User Manual for

m r M L M

multi-locus random-SNP-effect Mixed Linear Model tools for genome-wide association study

(**version 3.0**)

Zhang Ya-Wen, Li Pei, Ren Wen-Long, Ni Yuan-Li

Zhang Yuan-Ming (soyzhang@mail.hzau.edu.cn)

Last updated on March 10, 2018

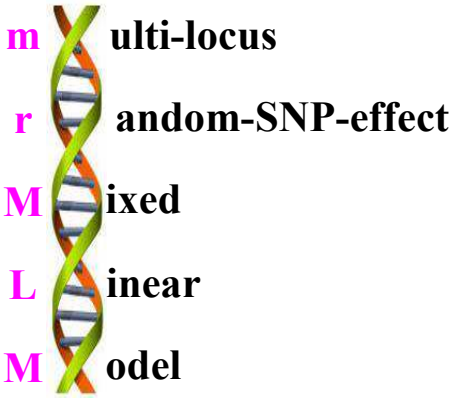
Disclaimer: While extensive testing has been performed by Yuan-Ming Zhang's Lab (Statistical Genomics Lab) at Huazhong Agricultural University, the results are, in general, reliable, correct or appropriate. However, results are not guaranteed for any specific datasets. We strongly recommend that users validate the mrMLM results with other software packages, such as GEMMA, EMMAX, GAPIT v2 and PLINK.

Download website:
https://cran.r-project.org/web/packages/mrMLM/index.html

Citation:

Method	References
mrMLM	Wang et al. *Scientific Reports* 2016, **6**:19444
FASTmrEMMA	Wen et al. *Briefings in Bioinformatics* 2017, bbw145, DOI: 10.1093/bib/bbw145
ISIS EM-BLASSO	Tamba et al. *PLoS Computational Biology* 2017, 13(1): e1005357.
pLARmEB	Zhang et al. *Heredity* 2017, 118: 517–524
pKWmEB	Ren et al. *Heredity* 2018, 120(3): 418-428
FASTmrMLM	Tamba. Nanjing Agricultural University 2017 Ph D Dissertation

Note: These references are listed in section of Reference.



This work was supported by the National Natural Science Foundation of China (31571268 and U1602261), Huazhong Agricultural University Scientific & Technological Self-innovation Foundation (Program No. 2014RC020), and State Key Laboratory of Cotton Biology Open Fund (CB2017B01).

INTRODUCTION

1.1 Why mrMLM?

mrMLM (**m**ulti-locus **r**andom-SNP-effect **M**ixed **L**inear **M**odel) program is an R package for multi-locus genome-wide association study (GWAS). At present this program (v3.0) includes six methods: 1) mrMLM, 2) FASTmrMLM (Fast multi-locus random-SNP-effect EMMA), 3) ISIS EM-BLASSO (Iterative Sure Independence Screening EM-Bayesian LASSO), 4)pLARmEB (polygene-background-control-based least angle regression plus empirical Bayes), 5) pKWmEB (polygene-background-control-based Kruskal-Wallis test plus empirical Bayes); and 6) FASTmrMLM (fast mrMLM).

In mrMLM, FASTmrMLM, FASTmrEMMA and pKWmEB, their visualization are based on package qqman, which is helpful to draw the Manhattan and QQ plots. In pLARmEB and ISIS EM-BLASSO, their visualizations are based on package ggplot2, which is helpful to draw the LOD score plot.

mrMLM 3.0 is able to work on the popular platforms, like Windows, Linux (desktop) and MacOS.

1.2 Getting started

mrMLM is a package that runs in the R software environment, which can be freely downloaded from https://cran.r-project.org/web/packages/mrMLM/index.html, or request from the maintainer, Dr Yuan-Ming Zhang at Huazhong Agricultural University (soyzhang@mail.hzau.edu.cn or soyzhang@hotmail.com).

1.2.1 One-Click installation

Within R environment, the mrMLM software can be installed directly using the below command:

```
install.packages(pkgs="mrMLM")
```

1.2.2 Step-by-step installation

1.2.2.1 Install the add-on packages

Online installation Within R environment on the internet, the mrMLM package can be installed online, using the below command:

install.packages(pkgs=c("qqman","stringr","MASS","lars","ncvreg","ggplot2","coin", "data.table","doParallel","sampling","openxlsx"))

Offline installation Users should download the below 47 packages from CRAN, github (https://github.com/), or google search:

assertthat, calibrate, cli, codetools, coin, colorspace, crayon, data.table, dichromat, digest, doParallel, foreach, ggplot2, glue, gtable, iterators, labeling, lars, lazyeval, lpsolve, magrittr, MASS, modeltools, multcomp, munsell, mvtnorm, ncvreg, openxlsx, pillar, plyr, qqman, R6, RColorBrewer, Rcpp, reshape2, rlang, sampling, sandwich, scales, sourcetools, stringi, stringr, TH.data, tibble, utf8, viridisLite, zoo.

Then, install them offline (under the R environment, select all the 47 packages and install them offline).

1.2.2.2 Install mrMLM

Open R GUI, select "Packages"—"Install package(s) from local files…" and then find the mrMLM package which you have downloaded on your desktop.

Within R environment, launch the mrMLM by command: library(mrMLM).

User Manual file Users can decompress the mrMLM package and find the User Manual file (name: **Instruction.pdf**) in the folder of "…/mrMLM/inst/doc".

2. Parameter settings

Parameter	Meaning	File format	Note
filePhe	File path & name in your computer, i.e., filePhe="D:\\Users\\Phenotype.csv"	*.csv; *.txt (Phenotypic values. **Row**: individual; **Column**: traits)	Table 1
fileGen	File path & name in your computer, i.e., fileGen="D:\\Users\\Genotype_num.csv"	*.csv; *.txt (Genotypic values. **Row**: markers; **Column**: individuals)	Tables 2~4
fileKin	File path & name in your computer, i.e., fileKin="D:\\Users\\Kinship.csv" or fileKin=NULL	*.csv; *.txt (Kinship matrix. **Row** & **Column**: individuals)	Table 5
filePS	File path & name in your computer,i.e., filePS="D:\\Users\\PopStr.csv" or filePS=NULL	*.csv; *.txt [Population structure. **Row**: individual; **Column**: sub-populations 1, 2, ---, *k* (No. of sub-populations)]	Table 6
Genformat	Format for genotypic codes: Num (number), Cha (character) & Hmp (Hapmap), i.e., Genformat="Num"		
method	Six multi-locus GWAS methods. Users may select one to six methods. For example, method=c("mrMLM","FASTmrMLM","FASTmrEMMA","pLARmEB","pKWmEB","ISIS EM-BLASSO")		

Likelihood	This parameter is only for FASTmrEMMA, including restricted maximum likelihood (REML) and maximum likelihood (ML). Likelihood="REML" or Likelihood="ML"
trait	Traits analyzed from number 1 to number 2. For example, **trait=1:3** indicates that users analyze the first to third traits.
SearchRadius	This parameter is only for mrMLM and FASTmrMLM, indicating Search Radius in search of potentially associated QTN. **SearchRadius=20** indicates that only one potentially associated QTN was selected within 20 kb.
CriLOD	Critical LOD score for significant QTN. **CriLOD=3** indicates that the critical LOD score for significant QTN is set at 3.0
SelectVariable	This parameter is only for pLARmEB. **SelectVariable=50** indicates that 50 potentially associated variables are selected from each chromosome. Users may change this number in real data analysis in order to obtain the best results as final results.
Bootstrap	This parameter is only for pLARmEB, including FALSE & TRUE. **Bootstrap=FALSE** indicates the analysis of only real dataset; **Bootstrap=TRUE** indicates the analysis of both real dataset and four resampling datasets.
DrawPlot	This parameter is for all the six methods, including FALSE and TRUE. **DrawPlot=FALSE** indicates no figure output; **DrawPlot=TRUE** indicates the output of the Manhattan, QQ and LOD score against genome position figures.
Plotformat	This parameter is for all the figure files, including *.jpeg, *.png, *.tiff and *.pdf. **Plotformat="jpeg"** indicates the *.jpeg format of plot file.
Resolution	This parameter is for all the figure files, including Low and High. **Resolution="Low"** indicates low figure resolution.

Example

```
mrMLM(fileGen="D:\\Users\\Genotype_num.csv",filePhe="D:\\Users\\Phenotype.csv",fileKin=NULL,filePS=NULL,Genformat="Num",method=c("mrMLM","FASTmrMLM","FASTmrEMMA","pLARmEB","pKWmEB","ISIS EM-BLASSO"),Likelihood="REML",trait=1:3,SearchRadius=20,CriLOD=3,SelectVariable=50,Bootstrap=FALSE,DrawPlot=FALSE,Plotformat ="jpeg",Resolution="Low")
```

Dataset format

Format for the filePhe dataset The first column presents individual ID. Note that "<Phenotype>" should be showed in the first row of the first column. Each of the following columns stands for observations of one trait, and the trait name is showed in the first row.

<Phenotype>	trait1	trait2	trait3
B46	42	43.02	44.32
B52	72.5	71.88	72.8
B57	41	41.7	41.42
B64	74.5	74.43	74.5
B68	65	66.4	65.33
B73	83.25	83.72	85.2
B73HTRHM	73	74.53	74.43
B75	56	57.24	58.01

Table 1. The format of the filePhe dataset

Numeric format for fileGen dataset The first column, named "**rs#**", stands for

marker ID. The second column, named "**chrom**", stands for chromosome. The third column, named "**pos**", stands for the position (bp) of SNP on the chromosome. The fourth column, named "**genotype for code 1**", indicates reference base for number 1. If the base for the first individual is missing, the base firstly observed in the next individual is what we list. Among the remaining columns, each column lists all the genotypes for one individual while the first row shows the individual names. For each marker, homozygous genotypes are expressed by 1 and -1, respectively, and the heterozygous and missing genotypes are indicated by zero. Note that the genotypes with code **1** will be listed in the **Result** files.

rs#	chrom	pos	genotype for code 1	33-16	Nov-38	A4226	A4722
PZB00859.1	1	157104	C	1	1	1	1
PZA01271.1	1	1947984	C	1	-1	1	-1
PZA03613.2	1	2914066	G	1	1	1	1
PZA03613.1	1	2914171	T	1	1	1	1
PZA03614.2	1	2915078	G	1	1	1	1
PZA03614.1	1	2915242	T	1	1	1	1
PZA02117.1	1	223466480	A	1	1	1	-1
PZA00403.5	1	223466873	T	1	1	1	0
PZB01979.2	1	224421551	A	1	-1	1	-1

Table 2. The numeric format of the fileGen dataset

Character format for fileGen dataset The first three columns are same as those in Table 2. The differences are that the marker values are character, such as **A, T, C, G** and **N**, and the other notations are heterozygous genotypes. The "**N**" indicates missing. The first rows from the fourth to last columns are individual name.

rs#	chrom	pos	33-16	Nov-38	A4226	A4722
PZB00859.1	1	157104	C	C	C	C
PZA01271.1	1	1947984	C	G	C	G
PZA03613.2	1	2914066	G	G	G	G
PZA03613.1	1	2914171	T	T	T	T
PZA03614.2	1	2915078	G	G	G	G
PZA03614.1	1	2915242	T	T	T	T

Table 3. The character format of the fileGen dataset

Hapmap format for fileGen dataset Please see the TASSEL software in details. Here we introduce simply. The first eleven columns describe the specific information of markers and individuals, and their column names must be **"rs#"**, **"alleles"**, **"chrom"**, **"pos"**, **"strand"**, **"assembly#"**, **"center"**, **"protLSID"**, **"assayLSID"**, **"panel"** and **"QCcode"**. In the **"rs#"** (1)**, "chrom"** (3) **and "pos"** (4) columns, their information is described as the above. The values for marker genotypes should be character, such as **AA, TT, CC, GG, NN, AC** and **AG**, where the "**NN**" indicates missing or unknown genotypes. In the 2 and 5 to 11 columns, **"NA"** indicates **no information** available. All the individual genotypic information will be showed from the 12 to last columns. In each column, individual name is listed in the first row, i.e., "33-16", and the others are the genotypes (character).

rs#	alleles	chrom	pos	strand	assembly#	center	protLSID	assayLSID	panel	QCcode	33-16
PZB00859.1	A/C	1	157104	+	AGPv1	Panzea	NA	NA	maize282	NA	CC
PZA01271.1	C/G	1	1947984	+	AGPv1	Panzea	NA	NA	maize282	NA	CC
PZA03613.2	G/T	1	2914066	+	AGPv1	Panzea	NA	NA	maize282	NA	GG
PZA03613.1	A/T	1	2914171	+	AGPv1	Panzea	NA	NA	maize282	NA	TT
PZA03614.2	A/G	1	2915078	+	AGPv1	Panzea	NA	NA	maize282	NA	GG
PZA03614.1	A/T	1	2915242	+	AGPv1	Panzea	NA	NA	maize282	NA	TT
PZA02117.1	A/G	1	223466480	+	AGPv1	Panzea	NA	NA	maize282	NA	AA
PZA00403.5	C/T	1	223466873	+	AGPv1	Panzea	NA	NA	maize282	NA	TT
PZB01979.2	A/G	1	224421551	+	AGPv1	Panzea	NA	NA	maize282	NA	AA

Table 4. The hapmap format of the fileGen dataset

The format for fileKin dataset The dataset consists of the $(n+1) \times (n+1)$ matrix. In the first column, the first number indicates sample size n, i.e., 263; the others are individual ID, i.e., 33-16, Nov-38, and A4226. The number n is the common individuals between the phenotypic and genotypic datasets.

fileKin=NULL indicates that the Kinship matrix is calculated by software mrMLM. Note that only the common individuals are used to calculate the Kinship matrix. fileKin="D:\\Users\\Kinship.csv" means that the K matrix with name Kinship.csv is

uploaded from the folder "D:\\Users\\". Note that the number and order of individuals in Kinship.csv may be not consistent with those of the above common individuals. However, our software may change the K matrix in order that the number and order of new K matrix matches the number and order of the above common individuals.

If the number of markers is very large, i.e., 50,000, we recommend that users calculate the K matrix using the other programs, especially for FASTmrEMMA.

263					
33-16	1.00809	0.45954	0.50677	0.42503	0.45591
Nov-38	0.45954	1.03352	0.43048	0.47044	0.39597
A4226	0.50677	0.43048	1.01717	0.45409	0.43775
A4722	0.42503	0.47044	0.45409	0.89002	0.34874
A188	0.45591	0.39597	0.43775	0.34874	1.0099
A214N	0.34693	0.33421	0.39779	0.29244	0.33058
A239	0.43593	0.46499	0.40323	0.36691	0.39597
A272	0.34874	0.40505	0.31423	0.3887	0.44138
A441-5	0.47952	0.44138	0.47226	0.47952	0.49224
A554	0.39779	0.45954	0.5431	0.48679	0.4214
A556	0.50858	0.40505	0.45954	0.40142	0.40687

Table 5. The format of the fileKin dataset

The format for filePS dataset The dataset consists of the $(n+1) \times (k+1)$ matrix, where n is the number of the common individuals and k is the number of sub-populations. In the first column, "<Covariate>" and "<Trait>" should present in the first and second rows, respectively. The following two to $(k+1)$ columns indicate the population structure. Note that the $\boldsymbol{Q_i}$ is listed in the second row.

filePS=NULL indicates that population structure isn't included in the genetic model. filePS="D:\\Users\\PopStr.csv" means that population structure with name PopStr.csv is uploaded from the folder "D:\\Users\\". Note that the number and order of individuals in PopStr.csv may be not consistent with those of the above common individuals. However, our software may change the population structure matrix in order that the number and order of new matrix matches the number and order of the above common individuals.

<Covariate>			
<Trait>	Q1	Q2	Q3
33-16	0.014	0.972	0.014
Nov-38	0.003	0.993	0.004
A4226	0.071	0.917	0.012
A4722	0.035	0.854	0.111
A188	0.013	0.982	0.005
A214N	0.762	0.017	0.221
A239	0.035	0.963	0.002
A272	0.019	0.122	0.859
A441-5	0.005	0.531	0.464
A554	0.019	0.979	0.002

Table 6. The format of the filePS dataset

3. Result

At the work directory of your R, two files of result for the first trait, "1_intermediate result.csv" and "1_Final result.csv", will appear.

In the **intermediate result** from the method mrMLM, the result table includes: Trait ID, Trait name, method, reference sequence number (rs#, marker name), chromosome, marker's position (bp) in the chromosome, SNP effect (γ_k , Effect), -$\log_{10}$(p), genotype for code 1.

In the **Final result** from the method mrMLM, the result table includes: Trait ID, Trait name, method, reference sequence number (rs#, marker names), chromosome, marker's position (bp) in the chromosome, QTN effect, LOD score, -$\log_{10}$(P), the proportion of phenotypic variance explained by significant QTN (r^2), minor allelic frequency, genotype for code 1, residual error variance, and total phenotypic variance.

4. References

1. Wang Shi-Bo, Feng Jian-Ying, Ren Wen-Long, Huang Bo, Zhou Ling, Wen Yang-Jun, Zhang Jin, Jim M. Dunwell, Xu Shizhong*, Zhang Yuan-Ming*. Improving power and accuracy of genome-wide association studies via a multi-locus mixed linear model methodology. ***Scientific Reports*** 2016, **6**: 19444.
2. Wen Yang-Jun, Zhang Hanwen, Ni Yuan-Li, Huang Bo, Zhang Jin, Feng Jian-Ying, Wang Shi-Bo, Jim M. Dunwell, Zhang Yuan-Ming*, Wu Rongling*. Methodological implementation of mixed linear models in multi-locus genome-wide association studies. ***Briefings in Bioinformatics*** 2017, bbw145, https://doi.org/10.1093/bib/bbw145

3. Tamba Cox Lwaka, Ni Yuan-Li, Zhang Yuan-Ming*. Iterative sure independence screening EM-Bayesian LASSO algorithm for multi-locus genome-wide association studies. ***PLoS Computational Biology*** 2017, 13(1): e1005357, DOI: 10.1371/journal.pcbi.1005357
4. Zhang Jin[#], Feng Jian-Ying[#], Ni Yuan-Li, Wen Yang-Jun, Niu Yuan, Tamba Cox Lwaka, Yue Chao, Song Qi-Jian, Zhang Yuan-Ming*. pLARmEB: Integration of least angle regression with empirical Bayes for multi-locus genome-wide association studies. ***Heredity*** 2017, **118**: 517–524.
5. Ren Wen-Long[#], Wen Yang-Jun[#], Jim M. Dunwell, Zhang Yuan-Ming *. pKWmEB: Integration of Kruskal-Wallis test with empirical Bayes under polygenic background control for multi-locus genome-wide association study. ***Heredity*** 2018, **120**: 208–218.
6. Tamba Cox Lwaka. A fast mrMLM algorithm improves statistical power, accuracy and computational efficiency of multi-locus GWAS (Ph D Dissertation). ***Nanjing Agricultural University*** 2017
7. Zhang YM, Mao Y, Xie C, Smith H, Luo L, Xu S. Mapping quantitative trait loci using naturally occurring genetic variance among commercial inbred lines of maize (*Zea mays* L.). ***Genetics*** 2005; **169**:2267–2275.