `make`.

- **MZ twins** The current version of the program should handle both members of a MZ twin pair to be genotyped, however whether this works has not been tested. If slight changes are made to the input of the MZ twin data, MZ twins can still be analyzed correctly without depending on the program correctly handling the MZ pairs:

  1. If twin 1 is genotyped and phenotyped and twin 2:
     (a) genotyped and phenotyped, set phenotype of twin 1 to be the sum of the phenotype values of the twins and use data from twin 1 and delete the data from twin 2.
     (b) genotyped but not phenotyped, use the data from twin 1 and delete twin 2.
     (c) phenotyped but not genotyped, set the phenotype of twin 1 to be the sum of the phenotype values of the twins and use data from twin 1 and delete twin 2.
  2. If twin 1 is genotyped and not phenotyped and twin 2:
     (a) genotyped but not phenotyped, use the data from twin 1 and delete twin 2.
     (b) phenotyped but not genotyped, set the phenotype of twin 1 to the phenotype of twin 2 and use data from twin 1 and delete twin 2.
  3. If twin 1 is phenotyped and not genotyped and twin 2:
     (a) phenotyped but not genotyped, set the phenotype of twin 1 to the sum of the phenotypes of twin 1 and 2 and use data from twin 1 and delete twin 2.

- **Segmentation faults (and other bugs)** Since there may still be bugs related to format checking the input, there are opportunities for the program to result in a segmentation fault without any message about what may be causing the error. Below is an incomplete list of things that could be wrong (that is still not checked for but the software):

  1. Parents in the wrong family.
  2. Counts are wrong (see previous point above).
  3. Other allele labels than '0', '1', and '2'.

8 The statistics in the MASTOR program

We briefly describe the statistics in the MASTOR program. More details on the derivations are in the paper and technical report, available on the programs website. Let $\mathbf{X} = (X_1, \ldots, X_n)^T$ denote the n-length vector of genotype data for individuals in the set N. Let $\mathbf{Y} = (Y_1, \ldots, Y_r)^T$ denote the r-length vector of phenotype data, where Y_i is the quantitative trait value for the ith individual in the set R, where R denotes the set of individuals in the sample who satisfy the following three criteria:

1. their phenotype value is not missing
2. none of their covariates are missing
3. they have either non-missing genotype at the variant or they have a relative in the study with non-missing genotype at the variant of interest

Let $\mathbf{W}$ be the $r \times (c+1)$ matrix of covariates with (i,j)th entry W_{ij} equal to the value of the jth covariate for the ith individual in set R. We assume that $\mathbf{W}$ always includes an intercept (i.e. column of 1's). Thus, the number of rows of $\mathbf{W}$ is n, and the number of columns is 1 more than the number of covariates to be included in the analysis.

8.1 The GTAM statistic

In the GTAM analysis individuals in the set $S = N \cap R$ are included. We assume the trait model

$$\mathbf{Y}_S = \mathbf{W}_S\boldsymbol{\beta} + \mathbf{X}_S\gamma + \boldsymbol{\epsilon}_S \tag{1}$$

where we let the subscript S represent that the vectors and matrix $\mathbf{Y}_S$, $\mathbf{X}_S$, and $\mathbf{W}_S$ are sub-vectors and sub-matrix of $\mathbf{Y}$, $\mathbf{X}$, and $\mathbf{W}$, $\boldsymbol{\beta}$ is a $(c+1)$-length vector of fixed effects of the covariates, including intercept and possible major genes, γ is the (scalar) association parameter, measuring the effect of the genotype of interest on phenotype, and $\boldsymbol{\epsilon}_S \sim N_n(\mathbf{0}, \boldsymbol{\Omega}_S)$. We assume a typical model for $\boldsymbol{\Omega}_S$ namely

$$\boldsymbol{\Omega}_S \equiv \sigma_a^2\boldsymbol{\Phi}_{SS} + \sigma_e^2\mathbf{I}_s \tag{2}$$

where σ_a^2 and σ_e^2 are additive and environmental variances, $\mathbf{I}_S$ is the $s \times s$ identity matrix, and $\boldsymbol{\Phi}_{SS}$ is the $s \times s$ kinship matrix given by

$$\boldsymbol{\Phi} = \begin{pmatrix} 1+h_1 & 2\phi_{12} & \ldots & 2\phi_{1s} \\ 2\phi_{12} & 1+h_2 & \ldots & 2\phi_{2s} \\ \vdots & \ldots & \ldots & \vdots \\ 2\phi_{s1} & 2\phi_{s2} & \ldots & 1+h_s \end{pmatrix}, \tag{3}$$

where ϕ_{ij} is the kinship coefficient between the ith and jth individuals in S, and h_i is the inbreeding coefficient of the ith individual in S.

To calculate the GTAM statistic to test the null hypothesis of no association, we first estimate the vector of variance components, $\boldsymbol{\psi} = (\sigma_a^2, \sigma_e^2)$, by its null maximum likelihood estimate (MLE), $\widehat{\boldsymbol{\psi}}_{0S}$, and we only compute it once per genome screen using the individuals who are genotyped and phenotyped, where those individuals in the genotype data file are assumed to be genotyped.

For each marker, the calculation of the GTAM statistic can be expressed in terms of a generalized regression based on equation (1), where we take $\boldsymbol{\epsilon}_S \sim N_s(\mathbf{0}, \xi^2\widehat{\boldsymbol{\Omega}}_S)$, with $\widehat{\boldsymbol{\Omega}}_S \equiv \boldsymbol{\Omega}_S|_{\boldsymbol{\psi}=\widehat{\boldsymbol{\psi}}_{0S}}$ treated as fixed and known and $\boldsymbol{\beta}$, γ and ξ^2 as unknown. Then the parameter estimates, $\widehat{\boldsymbol{\beta}}$, $\widehat{\gamma}$ and $\widehat{\xi^2}$, can be obtained by generalized regression under this model, and the GTAM statistic is equal to the generalized-regression t-statistic.

8.2 The MASTOR statistic

In contrast to GTAM, MASTOR uses larger set of individuals $N \cup R$, rather then $S = N \cap R$. MASTOR is a quasi-likelihood score test based on a retrospective mean model. The mean model for genotyped individual i is expressed as

$$E(X_i|\mathbf{W}, \mathbf{Y}) = p + \delta \sum_{j \in R} 2\phi_{ij}[\widehat{\boldsymbol{\Omega}}^{-1}(\mathbf{Y} - \mathbf{W}\hat{\boldsymbol{\beta}})]_j, \tag{4}$$

From this mean model the statistic can be derived to have a general form

$$\text{MASTOR} = \frac{(\mathbf{V}^T\mathbf{X})^2}{\widehat{\text{Var}}_0(\mathbf{V}^T\mathbf{X}|\mathbf{W}, \mathbf{Y})} = \frac{(\mathbf{V}^T\mathbf{X})^2}{\hat{\sigma}_X^2\mathbf{V}^T\boldsymbol{\Phi}_{NN}\mathbf{V}}, \tag{5}$$

where we assume:

- $E_0(\mathbf{V}^T\mathbf{X}|\mathbf{W}, \mathbf{Y}) = 0$
- $\text{Var}_0(\mathbf{X}|\mathbf{W}, \mathbf{Y}) = \sigma_X^2\boldsymbol{\Phi}_{NN}$
- $\boldsymbol{\Phi}_{NN}$ is the $n \times n$ kinship matrix for the individuals in N (similar to equation (3))
- σ_X^2 is an unknown scalar

In practice we estimate σ_X^2 is by $\hat{\sigma}_X^2 = \mathbf{X}^T\mathbf{U}\mathbf{X}(n-1)^{-1}$, where $\mathbf{U} = \boldsymbol{\Phi}_{NN}^{-1} - \boldsymbol{\Phi}_{NN}^{-1}\mathbf{1}(\mathbf{1}^T\boldsymbol{\Phi}_{NN}^{-1}\mathbf{1})^{-1}\mathbf{1}^T\boldsymbol{\Phi}_{NN}^{-1}$ and then get $\widehat{\text{Var}}_0(\mathbf{V}^T\mathbf{X}|\mathbf{W}, \mathbf{Y}) = \hat{\sigma}_X^2\mathbf{V}^T\boldsymbol{\Phi}_{NN}\mathbf{V}$ where

$$\mathbf{V} = \mathbf{U}\boldsymbol{\Phi}_{NR}\widehat{\boldsymbol{\Omega}}^{-1}(\mathbf{Y} - \mathbf{W}\hat{\boldsymbol{\beta}}) \tag{6}$$

where $\boldsymbol{\Phi}_{NR}$ is the $n \times r$ cross-kinship matrix having (i, j)th entry equal to twice the kinship coefficient between the ith individual in set N and the jth individualin set R and $\widehat{\boldsymbol{\Omega}}$ and $\hat{\boldsymbol{\beta}}$ are the MLEs in the model

$$\mathbf{Y} = \mathbf{W}\boldsymbol{\beta} + \boldsymbol{\epsilon} \tag{7}$$

Note that equation (7) is similar to equation (1) expect that it includes the larger set of individuals in R and is a null model.

8.3 The ASTOR statistic

The calculation of the MASTOR statistic involves estimation of variance components. Although, this can be done only once per genome screen, at least initially, quik and dirty alternative is assume that the additive variance component is zero so that $\mathbf{\Omega} = \sigma_e^2 \mathbf{I}_r$ and eliminating the variance component estimation step. The statistic calculated this way is called ASTOR.

9 Estimation of null MLEs: Theoretical and technical details

9.1 Model

Let $\mathbf{Y} = (Y_1, \ldots, Y_n)^T$ denote the n-length vector of phenotype data, where Y_i is the quantitative trait value for the ith individual, and let $\mathbf{W}$ be the $n \times (c+1)$ matrix of covariates with (i,j)th entry W_{ij} equal to the value of the jth covariate for the ith individual in set R. We assume that $\mathbf{W}$ always includes an intercept (i.e. column of 1's). Thus, the number of rows of $\mathbf{W}$ is n, and the number of columns is 1 more than the number of covariates to be included in the analysis.

We assume the trait model

$$\mathbf{Y} = \mathbf{W}\boldsymbol{\beta} + \boldsymbol{\epsilon} \tag{8}$$

where $\boldsymbol{\beta}$ is a $(c+1)$-length vector of fixed effects of the covariates, including intercept and possible major genes, and $\boldsymbol{\epsilon} \sim N_n(\mathbf{0}, \mathbf{\Omega})$. We assume a typical model for $\mathbf{\Omega}$ namely

$$\mathbf{\Omega} \equiv \sigma_a^2 \mathbf{\Phi} + \sigma_e^2 \mathbf{I} \tag{9}$$

where σ_a^2 and σ_e^2 are additive and environmental variances, $\mathbf{I}$ is the $n \times n$ identity matrix, and $\mathbf{\Phi}$ is the $n \times n$ kinship matrix given by

$$\mathbf{\Phi} = \begin{pmatrix} 1+h_1 & 2\phi_{12} & \ldots & 2\phi_{1n} \\ 2\phi_{12} & 1+h_2 & \ldots & 2\phi_{2n} \\ \vdots & \ldots & \ldots & \vdots \\ 2\phi_{n1} & 2\phi_{n2} & \ldots & 1+h_n \end{pmatrix}, \tag{10}$$

where ϕ_{ij} is the kinship coefficient between the ith and jth individuals, and h_i is the inbreeding coefficient of the ith individual.

9.2 Estimating $\boldsymbol{\beta}$, σ_a^2, and σ_e^2

Now, we want to estimate, $\boldsymbol{\beta}$, σ_a^2, and σ_e^2 by maximizing the log-likelihood:

$$\log(\mathcal{L}) = \ell(\boldsymbol{\beta}, \alpha, \sigma_e^2) = -\frac{1}{2}\log(2\pi) - \frac{1}{2}\log|\mathbf{\Omega}| - \frac{1}{2}(\mathbf{Y} - \mathbf{W}\boldsymbol{\beta})^T \mathbf{\Omega}^{-1} (\mathbf{Y} - \mathbf{W}\boldsymbol{\beta})$$

First we show how the the maximization can be done in one dimension and how the computational burden of numerical maximization can minimized.

We rewrite

$$\begin{aligned}\boldsymbol{\Omega} &= \sigma_a^2\boldsymbol{\Phi} + \sigma_e^2\mathbf{I} && (11)\\ &= \sigma_e^2(\alpha\boldsymbol{\Phi} + \mathbf{I}) && (12)\\ &= \sigma_e^2\mathbf{K} && (13)\end{aligned}$$

Because $\boldsymbol{\Phi}$ is positive definite (p.d.) matrix its singular value decomposition (s.v.d) exists, namely there exists an orthogonal matrix $n \times n$ $\mathbf{U}$ and diagonal $n \times n$ matrix $\boldsymbol{\Lambda}$, with the diagonal elements being the eigenvalues $(\lambda_1, \ldots, \lambda_n)$ of $\boldsymbol{\Phi}$, such that $\boldsymbol{\Phi} = \mathbf{U}\boldsymbol{\Lambda}\mathbf{U}^T$. We may write:

$$\begin{aligned}\mathbf{K} &= \alpha\boldsymbol{\Phi} + \mathbf{I} && (14)\\ &= \alpha\mathbf{U}\boldsymbol{\Lambda}\mathbf{U}^T + \mathbf{U}\mathbf{U}^T && (15)\\ &= \mathbf{U}(\alpha\boldsymbol{\Lambda} + \mathbf{I})\mathbf{U}^T && (16)\end{aligned}$$

Since $\mathbf{K}$ is p.d. it can be Cholesky decomposed as $\mathbf{K} = \mathbf{C}^T\mathbf{C}$, where $\mathbf{C}$ is an $n \times n$ upper triangular matrix. Assuming $\mathbf{K}$ is known, regression of $\mathbf{C}^{-1}\mathbf{W}$ on $\mathbf{C}^{-1}\mathbf{Y}$ gives MLEs:

$$\begin{aligned}\widehat{\boldsymbol{\beta}} &= (\mathbf{W}^T\mathbf{K}^{-1}\mathbf{W})^{-1}\mathbf{W}^T\mathbf{K}^{-1}\mathbf{Y} && (17)\\ \widehat{\sigma_e^2} &= \frac{(\mathbf{C}^{-T}\mathbf{Y} - \mathbf{C}^{-T}\mathbf{W}\widehat{\boldsymbol{\beta}})^T(\mathbf{C}^{-T}\mathbf{Y} - \mathbf{C}^{-T}\mathbf{W}\widehat{\boldsymbol{\beta}})}{n} && (18)\\ &= \frac{(\mathbf{Y} - \mathbf{W}\widehat{\boldsymbol{\beta}})^T\mathbf{K}^{-1}(\mathbf{Y} - \mathbf{W}\widehat{\boldsymbol{\beta}})}{n} && (19)\end{aligned}$$

If we plug $\widehat{\boldsymbol{\beta}}$ and $\widehat{\sigma_e^2}$ into $\ell(\boldsymbol{\beta}, \alpha, \sigma_e^2)$ and get the log profile likelihood for α as:

$$\begin{aligned}\log(\mathcal{L}_p) &= \ell(\widehat{\boldsymbol{\beta}}, \alpha, \widehat{\sigma_e^2}) && (20)\\ &= -\frac{1}{2}\log(2\pi) - \frac{1}{2}\log|\widehat{\sigma_e^2}\mathbf{K}| - \frac{1}{2}\frac{(\mathbf{Y} - \mathbf{W}\widehat{\boldsymbol{\beta}})^T\mathbf{K}^{-1}(\mathbf{Y} - \mathbf{W}\widehat{\boldsymbol{\beta}})}{\widehat{\sigma_e^2}} && (21)\\ &= -\frac{1}{2}\log(2\pi) - \frac{1}{2}\log|\widehat{\sigma_e^2}\mathbf{K}| - \frac{n}{2} && (22)\end{aligned}$$

Therefore to maximize $\log(\mathcal{L}_p)$ we minimize $-\log|\widehat{\sigma_e^2}\mathbf{K}| = -n\log\widehat{\sigma_e^2} - \log|\mathbf{K}|$. Now $(\alpha\boldsymbol{\Lambda} + \mathbf{I})$ is a diagonal matrix and thus we get

$$\begin{aligned}-\log(\mathcal{L}_p) &= -n\log\widehat{\sigma_e^2} - \log|\mathbf{K}| && (23)\\ &= -n\log\widehat{\sigma_e^2} - \log|\mathbf{U}(\alpha\Lambda + I)\mathbf{U}^T| && (24)\\ &= -n\log\widehat{\sigma_e^2} - \log\Pi_{i=1}^n(\alpha\lambda_i + 1) && (25)\end{aligned}$$

We note that for the n-length vector $\mathbf{Y}$ and $n \times (c+1)$ matrix $\mathbf{W}$ the MLE of the error variance can be written as:

$$\widehat{\sigma_e^2} = \frac{\mathbf{Y}^T\mathbf{K}^{-1}\mathbf{Y} - \mathbf{Y}^T\mathbf{K}^{-1}\mathbf{W}(\mathbf{W}^T\mathbf{K}^{-1}\mathbf{W})^{-1}\mathbf{W}^T\mathbf{K}^{-1}\mathbf{Y}}{n} \tag{26}$$

where

$$\mathbf{Y}^T\mathbf{K}^{-1}\mathbf{Y} = \sum_{i=1}^{n} \frac{1}{\alpha\lambda_i + 1}\left[(\mathbf{Y}^T\mathbf{U})_i\right]^2 \tag{27}$$

$$\left[\mathbf{Y}^T\mathbf{K}^{-1}\mathbf{W}\right]_{i_{cov}} = \sum_{i=1}^{n} \frac{1}{\alpha\lambda_i + 1}\left[(\mathbf{Y}^T\mathbf{U})_i(\mathbf{W}^T\mathbf{U})_{i_{cov},i}\right] \tag{28}$$

$$\left[\mathbf{W}^T\mathbf{K}^{-1}\mathbf{W}\right]_{i_{cov},j_{cov}} = \sum_{i=1}^{n} \frac{1}{\alpha\lambda_i + 1}\left[(\mathbf{W}^T\mathbf{U})_{i_{cov},i}(\mathbf{W}^T\mathbf{U})_{j_{cov},i}\right] \tag{29}$$

Therefore, $-\log(\mathcal{L}_p)$ can be written in such a way that the minimization is done over α only, that is in one dimension. The vector $\mathbf{Y}^T\mathbf{U}$, matrix $\mathbf{W}^T\mathbf{U}$ and eigenvalues λ's are the same for all α's. Therefore the s.v.d. needs to be performed only once per analysis independent of number of iterations needed to find the MLE of α. These two components that 1) the minimization is done in one dimension and 2) the s.v.d. is done once, greatly reduces the computational burden of estimating the MLEs.

9.2.1 Variances of $\boldsymbol{\beta}$, σ_a^2, and σ_e^2

When the parameters are not estimated at or outside the boundary of the parameter space, then the variance-covariance matrix of the parameter estimates can be derived analytically. We can show that the expected Fisher information matrix is:

$$\mathcal{I} = \begin{bmatrix} \mathbf{W}^T\boldsymbol{\Omega}^{-1}\mathbf{W} & 0 & 0 \\ 0 & \frac{1}{2}\sum(\sigma_a^2\lambda_i + \sigma_e^2)^{-2}\lambda_i^2 & \frac{1}{2}\sum(\sigma_a^2\lambda_i + \sigma_e^2)^{-2}\lambda_i \\ 0 & \frac{1}{2}\sum(\sigma_a^2\lambda_i + \sigma_e^2)^{-2}\lambda_i & \frac{1}{2}\sum(\sigma_a^2\lambda_i + \sigma_e^2)^{-2} \end{bmatrix}$$

The inverse of this matrix is the variance-covariance matrix:

$$\mathcal{S} = \mathcal{I}^{-1} = \begin{bmatrix} (\mathbf{W}^T\boldsymbol{\Omega}^{-1}\mathbf{W})^{-1} & 0 & 0 \\ 0 & \mathrm{Var}(\sigma_a^2) & \mathrm{Cov}(\sigma_a^2, \sigma_e^2) \\ 0 & \mathrm{Cov}(\sigma_a^2, \sigma_e^2) & \mathrm{Var}(\sigma_e^2) \end{bmatrix}$$

9.2.2 Heritability

We estimate the (narrow sense) heritability, h^2 as:

$$\widehat{h^2} = \frac{\widehat{\sigma_a^2}}{\widehat{\sigma_a^2} + \widehat{\sigma_e^2}}$$

The variance of the heritability can be derived using the multivariate delta method and is:

$$\mathrm{Var}(h^2) = \frac{(\sigma_e^2)^2\mathrm{Var}(\sigma_a^2) + (\sigma_a^2)^2\mathrm{Var}(\sigma_e^2) - 2\sigma_a^2\sigma_e^2\mathrm{Cov}(\sigma_a^2, \sigma_e^2)}{(\sigma_a^2 + \sigma_e^2)^4} \tag{30}$$

9.2.3 Numerical minimization

We want to numerically minimize the negative log profile likelihood for α: $-\log(\mathcal{L}_p)$. To do that we use the Brent's method. The method requires starting values that bracket the minimum.

While two points can bracket a root of a function, a triplet of points is needed to bracket a minimum. The `gsl_min_find_bracket` function from GSL (GNU Scientific Library) returns such a triplet, which in turn is supplied to routine that implements Brent's method (e.g. `gsl_min_fminimizer_brent` from GSL). Our `brak` and `brent` functions are adaptations of `gsl_min_find_bracket` and `gsl_min_fminimizer_brent` routines, respectively.

Two starting values for α are needed for `brak`. For one starting value, we use the ordinary least square regression for $\boldsymbol{\beta}$ and σ_e^2 (i.e. we essentially ignore the relatedness) and the average over the outer product of adjusted trait values of related individuals: $(\mathbf{Y} - \mathbf{W}\widehat{\boldsymbol{\beta}})_i(\mathbf{Y} - \mathbf{W}\widehat{\boldsymbol{\beta}})_j/\phi_{ij}$ for the polygenic variance, σ_a^2, and then set the starting value to be $\alpha_1 = \max(\sigma_a^2/\sigma_e^2, 10^{-6})$. For the other starting value, α_2, we do the following: if $\alpha_1 < 1$ then $\alpha_2 = 10 + 1000\alpha_1$ else if $\alpha_1 \geq 1$ then $\alpha_2 = 1000\alpha_1$. Although the second starting value, α_2 is chosen somewhat arbitrary, this procedure has worked well.

10 Bug reports and feedback

We appreciate comments and suggestions and if you do encounter a bug in the MASTOR software please send us a message. Please include in your message the program version (printed out when the program is run), platform (windows, mac, linux, etc.), description of your problem, and if possible example files (in a zip folder) that caused the problem.

11 Acknowledgements

- The Matrix cookbook.
- Valgrind, an instrumentation framework for building dynamic analysis tools. Specifically we used Valgrind tools to detect many memory management bugs and to profile our program.
- khash.h a fast and light-weighted hash table library in C.
- Numerical Recipes in C. We use the utility functions in nrutil.h
- GNU Scientific library. We adapted the brent and brak routines into our code. Also our implementations of the T test and χ_1^2 test are based on GNU GSL.
- LAPACK. We used code from CLAPACK for singular value decomposition.
- Mark Abeny for sharing Cholesky routine and other help with the programming.

- Pall Melsted for help implementing the hash library and other help with the programming. C code for parsing long options is from the BFcounter program written by Pall Melsted.

- Sheng Zhong, Timothy Thornton, Albert V. Smith for bug discoveries and fixes.

References

[1] Johanna Jakobsdottir and Mary Sara McPeek. MASTOR: Mixed-model association mapping of quantitative traits in samples with related individuals. *American Journal of Human Genetics*, 92(5):625–666, 2013.

[2] Timothy Thornton and Mary Sara McPeek. Case-control association testing with related individuals: A more powerful quasi-likelihood score test. *American Journal of Human Genetics*, 81(2):321–337, 2007.