User Manual for

mrMLM.GUI

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

(version 3.2)

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

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

Last updated on August 25, 2018

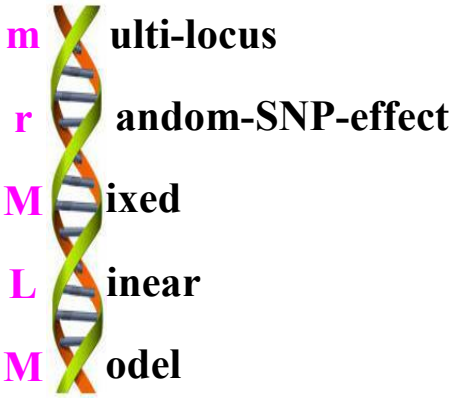
Disclaimer: While extensive testing has been performed by Yuan-Ming Zhang's Lab at the Crop Information Center of Huazhong Agricultural University, the results are, in general, reliable, correct or appropriate. However, results are not guaranteed for any specific datasets. We strongly recommend that users validate the mrMLM.GUI results with other software packages, i.e., GEMMA, EMMAX, GAPIT v2 & PLINK.

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

Method	References
mrMLM	Wang et al. *Scientific Reports* 2016, **6**:19444
FASTmrEMMA	Wen et al. *Briefings in Bioinformatics* 2018, 19(4): 700–712. 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 & Zhang, bioRxiv preprint first posted online 2018, doi: https://doi.org/10.1101/341784

Citation:

Note: These references are listed in section of Reference.



This work was supported by the National Natural Science Foundation of China (31571268, 31871242 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.GUI?

mrMLM.GUI (**m**ulti-locus **r**andom-SNP-effect **M**ixed **L**inear **M**odel with **G**raphical **U**ser **I**nterface) program is an R package for multi-locus genome-wide association study (GWAS). At present this program (v3.2) 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 (polygenic-background-control-based least angle regression plus empirical Bayes), 5) pKWmEB (polygenic-background-control-based Kruskal-Wallis test plus empirical Bayes); and 6) FASTmrMLM (fast mrMLM).

In mrMLM.GUI, 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.GUI 3.2 is able to work on the popular platforms, like Windows, Linux (desktop) and MacOS.

1.2 Getting started

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

1.2.1 One-Click installation

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

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

1.2.2 Step-by-step installation

1.2.2.1 Install the add-on packages

Offline installation Users should download the below 61 packages from CRAN, github (https://github.com/), or google search:

assertthat, bigmemory, bigmemory.sri, calibrate, cli, codetools, coin, colorspace, crayon, data.table, dichromat, digest, doParallel, foreach, ggplot2, glue, gtable, htmltools, httpuv, iterators, jsonlite, labeling, lars, later, lazyeval, lpsolve, magrittr, MASS, mime, miniUI, modeltools, mrMLM, multcomp, munsell, mvtnorm, ncvreg, openxlsx, pillar, plyr, promises, qqman, R6, RColorBrewer, Rcpp, reshape2, rlang, sampling, sandwich, scales, shiny, shinyjs, sourcetools, stringi, stringr, TH.data, tibble, utf8, viridisLite, xtable, zip, zoo.

Then, install them offline (under the R environment, select all the 61 packages and install them offline).

1.2.2.2 Install mrMLM.GUI

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

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

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

2. Dataset input

2.1 Phenotypic dataset

The **Phenotypic** file should be a ***.csv** or ***.txt** format file. 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 the observations of each 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

Table 1. The format of Phenotypic dataset

2.2 Genotypic dataset

The **Genotypic** file should be a ***.csv** or ***.txt** format file.

Numeric format for Genotypic 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 variable $x = 1$. 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 also 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 genotypic dataset

Character format for Genotypic 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

Table 3. The character format of the genotypic dataset

Hapmap format for Genotypic 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 genotypic dataset

2.3 Kinship dataset

The Kinship file should be a ***.csv** or ***.txt** format file.

The dataset consists of the (n+1) × (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. n is the number of common individuals between the phenotypic and genotypic datasets.

If you select "Calculate kinship (K) matrix by this software", it will be calculated by software mrMLM.GUI. Note that only the common individuals are used to calculate

the Kinship matrix. If you select “Input Kinship (K) matrix file” and upload it, note that the number and order of individuals in the uploaded file may be not consistent with those of the above common individuals. However, our software may match the uploaded K matrix in order that the number and order of new K matrix are consistent with those in 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 Kinship dataset

2.4 Population Structure dataset

The dataset consists of the $(n+2) \times (k+1)$ 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.

If you select “Not included in the model” indicates that population structure isn’t considered in the genetic model. If you select “included” and upload the population structure file, note that the number and order of individuals in the uploaded file may be not consistent with those in the above common individuals. However, our software may change the population structure matrix in order that the number and order of new matrix are consistent with those in 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 Population Structure dataset

3. Operation process

3.1 The Graphical User Interface of mrMLM.GUI

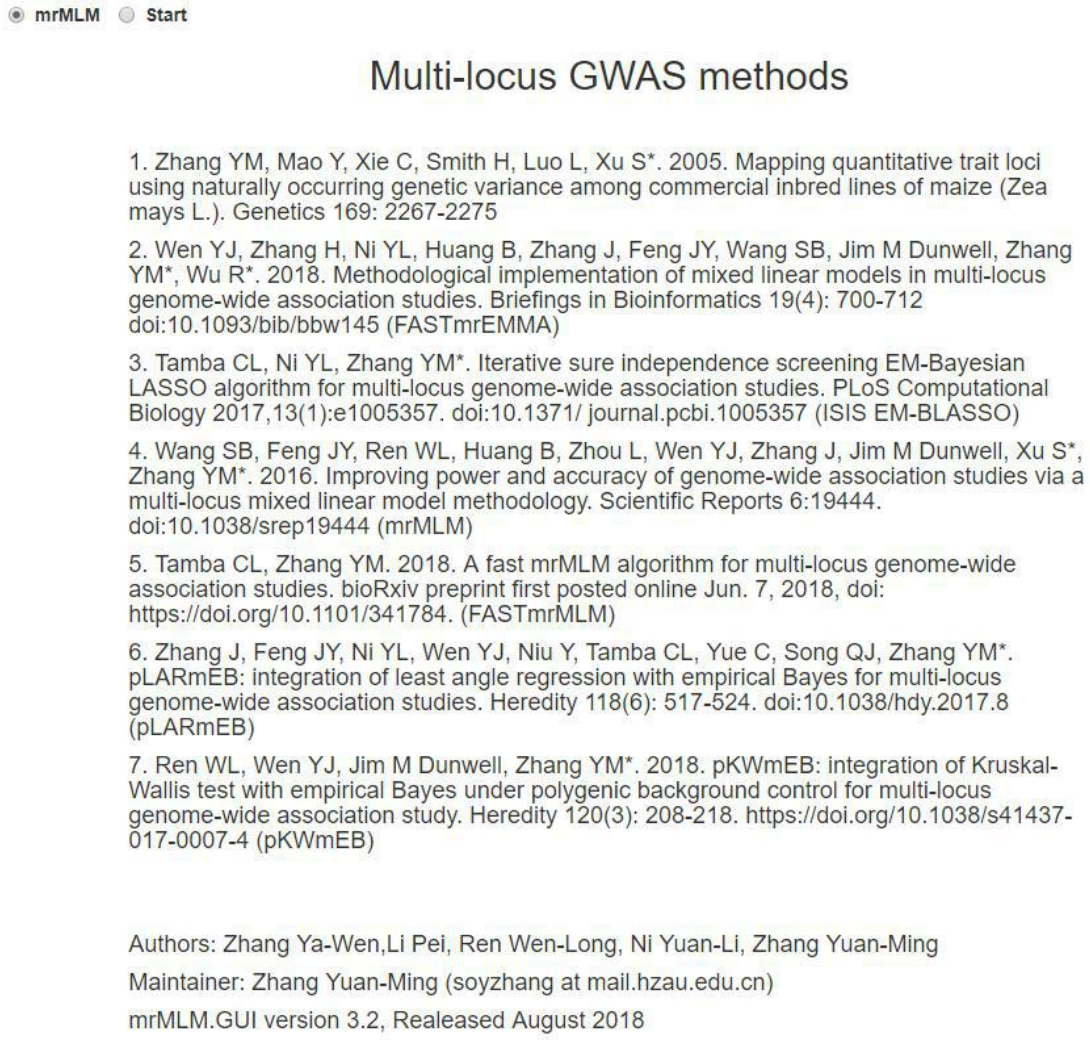


Figure 1. The Graphical User Interface of mrMLM.GUI

3.2 Input dataset

Users must upload the genotypic and phenotypic files (**Figs 2 & 3**) while the Kinship and Population-Structure files are optional. In Kinship module, users should upload the Kinship matrix if you select "**Input Kinship (K) matrix file**" (**Fig 4**); users don't need to upload this file, which will be calculate automatically, if you select "**Calculate Kinship (K) matrix by this software**". In Population Structure module, users should upload the Population Structure file if you select "**Included**" (**Fig 5**); the Population Structure will be not considered in the model if you select "**Not included in the model**".

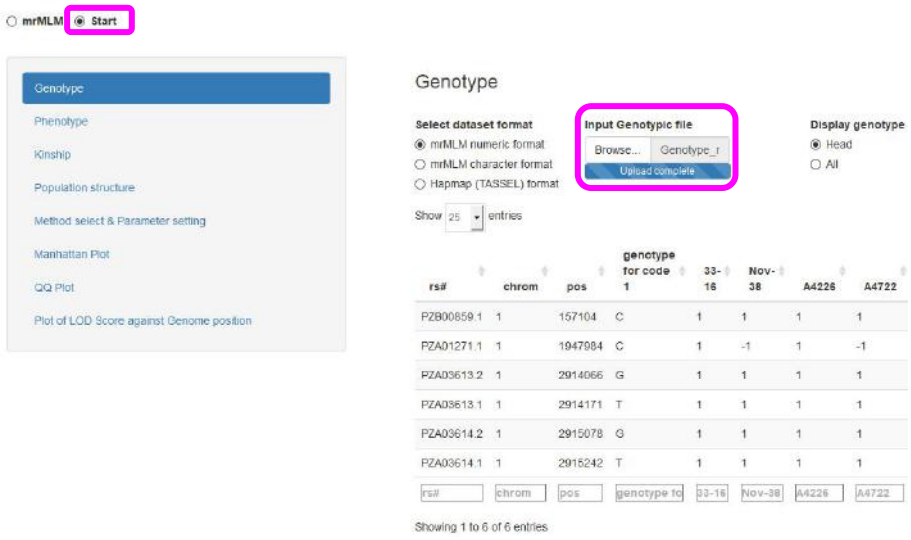


Figure 2. Input genotypic dataset

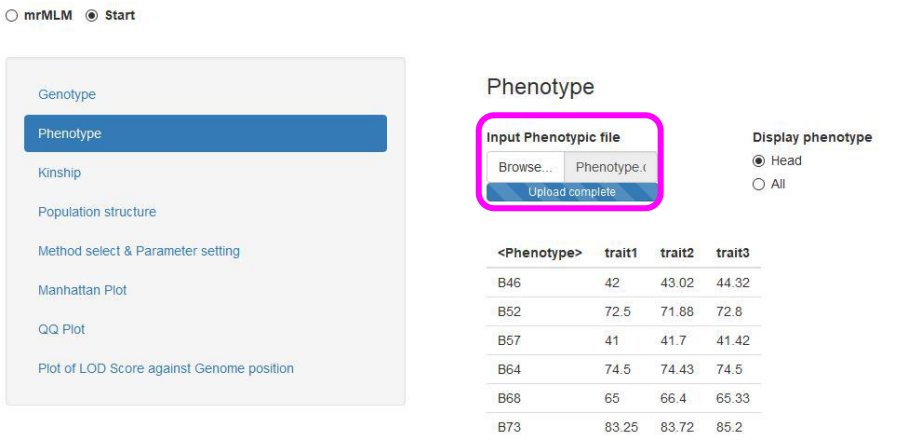


Figure 3. Input Phenotypic dataset

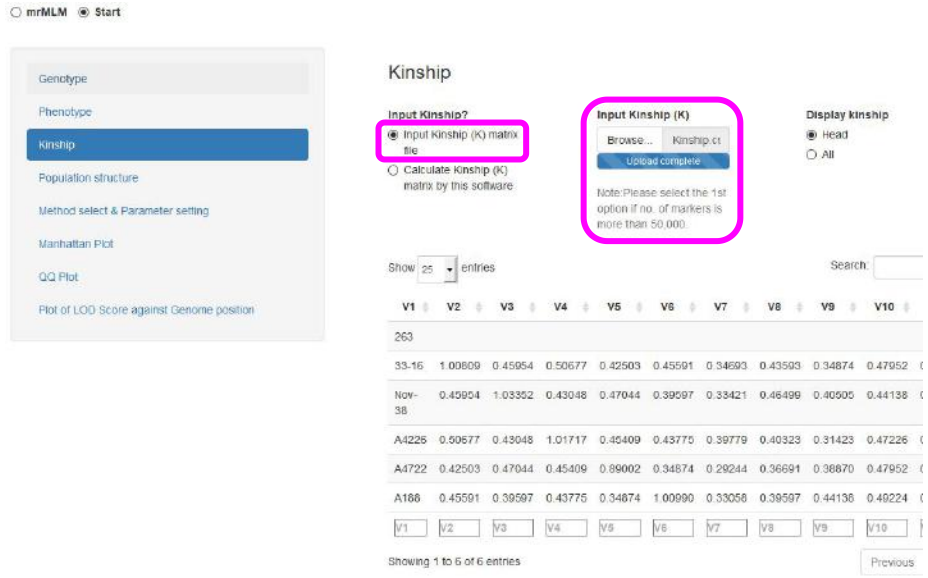


Figure 4. Input kinship dataset

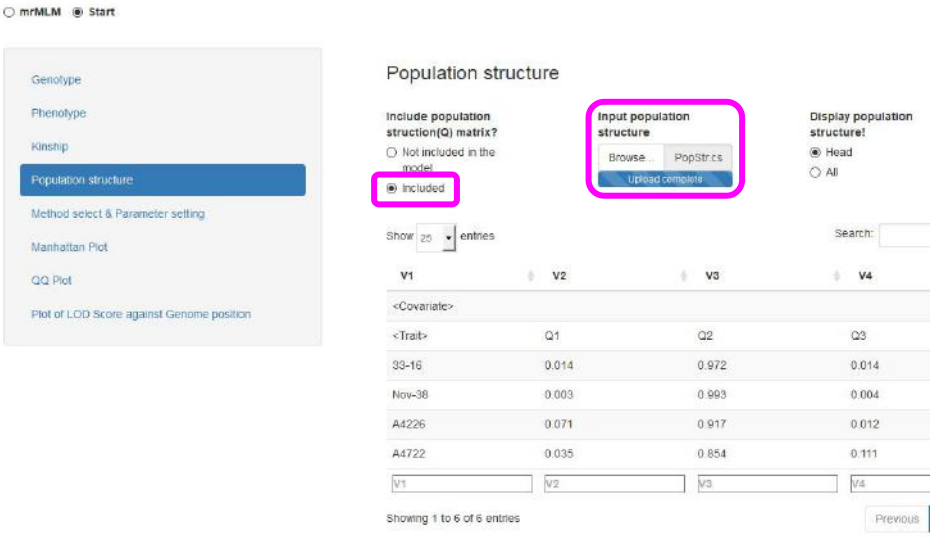


Figure 5. Input Population Structure dataset

3.3 Method select & Parameter setting (Fig 6)

Method select: There are six multi-locus GWAS methods in the mrMLM.GUI. Users may select one to six methods.

Critical LOD score (All method): Critical LOD score for significant QTN. If users set this value as 3, all the QTNs with their LOD scores ≥ 3.0 are viewed as true.

Search radius of candidate gene (kb) (mrMLM & FASTmrMLM): This parameter is only for mrMLM and FASTmrMLM, indicating Search Radius in search of potentially associated QTN. If users set it as 20 kb, only one potentially associated QTN within the radius of 20 kb was selected into multi-locus model.

No. of potentially associated variables selected by LARS (pLARmEB): This parameter is only for pLARmEB. If users set it as 50, 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 ones.

Likelihood function (FASTmrEMMA): This parameter is only for FASTmrEMMA, including restricted maximum likelihood (REML) and maximum likelihood (ML).

Bootstrap (pLARmEB): This parameter is only for pLARmEB, including FALSE & TRUE. **FALSE** indicates real dataset analysis only; **TRUE** indicates the analyses of both real dataset and four resampling datasets.

Draw plot or not (All methods): This parameter is for all the six methods, including FALSE and TRUE. **FALSE** indicates no figure output; **TRUE** indicates the output of the Manhattan, QQ and LOD score against genome position figures.

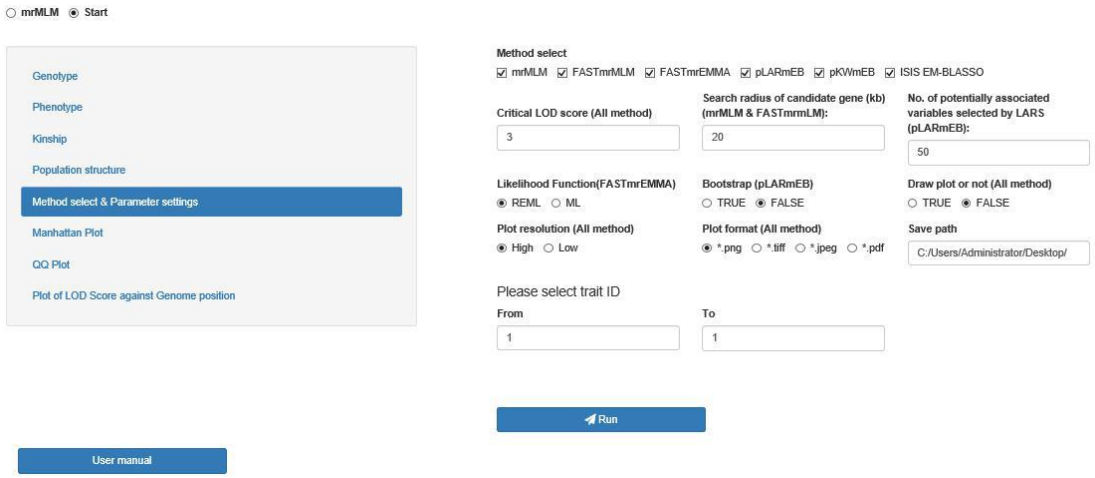


Figure 6. Method select & Parameter setting

Plot format (All methods): This parameter is for all the figure files, including *.jpeg, *.png, *.tiff and *.pdf.

Plot Resolution (All methods): This parameter is for all the figure files, including Low and High. Their parameters were showed at Page 14.

Save path: save path in your computer

3.4 Run the software (Fig 7)

After uploading all the needed files and setting the parameters, users can run the

software. The result files will be saved to the path that users set.

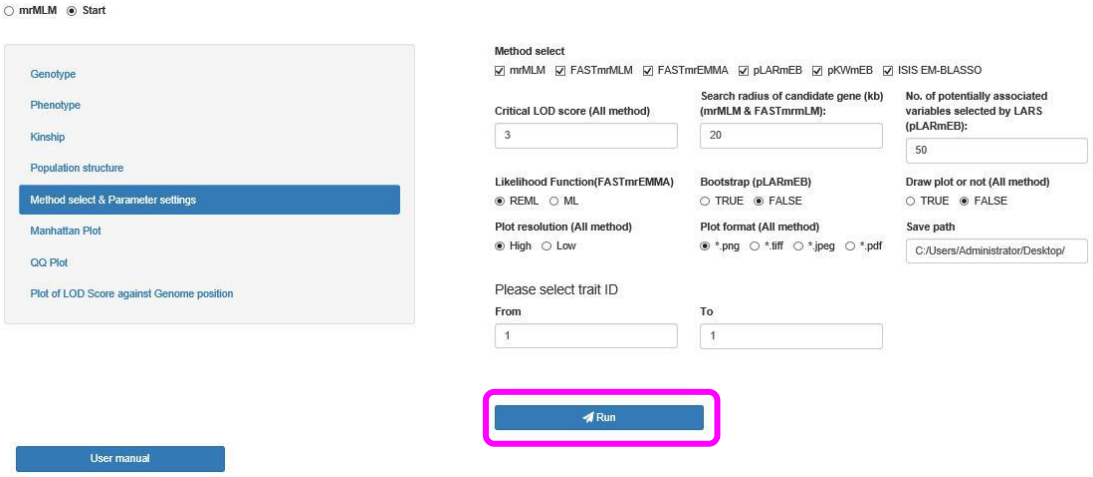


Figure 7. Run the software mrMLM.GUI

3.5 Re-draw the plot according to your own requirement

When you finish the calculating, you will get a file called **resultforplot.xlsx,** it can be used to redraw the plot.

3.5.1 Manhattan plot

Use Manhattan plot module to preview Manhattan plot in an independent dialog window. Before saving this Figure, please set the related parameters: **width** and **height** [with the unit of pixel (px)], **word resolution** [with the unit of 1/72 inch, being pixels per inch (ppi)], and **figure resolution** [with the unit of pixels per inch (ppi)]. Users may set the colors for the adjacent chromosomes, with a drop-down option. The critical value for $-\log_{10}$(P-value) is defaulted as the value of $0.05/m_e$, where m_e is the effective number of markers (please see Wang et al. *Scientific Reports* 2016, 6: 19444). Use **Save Manhattan plot** button to choose a path and to save the Figure, with four frequently used image formats: *.png, *.tiff , *.jpeg and *.pdf (**Fig 8**).

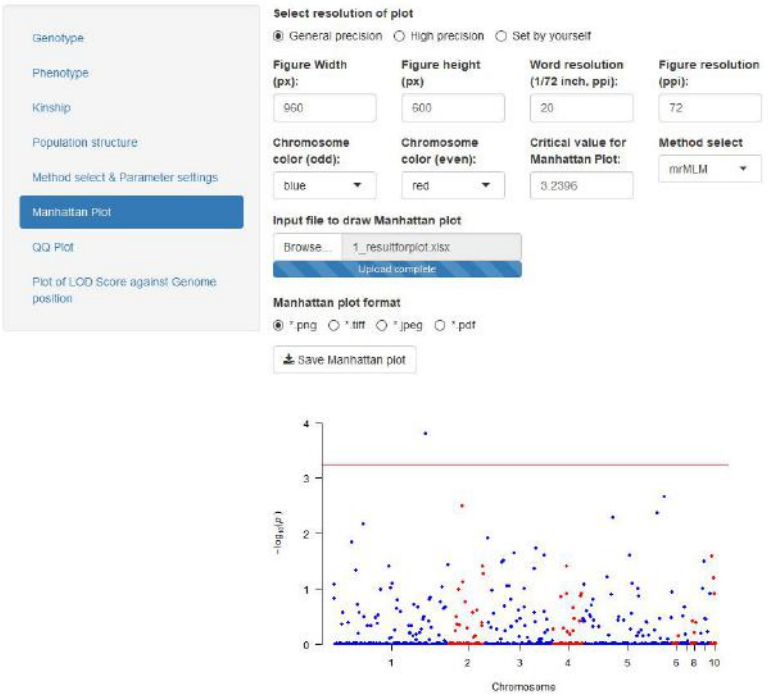


Figure 8. Manhattan plot module

3.5.2 QQ plot

Use **QQ plot** module to redraw the QQ plot in an independent dialog window. The parameter settings are the same as those in the Manhattan plot. Use **Save QQ plot** button to choose a path and to save the Figure, with four frequently used image formats: *.png, *.tiff , *.jpeg and *.pdf (**Fig 9**).

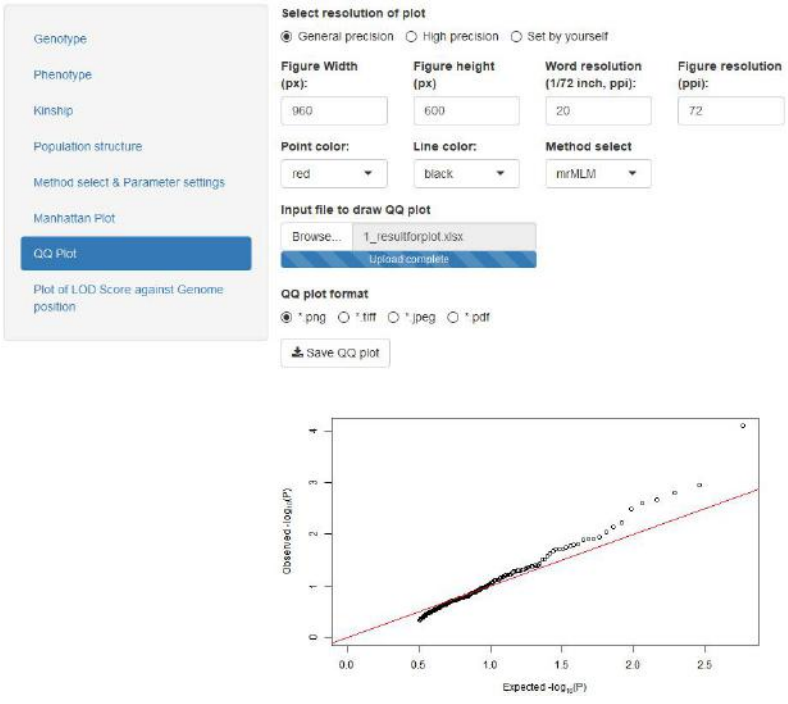


Figure 9. QQ plot module

3.5.3 Plot of LOD Score against Genome Position

Use **Plot of LOD Score against Genome Position** module to redraw the plot. The parameter settings are the same as those in the Manhattan plot. Users may set the color of LOD line. Use **Save plot** button to choose a path and to save the Figure, with four frequently used image formats: *.png, *.tiff, *.jpeg and *.pdf (**Fig 10**).

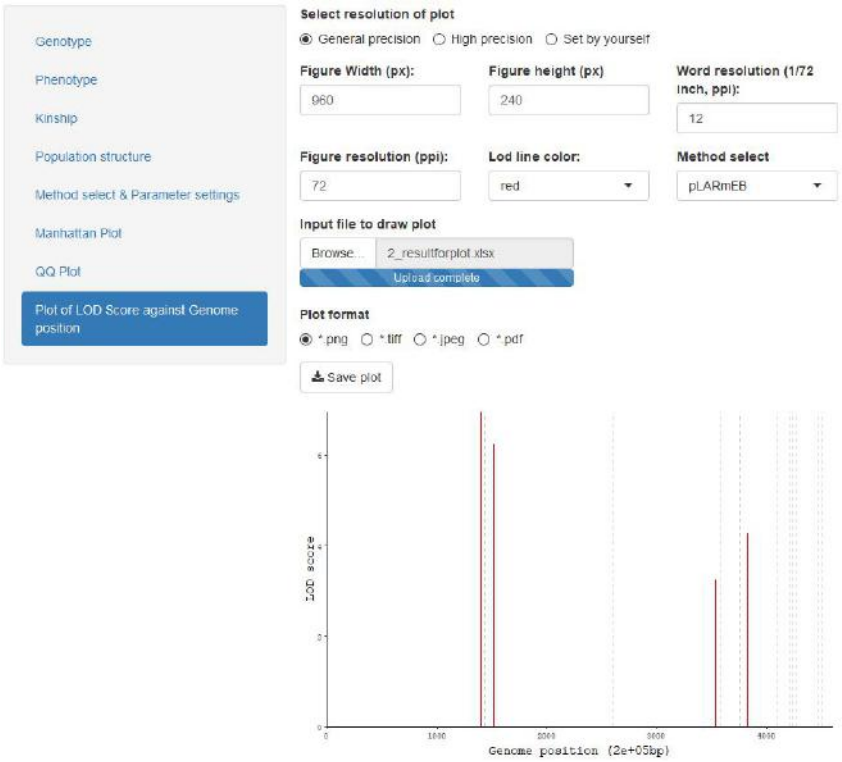


Figure 10. Plot of LOD score against Genome Position

4. 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.

In the **plot results**, there are ten sheets, including "Manhattan mrMLM", "qq mrMLM", "Manhattan FASTmrMLM", "qq FASTmrMLM", "Manhattan FASTmrEMMA", "qq FASTmrEMMA", "Plot pLARmEB", "Manhattan pKWmEB", "qq pKWmEB", "Plot ISIS EM-BLASSO". These plot results will be saved to the ten sheets if users selected all the methods. Users may upload this file into mrMLM.GUI so that all the figures may be adjusted based on your opinions.

5. 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*** 2018, **19**(4): 700–712. 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, Zhang Yuan-Ming*. A fast mrMLM algorithm for multi-locus genome-wide association studies. ***bioRxiv*** 341784; doi: https://doi.org/10.1101/341784, Posted June 7, 2018
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.