MQLS-XM Software Documentation

Version 1.0

Timothy Thornton[1] and Mary Sara McPeek[2,3]

Department of Biostatistics[1]
The University of Washington

Departments of Statistics[2] and Human Genetics[3]
The University of Chicago

MQLS-XM

A C program for case-control association testing on the autosomal chromosomes and the X-chromosome in samples that contain related individuals.

Homepage: http://galton.uchicago.edu/∼mcpeek/software/index.html

Release 1.0 July 30, 2012

==========

License

We request that use of this software be cited in publications as follows:

> Thornton T., McPeek M. S. (2007) "Case-Control Association Testing with Related Individuals: A More Powerful Quasi-Likelihood Score Test" American Journal of Human Genetics, vol 81, pp. 321-337.
>
> Thornton T, Zhang Q, Cai X, Ober C, and McPeek MS (2012) "XM: Association Testing on the X-Chromosome in Case-Control Samples with Related Individuals" Genetic Epidemiology, vol 36, pp. 438-450.

To contact the first author:

Timothy A. Thornton
Department of Biostatistics
University of Washington
Health Sciences Building F-600
Box 357232
Seattle, WA 98195-7232

email: tathornt@u.washington.edu

Contents

1 Overview of MQLS-XM

MQLS-XM is a program, written in C, that performs single-SNP, case-control association testing on the autosomal chromosomes and the X-chromosome in samples with related individuals. The program is applicable to association studies with completely general combinations of related and unrelated individuals, where the relationships among the sampled individuals are assumed to be known. For instance, the program allows cases to be related to controls, and it is equally applicable to complex inbred pedigrees and to simpler study designs consisting of unrelated individuals and small outbred families.

The MQLS-XM program can be considered a significantly enhanced version of the MQLS program of Thornton and McPeek (2007) where some of the new features of the program include:

(1) implementation of the X-chromosome association statistics of Thornton et al. (2012) for single-SNP case-control association testing on the X chromosome;

(2) user option for calculating the association test statistics with a robust variance estimator (the default) or a variance estimator that assumes Hardy-Weinberg Equilibrium (HWE);

(3) allows for males and females to have different prevalence values for a trait;

(4) supports PLINK's (Purcell et al. 2007) transposed PED file as the input SNP genotype data file, to allow for the analysis of millions of SNPs without excessive memory allocation.

Additional features of MQLS that are retained in the new MQLS-XM include:

(1) improved power by taking advantage of the principle that there is an enrichment for predisposing variants in affected individuals with affected relatives;

(2) the ability to incorporate both unaffected controls and controls of unknown phenotype (e.g., general population controls) in the analysis;

(3) appropriate handling of missing genotype data to construct valid tests by taking into account the particular missing genotype pattern at each SNP;

(4) incorporation of phenotype information, for individuals with missing genotype data at a SNP being tested, by BLUP imputation of the missing genotype based on information from genotyped relatives (McPeek 2012).

For each SNP, the program computes three different test statistics for association. For autosomal SNPs, the three test statistics computed are the M_{QLS} of Thornton and McPeek (2007), the W_{QLS} of Bourgain et al. (2003), and the corrected χ^2 of Bourgain et al. (2003), except that the robust variance estimator of Thornton and McPeek (2010) is used for all 3

statistics. (The HWE variances proposed in the original papers can be substituted at the user's option.) Also, the M_{QLS} test has been generalized to allow sex-specific prevalences. See the next section for details on these changes. For X-chromosome SNPs, the X_M, X_χ, and X_W test statistics of Thornton et al. (2012) are computed. For each test, a p-value is calculated based on a χ^2_1 asymptotic null distribution.

In most applications involving autosomal and/or X-chromosome SNP association testing, we recommend using the M_{QLS} and/or X_M tests. In the context of complex trait mapping in samples of related individuals, the M_{QLS} test has been shown to generally have more power than both the W_{QLS} and corrected χ^2 tests for autosomal variants; and similarly, the X_M test has been shown to perform approximately as well as or significantly better than the $X\chi$ and X_W tests for X-chromosome variants. For a more detailed comparison of the autosomal test statistics, see Thornton and McPeek (2007), and for a more detailed comparison of the X-chromosome test statistics, see Thornton et al. (2012).

2 Detailed Descriptions of Some of the New Features

In this section, we give more details on some of the new features in MQLS-XM, specifically, the options for (1) sex-specific prevalences and (2) robust vs. HWE variance calculation.

The MQLS-XM calculates three autosomal test statistics (M_{QLS}, W_{QLS}, and corrected χ^2) and three X-chromosome test statistics (X_M, X_W, and X_χ). Note that the M_{QLS} and X_M are the preferred statistics for most applications.

Extension of M_{QLS} to Include Sex-Specific Prevalences

The X_M test includes sex-specific prevalences (see Thornton et al. 2012 for details), and this is implemented in the MQLS-XM. In addition, the MQLS-XM includes a new extension of the M_{QLS} to also include sex-specific prevalences. This extension is closely related to the work of Thornton et al. (2012), but we spell it out in a little more detail here.

The original M_{QLS} test (Thornton and McPeek 2007) involves a population prevalence estimate, k, that is not sex-specific. To describe the extension to sex-specific prevalences, it is convenient to use the formulation of the M_{QLS} statistic given by equations (2) and (3) on p. 440 of Thornton et al. (2012). This formulation is in terms of a phenotypic residual vector, $\mathbf{R}$, having ith entry $R_i = 0$ if individual i has unknown phenotype and $R_i = 1_{\{i \text{ case}\}} - k$ if i has known phenotype, where $1_{\{i \text{ case}\}}$ is the indicator function for the event that i is affected. The extension to sex-specific prevalences is obtained by replacing $\mathbf{R}$ by a different phenotypic residual vector, $\mathbf{A}$ (given on p. 443 of Thornton et al. 2012), where $A_i = 0$ for i of unknown phenotype, $A_i = 1_{\{i \text{ case}\}} - k_f$ for i female with known phenotype and $A_i = 1_{\{i \text{ case}\}} - k_m$ for i male with known phenotype, where k_f and k_m are estimates of the

population prevalence of the trait in, respectively, females and males, with $0 < k_f, k_m < 1$.

What should I plug in for the sex-specific prevalences used in M_{QLS} and X_M?

To calculate the M_{QLS} and X_M statistics, estimates of the prevalences of the trait among males and among females in a suitable reference population must be specified by the user. These sex-specific prevalences can be replaced by a common, pooled estimate of the population prevalence if sex-specific prevalence information is not available or if there is no evidence of a sex difference in prevalence. We recommend using prevalence estimates from previous studies or registry data from the population, when available. For studies with random ascertainment, the sex-specific prevalences could be estimated by the sample case frequencies in females and in males, or, if the male and female prevalences are assumed to be equal, then the overall sample case frequency could be used. However, when ascertainment is phenotype-based, prevalence estimates should ideally be obtained from external, population-based data, rather than from the sample case frequencies in the data set. We emphasize that the M_{QLS} and X_M tests will be valid regardless of the prevalence values used, but accurate prevalence values would be expected to increase power (see Thornton and McPeek 2007 and Thornton et al. 2012 for details).

Robust vs. HWE variance calculation

The MQLS-XM program allows for all six test statistics to be calculated using either a robust variance estimator or a variance estimator that assumes HWE. The robust variance estimator is the default option. For the autosomal tests, the robust variance estimator is given by equation (3) of Thornton and McPeek (2010) with $\mathbf{\Psi}$ taken to be the known kinship matrix. For the X-chromosome tests, the robust variance estimator is given by equation (7) of Thornton et al. (2012).

Alternatively, the user has the option of specifying that the HWE variance estimator be used. For the autosomal tests, this is the variance estimator used in the original papers (Bourgain et al. 2003, Thornton and McPeek 2007) and is given, for example, by the formula for $\hat{\xi}_1^2$ near the bottom of p. 439, column 2 in Thornton et al. (2012). For the X-chromosome tests, the HWE variance estimator is given by equation (6) of Thornton et al. (2012).

3 Running MQLS-XM

Installation instructions:

1. Download the MQLS-XM package. This package contains documentation, source code, example input and output files, and a precompiled executable for Linux platforms.

2. Read the file MQLS_XM_Documentation.pdf carefully to understand the purpose of this program and how it works.

3. Edit the Makefile as necessary according to the instructions in the Makefile. You should only need to make sure that the correct compiler and compiler options for your machine are chosen.

4. Type "make". This will build an executable program called "MQLS-XM". If the message "make: 'MQLS-XM' is up to date" appears after typing "make", then to build the executable program you must first delete the precompiled binary MQLS-XM program that comes with the software by typing "rm MQLS-XM", and then type "make" to build the executable program MQLS-XM.

5. MQLS-XM is run from the command line via the command 'MQLS-XM' with all information, including the type of anlaysis, specified by command line options. To run the executable program MQLS-XM:

   First, prepare the input files, e.g., genofile, phenofile, kinfile, prevalence (see Section 4 for more details).

   Then, to run MQLS-XM with the default input filenames and settings, one need only type

   ./MQLS-XM

   Alternatively, to change input filenames or settings, use flags in the command line. The following flags are available:

   ./MQLS-XM -g genofile -p phenofile -k kinfile -r prevalence -x -u -m -h

   We briefly summarize the meanings of the flags below. More details can be found in section 4:

   -g genofile Allows the user to specify the name of the SNP genotype data input file. Filename defaults to "genofile" if this flag is not used. To specify a different filename, replace "genofile" with the appropriate filename.

   -p phenofile Allows the user to specify the name of the phenotype information input file. (This file also includes family ID numbers, individual ID numbers, and sex, in addition to phenotype.) The filename defaults to "phenofile".

   -k kinfile Allows the user to specify the name of the kinship coefficient input file, which contains the kinship and inbreeding coefficients for all possible pairs of individuals within each family. Filename defaults to "kinfile".

   -r prevalence Allows the user to specify the name of the prevalence input file, which contains estimates of the prevalence of the binary trait in males and females from a suitable reference population. Filename defaults to "prevalence".

   -x Allows the user to perform an X-chromosome association analysis. An autosomal association analysis will be performed if this option is not used.

   -u Allows the user to exclude individuals with unknown phenotype from the analysis for the three test statistics. All individuals will be included in the analysis if this option is not used.

-m Allows the user to specify that only individuals who have non-missing genotypes at a marker will be included in calculating the M_{QLS} and X_M statistics at that SNP, i.e., phenotype information for individuals with missing genotype data at a SNP will not be used. If this option is not used, the M_{QLS} and X_M statistics will incorporate phenotype information for individuals with missing genotype data at a SNP being tested, provided that those individuals have a sampled relative who is genotyped at the marker.

-h Allows the user to specify that the association test statistics will be calculated using a variance estimator that assumes HWE. The association test statistics will be calculated using a robust variance estimator if this option is not used.

6. You can test the executable program MQLS-XM by running it with the sample input files: genofile, phenofile, pedinfo, and prevalence. You can then compare the resulting output, which will be printed to the files MQLStest.out, MQLStest.top and MQLStest.pvalues, with the correct output provided in the sample output files MQLStest.out.ex, MQLStest.top.ex, and MQLStest.pvalues.ex, respectively.

7. The program stops if any errors are detected in the format of the input files.

4 Input

Required Input Files:

1. **genotype data file**

   The genotype data file is a transposed genotype file containing the SNP names and locations and the genotypes of the sampled individuals. The genotype data file is in the PLINK tped file format, with some additional restrictions, namely:

   1. genotypes for individuals from the same family must be listed consecutively;
   2. the order of individuals must be the same in the genotype data file and in the phenotype information file described in the next subsection.
   3. If an autosomal analysis is to be conducted, then the genotype data file should contain genotypes only for autosomal SNPs. Similarly, if an X-chromosome analysis is to be conducted, then the file should contain genotypes only for X-chromsome SNPs. In other words, autosomal and X-chromosome SNPs should not be in the same genotype data file.
   4. The two alleles of a SNP must be coded as 1 and 2, and missing alleles must be coded as 0.
   5. For an X-chromosome analysis, males should be coded to have homozygous genotypes at the X-linked SNPs. For example, a male with the 1 allele at an X-linked SNP should have genotype "1 1" in this file.

To illustrate the format of the genotype data file, consider a study sample with a total of 8 individuals. The first few rows of the genotype data file for this sample could be as follows:

1	rs3094315	0	742429	1	2	2	2	1	1	0	0	1	1	1	2	1	1	1	2
1	rs2286139	0	751595	1	1	1	1	1	1	0	0	1	2	0	0	1	1	1	2
1	rs11240776	0	755132	2	1	2	1	1	2	1	2	1	1	1	1	1	1	1	2
1	rs2980300	0	775852	0	0	2	1	1	1	2	1	2	1	2	2	1	1	1	2
.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)	(17)	(18)	(19)	(20)

Column (1) contains the chromosome name (1-22, X). This information will be ignored by the program (but some string must be present in the column).
Column (2) contains the rs number or SNP identifier.
Column (3) contains the genetic distance in Morgans (0 if genetic distance is unknown). This information will be ignored by the program (but some string must be present in the column).
Column (4) contains the base-pair position in bp units (0 if base-pair position is unknown). This information will be ignored by the program (but some string must be present in the column).
Columns (5) and (6) contain the marker genotype (one allele in each column) for the 1st individual. (Note restrictions 1-5 listed above on how the individuals should be ordered and how the genotypes should be coded.)
Columns (7) and (8) contain the marker genotype for the 2nd individual.
⋮
Columns (17) and (18) contain the marker genotype for the 7th individual.
Columns (19) and (20) contain the marker genotype for the 8th individual.

For more details on the tped file format, you could consult the PLINK website (http://pngu.mgh.harvard.edu/ purcell/plink/data.shtml#tr). PLINK provides a convenient venue to convert from many different file formats. For example, supposed you had a PLINK ped file named "mydata.ped" which was coded as A/C/G/T or 1/2/3/4 for the four alleles. Then, assuming that restrictions 1 and 3 above were met by mydata.ped, i.e. individuals from the same family were listed consecutively and the file did not mix autosomal SNPs with X-chromosome SNPs, you could generate the desired MQLS-XM genotype input file with the PLINK command:

./plink --file mydata --recode12 --output-missing-genotype 0 --transpose --out newfile

The PLINK software would then create the two files "newfile.tped" and "newfile.tfam". The file "newfile.tped" is a genotype data file that is in the appropriate format for the MQLS-XM software. **If, in addition, the phenotype was coded as 2=affected, 1=unaffected, and 0=unknown** in the file mydata.ped, then the tfam file "newfile.tfam" could be converted to the required phenotype information file as discussed in item 2 below.

The default filename for the genotype data file is "genofile". To specify a different filename, use the command-line flag -g followed by the filename. For example, to use the PLINK SNP genotype data file named "newfile.tped", you could type the command

./MQLS-XM -g newfile.tped

2. **phenotype information file**

This file contains the phenotype data as well as family ID and individual ID numbers for the study individuals. Important restrictions to keep in mind include:

1. Individuals must be listed in the same order as in the genotype data file;
2. Individuals from the same family must appear in a single cluster, and families must be numbered consecutively in the file from 1 to F, where F is the total number of families in the sample. In other words, the first family listed in the file must have family ID number 1, the second family listed in the file must have family ID number 2, and so on.
3. Individual IDs must be positive integers.
4. The phenotype must be coded as 2=affected, 1=unaffected, 0=unknown.

To illustrate the format of the input phenotype information file, consider a study sample with a total of 8 individuals from 3 families. The columns in the phenotype information file should be organized as follows:

1	11	0	0	1	1
1	23	0	0	2	2
1	13	11	23	1	2
2	2	0	0	2	1
2	51	0	0	1	2
2	3	51	2	1	0
2	41	51	2	2	1
3	1	0	0	2	0
(1)	(2)	(3)	(4)	(5)	(6)

Column (1) contains family ID (consecutive positive integers from 1 to F).
Column (2) individual ID (positive integer)
Column (3) father's ID (0=founder)

Column (4) mother's ID (0=founder)

Column (5) sex (1=male, 2=female)

Column (6) affection status (0=unknown, 1=unaffected, 2=affected)

Sampled individuals who are unrelated to anyone else in the sample should be included in this file by giving each such person their own unique family ID. There is no limit on the number of individuals, but the number of families is set to be smaller than 10,000.

To increase this limit, just change the value of MAXFAM in the MQLS_XM_SOURCE.c source file and recompile the program.

As mentioned in the previous sub-section, a PLINK tfam file that meets certain restrictions can easily be converted to a phenotype information file by using the FORMAT_PED_PHENO software, which can be found at
http://galton.uchicago.edu/∼mcpeek/software/index.html. The restrictions are:

1. phenotype values must be coded as 2=affected, 1=unaffected, and 0=unknown;
2. individuals from the same family must be listed consecutively;
3. the order of individuals must be the same as that in the genotype data file (previous subsection).

For example, to use the FORMAT_PED_PHENO software to convert the PLINK file "newfile.tfam" discussed in the previous subsection, the following command can be used:

./FORMAT -f newfile.tfam -o newfile

This command will generate the file "newfile.pedpheno", which is in the appropriate input format for the MQLS-XM software.

The default filename for the phenotype information file used by the MQLS-XM is "phenofile". To specify a different filename, use the command-line flag -p followed by the filename. For example, to use the file "newfile.pedpheno", you could type the command

./MQLS-XM -p newfile.pedpheno

3. **kinship coefficient file**

If an autosomal analysis is to be performed, then the kinship coefficient file should contain autosomal kinship and inbreeding coefficients. If an X-chromosome analysis is to be performed, then the kinship coefficient file should instead contain X-chromosome kinship and inbreeding coefficients. In what follows, we will use the terms "kinship coefficient" and "inbreeding coefficient" generically to refer to the autosomal kinship and inbreeding coefficients in case of an autosomal analysis and to the X-chromosome kinship and inbreeding coefficients in the case of an X-chromosome analysis.

A kinship coefficient should be provided for every pair of individuals who (1) are both in the phenotype information file and (2) have the same family ID. If individuals have different family IDs, then no kinship coefficient between them should be given. An inbreeding coefficient should be provided for every individual in the phenotype information file (even if the person is outbred). A sampled individual who does not share a family ID with anyone else in the phenotype information file should be represented in the kinship coefficient file by a single line that specifies the individual's inbreeding coefficient value (in the format below). Family and individual ID's should match exactly with those in the phenotype information file.

The autosomal inbreeding coefficient is 0 for all outbred individuals. The X-chromosome inbreeding coefficient is 0 for all outbred females and is 1 for **all** males, outbred or inbred (because males have only 1 allele at an X-linked SNP).

Following is an example of the format of the autosomal kinship coefficient file (note that this is not meant to correspond to the example pedigree in the previous subsection):

1	11	11	0
1	11	23	0
1	11	13	0.25
1	23	23	0
1	23	13	0.25
1	13	13	0
2	2	2	0.01251
2	2	51	0.26124
.	.	.	.
.	.	.	.
(1)	(2)	(3)	(4)

Column (1) family ID

Column (2) individual 1 ID (Id1)

Column (3) individual 2 ID (Id2)

Column (4) autosomal kinship coefficient between individuals Id1 and Id2 if Id1 $\neq$ Id2, and autosomal inbreeding coefficient of individual Id1 if Id1 = Id2

Following is an example of the format of the X-chromosome kinship coefficient file (note that this is not meant to correspond to the example pedigree in the previous subsection):

1	11	11	1
1	11	23	0
1	11	13	0
1	23	23	0
1	23	13	0.5
1	13	13	1
2	2	2	.0154
2	2	51	0
.	.	.	.
.	.	.	.
(1)	(2)	(3)	(4)

Column (1) family ID

Column (2) individual 1 ID (Id1)

Column (3) individual 2 ID (Id2)

Column (4) X-kinship kinship coefficient between individuals Id1 and Id2 if Id1 $\neq$ Id2, and X-chromsome inbreeding coefficient of individual Id1 if Id1 = Id2

We provide two software programs, one for autosomes and one for X, that can be used to calculate the required coefficients and generate the kinshp coefficient input file. The output files of these programs have the exact format required for the MQLS-XM kinship coefficient input file:

1. The KinInbcoef software for calculating autosomal kinship and inbreeding coefficients.
2. The KinInbcoefX software for calculating X-chromosome kinship and inbreeding coefficients.

Both programs can be found at
http://galton.uchicago.edu/∼mcpeek/software/index.html

The FORMAT_PED_PHENO software, discussed in the previous subsection, for converting a phenotype information file (e.g., a PLINK tfam file) also creates input files that are in the appropriate format for the KinInbcoef and KinInbcoefX software programs. For example, the MQLS-XM input phenotype file "newfile.pedpheno" mentioned in the previous subsection will be created by the FORMAT_PED_PHENO software with the following command:

./FORMAT -f newfile.tfam -o newfile

This command also creates the output files "newfile.kinpedigree", "newfile.kinpedigreeX", and "newfile.kinlist", which can be used as input to the KinInbcoef and KinInbcoefX software.

To obtain autosomal kinship coefficients using the KinInbcoef software, type the following command:

./KinInbcoef newfile.kinpedigree newfile.kinlist newfile.kinship

This command creates the output file "newfile.kinfile" which is in the exact format required for the MQLS-XM software.

Similarly, to obtain X-chromosome kinship coefficients using the KinInbcoefX software, type the following command:

./KinInbcoefX newfile.kinpedigreeX newfile.kinlist newfile.kinshipX

This command creates the output file "newfile.kinfileX" which is in the exact format required for the MQLS-XM software.

The default filename for the kinship coefficient file used by MQLS-XM is "kinfile". To specify a different filename, use the command-line flag -k followed by the filename. For example, to use the file "newfile.kinfile" for an autosomal analysis, you could type the command

./MQLS-XM -k newfile.kinfile

or, to use the file "newfile.kinfileX" for an X-chromsome analysis, you could type the command

./MQLS-XM -k newfile.kinfileX -x

4. **prevalence file**

   This file contains estimates of the male and female prevalence values for the binary trait in an appropriate reference population. These values are used in the calculation of the M_{QLS} and X_M statistics. The two estimates can either be in a single row or a single column, where the first estimate is for the male prevalence and the second estimate is for the female prevalence. If the prevalence estimates are in a single row, then the values must be separated by a blank space. If only one value is given in the prevalence file, then both the male and female prevalences will be set to this value in the analysis. Please read the subsection **What should I plug in for the sex-specific prevalences used in M_{QLS} and X_M?** in section 2 of this document.

   The default filename is "prevalence". To specify a different filename, use the command-line flag -r followed by the filename. For example, to use a prevalence file called "myprevalence", you could type the command

   ./MQLS-XM -r myprevalence

Optional Input:

5. **Perform an X-chromosome association analysis**

   The command-line flag -x would be used to perform an association analysis for X-linked SNPs. For example, to perform an X-chromosome association analysis, you could type the command

   ./MQLS-XM -x

5. **Exclude individuals with unknown phenotype**

   The command-line flag -u can be used to exclude all individuals with unknown phenotype from the analysis. For example, to exclude individuals with unknown phenotype, you could type the command

   /MQLS-XM -u

   The M_{QLS} and X_M tests explicitly allow for individuals of unknown phenotype and handle them appropriately in the analysis if this flag is not used. In contrast, the W_{QLS}, corrected χ^2, X_W, and X_χ tests classify individuals of unknown phenotype as controls (same as unaffected) if this flag is not used.

7. **Exclude phenotyped individuals with missing genotypes for the M_{QLS} and X_M tests**

   The M_{QLS} and X_M test statistics can allow phenotyped individuals with missing genotypes at a SNP to contribute to the statistic, provided that those individuals have a sampled relative who is genotyped at the SNP. The command-line flag -m can be used

to exclude phenotyped individuals with missing genotypes at a SNP from contributing to the statistics. For example, to exclude phenotyped individuals with missing genotype from the analysis, you could type the command

/MQLS-XM -m

This flag has no effect on the W_{QLS}, corrected χ^2, X_W, and X_χ tests, which always exclude individuals with missing genotypes.

7. **Calculate Test Statistics Assuming HWE for the Variance**

   The command-line flag -h can be used to calculate all test statistics using a variance estimator that assumes HWE, instead of the robust variance estimator (the default). For example, to perform an analysis using the HWE variance estimator, you could type the command

   /MQLS-XM -h

   For more details, see the subsection **Robust vs. HWE variance calculation** in section 2 of this document.

5 Output

1. **MQLStest.out** is the primary output file for an autosomal analysis (the default) and **XMtest.out** is the primary output file when the "-x" command line flag is used for an X-chromosome analysis. The files contain
   - Summary of the phenotype file information: total number of individuals in the phenotype file, number of independent families, number of individuals in each phenotype class (affected/unaffected/unknown)
   - Prevalence values used in the M_{QLS} (or X_M when the "-x" command line flag is used) calculations.
   - For each marker
     - SNP identifier/rs number
     - among those genotyped at the marker, the numbers who are affected, unaffected, and of unknown phenotype, respectively.
     - value of the M_{QLS} statistic (or the X_M statistic when the "-x" command line flag is used) and corresponding p-value using the chi-squared null distribution.
     - value of the corrected χ^2 statistic (or the $X\chi$ statistic when the "-x" command line flag is used) and corresponding p-value using the chi-squared null distribution
     - value of the W_{QLS} statistic (or the X_W statistic when the "-x" command line flag is used) and corresponding p-value using the chi-squared null distribution.
     - the signs of the M_{QLS} and W_{QLS} (or the X_M and X_W when the "-x" command line flag is used) quasi-scores associated to each allele when the p-value is smaller than 0.05, in order to know the direction of the change in allele frequency associated with the M_{QLS}, W_{QLS}, X_M , or X_W result.

- a warning message is printed when some allele counts are small, a situation in which the χ^2 asymptotic null distribution might not provide accurate p-values
- allele frequencies and s.d.'s estimated using the quasi-likelihood score function proposed by McPeek, Wu and Ober (2003) for autosomal markers and Thornton et al. (2012) for X-chromosome markers in the case sample, the unaffected control sample, the unknown phenotype control sample, and the entire sample (cases, unaffected controls, and unknown phenotype controls).
- allele frequencies estimated by naive counting in the case sample, the unaffected control sample, the unknown phenotype control sample, and the entire sample (cases, unaffected controls, and unknown phenotype controls).

2. **MQLStest.top** (or **XMtest.top** when the "-x" command line flag is used) lists the top 20 SNPs with the smallest p-values for each of the 3 tests. The number of markers output to this file can be changed by changing the value of MAXTOP (currently set to 20) in the MQLS_XM_SOURCE.c file.

3. **MQLStest.testvalues** (or **XMtest.testvalues** when the "-x" command line flag is used) lists, for every SNP, the values of each of the three test statistics.

4. **MQLStest.pvalues** (or **XMtest.pvalues** when the "-x" command line flag is used) lists the p-values for every SNP for each of the three statistics.

5. **MQLStest.err** (or **XMtest.err** when the "-x" command line flag is used) is an error file that may contain warnings

   - when a line has an incorrect number of fields in the genotype data file
   - when an individual from the kinship coefficient file is not listed in the pedigree data file
   - when an X-chromosome analysis is performed (i.e. "-x" command line flag is used) and a male does not have an X-chromosome inbreeding coefficient of 1
   - when an X-chromosome analysis is performed and a male has a heterozygous genotype. The MQLS-XM program will screen the first 500 SNPs with minor allele frequency greater than .1 in the input genotype file to identify males with heterozygous genotypes.

6 Tips

1. Input

   The program will stop if errors are detected in the formats of any of the input files. Please read Section 4 carefully and make sure the input files are in the correct format and have concordant information.

7 Example

1. Consider a PLINK ped file named "mydata.ped". Suppose that the following two conditions are met: (1) individuals from the same family are listed consecutively in the file; and (2) the file contains either autosomal or X-chromosome genotypes, but not both in the same file. Then the PLINK command below can be used to obtain tped and tfam output files:

   ./plink --file mydata --recode12 -- output-missing-genotype 0 --transpose --out newfile

   This command creates the two files "newfile.tped" and "newfile.tfam". The file "newfile.tped" is a genotype data file that is in the appropriate format for the MQLS-XM software package.

2. If, in addition to the two requirements above, the phenotype values are coded as 2=affected, 1=unaffected, and 0=unknown in the "newfile.tfam" file, then the FORMAT_PED_PHENO software can be used to obtain a phenotype information file that is in the required format for the MQLS-XM software. To convert the "newfile.tfam" file with the FORMAT_PED_PHENO software, the following command can be used:

   ./format -f newfile.tfam -o formattedfile

   This command will create a phenotype information file named "formattedfile.pedpheno" file that is in the appropriate format for the MQLS-XM software.

3. As mentioned in the previous subsection, the "formattedfile.pedpheno" file will be created by the FORMAT_PED_PHENO software with the command

   ./format -f newfile.tfam -o formattedfile

   This commond also creates the three files "formattedfile.kinpedigree", "formattedfile.kinpedigreeX", and "formattedfile.kinlist" which are in the appropriate format for the KinInbcoef software and the KinInbcoefX software. To obtain autosomal kinship and inbreeding coefficients with the KinInbcoef software and the input files "formattedfile.kinpedigree" and "formattedfile.kinlist", the following command can be used:

   ./KinInbcoef formattedfile.kinpedigree formattedfile.kinlist final.kinship

   This command creates the output file "final.kinship" which is in the exact format required by the MQLS-XM software for the autosomal kinship coefficient file.

4. Similarly, to obtain X-chromosome kinship coefficients using the KinInbcoefX software, the following command can be used:

   ./KinInbcoefX formattedfile.kinpedigreeX formattedfile.kinlist final.kinshipX

   This command creates the output file "final.kinshipX" which is in the exact format required by the MQLS-XM software for the X-chromosome kinship coefficient file.

5. Now, to run the MQLS-XM software for association testing of autosomal SNPs using the genotype input file from the PLINK software, the phenotype information file from

that FORMAT software, the output autosomal kinship file from the KinInbcoef software, and a file called "myprev" that contains the male and female prevelances, the following command can be used:

./MQLX-XM -g newfile.tped -p formattedfile.pheno -k final.kinship -r myprev

Similarly, if the newfile.tped contains only X-chromsome SNPs, the following command can be used for association testing of SNPs on the X-chromosome with the MQLS-XM software:

./MQLX-XM -g newfile.tped -p formattedfile.pheno -k final.kinshipX -r myprev -x

8 Acknowledgements

We gratefully acknowledge Jerry Halpern (funn@stanford.edu) for his contribution in implementing an algorithm for calculating p-values based on a χ^2_1 asymptotic null distribution.

9 References

1. Bourgain, C., Hoffjan, S., Nicolae, R., Newman, D., Steiner, L., Walker, K., Reynolds, R., Ober, C., McPeek, M.S. (2003). Novel case-control test in a founder population identifies P-selectin as an atopy-susceptibility locus. Am. J. Hum. Genet. 73, 612-626.

2. McPeek, M.S. (2012). BLUP genotype imputation for case-control association testing with related individuals and missing data. J. Comp. Biol. 19, 756-765.

3. McPeek, M.S., Wu, X., Ober, C. (2004). Best linear unbiased allele-frequency estimation in complex pedigrees. Biometrics 60, 359-367.

4. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, Maller J, Sklar P, de Bakker PIW, Daly MJ, Sham PC (2007). PLINK: a toolset for whole-genome association and population-based linkage analysis. Am. J. Hum. Genet. 81, 559-575. with Partially or Completely Unknown Population and Pedigree Structure. Am. J. Hum. Genet. 86, 172-184

5. Thornton, T., McPeek, M.S. (2007). Case-control association testing with related individuals: a more powerful quasi-likelihood score test. Am. J. Hum. Genet. 81, 321-337.

6. Thornton T., McPeek M. S. (2010) ROADTRIPS: Case-Control Association Testing with Partially or Completely Unknown Population and Pedigree Structure. Am. J. Hum. Genet. 86, 172-184

7. Thornton T, Zhang Q, Cai X, Ober C, and McPeek MS (2012) XM: Association Testing on the X-Chromosome in Case-Control Samples with Related Individuals. Genet Epidemiol 36, 438-450.