

# User Manual for

# m r MLM

multi-locus random-SNP-effect Mixed Linear Model tools for  
multi-locus GWAS and linkage analysis

([version 2.1](#))

**Ren Wen-Long, Ni Yuan-Li, Wen Yang-Jun, Wang Shi-Bo,  
Huang Bo, Cox Lwaka Tamba, Zhang Jin,  
Zhang Yuan-Ming ([soyzzhang@mail.hzau.edu.cn](mailto:soyzzhang@mail.hzau.edu.cn))**

**Last updated on January 18, 2017**

**Disclaimer:** While extensive testing has been performed by Yuan-Ming Zhang's Lab (Statistical Genomics Lab) at Huazhong Agricultural University and Nanjing 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 EMMA, GAPIT v2 and PLINK.

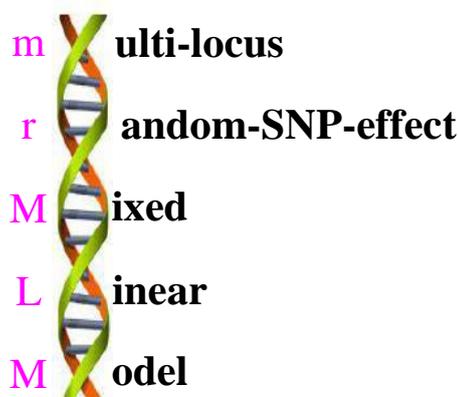
**Download website:**

<https://cran.r-project.org/web/packages/mrMLM/index.html>

**Citation:**

Method	References
mrMLM	Wang et al. <i>Scientific Reports</i> 2016, <b>6</b> :19444 <sup>[1]</sup>
GCIM	Wang et al. <i>Scientific Reports</i> 2016, <b>6</b> :29951 <sup>[2]</sup>
FASTmrEMMA	Wen et al. <i>Briefings in Bioinformatics</i> 2017, Accepted, DOI: 10.1093/bib/bbw145 <sup>[3]</sup>
ISIS EM-BLASSO	Tamba et al. <i>PLoS Computational Biology</i> 2017, Accepted, DOI: 10.1371/journal.pcbi.1005357 <sup>[4]</sup>
pLARmEB	Zhang et al. <i>Heredity</i> 2017, Accepted <sup>[5]</sup>

Note: These references are listed in section of Reference.



This work was supported by the National Natural Science Foundation of China (31571268), Huazhong Agricultural University Scientific & Technological Self-innovation Foundation (Program No. 2014RC020) and Nanjing Agricultural University.



# Contents

<b>1 INTRODUCTION</b>	<b>4</b>
<b>1.1 Why mrMLM ?</b>	<b>4</b>
<b>1.2 Getting started</b>	<b>5</b>
<b>1.2.1 Online install</b>	<b>5</b>
<b>1.2.2 Offline install</b>	<b>5</b>
1.2.2.1 Install GTK+	5
1.2.2.2 Install the add-on packages	6
1.2.2.3 Install mrMLM	6
<b>2 DATASETS</b>	<b>7</b>
<b>2.1 GWAS datasets</b>	<b>7</b>
<b>2.1.1 Phenotypic data</b>	<b>7</b>
<b>2.1.2 Genotypic data</b>	<b>7</b>
2.1.2.1 Numeric format	7
2.1.2.2 Character format	8
2.1.2.3 Hapmap format	9
<b>2.1.3 Kinship</b>	<b>9</b>
<b>2.1.4 Population structure</b>	<b>10</b>
<b>2.2 Linkage analysis datasets</b>	<b>10</b>
<b>2.2.1 GCIM format</b>	<b>11</b>
2.2.1.1 Phenotypic data	11
2.2.1.2 Genotypic data	11
2.2.1.3 Linkage map data	12
<b>2.2.2 WinQTLCart format</b>	<b>12</b>
<b>2.2.3 QTLIciMapping format</b>	<b>13</b>
2.2.3.1 Phenotypic data	13
2.2.3.2 Genotypic data	13
2.2.3.3 Linkage map data	14
<b>2.2.4 Covariate format</b>	<b>14</b>

<b>3 ANALYSIS AND RESULTS</b>	<b>15</b>
<b>3.1 mrMLM module</b>	<b>15</b>
3.1.1 Input dataset	15
3.1.2 Run program	17
3.1.3 Output results	19
3.1.4 Manhattan plot	20
3.1.5 QQ plot	21
<b>3.2 FASTmrEMMA module</b>	<b>23</b>
3.2.1 Input data	23
3.2.2 Run program	25
3.2.3 Output results	27
3.2.4 Manhattan plot	29
3.2.5 QQ plot	30
<b>3.3 ISIS EM-BLASSO module</b>	<b>31</b>
3.3.1 Input data	31
3.3.2 Run program	33
3.3.3 Output results	34
3.3.4 LOD Scores plot	35
<b>3.4 pLARmEB module</b>	<b>36</b>
3.4.1 Input Data	36
3.4.2 Run program	37
3.4.3 Output results	39
3.3.4 LOD Scores plot	40
<b>3.4 GCIM module</b>	<b>41</b>
3.4.1 Input data and Parameter setting	41
3.3.2 Run program	44
3.3.3 Output results	46
3.3.4 Draw plot	47
<b>4 REFERENCE</b>	<b>48</b>

# 1 INTRODUCTION

## 1.1 Why mrMLM?

**mrMLM** (**m**ulti-locus **r**andom-SNP-effect **M**ixed **L**inear **M**odel) software package is an R package with interactive graphic user interface (GUI) for multi-locus genome-wide association study (GWAS) and multi-locus linkage analysis. At present it (version 2.1) includes five modules, which are 1) mrMLM, 2) FASTmrMLM (**F**AST **m**ulti-locus **r**andom-SNP-effect **E**MMA), 3) ISIS EM-BLASSO (**I**terative **S**ure **I**ndependence **S**creening **E**M-**B**ayesian **L**ASSO), 4) pLARmEB (polygene-background-control-based Least Angle Regression plus Empirical Bayes) and 5) GCIM (**G**enome-wide **C**omposite **I**nterval **M**apping). The first four modules (mrMLM, FASTmrMLM, ISIS EM-BLASSO and pLARmEB) are used to conduct multi-locus GWAS, and the last one (GCIM) is used to implement multi-locus linkage analysis.

(i) **mrMLM**, aims to provide a user-friendly interface to conduct multi-locus GWAS via mrMLM methodology and to visualize its results. Its visualization is based on package [qqman](#), which can help draw figures such as Manhattan and QQ plots.

(ii) **FASTmrEMMA**, aims to provide a user-friendly interface to conduct multi-locus GWAS via FASTmrEMMA methodology and to visualize its results. Its visualization is also based on package [qqman](#), which can help draw figures such as Manhattan and QQ plots.

(iii) **ISIS EM-BLASSO**, aims to provide a user-friendly interface to conduct multi-locus GWAS via ISIS EM-BLASSO methodology and to visualize its results. Its visualization is based on package [ggplot2](#), which can help draw plot of LOD Scores.

(iv) **pLARmEB**, aims to provide a user-friendly interface to conduct multi-locus GWAS via pLARmEB methodology and to visualize its results. Its visualization is also based on package [ggplot2](#), which can help draw plot of LOD Scores.

(v) **GCIM**, aims to provide a user-friendly interface to conduct multi-locus linkage

analysis via GCIM methodology and to visualize its results. Its visualization is based on package [graphics](#), which can help draw figures about the plot of LOD score ( $y_1$  axis) and  $-\log_{10}(\text{P-value})$  ( $y_2$  axis) against genome position (cM).

mrMLM 2.1 is able to work on the popular platforms, like Windows, Linux (desktop) and MacOS. And the interactive GUI is based on available add-on package [RGtk2](#), via the aid of another package [gWidgetsRGtk2](#).

## 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 ([soyzzhang@mail.hzau.edu.cn](mailto:soyzzhang@mail.hzau.edu.cn) or [soyzzhang@hotmail.com](mailto:soyzzhang@hotmail.com)).

### 1.2.1 Online install

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

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

### 1.2.2 Offline install

#### 1.2.2.1 Install GTK+

You may need to install GTK+ before installing RGtk2, because RGtk2 depends on GTK+.

For **Windows** user, you do as below:

Download GTK+ here

(<http://sourceforge.net/projects/gladewin32/files/gtk%2B-win32-runtime/2.10.11/gtk-2.10.11-win32-1.exe>).

Run the resulting file ([gtk-2.10.11-win32-1.exe](#)), which is an automated installer that will help you to complete the installation of Gtk2 libraries. If you use **64-bit** R software, please download corresponding GTK+ version.

For **Mac OS** users, you do as below:

Download GTK+ here (<http://sourceforge.net/projects/gtk-osx/files/latest/download>).

Extract and run the resulting file ([gtk-osx-docbook-1.2.tar.gz](http://sourceforge.net/projects/gtk-osx/files/latest/download)).

For **Linux** users, you do as below:

You may or may not upgrade the GTK libraries depending on your distribution.

There are more details on RGtk2 at RGtk2's home page (<http://www.ggobi.org/rgtk2/>).

### 1.2.2.2 Install the add-on packages

The following R packages are needed: RGtk2, cairoDevice, gWidgets, gWidgetsRGtk2, RGtk2Extras and qqman, which can be downloaded from CRAN (<https://cran.r-project.org/>). To install them in order, as some depend on others. Within R environment, these packages can be installed directly using the below command:

```
install.packages(pkgs=c("RGtk2","cairoDevice","gWidgets","gWidgetsRGtk2","RGtk2Extras","qqman","openxlsx","stringr","MASS","lars","ncvreg","ggplot2"))
```

### 1.2.2.3 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)`, then the following dialog will appear.

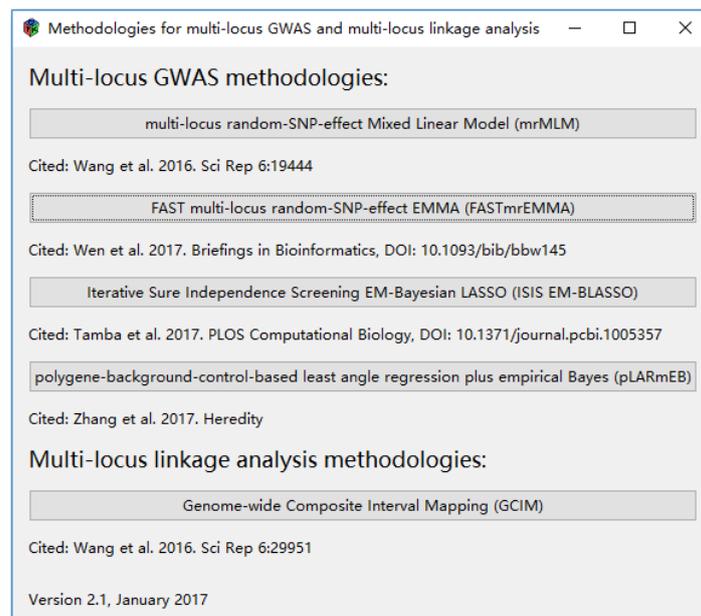


Figure 1.1. Screenshot of **mrMLM** package GUI

To restart the GUI, the command `mrMLM()` can be issued.

**Note:** Users should install RStudio software (<https://www.rstudio.com/>) in order to obtain the [User Manual.pdf](#) file from the **User Manual** button. If the RStudio package is not installed, users can decompress the mrMLM package and find the User Manual file (name: [Instruction.pdf](#)) in the folder of ".../mrMLM/inst/doc".

## 2 DATASETS

### 2.1 GWAS datasets

#### 2.1.1 Phenotypic data

The **Phenotypic** file should be a `*.csv` format file. The first column stands for individual ID, such as 33-16, Nov-38 and 4226. The second column is phenotypic values for the trait. Note that the phenotypic file includes only one trait. The first element in the first column must be "**<Phenotype>**" or "**<Trait>**".

	A	B
1	<Phenotype>	
2	B46	42
3	B52	72.5
4	B57	41
5	B64	74.5
6	B68	65
7	B73	83.25
8	B73HTRHM	73
9	E75	56
10	E76	53
11	E77	50
12	E79	66.5
13	CM7	28
14	CML10	110
15	CML103	88.5
16	CML108	55.5

Figure 2.1. The Phenotypic file

#### 2.1.2 Genotypic data

The **Genotypic** file should be a `*.csv` format file.

##### 2.1.2.1 Numeric format

The first column, named "**rs#**" in the first row, stands for marker ID. The second column, named "**chrom**" in the first row, stands for chromosome. The third column, named "**pos**" in the first row, stands for the position (bp) of SNP in the chromosome. The fourth column, named "**genotype for code 1**" in the first row, stands for reference bases. If the base is missing, the observed base for this marker in the first individual is what we list. And each of the remaining columns is for one individual. In their first rows, the individual names are appeared. 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. For the **FASTmrEMMA** module, homozygous genotypes are expressed by 1 (AA) and 0 (aa), respectively, and the heterozygous and missing genotypes are indicated by 0.5.

	A	B	C	D	E	F	G	H
1	rs#	chrom	pos	genotype	33-16	Nov-38	A4226	A4722
2	PZB00859.1	1	157104	C	1	1	1	1
3	PZA01271.1	1	1947984	C	1	-1	1	-1
4	PZA03613.2	1	2914066	G	1	1	1	1
5	PZA03613.1	1	2914171	T	1	1	1	1
6	PZA03614.2	1	2915078	G	1	1	1	1
7	PZA03614.1	1	2915242	T	1	1	1	1
8	PZA02117.1	1	223466480	A	1	1	1	-1
9	PZA00403.5	1	223466873	T	1	1	1	0
10	PZB01979.2	1	224421551	A	1	-1	1	-1
11	PZB01979.3	1	224421962	C	1	1	-1	1
12	PZA00658.21	1	224518409	T	0	1	1	-1

Figure 2.2 The genotypic file with numeric format in **mrMLM** module

	A	B	C	D	E	F	G	H
1	rs#	chrom	pos	genotype	33-16	Nov-38	A4226	A4722
2	PZB00859.1	1	157104	C	1	1	1	1
3	PZA01271.1	1	1947984	C	1	0	1	0
4	PZA03613.2	1	2914066	G	1	1	1	1
5	PZA03613.1	1	2914171	T	1	1	1	1
6	PZA03614.2	1	2915078	G	1	1	1	1
7	PZA03614.1	1	2915242	T	1	1	1	1
8	PZA02117.1	1	223466480	A	1	1	1	0
9	PZA00403.5	1	223466873	T	1	1	1	0.5
10	PZB01979.2	1	224421551	A	1	0	1	0
11	PZB01979.3	1	224421962	C	1	1	0	1
12	PZA00658.21	1	224518409	T	0.5	1	1	0

Figure 2.3 The genotypic file with numeric format in **FASTmrEMMA** module

### 2.1.2.2 Character format

The first three columns are same as those in the “**Figure 2.2 The genotypic file with numeric format in mrMLM module**”. 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 code.

	A	B	C	D	E	F	G
1	rs#	chrom	pos	33-16	Nov-38	A4226	A4722
2	PZB00859.1	1	157104	C	C	C	C
3	PZA01271.1	1	1947984	C	G	C	G
4	PZA03613.2	1	2914066	G	G	G	G
5	PZA03613.1	1	2914171	T	T	T	T
6	PZA03614.2	1	2915078	G	G	G	G
7	PZA03614.1	1	2915242	T	T	T	T
8	PZA02117.1	1	223466480	A	A	A	G
9	PZA00403.5	1	223466873	T	T	T	N
10	PZB01979.2	1	224421551	A	G	A	G
11	PZB01979.3	1	224421962	C	C	G	C
12	PZA00658.21	1	224518409	N	T	T	C

Figure 2.4 The genotypic file with character format

### 2.1.2.3 Hapmap format

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, the **no available** information must be marked by "NA". All the individual genotypic information will be showed from the 12 to last columns, the first element in each column is individual ID (name) and the others are the genotypes (character).

	A	B	C	D	E	F	G	H	I	J	K	L
1	rs#	alleles	chrom	pos	strand	assembly#	center	protLSID	assayLSID	panel	QCcode	33-16
2	PZB00859.1	A/C	1	157104	+	AGPv1	Panzea	NA	NA	maize282	NA	CC
3	PZA01271.1	C/G	1	1947984	+	AGPv1	Panzea	NA	NA	maize282	NA	CC
4	PZA03613.2	G/T	1	2914066	+	AGPv1	Panzea	NA	NA	maize282	NA	GG
5	PZA03613.1	A/T	1	2914171	+	AGPv1	Panzea	NA	NA	maize282	NA	TT
6	PZA03614.2	A/G	1	2915078	+	AGPv1	Panzea	NA	NA	maize282	NA	GG
7	PZA03614.1	A/T	1	2915242	+	AGPv1	Panzea	NA	NA	maize282	NA	TT
8	PZA02117.1	A/G	1	223466480	+	AGPv1	Panzea	NA	NA	maize282	NA	AA
9	PZA00403.5	C/T	1	223466873	+	AGPv1	Panzea	NA	NA	maize282	NA	TT
10	PZB01979.2	A/G	1	224421551	+	AGPv1	Panzea	NA	NA	maize282	NA	AA
11	PZB01979.3	C/G	1	224421962	+	AGPv1	Panzea	NA	NA	maize282	NA	CC
12	PZA00658.21	C/T	1	224518409	+	AGPv1	Panzea	NA	NA	maize282	NA	NN

Figure 2.5 The genotypic file with Hapmap format

### 2.1.3 Kinship

The **Kinship** file should be a \*.csv file. The number of rows (or columns) equals to the number of the common individuals between the phenotypic and genotypic datasets. If the Kinship matrix is calculated by this software, we calculate only the Kinship matrix between the common individuals. If the Kinship matrix has been obtained and uploaded from a known file, it is possible that the number and order of individuals in the known Kinship file are not consistent with those of the common (valid) individuals in the Phenotype and Genotype files. At this situation, the software will change the known K matrix in order that the number and order of new K matrix matches the number and order of common (valid) individuals in the Phenotypic and Genotypic files. In the known K matrix, the **first element** in the first column must be **the number of valid individuals**, such as 263; and the other elements in the first column are individual ID (names). If the number of markers is very large, i.e., 50,000, we recommend that users calculate the K matrix using the other softwares, especially

for FASTmrEMMA.

	A	B	C	D	E	F
1	263					
2	33-16	1.00809	0.45954	0.50677	0.42503	0.45591
3	Nov-38	0.45954	1.03352	0.43048	0.47044	0.39597
4	A4226	0.50677	0.43048	1.01717	0.45409	0.43775
5	A4722	0.42503	0.47044	0.45409	0.89002	0.34874
6	A188	0.45591	0.39597	0.43775	0.34874	1.0099
7	A214N	0.34693	0.33421	0.39779	0.29244	0.33058
8	A239	0.43593	0.46499	0.40323	0.36691	0.39597
9	A272	0.34874	0.40505	0.31423	0.3887	0.44138
10	A441-5	0.47952	0.44138	0.47226	0.47952	0.49224
11	A554	0.39779	0.45954	0.5431	0.48679	0.4214
12	A556	0.50858	0.40505	0.45954	0.40142	0.40687

Figure 2.6. The Kinship file

### 2.1.4 Population structure

The **Population Structure** file is a \*.csv file. Using the **Structure** software, the population structure matrix may be calculated. The first column stands for the valid individual ID (names). The first and second elements in the first column must be "<Covariate>" and "<Trait>" respectively. If population structure matrix has  $k$  columns, please input all the  $k$  columns. In the second row, it must be "Q1", "Q2", ... , "Qk" following the "<Trait>".

If the **Population Structure** file has been obtained and uploaded from a known file, it is possible that the number and order of individuals in the known file are not consistent with those of the common (valid) individuals in the further analysis. At this situation, the software will change the known matrix in order that the number and order of new **Population Structure** matrix match the number and order of common (valid) individuals in the Phenotypic and Genotypic files.

	A	B	C
1	<Covariate>		
2	<Trait>	Q1	Q2
3	33-16	0.972	0.014
4	Nov-38	0.993	0.004
5	A4226	0.917	0.012
6	A4722	0.854	0.111
7	A188	0.982	0.005
8	A214N	0.017	0.221
9	A239	0.963	0.002
10	A272	0.122	0.859
11	A441-5	0.531	0.464
12	A554	0.979	0.002

Figure 2.7. The Population Structure file

## 2.2 Linkage analysis datasets

## 2.2.1 GCIM format

### 2.2.1.1 Phenotypic data

The **Phenotypic** file should be a \*.csv format file. In the first column, “phenotype” is appeared in the first row and **individual ID or individual names** are listed in the other rows, such as DH6-10, DH6-101 and DH6-101. In the second column, the phenotypic values for the first trait are listed. The trait name is appeared in the first row, such as DS1–BLUEs, and observations are listed in the other rows. In the other columns, it is similar to the column 2. Missing phenotypic values is indicated by **NA**.

	A	B
1	phenotype	T19
2	DH6-10	75.33
3	DH6-101	105
4	DH6-102	96.33
5	DH6-104	81
6	DH6-105	101.67
7	DH6-108	89
8	DH6-11	82.67
9	DH6-111	133.33
10	DH6-112	105
11	DH6-114	NA
12	DH6-119	104.67
13	DH6-124	75.67
14	DH6-125	113.33

Figure 2.8. The Phenotypic file with **GCIM** format

### 2.2.1.2 Genotypic data

The **Genotypic** file should be a \*.csv format file. In the first column, "**genotype**" is showed in the first row and marker names are appeared from row 2 to row  $m+1$  ( $m$  is the number of markers). In the column 2, **individual ID or individual names** is appeared in the first row and marker genotypes for all the markers are listed in the other rows. For the other rows, it is similar to the column 2.

**Population Type.** At present, GCIM can analyze five population types derived from two parental lines.  $F_1 = P_1 \times P_2$ .

1.  $BC_1: F_1 \times P_1$ . 1 stands for AA, -1 stands for Aa, and 99 stands for missing genotypes.

2.  $BC_2: F_1 \times P_2$ . 1 stands for Aa, -1 stands for aa, and 99 stands for missing genotypes.

3. **DH:** doubled haploid lines derived from  $F_1$ . 1 stands for AA, -1 stands for aa,

and 99 stands for missing genotypes.

4. RIL: recombinant inbred lines derived from repeatedly selfing since F<sub>1</sub>. 1 stands for AA, -1 stands for aa, and 99 stands for missing genotypes.

5. Chromosome Segment Substitution Line (CSSL): 1 stands for AA, -1 stands for aa, 99 stands for missing genotypes.

	A	B	C	D	E	F	G	H
1	genotype	DH6-10	DH6-101	DH6-102	DH6-104	DH6-105	DH6-108	DH6-11
2	RGA3(1)	-1	99	-1	1	-1	-1	1
3	wPt-6358	-1	99	99	99	99	-1	1
4	Hplc2	1	1	-1	1	-1	-1	1
5	wPt-9752	1	99	99	99	99	99	1
6	abc156a	1	1	-1	1	-1	-1	-1
7	RGA36b(2)	1	99	-1	1	99	-1	-1
8	bcd98	-1	1	-1	1	-1	-1	-1
9	wmc24	-1	1	-1	1	-1	-1	-1
10	ksuG9c	-1	1	-1	1	-1	-1	-1
11	wPt-2436	-1	99	99	99	99	-1	-1
12	wPt-4886	-1	99	99	99	99	-1	-1

Figure 2.9. The genotypic file with **GCIM** format

### 2.2.1.3 Linkage map data

The **Linkage Map** file should be a \*.csv format file. In the first column, “marker” is appeared in the first row, and marker names are listed in the other rows. In the second column, “chr” is chromosome and appeared in the first row, and chromosome information for all the markers are listed in the other rows. In the third column, “pos” is marker position information, marker position (cM) are listed in the other rows. Note that the marker position is not interval length.

	A	B	C
1	marker	chr	pos
2	RGA3(1)	1	0
3	wPt-6358	1	3.034
4	Hplc2	1	8.8291
5	wPt-9752	1	10.1452
6	abc156a	1	41.3408
7	RGA36b(2)	1	43.8429
8	bcd98	1	51.9122
9	wmc24	1	54.5814
10	ksuG9c	1	55.3333
11	wPt-2436	1	59.1871
12	wPt-4886	1	61.7209
13	wmc120	1	62.4012
14	cdo105	1	63.0679

Figure 2.10. The linkage map file with **GCIM** format

### 2.2.2 WinQTLCart format

Please see the WinQTLCart software in details. Here we introduce simply. Between the first “-start” and “-stop” information, marker information is listed, including **chromosome name, marker name and position (interval length)**. Between the second “-start” and “-stop”, genotypic information is showed. Between the third “-start” and “-stop”, phenotypic information is appeared. If variable information is included, it will be **otrait** information between the fourth “-start” and “-stop”.

```
#FileID 1438801284
#bychromosome
-type position
-function 1
-Units cM
-chromosomes 3
-maximum 13
-named yes
-start
-Chromosome C1
AXR-1 0.0000
HH.335C-COL 9.3000
EC.480C 17.2000
EC.650C 29.9000
GC.845L 38.7000
G23554 52.8000
CH.160L-COL 57.8000
DF.701L 72.4000
AD.333C 76.6000
GBB.80L 93.2000
GD.97L 97.0000
FD.8411L 115.5000
MXF-F-001 116.5000
```

Figure 2.11. The mcd file in the **WinQTLCart** software

## 2.2.3 QTLIciMapping format

### 2.2.3.1 Phenotypic data

All the information for each trait is listed in one row. The first column stands for trait ID. And each of the remaining columns stands for the phenotypic value of every individual. Note that -100 stands for missing phenotypes.

	A	B	C	D	E	F	G	H	I	J	K	L
1	T15	103	111.7	106.7	92.67	122	93.67	108	118.3	102.3	-100	130
2	T19	75.33	105	96.33	81	101.7	89	82.67	133.3	105	-100	104.7

Figure 2.12. The phenotypic format in the QTL **IciMapping** software

### 2.2.3.2 Genotypic data

All the information for each marker is listed in one row. The first column stands for marker ID. And each of the remaining columns stands for one individual. In QTL IciMapping, 2 was used to represent the marker genotype of the first parent (P<sub>1</sub>), 0 for the second parent (P<sub>2</sub>), 1 for the F<sub>1</sub>, and -1 for missing genotypes.

	A	B	C	D	E	F	G	H	I	J	K
1	RGA3(1)	0	-1	0	2	0	0	2	2	2	-1
2	wPt-6358	0	-1	-1	-1	-1	0	2	2	2	-1
3	Hplc2	2	2	0	2	0	0	2	2	2	0
4	wPt-9752	2	-1	-1	-1	-1	-1	2	2	2	-1
5	abc156a	2	2	0	2	0	0	0	0	2	0
6	RGA36b(2)	2	-1	0	2	-1	0	0	0	2	-1
7	bcd98	0	2	0	2	0	0	0	0	2	0
8	wmc24	0	2	0	2	0	0	0	0	2	0
9	ksuG9c	0	2	0	2	0	0	0	0	2	0
10	wPt-2436	0	-1	-1	-1	-1	0	0	0	0	-1
11	wPt-4886	0	-1	-1	-1	-1	0	0	0	0	-1
12	wmc120	0	2	0	2	0	0	0	0	0	2

Figure 2.13. The genotypic file in the QTL **IciMapping** software

### 2.2.3.3 Linkage map data

The linkage map file has three columns. The first column stands for marker ID, the second column stands for chromosome, and the third column is marker position (cM) in the chromosome, which is not the length of marker interval.

	A	B	C
1	RGA3(1)	1	0
2	wPt-6358	1	3.034
3	Hplc2	1	8.8291
4	wPt-9752	1	10.1452
5	abc156a	1	41.3408
6	RGA36b(2)	1	43.8429
7	bcd98	1	51.9122
8	wmc24	1	54.5814
9	ksuG9c	1	55.3333
10	wPt-2436	1	59.1871
11	wPt-4886	1	61.7209
12	wmc120	1	62.4012

Figure 2.14. The Linkage map format in the QTL **IciMapping** software

### 2.2.4 Covariate format

The **Covariate** file should be a \*.csv format file. In the first column, “covariate” is appeared in the first row and **individual names** or **individual ID**, such as DH6-10, DH6-101 and DH6-101, are listed in the other rows. In the second column, the name for the first covariate is appeared in the first row and covariate information is listed in the other row. The other columns are similar to the second column if there are several covariates.

	A	B
1	covariate	EnvironName
2	DH6-10	A
3	DH6-101	A
4	DH6-102	A
5	DH6-104	A
6	DH6-105	A
7	DH6-108	A
8	DH6-11	A
9	DH6-111	A
10	DH6-112	A
11	DH6-114	A
12	DH6-119	A

Figure 2.15. The covariate file with **GCIM** format

### 3 ANALYSIS AND RESULTS

#### 3.1 mrMLM module

Click the button “[Multi-locus Random-SNP-effect Mixed Linear Model \(mrMLM\)](#)”, then the following dialog will appear.

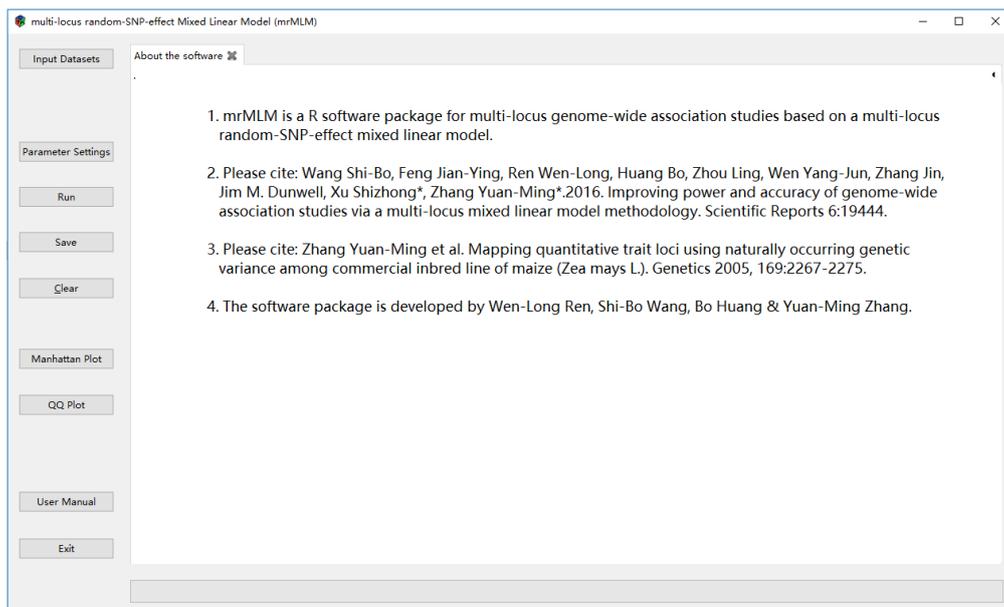


Figure 3.1.1. Screenshot of **mrMLM** module GUI

##### 3.1.1 Input dataset

Use the **Input Dataset** button to input dataset files, and then a dialog box will be appeared. In the dialog box, there are four steps. First, users select the dataset formats, which include **mrMLM numeric** format, **mrMLM character** format and **hapmap** format used in the TASSEL software. Then, use the **Genotype** and **Phenotype** buttons to input the genotypic and phenotypic datasets, respectively. Once one file is

successfully uploaded, one tabbed page is added to the software notebook. Third, two things will be implemented in this step. One is to sort the individuals between the genotypic and phenotypic files and all the common individuals between the two files are selected to be analyzed in the further analyses. Another is to transfer the character genotypes into the numeric genotypes if the genotypes are character. Once users press the **DO** button, the two things will be conducted. Once the two files will be successfully uploaded, two tabbed pages (Genotype and Phenotype) will be added to the software notebook. Finally, use the **Kinship** and **Population Structure** buttons to input the kinship and population structure matrices, respectively. If one file is successfully uploaded, the corresponding data page will be added to the notebook. Note that the **Kinship** and **Population Structure** buttons have two options. For the **kinship** button, one is to directly upload the kinship matrix and another is to calculate the kinship matrix using this software. For the **Population Structure** button, the population structure matrix may be not included in the mixed linear model of the GWAS if it has no effect on GWAS. If not, it should be included in the mixed model. The population structure matrix in your uploaded file will be deleted one column if the sum of all the Q-matrix columns for one individual (one row) equal to 1. In the filter dialog, you should choose one column that should be deleted.

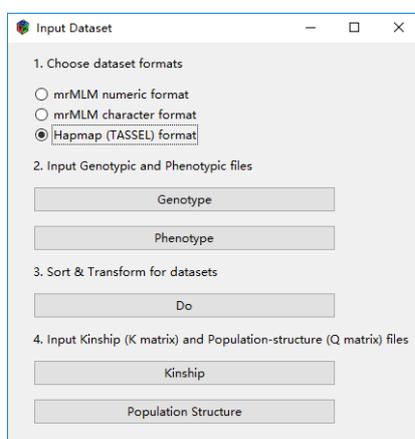


Figure 3.1.2. The Input Dataset dialog of **mrMLM** module

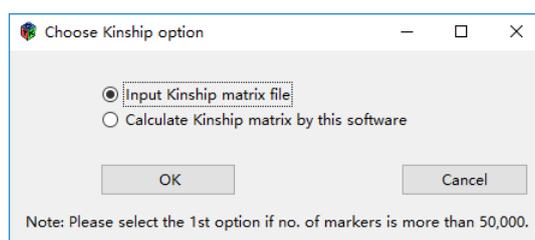


Figure 3.1.3. The Kinship dialog of **mrMLM** module

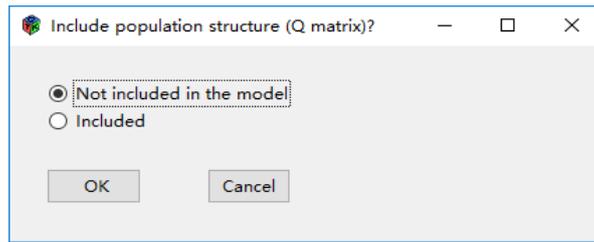


Figure 3.1.4. The population structure dialog of **mrMLM** module

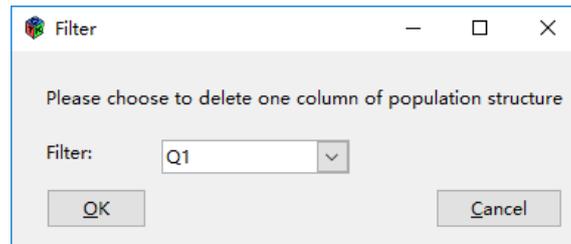


Figure 3.1.5. Filter dialog of the population structure of **mrMLM** module

### 3.1.2 Run program

Use the **Parameter Setting** button to set parameters before running the program. “Search radius of candidate gene (kb)” means to keep the one marker with having the least P-value, and to delete all the other markers within the radius of the associated marker with the least P-value. Use the **Run** button to execute the software. If the program runs, a progress bar with the “**Please be patient...**” words will appear in the bottom of the interface. If the program finished, a bar with the “**All done.**” will appear.

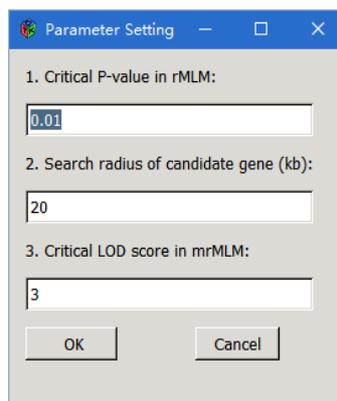


Figure 3.1.6. The Parameter Setting dialog of **mrMLM** module

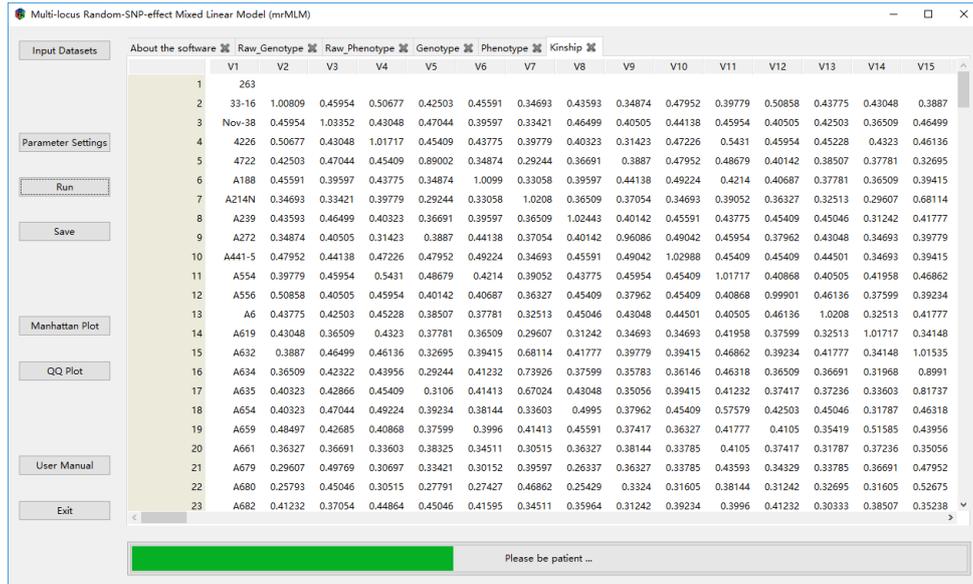


Figure 3.1.7. A running program interface of mrMLM module

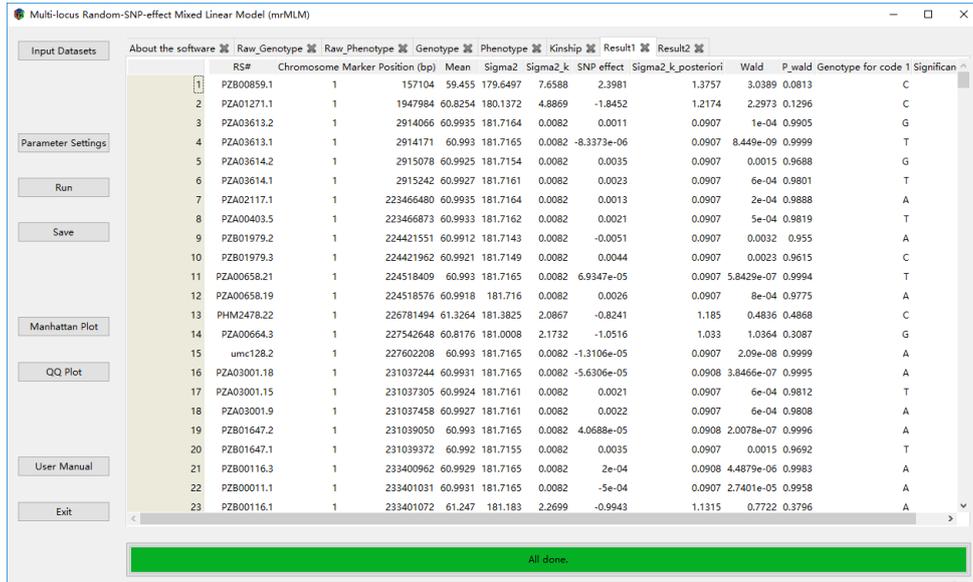


Figure 3.1.8. A finished program interface (the rMLM Results: Result1)

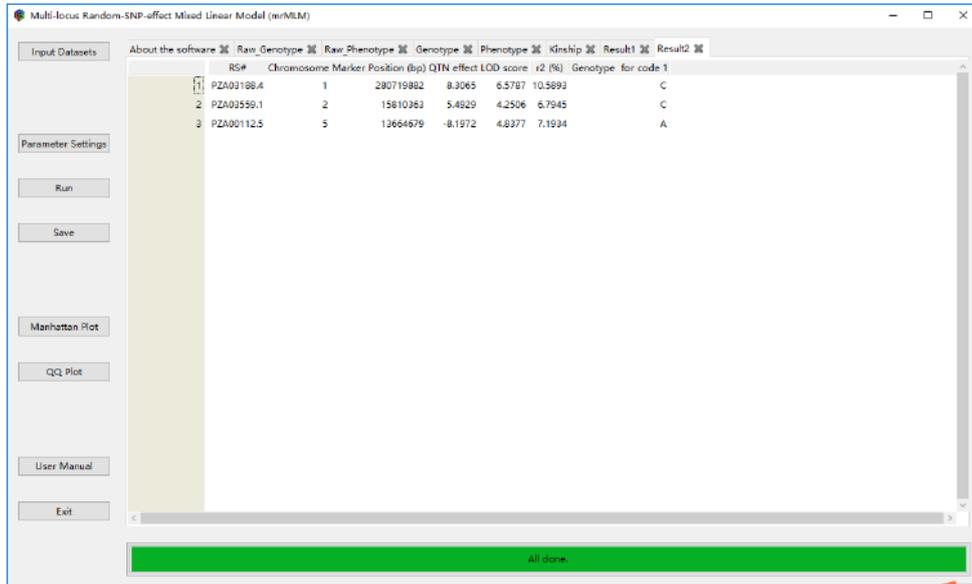


Figure 3.1.9. A finished program interface (the **mrMLM** Results: [Result2](#))

### 3.1.3 Output results

Use **Save** button to save the results as \*.csv files. The **Results in the rMLM** are saved as [Result1.csv](#) and the **Results in the mrMLM** are saved as [Result2.csv](#). If click **OK** button (after the **Save** button), a dialog is used to choose the pathway and the file name for the saving files.

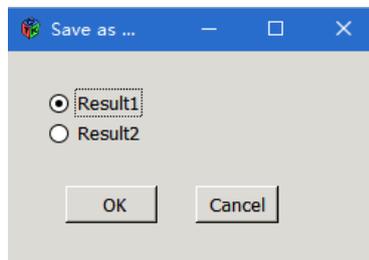


Figure 3.1.10. The result **Save** dialog for the **rMLM** and **mrMLM** methods

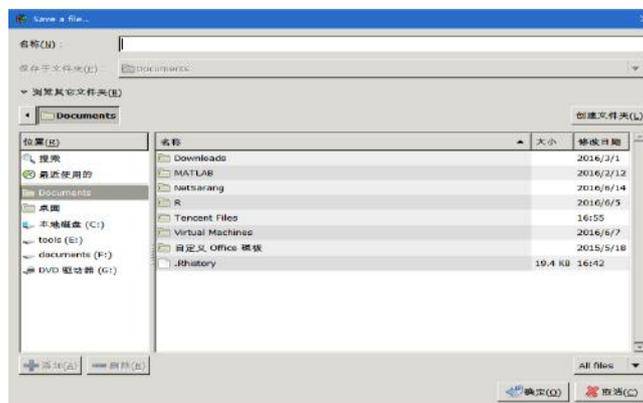


Figure 3.1.11. The **Save** dialog

The **Result1** table with twelve columns shows the results from the rMLM (random-SNP-effect mixed linear model) method. The corresponding column names are as follows: reference sequence number (rs#, marker name), chromosome, marker's position (bp) in the chromosome, population mean value (Mean), residual variance ( $\sigma^2$ , Sigma2), priori variance of the  $k$ th SNP effect ( $\phi_k^2$ , Sigma2\_k), SNP effect ( $\gamma_k$ , Effect), posteriori variance of SNP effect ( $\text{var}(\gamma_k)$ , Sigma2\_k\_posteriori), Wald test statistic value, the P-value of Wald test, significance and genotype for code 1, respectively. In the significance column, only significant markers under the critical value  $0.05/m_e$  are marked.

	A	B	C	D	E	F	G	H	I	J	K	L	
1	RS#	Chromosome	Marker	Position (bp)	Mean	Sigma2	Sigma2_k	SNP effect	Sigma2_k_posteriori	Wald	P_wald	Genotype for code 1	Significance
2	PZB00859.1	1		157194	59.455	179.0497	7.0588	2.3981		1.3707	3.0589	0.0813	C
3	PZA01371.1	1		1947994	60.6254	180.1372	6.8899	-1.8452		1.2174	2.2973	0.1290	C
4	PZA03513.2	1		2914096	60.9935	181.7164	0.0082	0.0011		0.0907	1.00E-04	0.9995	C
5	PZA03813.1	1		2914171	60.993	181.7165	0.0082	-8.34E-09		0.0907	8.45E-09	0.9999	T
6	PZA03814.2	1		2915078	60.9925	181.7154	0.0082	0.0035		0.0907	0.0015	0.9688	C
7	PZA03814.1	1		2915242	60.9927	181.7161	0.0082	0.0023		0.0907	6.00E-04	0.9801	T
8	PZA02117.1	1		223484480	60.9935	181.7164	0.0082	0.0013		0.0907	2.00E-04	0.9888	A
9	PZA00403.5	1		233466873	60.9933	181.7162	0.0082	0.0021		0.0907	5.00E-04	0.9819	T
10	PZB01979.2	1		224421551	60.9912	181.7143	0.0082	-0.0051		0.0907	0.0032	0.955	A
11	PZB01979.3	1		224421962	60.9921	181.7149	0.0082	0.0044		0.0907	0.0023	0.9615	C
12	PZA00858.21	1		224518409	60.993	181.7165	0.0082	6.93E-05		0.0907	5.84E-07	0.9994	T
13	PZA00858.19	1		224518576	60.9918	181.716	0.0082	0.0026		0.0907	8.00E-04	0.9775	A
14	FW2478.22	1		226781494	61.3264	181.3825	2.0667	-0.8241		1.185	0.4836	0.4868	C

Figure 3.1.12. Results in the rMLM (Result1)

The **Result2** table with seven columns shows the final results of the mrMLM (multi-locus random-SNP-effect mixed linear model) method. The corresponding column names are as follows: reference sequence number (rs#, marker names), chromosome, marker's position (bp) in the chromosome, QTN effect, LOD score, the proportion of phenotypic variance explained by the putative QTN, and genotype for code 1, respectively.

	A	B	C	D	E	F	G	H	I
1	RS#	Chromosome	Marker	Position (bp)	QTN effect	LOD score	r2 (%)	Genotype	for code 1
2	PZA03188.4	1		280719882	8.3065	6.5787	10.5893	C	
3	PZA03559.1	2		15810363	5.4929	4.2506	6.7945	C	
4	PZA00112.5	5		13664679	-8.1972	4.8377	7.1934	A	
5									
6									
7									
8									

Figure 3.1.13. Results in the mrMLM (Result2)

### 3.1.4 Manhattan plot

Use **Manhattan plot** button to preview Manhattan plot in an independent dialog window. Before saving the Figure, please set [the width and height of the Figure](#), with the unit of pixel (px). And set [word resolution in the Figure](#), with the unit of 1/72 inch, being pixels per inch (ppi). And set [figure resolution in the Figure](#), with the unit of pixels per inch (ppi). The colors of the two adjacent chromosomes can be changed via

the combo box, with a drop-down option. Set the critical value for  $-\log_{10}(P\text{-value})$ , which is defaulted 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** button to choose a path and to save the Figure, with three frequently used image formats: \*.png, \*.tiff and \*.jpeg.

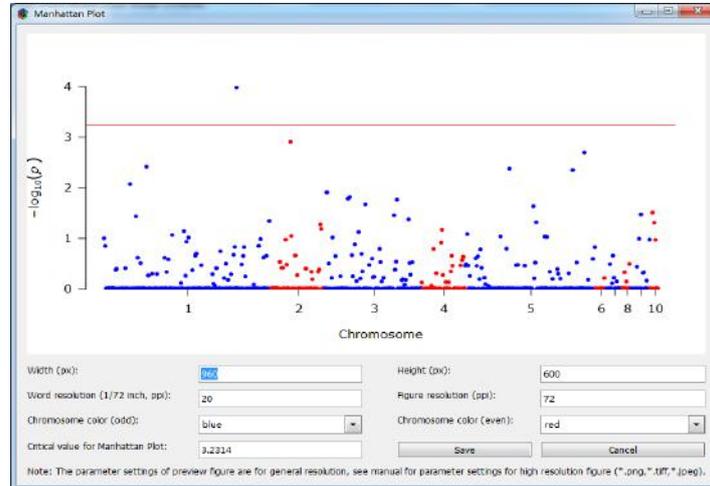


Figure 3.1.14. The Manhattan Plot of **mrMLM** module using example data

### 3.1.5 QQ plot

Use **QQ plot** button to preview the QQ plot in an independent dialog window. The settings of the width, height, word resolution and figure resolution are the same as those in the Manhattan plot. Set the critical P-value for QQ plot, which is defaulted the value of 0.95. This is because the P-values are a mixture of a  $\chi^2$  distribution with one degree of freedom and a point mass at one. Note that users may also change this 0.95 based on yourself results.

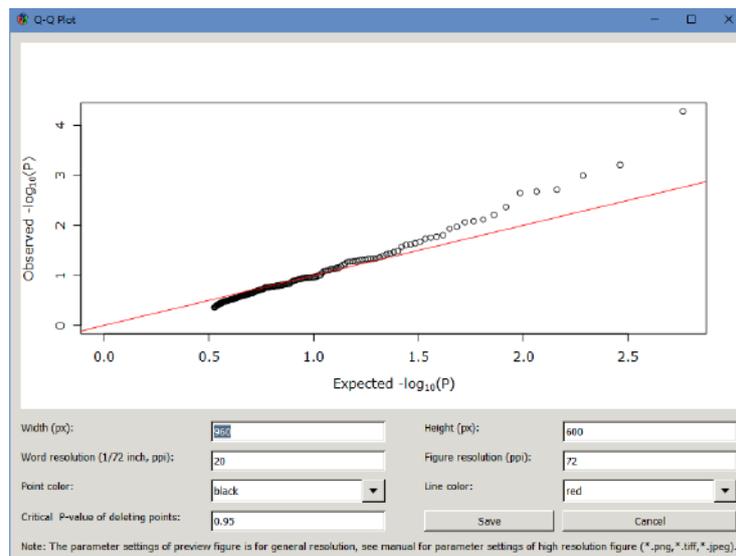


Figure 3.1.15. The QQ Plot of **mrMLM** module using example data

Using example data in this software, the preview Manhattan and QQ plots with the general resolution parameter settings were shown in **Figures 3.1.14** and **3.1.15**, respectively. Using Arabidopsis real data in Atwell et al. (2010) Nature 465: 627-631, the preview Manhattan and QQ plots with the general resolution parameter settings were shown in **Figures 3.1.16** and **3.1.17**, respectively.

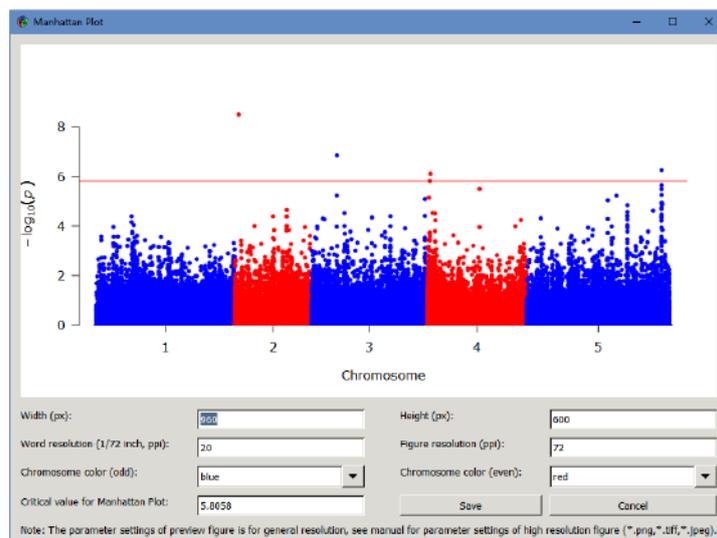


Figure 3.1.16. The Manhattan Plot of **mrMLM** module using Arabidopsis real data

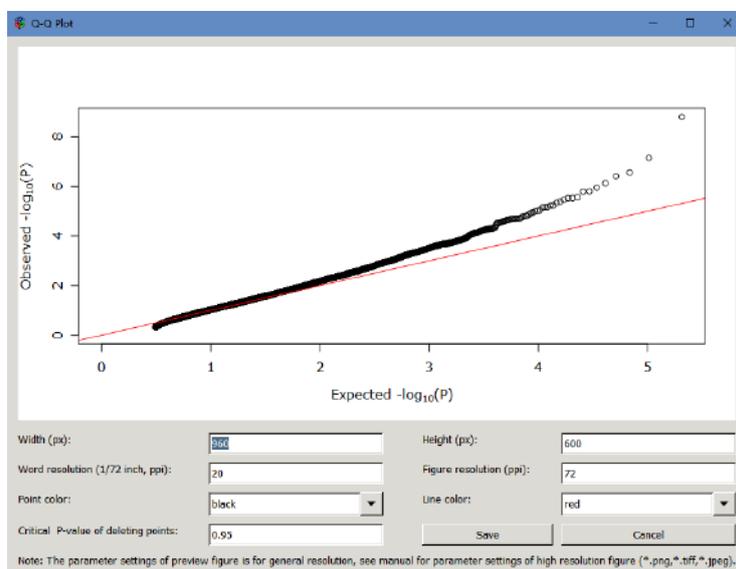
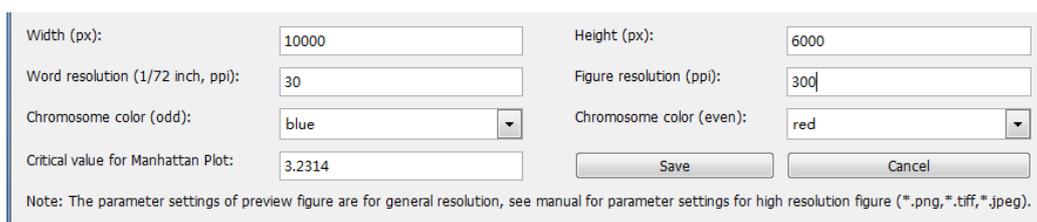


Figure 3.1.17. The QQ Plot of **mrMLM** module using Arabidopsis real data

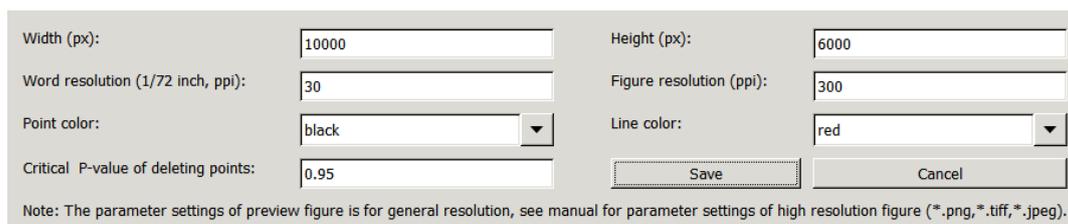
If users want to obtain **high resolution** figure, we recommend the parameter settings shown in **Figure 3.1.18** and **Figure 3.1.19** for Manhattan plot and QQ plot, respectively. It is emphasized that preview figure may be not the same with the **saved**

figure.



Width (px): 10000 Height (px): 6000  
Word resolution (1/72 inch, ppi): 30 Figure resolution (ppi): 300  
Chromosome color (odd): blue Chromosome color (even): red  
Critical value for Manhattan Plot: 3.2314 Save Cancel  
Note: The parameter settings of preview figure are for general resolution, see manual for parameter settings for high resolution figure (\*.png,\*.tiff,\*.jpeg).

Figure 3.1.18. Parameter settings for high resolution Manhattan plot of **mrMLM** Module



Width (px): 10000 Height (px): 6000  
Word resolution (1/72 inch, ppi): 30 Figure resolution (ppi): 300  
Point color: black Line color: red  
Critical P-value of deleting points: 0.95 Save Cancel  
Note: The parameter settings of preview figure is for general resolution, see manual for parameter settings of high resolution figure (\*.png,\*.tiff,\*.jpeg).

Figure 3.1.19. Parameter settings for high resolution QQ plot of **mrMLM** module

## 3.2 FASTmrEMMA module

Click the button “[Fast Multi-locus Random-SNP-effect EMMA \(FASTmrEMMA\)](#)”, then the following dialog will appear.

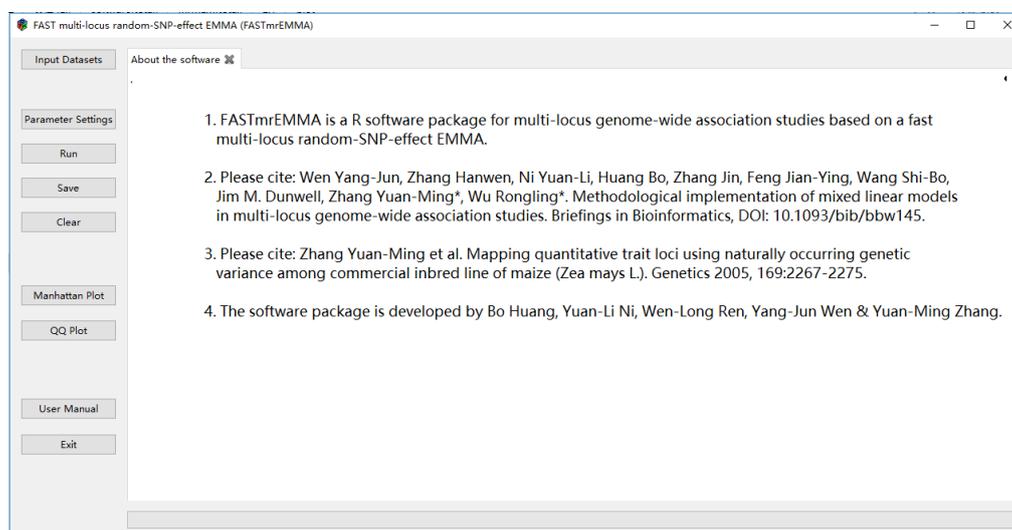


Figure 3.2.1. Screenshot of **FASTmrEMMA** module GUI

### 3.2.1 Input Data

Use the **Input Dataset** button to input dataset files, and then a dialog box will be appeared. In the dialog box, there are five steps. First, users select the dataset formats, which include **FASTmrEMMA numeric** format, **FASTmrEMMA character** format and **hapmap** format used in the TASSEL software. Second, use the **Genotype** and

**Phenotype** buttons to input the genotypic and phenotypic datasets, respectively. Once one file is successfully uploaded, one tabbed page is added to the software notebook. Third, users select one objective function: **Restricted Likelihood Function or Likelihood Function**. Fourth, some things will be implemented in this step: 1) firstly, to delete missing phenotypic data, then to sort the individuals between the genotypic and phenotypic files and all the common individuals between the two files are selected to be analyzed in the further analyses; 2) to transfer the character genotypes into the numeric genotypes if the genotypes are character; and 3) to select one objective function. Once users press the **DO** button, these things will be implemented. Once the two files will be successfully uploaded, two tabbed pages (Genotype and Phenotype) will be added to the software notebook. Finally, use the **Kinship** and **Population Structure** buttons to input the kinship and population structure matrices, respectively. If one file is successfully uploaded, the corresponding data page will be added to the notebook. Note that there are two options for the **Kinship** and **Population Structure** buttons. For the **kinship** button, one is to directly upload the kinship matrix and another is to calculate the kinship matrix in this software. For the **Population Structure** button, the population structure matrix is not included in FASTmrEMMA methodology if it has no effect on GWAS. If not, it should be included in FASTmrEMMA methodology. The population structure matrix in your uploaded file will be deleted one column if the sum of all the Q-matrix columns for one individual (one row) equal to 1. In the filter dialog, you should choose one column that should be deleted.

**Note:** If the number of SNP markers in genotypic dataset is large (number of SNP markers is more than 100,000), please input the kinship matrix file directly. About the **input file format**, please see **Direction 1** in the end of the manual.

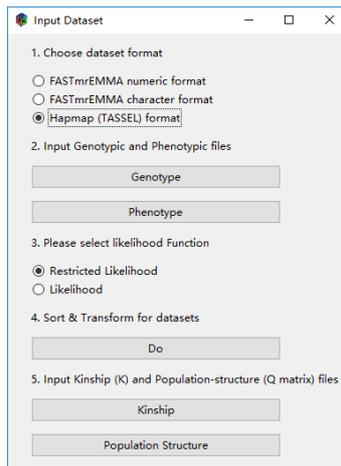


Figure 3.2.2. The Input Dataset dialog of **FASTmrEMMA** module

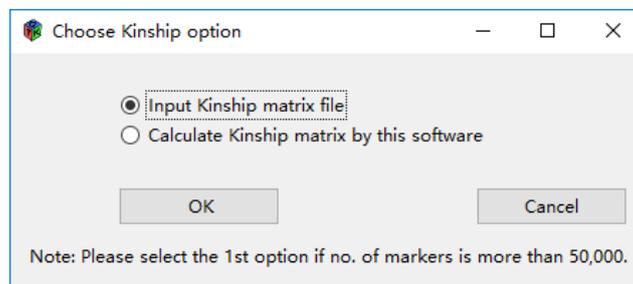


Figure 3.2.3. The Kinship dialog of **FASTmrEMMA** module

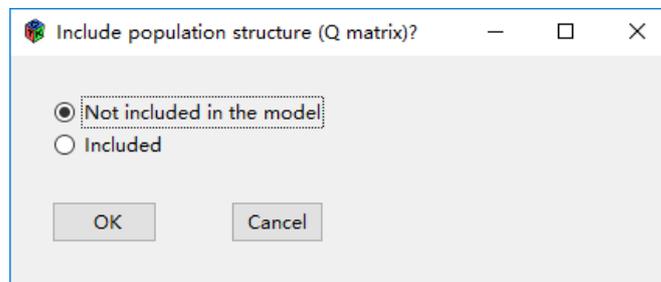


Figure 3.2.4. The population structure dialog of **FASTmrEMMA** module

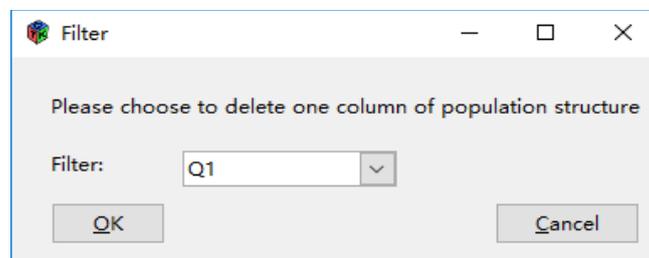


Figure 3.2.5. Filter dialog of the population structure of **FASTmrEMMA** module

### 3.2.2 Run Program

Use the **Parameter Setting** button to set parameters before run the program. Use the

**Run** button to execute the software. If the program runs, a progress bar with the “**Please be patient...**” words will appear in the bottom of the interface. If the program finished, a bar with the “**All done.**” will appear.

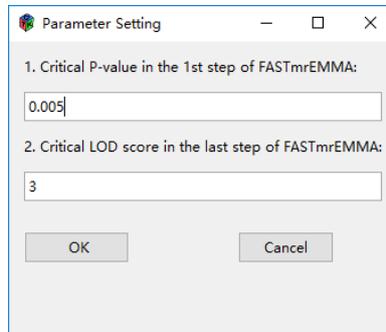


Figure 3.2.6. The Parameter Setting dialog of **FASTmrEMMA** module

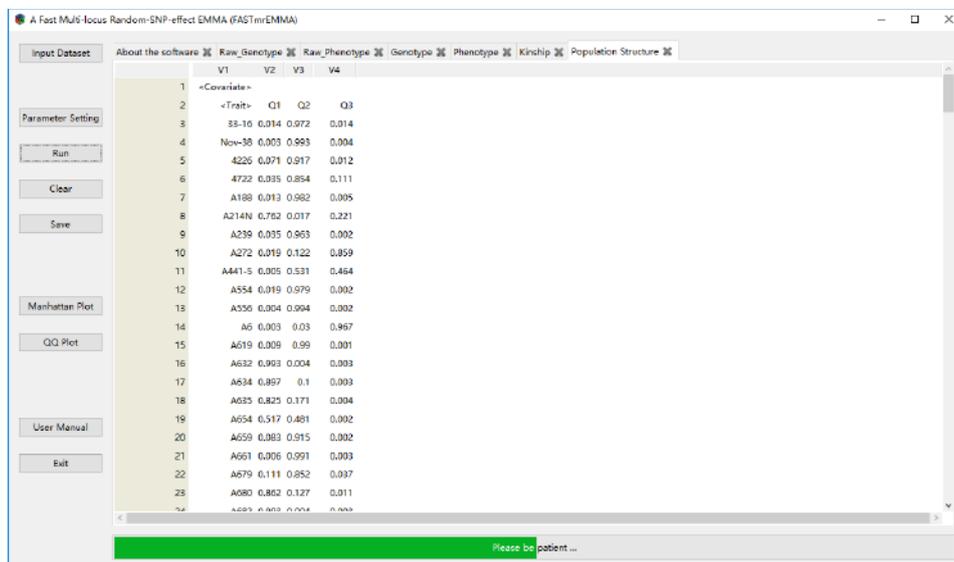


Figure 3.2.7. A running program interface of **FASTmrEMMA** module

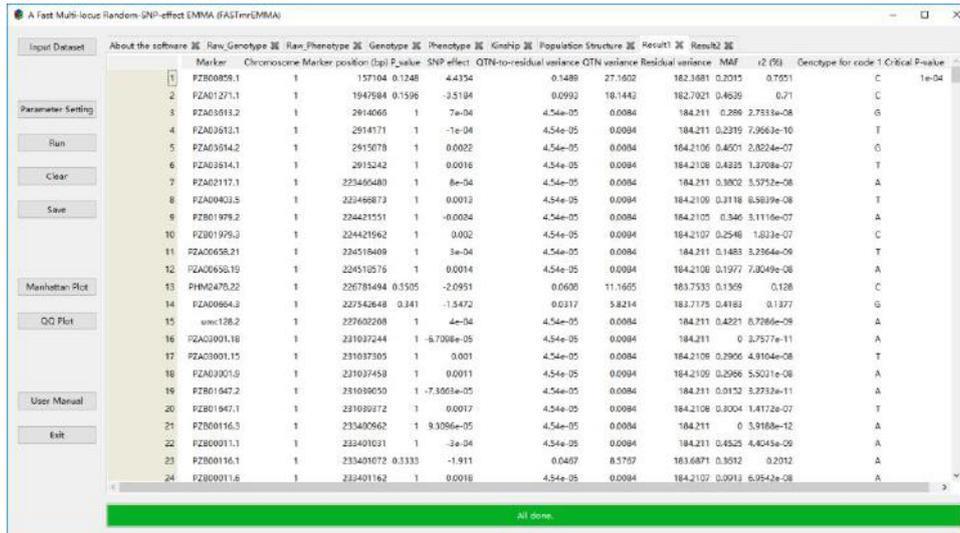


Figure 3.2.8. A finished program interface of FASTmrEMMA module (Results: Result1)

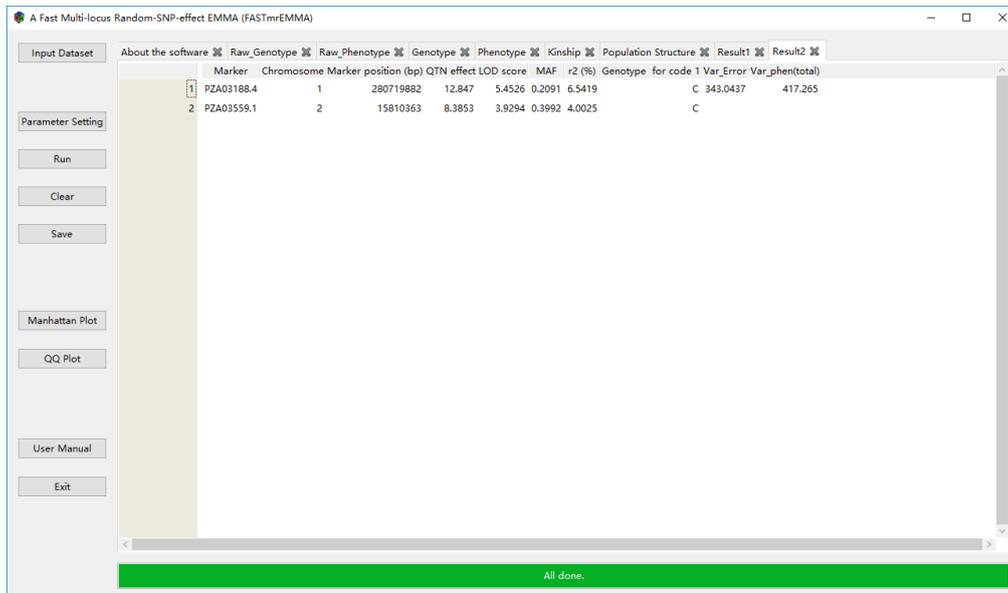


Figure 3.2.9. A finished program interface of FASTmrEMMA module (Results: Result2)

### 3.2.3 Output results

Use **Save** button to save the results as \*.csv files. The **Result1** are saved as **Result1.csv** and the **Result2** are saved as **Result2.csv**. If click **OK** button (after the **Save** button), a dialog is used to choose the pathway and the file name for the saving files.

**Note:** About the **explanation of Result1 and Result2** in details, please see **Direction 2** in the end of the manual.

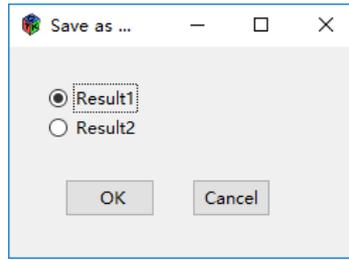


Figure 3.2.10. The result **Save** dialog of **FASTmrEMMA** module

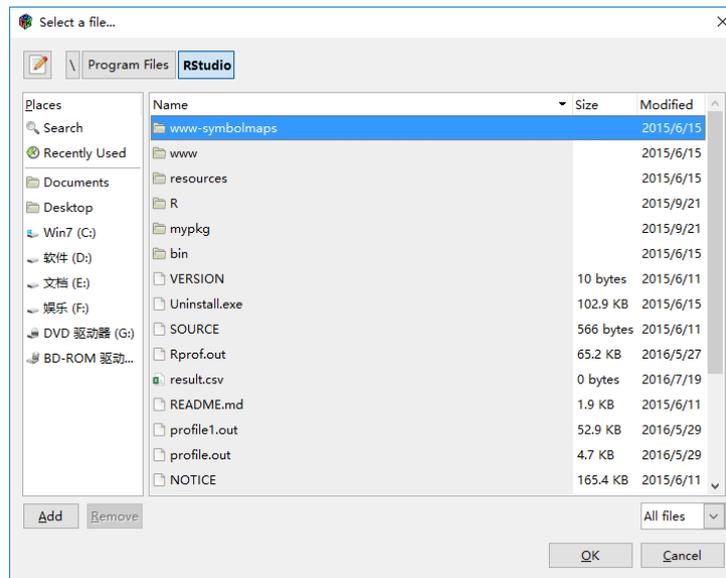


Figure 3.2.11. The **Save** dialog of **FASTmrEMMA** module

The **Result1** table with twelve columns shows the results. The corresponding column names are as follows: **Marker** (marker name or reference sequence number), **Chromosome**, **Marker position (bp)** in the chromosome, the **P\_value** for each QTN, **SNP effect**, **QTN-to-residual variance ratio** (ratio of QTN variance to residual variance), **QTN variance**, **Residual variance** for the current SNP, **MAF** (minor allele frequency), **r<sup>2</sup> (%)** (the proportion of phenotypic variance explained by the putative QTN), **genotype for code 1**, and **Critical P-value** (critical value for the significant test,  $0.05/m_e$ ), respectively.

#	A	B	C	D	E	F	G	H	I	J	K	L	M
1	Marker	Chromosome	Marker position (bp)	P_value	SNP effect	QTN-to-residual variance ratio	QTN variance	Residual variance	MAF	r <sup>2</sup> (%)	Genotype for code 1	Critical P-value	
2	F2B00859.1	3	197104	0.1161	4.558	0.1625	25.2184	175.5541	0.2015	0.8052	C	1.00E-04	
3	F2A03271.1	3	1949784	0.1554	-3.4511	0.1001	17.4347	174.1195	0.4639	0.6752	C		
4	F2A03013.2	3	2914996	1	5.93E-04	4.54E-05	0.008	175.5027	0.289	3.13E-09	G		
5	F2A03013.1	3	2914171	1	-0.00017597	4.54E-05	0.008	175.5028	0.2319	1.32E-09	T		
6	F2A03014.2	3	2915078	1	0.0021	4.54E-05	0.008	175.5024	0.4601	2.55E-07	G		
7	F2A03014.1	3	2915242	1	0.0015	4.54E-05	0.008	175.5026	0.4326	1.18E-07	T		
8	F2A02117.1	3	223466480	1	0.001	4.54E-05	0.008	175.5027	0.3802	5.20E-08	A		
9	F2A00403.5	3	223466873	1	0.0013	4.54E-05	0.008	175.5026	0.3118	3.98E-08	T		
10	F2B13979.2	3	224421531	1	-0.0024056	4.54E-05	0.008	175.5023	0.346	3.07E-07	A		
11	F2B13979.3	3	224421992	1	0.002	4.54E-05	0.008	175.5024	0.2548	1.38E-07	C		
12	F2A00958.21	3	224515495	1	2.40E-04	4.54E-05	0.008	175.5025	0.1465	1.40E-09	T		
13	F2A00958.19	3	224515570	1	0.0012	4.54E-05	0.008	175.5027	0.1977	5.03E-08	A		
14	F2M2473.22	3	226781494	0.3718	-1.8926	0.053	5.2878	175.1494	0.1368	0.0988	C		
15	F2A00954.3	3	227542648	0.3973	-1.0066	0.0197	3.4489	175.2234	0.4183	0.0583	C		
16	unc12B.2	3	227602208	1	4.40E-04	4.54E-05	0.008	175.5028	0.4221	1.05E-08	A		
17	F2A03001.18	3	231037244	1	5.30E-06	4.54E-05	0.008	175.5028	0	2.35E-11	A		
18	F2A03001.19	3	231037305	1	5.90E-04	4.54E-05	0.008	175.5027	0.2956	2.82E-08	T		

Figure 3.2.12.Results in FASTmrEMMA module (Result1)

The **Result2** table with ten columns shows the final results of the FASTmrEMMA (a fast multi-locus random-SNP-effect EMMA) method. The corresponding column names are as follows: **Marker** (marker name or reference sequence number), **Chromosome**, **Marker position (bp)** in the chromosome, **QTN effect**, **LOD score**, **MAF** (minor allele frequency), **r<sup>2</sup> (%)** (the proportion of phenotypic variance explained by the putative QTN), **genotype for code 1**, **Var\_Error** (residual error variance), and **Var\_phen (total)** (total phenotypic variance), respectively.

#	A	B	C	D	E	F	G	H	I	J	K
1	Marker	Chromosome	Marker position (bp)	QTN effect	LOD score	MAF	r <sup>2</sup> (%)	Genotype for code 1	Var_Error	Var_phen(total)	
2	F2A03188.4	1	28071882	12.847	5.4526	0.2091	6.5419	C	343.0437	417.265	
3	F2A03559.1	2	15810363	3.3853	3.9294	0.3992	4.0025	C			
4											
5											
6											
7											
8											
9											
10											

Figure 3.2.13.Results in FASTmrEMMA module (Result2)

### 3.2.4 Manhattan plot

Click **Manhattan plot** button to preview Manhattan plot in an independent dialog window. Before saving the Figure, please set **the width and height of the Figure**, with the unit of pixel (px). And set **word resolution in the Figure**, with the unit of 1/72 inch, being pixels per inch (ppi). And set **figure resolution in the Figure**, with the unit of pixels per inch (ppi). The colors of the two adjacent chromosomes can be changed via the combo box, with a drop-down option. Set the critical value for  $-\log_{10}(P)$ , which is defaulted the value of  $-\log(0.05/m_e)$ , where  $m_e$  is the effective number of markers. Of course, users may change this value. Use **Save** button to choose a path and to save the Figure, with three frequently used image formats: \*.png, \*.tiff and \*.jpeg.

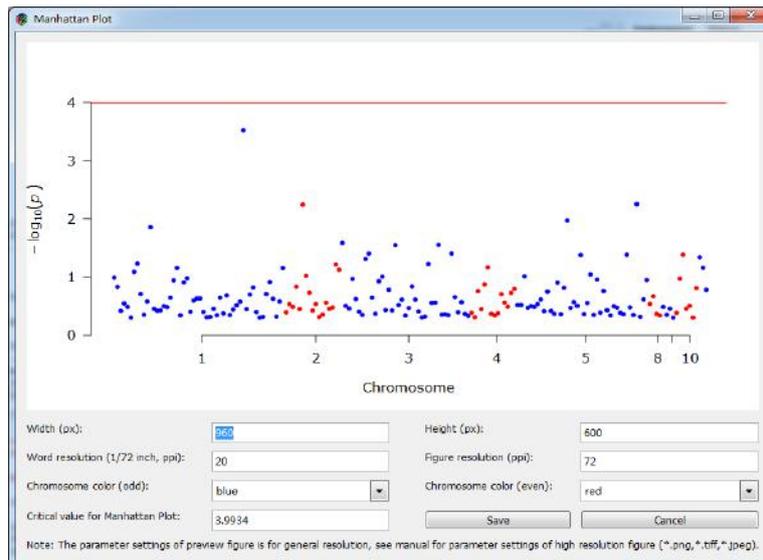


Figure 3.2.14. The Manhattan Plot of **FASTmrEMMA** module

### 3.2.5 QQ plot

Use **QQ plot** button to preview the QQ plot in an independent dialog window. The settings of the width, height, word resolution and figure resolution are the same as those in the Manhattan plot.

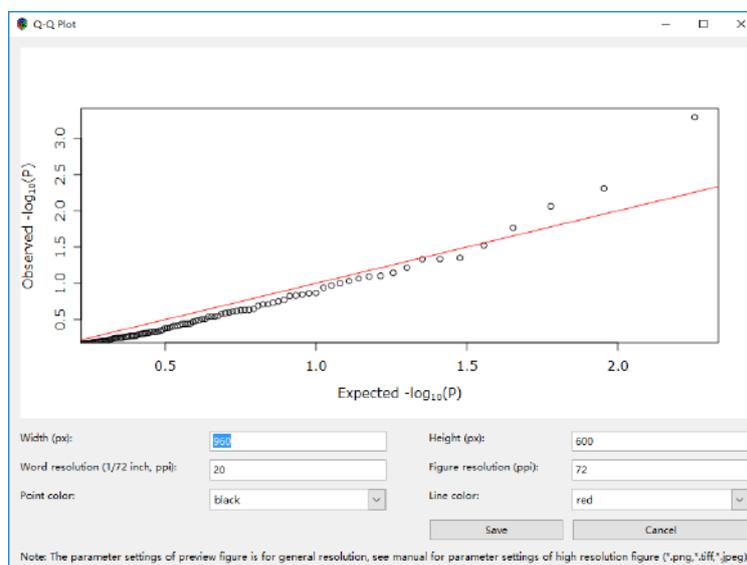


Figure 3.2.15. The QQ Plot of **FASTmrEMMA** module

**Figure 3.2.14** and **Figure 3.2.15** show the preview picture of Manhattan plot and QQ plot with the parameter settings **in general resolution**, respectively.

If you want to obtain **high resolution** figure, we recommend the parameter settings shown in **Figure 3.2.16** and **Figure 3.2.17** for Manhattan plot and QQ plot,

respectively. It is emphasized that preview figure may not seem the same with **saved figure**. To give priority to saved figure in application.

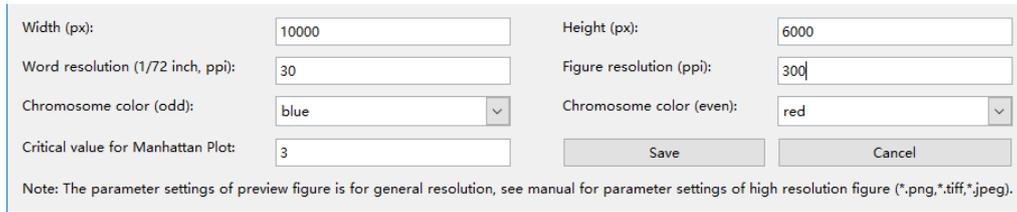


Figure 3.2.16. Parameter settings for high resolution Manhattan plot of **FASTmrEMMA** Module

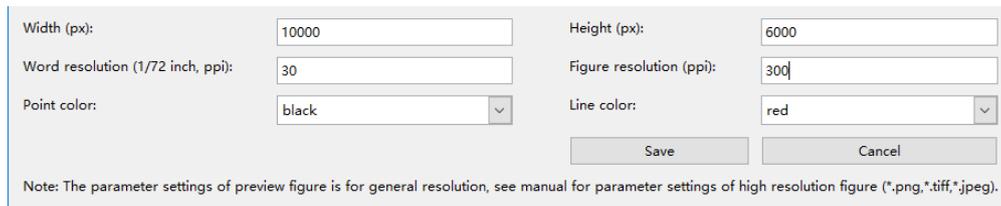


Figure 3.2.17. Parameter settings for high resolution QQ plot of **FASTmrEMMA** module

### 3.3 ISIS EM-BLASSO module

Click the button “[Iterative Sure Independence Screening EM-Bayesian LASSO \(ISIS EM-BLASSO\)](#)”, then the following dialog will appear.

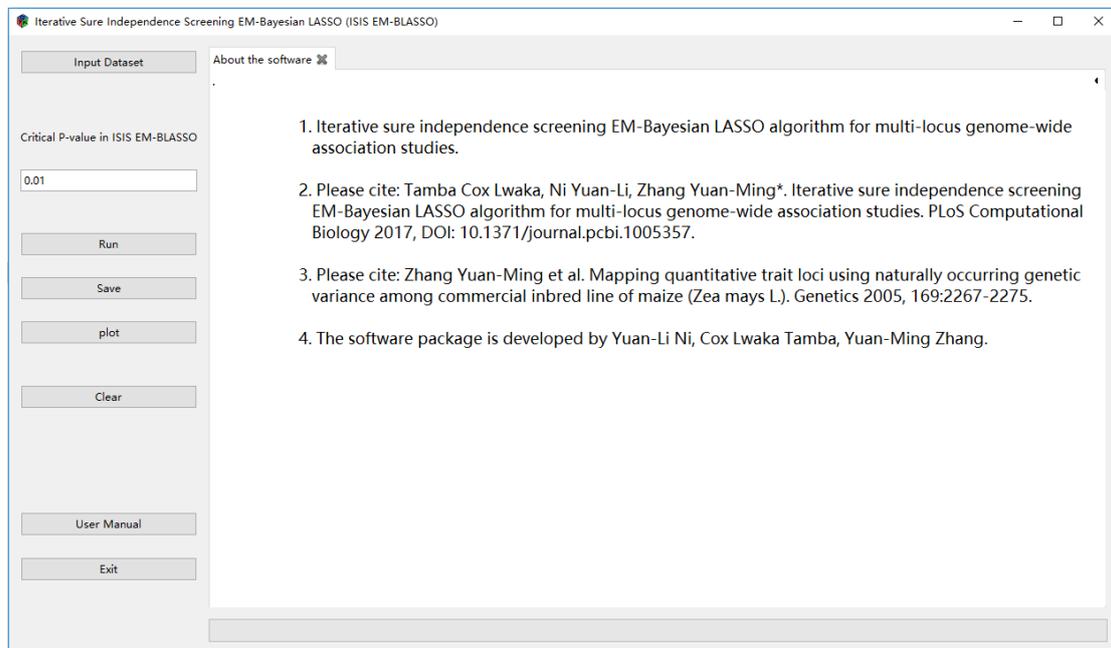


Figure 3.3.1. Screenshot of **ISIS EM-BLASSO** module GUI

#### 3.3.1 Input Data

Use the **Input Dataset** button to input dataset files, and then a dialog box will be

appeared. In the dialog box, there are four steps. First, users select the dataset formats, which include **mrMLM numeric** format, **mrMLM character** format and **hapmap** format used in the TASSEL software. Then, use the **Genotype** and **Phenotype** buttons to input the genotypic and phenotypic datasets, respectively. Once one file is successfully uploaded, one tabbed page is added to the software notebook. Third, two things will be implemented in this step. One is to sort the individuals between the genotypic and phenotypic files and all the common individuals between the two files are selected to be analyzed in the further analyses. Another is to transfer the character genotypes into the numeric genotypes if the genotypes are character. Once users press the **DO** button, the two things will be conducted. Once the two files will be successfully uploaded, two tabbed pages (Genotype and Phenotype) will be added to the software notebook. Finally, use the **Population Structure** buttons to input the population structure matrices, respectively. If one file is successfully uploaded, the corresponding data page will be added to the notebook. Note that the **Population Structure** buttons have two options. The population structure matrix may be not included in the mixed linear model of the GWAS if it has no effect on GWAS. If not, it should be included in the mixed model. The population structure matrix in your uploaded file will be deleted one column if the sum of all the Q-matrix columns for one individual (one row) equal to 1. In the filter dialog, you should choose one column that should be deleted .

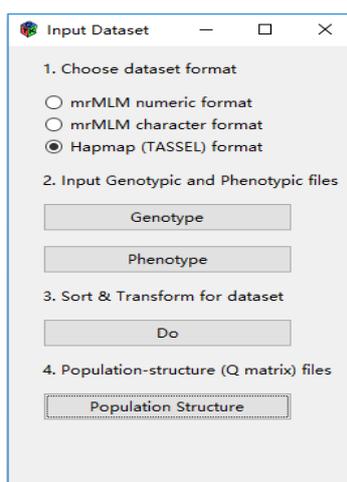


Figure 3.3.2. The Input Dataset dialog of **ISIS EM-BLASSO** module

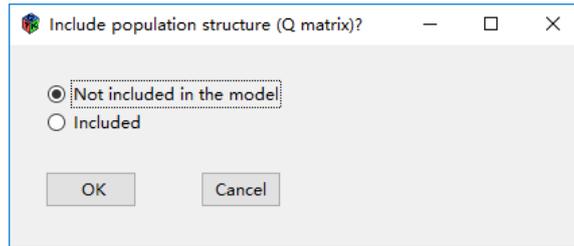


Figure 3.3.3. The population structure dialog of **ISIS EM-BLASSO** module

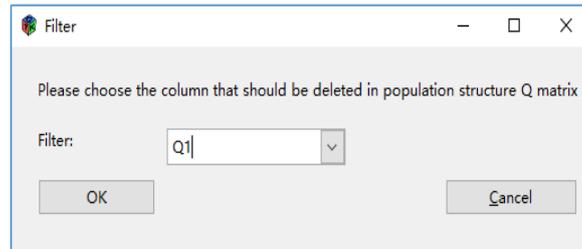


Figure 3.3.4. Filter dialog of the population structure of **ISIS EM-BLASSO** module

### 3.3.2 Run Program

Please set critical P-value (0.01 is default value) in **ISIS EM-BLASSO** before run the program. Use the **Run** button to execute the software. If the program runs, a progress bar with the “**Please be patient...**” words will appear in the bottom of the interface. If the program finished, a bar with the “**All done.**” will appear.

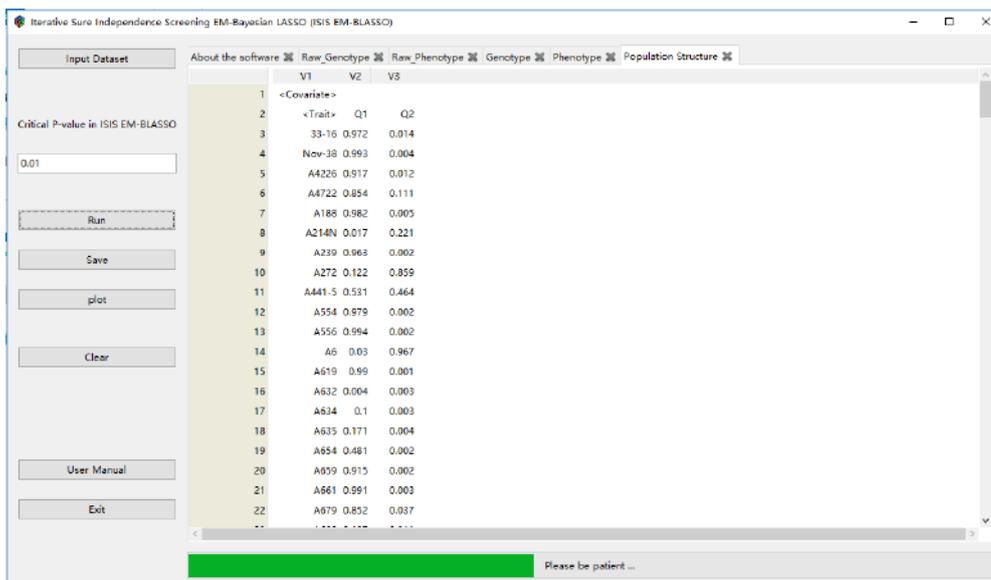


Figure 3.3.5. A running program interface of **ISIS EM-BLASSO** module

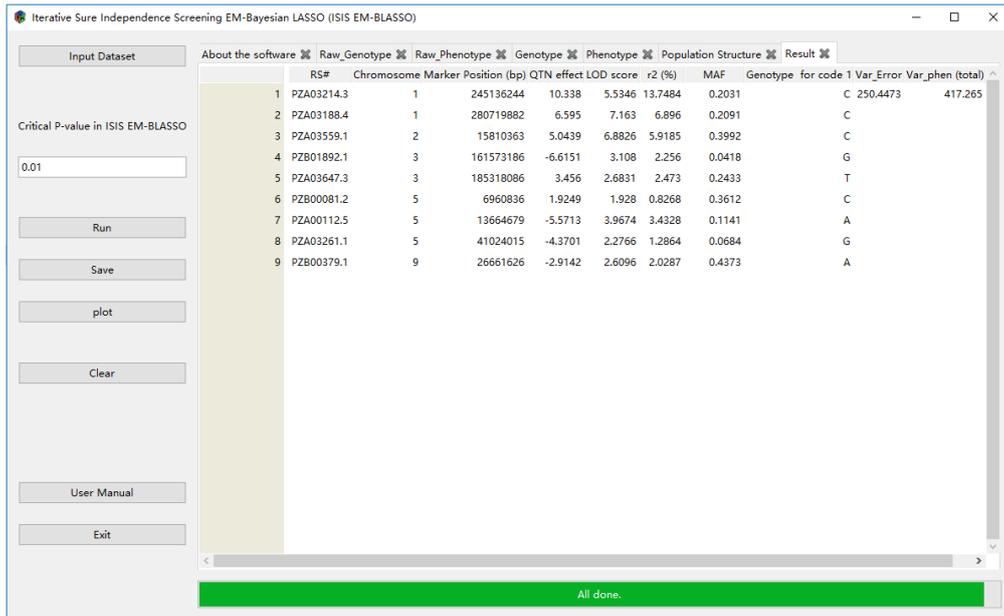


Figure 3.3.6. A finished program interface of **ISIS EM-BLASSO** module (**Results: Result**)

### 3.3.3 Output results

Use **Save** button to save the results as \*.csv files. If click **Save** button, a dialog is used to choose the pathway and the file name for the saving files.

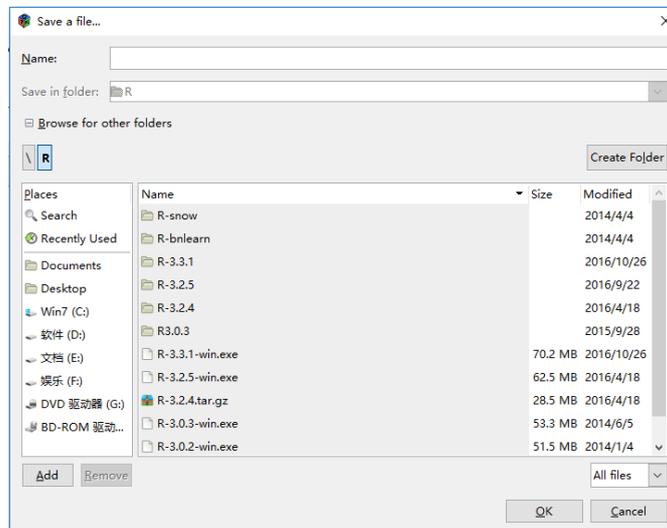


Figure 3.3.7. The **Save** dialog of **ISIS EM-BLASSO** module

The **Result** table with ten columns shows the final results of the ISIS EM-BLASSO (Iterative Sure Independence Screening EM-Bayesian LASSO) method. The corresponding column names are as follows: **RS#** (marker name or reference sequence number), **Chromosome**, **Marker position (bp)** in the chromosome, **QTN effect**, **LOD score**, **r<sup>2</sup> (%)** (the proportion of phenotypic variance explained by the putative QTN), **MAF**

(minor allele frequency) , **genotype for code 1**, **Var\_Error** (residual error variance), and **Var\_phen (total)** (total phenotypic variance), respectively.

	A	B	C	D	E	F	G	H	I	J	
1	RS#	Chromosome	Marker	Position	QTN effect	LOD score	r2 (%)	MAF	Genotype	Var_Error	Var_phen (total)
2	PZA03214.3	1	245136244		10.338	5.5346	13.7484	0.2031	C	250.4473	417.265
3	PZA03188.4	1	280719882		6.595	7.163	6.896	0.2091	C		
4	PZA03559.1	2	15810363		5.0439	6.8826	5.9185	0.3992	C		
5	PZB01892.1	3	161573186		-6.6151	3.108	2.256	0.0418	G		
6	PZA03647.3	3	185318086		3.456	2.6831	2.473	0.2433	T		
7	PZB00081.2	5	6960836		1.9249	1.928	0.8268	0.3612	C		
8	PZA00112.5	5	13664679		-5.5713	3.9674	3.4328	0.1141	A		
9	PZA03261.1	5	41024015		-4.3701	2.2766	1.2864	0.0684	G		
10	PZB00379.1	9	26661626		-2.9142	2.6096	2.0287	0.4373	A		
11											
12											

Figure 3.3.8. Results in ISIS EM-BLASSO module (**Result**)

### 3.3.4 LOD Scores plot

Click **plot** button to preview **Plot of LOD Score against Genome Position** dialog window. Before saving the Figure, please set the **width and height of the Figure**, with the unit of pixel (px). And set **word resolution in the Figure**, with the unit of 1/72 inch, being pixels per inch (ppi). And set **figure resolution in the Figure**, with the unit of pixels per inch (ppi). And Set **LOD line color**. Use **Save** button to choose a path and to save the Figure, with three frequently used image formats: \*.png, \*.tiff and \*.jpeg.

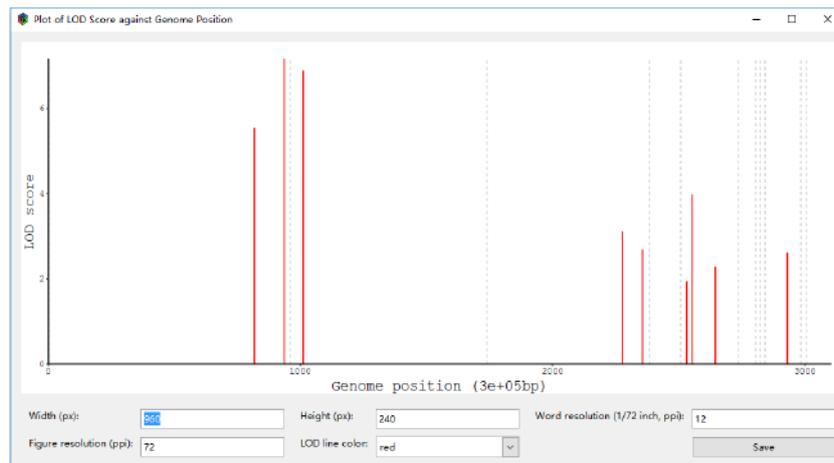


Figure 3.3.9. The LOD Scores Plot of ISIS EM-BLASSO module

**Figure 3.3.9** show the preview picture of LOD Scores plot with the parameter settings **in general resolution**, respectively.

If you want to obtain **high resolution** figure, we recommend the parameter settings shown in **Figure 3.3.10** for LOD Scores plot. It is emphasized that preview figure may not seem same with **saved figure**. To give priority to saved figure in application.

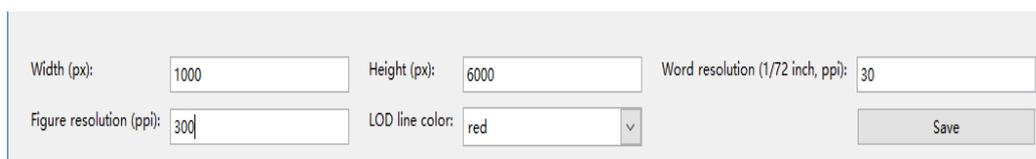


Figure 3.3.10. Parameter settings for high resolution LOD Scores plot of **ISIS EM-BLASSO**

### 3.4 pLARmEB module

Click the button “[polygene-background-control-based least angle regression plus empirical Bayes \(pLARmEB\)](#)”, then the following dialog will appear.

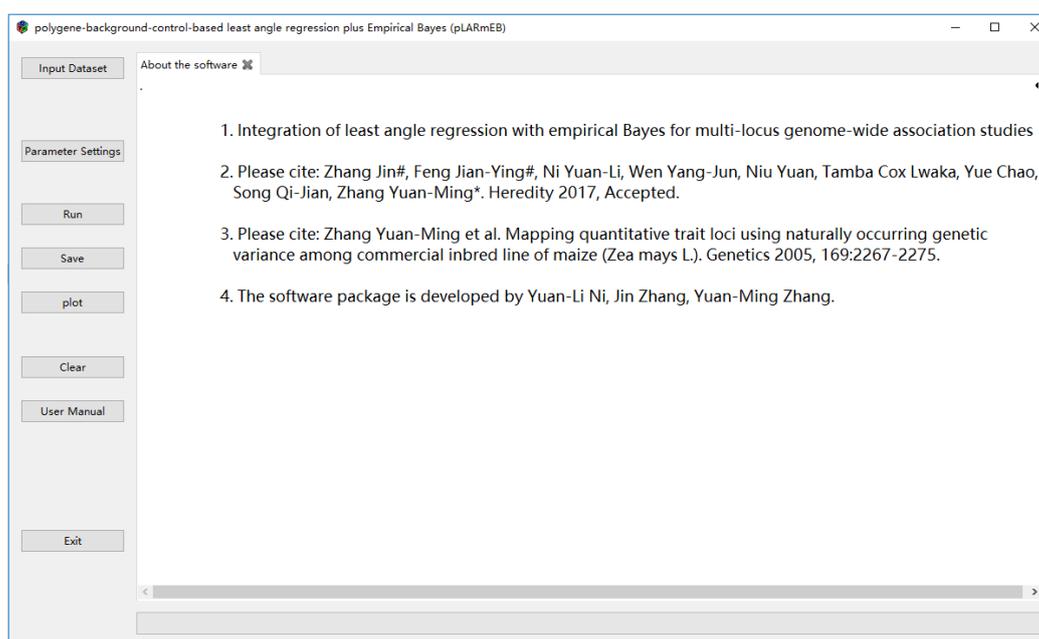


Figure 3.4.1. Screenshot of **pLARmEB** module GUI

#### 3.4.1 Input Data

Use the **Input Dataset** button to input dataset files, and then a dialog box will be appeared. In the dialog box, there are four steps. First, users select the dataset formats, which include **mrMLM numeric** format, **mrMLM character** format and **hapmap** format used in the TASSEL software. Then, use the **Genotype** and **Phenotype** buttons to input the genotypic and phenotypic datasets, respectively. Once one file is successfully uploaded, one tabbed page is added to the software notebook. Third, two things will be implemented in this step. One is to sort the individuals between the genotypic and phenotypic files and all the common individuals between the two files are selected to be analyzed in the further analyses. Another is to transfer the character genotypes into the numeric genotypes if the genotypes are character. Once users press

the **DO** button, the two things will be conducted. Once the two files will be successfully uploaded, two tabbed pages (Genotype and Phenotype) will be added to the software notebook. Finally, use the **Population Structure** buttons to input the population structure matrices, respectively. If one file is successfully uploaded, the corresponding data page will be added to the notebook. Note that the **Population Structure** buttons have two options. The population structure matrix may be not included in the mixed linear model of the GWAS if it has no effect on GWAS. If not, it should be included in the mixed model. The population structure matrix in your uploaded file will be deleted one column if the sum of all the Q-matrix columns for one individual (one row) equal to 1. In the filter dialog, you should choose one column that should be deleted.

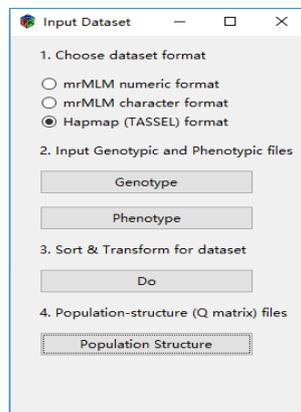


Figure 3.4.2. The Input Dataset dialog of **pLARmEB** module

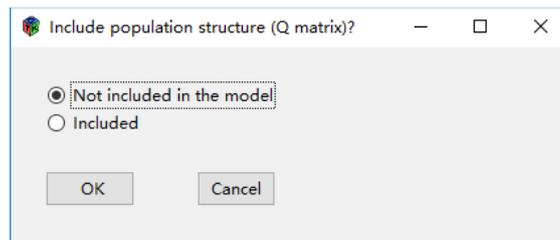


Figure 3.4.3. The population structure dialog of **pLARmEB** module

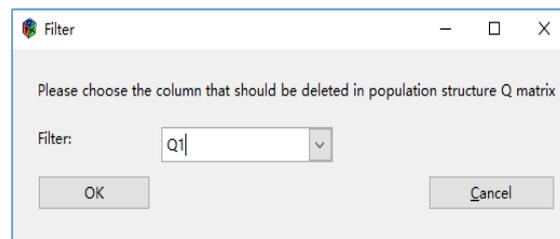


Figure 3.4.4. Filter dialog of the population structure of **pLARmEB** module

### 3.4.2 Run Program

Use the **Parameter Setting** button to set parameters before running the program. “The number of potentially associated variables selected by LARS”, the values must be less than the sample size. **In the analyses of real dataset, users may change this number in order to obtain the best result.** Use the **Run** button to execute the software. If the program runs, a progress bar with the “**Please be patient...**” words will appear in the bottom of the interface. If the program finished, a bar with the “**All done.**” will appear.

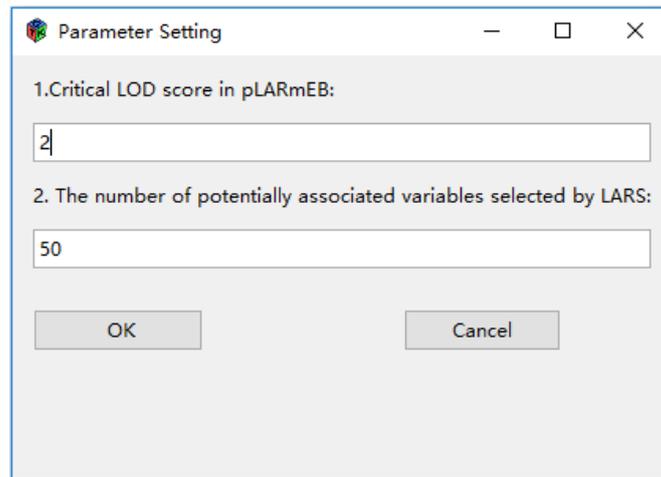


Figure 3.1.6. The Parameter Setting dialog of **pLARmEB** module

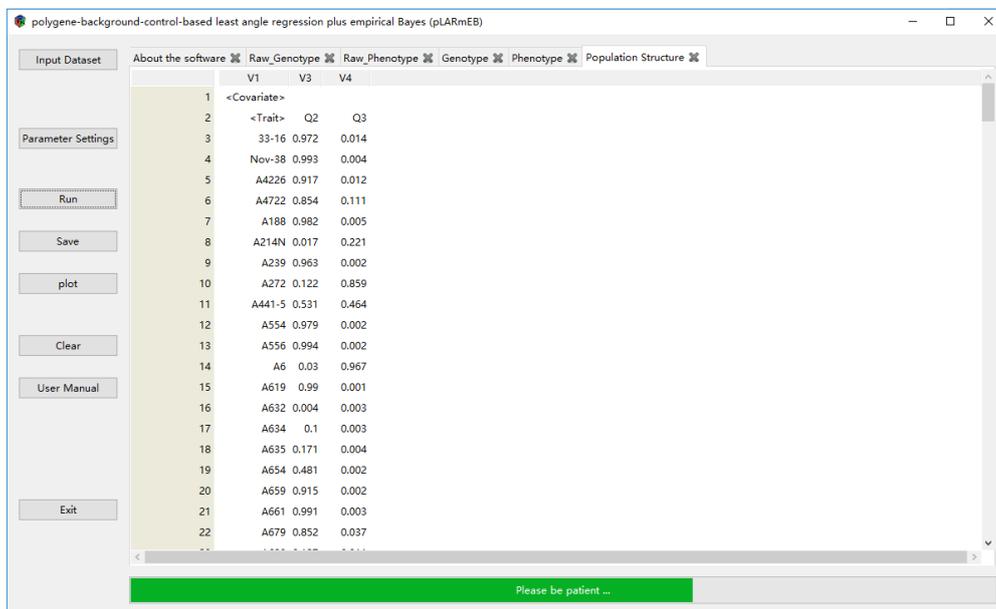


Figure 3.4.5. A running program interface of **pLARmEB** module

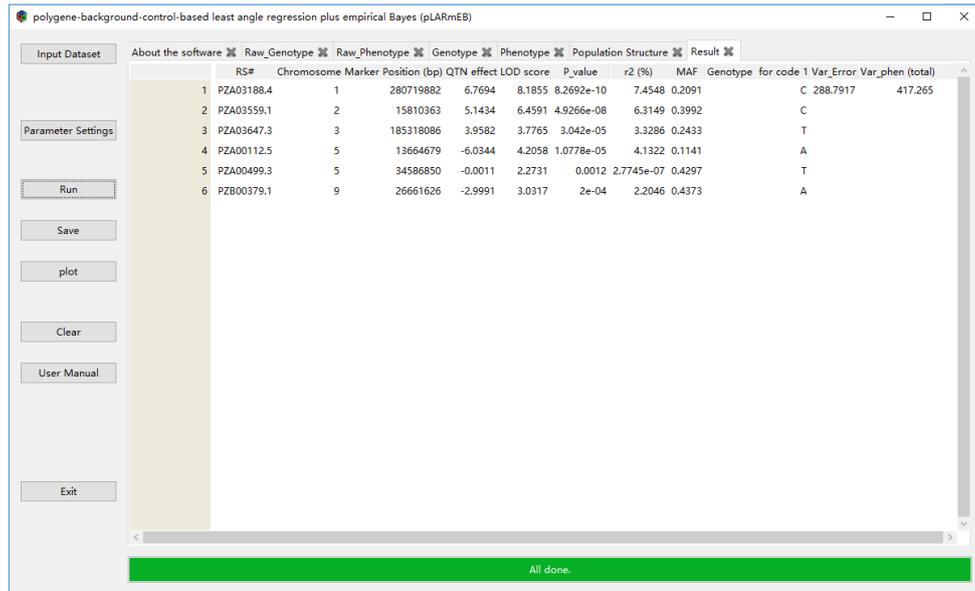


Figure 3.4.6. A finished program interface of pLARmEB module (**Results: Result**)

### 3.4.3 Output results

Use **Save** button to save the results as \*.csv files. If click **Save** button, a dialog is used to choose the pathway and the file name for the saving files.

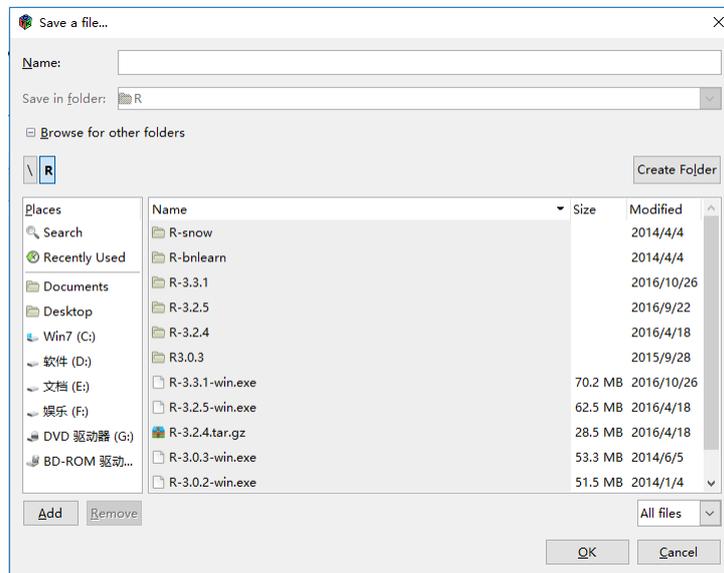


Figure 3.4.7. The **Save** dialog of pLARmEB module

The **Result** table with eleven columns shows the final results of the pLARmEB (polygene-background-control-based least angle regression plus empirical Bayes) method. The corresponding column names are as follows: **RS#** (marker name or reference sequence number), **Chromosome**, **Marker position (bp)** in the chromosome, **QTN effect**, **LOD score**, **P\_value**, **r<sup>2</sup> (%)** (the proportion of phenotypic variance explained

by the putative QTN), **MAF** (minor allele frequency), **genotype for code 1**, **Var\_Error** (residual error variance), and **Var\_phen (total)** (total phenotypic variance), respectively.

#	A	B	C	D	E	F	G	H	I	J	K
	RS#	Chromosome	Marker Position (bp)	QTN effect	LOD score	P-value	r <sup>2</sup> (%)	MAF	Genotype for code 1	Var_Error	Var_phen (total)
1	FZA03188.4	1	280719882	6.7694	8.1855	8.27E-10	7.4548	0.2091	C	288.7917	417.265
2	FZA03559.1	2	15810363	5.1434	6.4591	4.93E-08	6.3149	0.3992	C		
3	FZA03647.3	3	185318086	3.9582	3.7765	3.04E-05	3.3286	0.2433	T		
4	FZA00112.5	5	13664679	-6.0344	4.2058	1.08E-05	4.1322	0.1141	A		
5	FZA00499.3	5	34588850	-0.9011	2.2731	0.0012	2.77E-07	0.4287	T		
6	FZE00379.1	9	26661626	-2.9991	3.0317	2.00E-04	2.2046	0.4373	A		
7											
8											
9											

Figure 3.4.8. Results in pLARmEB module (Result)

### 3.4.4 LOD Scores plot

Click **plot** button to preview **Plot of LOD Score against Genome Position** dialog window. Before saving the Figure, please set the width and height of the Figure, with the unit of pixel (px). And set word resolution in the Figure, with the unit of 1/72 inch, being pixels per inch (ppi). And set figure resolution in the Figure, with the unit of pixels per inch (ppi). And Set LOD line color. Use **Save** button to choose a path and to save the Figure, with three frequently used image formats: \*.png, \*.tiff and \*.jpeg.

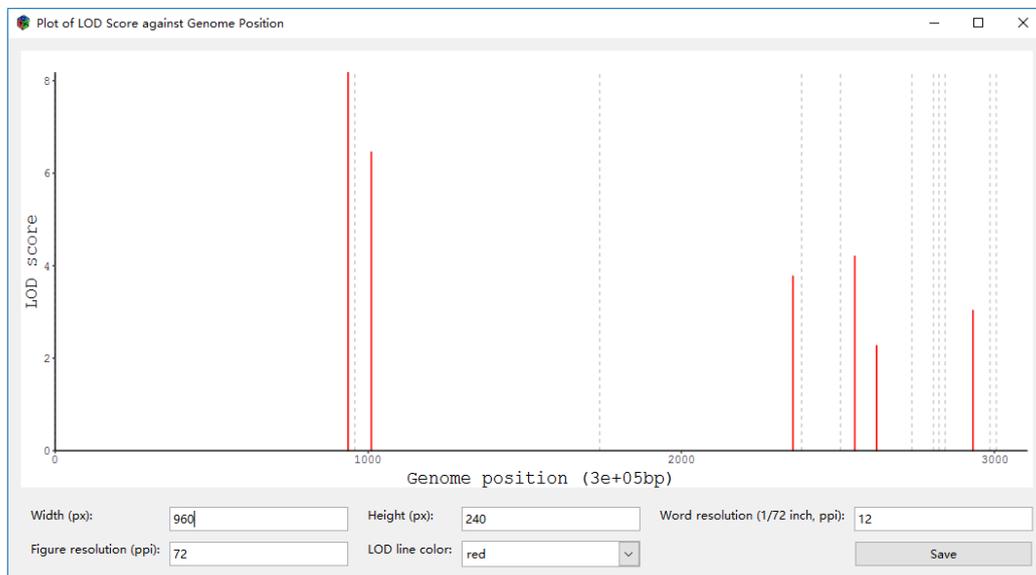


Figure 3.4.9. The LOD Scores Plot of pLARmEB module

**Figure 3.4.9** show the preview picture of LOD Scores plot with the parameter settings in general resolution, respectively.

If you want to obtain high resolution figure, we recommend the parameter settings shown in **Figure 3.4.10** for LOD Scores plot. It is emphasized that preview figure

may not seem the same with **saved figure**. To give priority to saved figure in application.



Figure 3.4.10. Parameter settings for high resolution LOD Scores plot of **pLARMmEB** module

### 3.5 GCIM module

Click the button “**Genome-wide Composite Interval Mapping (GCIM)**”, then the following dialog will appear.

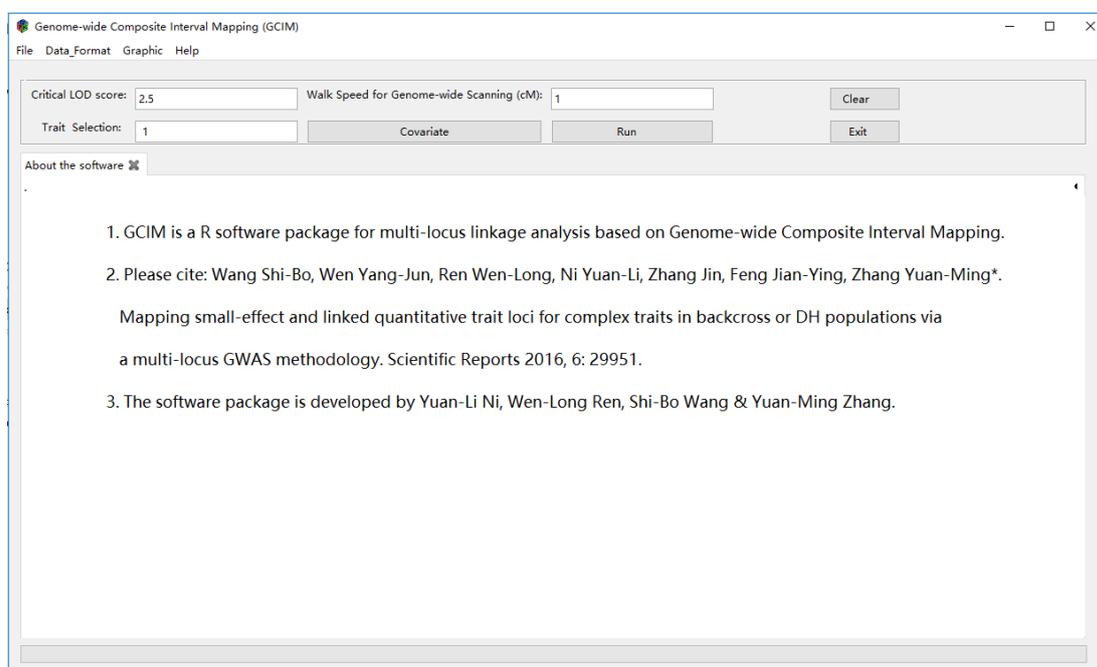


Figure 3.5.1. Screenshot of **GCIM** module GUI

#### 3.5.1 Input data and Parameter setting

##### 3.5.1.1 GCIM Format

To click GCIM will appear GCIM dialog box. In the dialog box, there are four steps. Firstly, the **Genotype**, **Phenotype** and **Linkage Map** buttons are used to input genotype, phenotype and linkage map datasets, respectively. Once one file is successfully uploaded, one tabbed page is added to the software notebook. Secondly, the user selects the required the population type, including **BC<sub>1</sub> (F<sub>1</sub> × P<sub>1</sub>)**, **BC<sub>2</sub> (F<sub>1</sub> ×**

**P<sub>2</sub>**), **DH**, **RIL** and **Chromosome Segment Substitution Lines (CSSL)**. Thirdly, users should give the symbols in the marker translation table that users indicate genotypes. Users can enter any alpha numeric character(s) as these symbols. In this software, these symbols will translate into numeric variables in our method. Fourthly, the user selects the required **QTL-effect model type**: **Random** and **Fixed** models. Finally, some things will be implemented in this step: 1) to upload the above three files; 2) to select population type; 3) to conduct marker-genotype translation; and 4) to select QTL-effect model. Once users press the **DO** button, these things may be executed, until GCIM Format dialog box disappear and **info** window is appeared. When click **Continue** button, the next work is to select covariate and to run the program. If click **Cancel** button, the next work is not to execute these above things and GCIM Format dialog box is also disappeared.

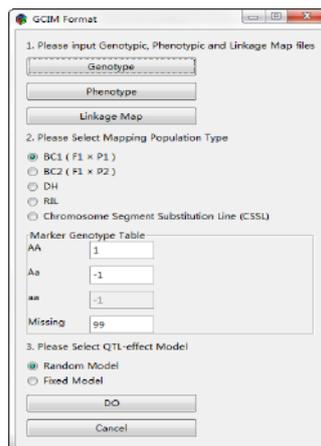


Figure 3.5.2. GCIM Format Dataset dialog of **GCIM** module

### 3.5.1.2 QTLIciMapping Format

To click QTLIciMapping Format will appear QTLIciMapping dialog box. In the dialog box, there are three steps. Firstly, **QTLIciMapping\_Format** button is used to input QTLIciMapping format datasets. Once the file is successfully uploaded, one tabbed page is added to the software notebook. Secondly, the user selects the required population type (see §3.5.1). Thirdly, the users select the required QTL-effect model type (see §3.5.1). Finally, two things will be implemented in this step: 1) to upload QTLIciMapping-formatted files; and 2) to select population and model types. Once

users press the **DO** button, the two things may be executed. Three tabbed pages (Genotype, Phenotype and Linkage map) will be added to the software notebook, and QTL Ici Mapping Format dialog box will be disappeared. To click **Cancel** button is not to execute the above things and QTL Ici Mapping Format dialog box is also disappeared.

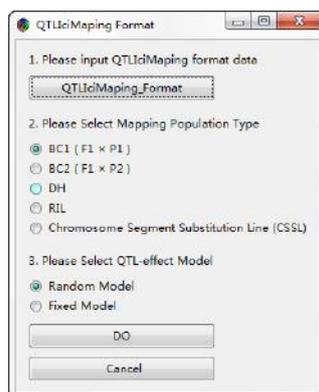


Figure 3.5.3. QTL Ici Mapping Format Dataset dialog of **GCIM** module

### 3.5.1.3 WinQTL Cart Format

To click WinQTL Cart will appear WinQTL Cart dialog box. In the dialog box, there are three steps. Firstly, **WinQTL Cart Format** button is used to input WinQTL Cart format datasets. Once the related files are successfully uploaded, the tabbed pages are added to the software notebook. Secondly, the users select the required population type (see §3.5.1), and the users select the required model type (see §3.5.1). Finally, two things will be implemented in this step: 1) to upload WinQTL Cart format files; and 2) to select population and model types. Once users press the **DO** button, the two things may be executed, three tabbed pages (Genotype, Phenotype and Linkage map) will be added to the software notebook, and WinQTL Cart Format dialog box will be disappeared. If to click **Cancel** button, the next work is not to execute the above two things and WinQTL Cart Format dialog box is also disappeared.

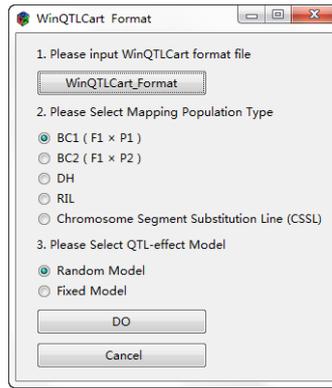


Figure 3.5.4. WinQTLCart Format Dataset dialog of **GCIM** module

### 3.5.1.4 Covariate

To click **Covariate** button will appear **Covariate** dialog box. In the dialog box, covariate(s) may be not included in the genetic model if there is no covariate. If not, the covariates should be included in the genetic model. Once the covariate file is successfully uploaded, the tabbed page will add to the software notebook.

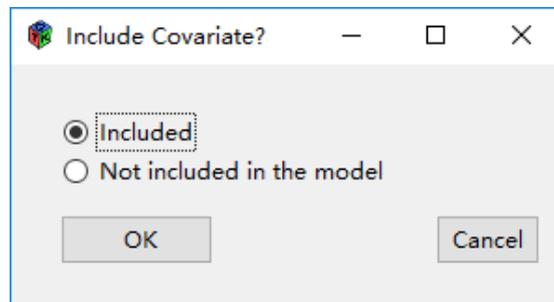


Figure 3.5.5. The Covariate Format Dataset dialog of **GCIM** module

### 3.5.2 Run Program

Before run the program, please set up **the Critical LOD score**, which is defaulted by the 2.5 (LOD) value. If a LOD score for a putative QTL is larger than the Critical LOD score, the putative QTL is viewed as true. If not, the putative QTL is viewed as false. Of course, the critical LOD score may be determined by permutation tests. If doing so, the software **WinQTLCart** is available. **Walk Speed for Genome-wide Scanning (cM)** is defaulted by the 1.0 (cM) value, which may be modified by users. **Trait Selection** is defaulted by the first trait. If user wants to analyze the third trait,

the value in Trait Select column is set up by “3”. In the GCIM, to click Run button will execute the software. If the program runs, a progress bar with the “**Please be patient...**” words will appear in the bottom of the interface. If the program finished, a bar with the “**All done.**” will appear.

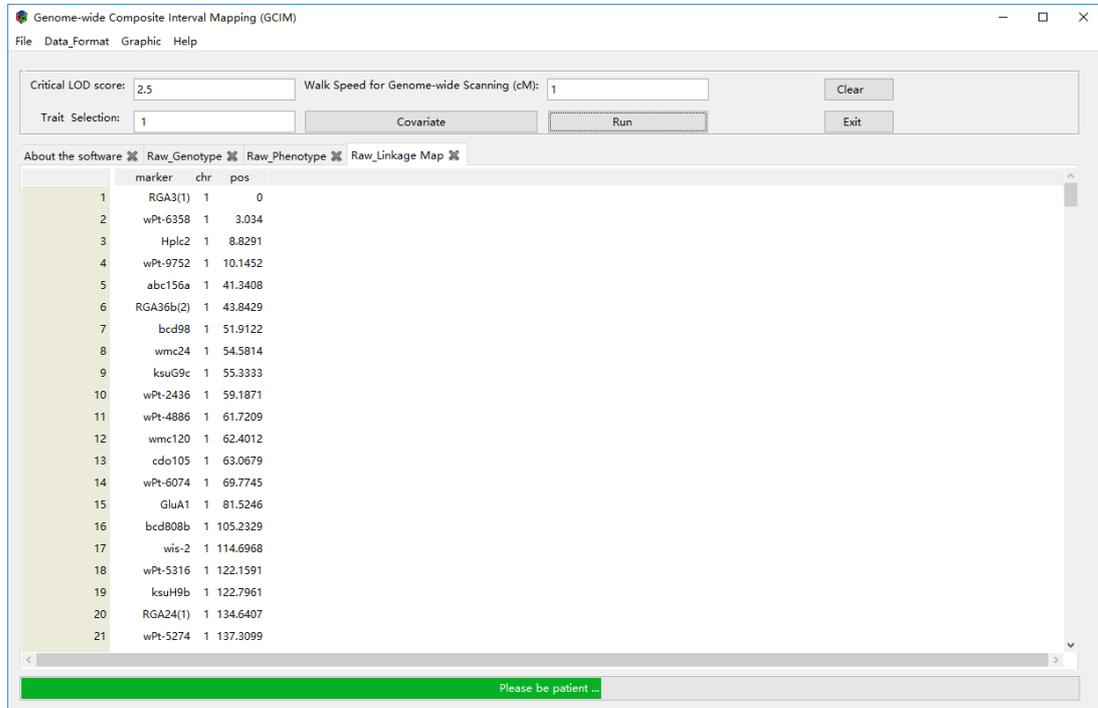


Figure 3.5.6. A running program interface of **GCIM** module

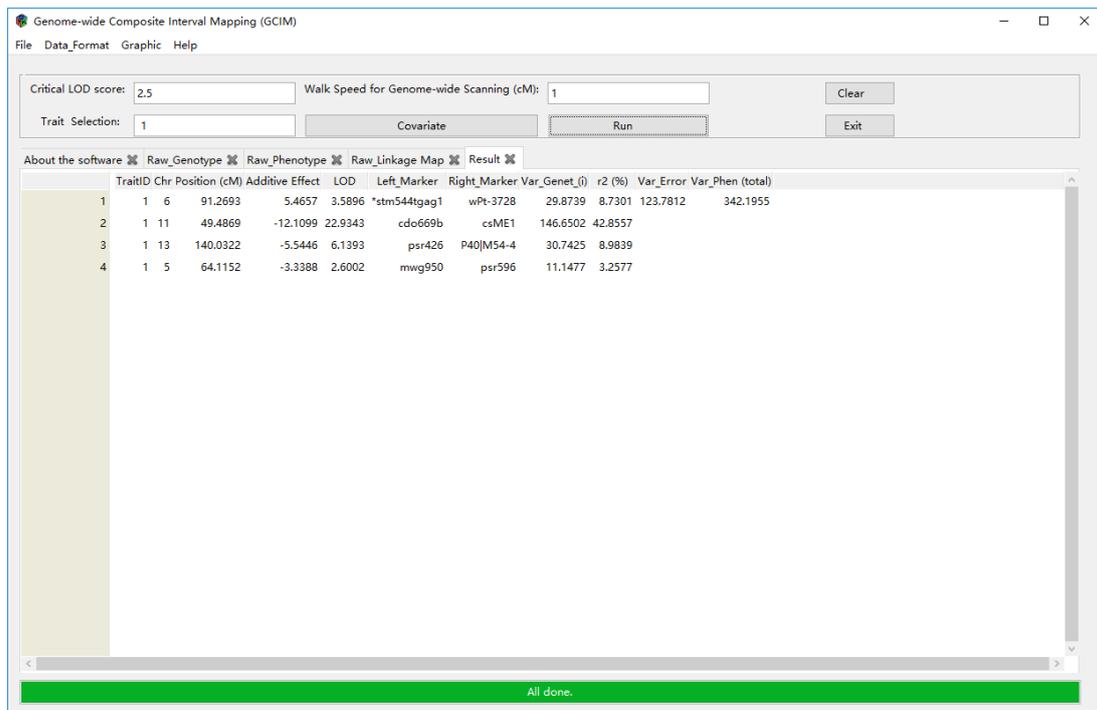


Figure 3.5.7. A finished program interface (the **GCIM** Results: [Trait 1](#))

### 3.5.3 Output results

To click **save** button in the **File** menu is used to save the result as \*.csv file. When a trait is analyzed, the results may be saved, a dialog is used to select the pathway and the file name for the saved file.

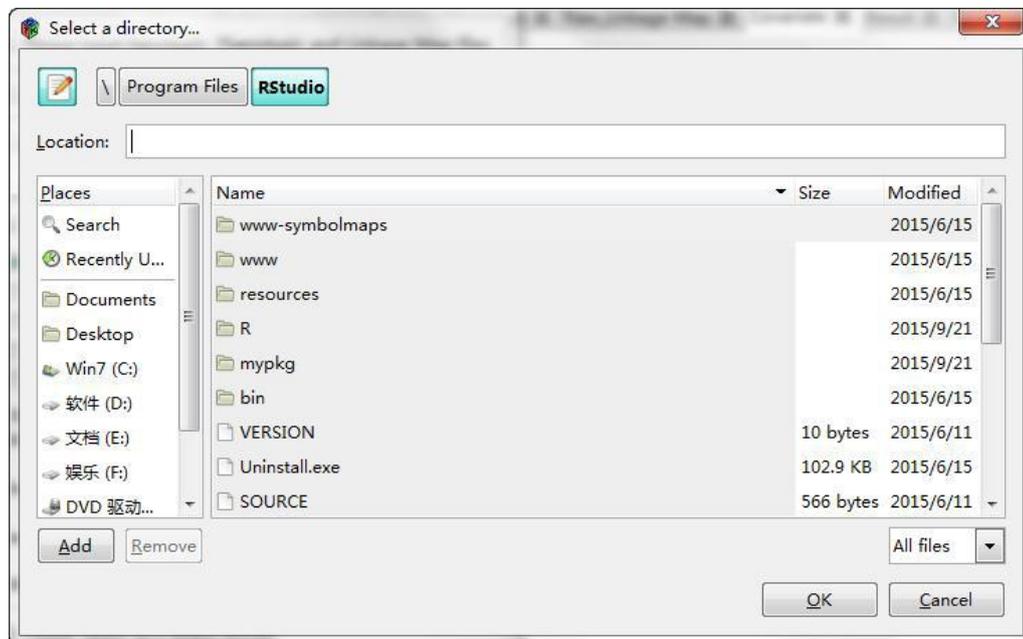


Figure 3.5.8. The Save dialog of **GCIM** module

The **Results** are listed in ten columns for the GCIM (Genome-wide Composite Interval Mapping) method. The corresponding column names are as follows:

TraitID: Trait ID represented by an integer number.

Chr: Raw name of chromosome or chromosome ID represented by an integer number.

Position (cM): The scanning position (cM) on the chromosome.

Additive Effect: Estimated additive effect of the putative QTL.

LOD: LOD score for the putative QTL.

Left\_Marker: Left flanking marker name of current scanning position (or putative QTL).

Right\_Marker: Right flanking marker name of current scanning position (or putative QTL).

Var\_Genet\_(i): Genetic variance for all the detected QTL

r2 (%): proportion of phenotypic variance explained by single QTL.

Var\_Error: residual variance for the full model.

Var\_Phen (total): Phenotypic variance.

	A	B	C	D	E	F	G	H	I	J	K
1	TraitID	Chr	Position (cM)	Additive Effect	LOD	Left_Marker	Right_Marker	Var_Genet_(i)	r2 (%)	Var_Error	Var_Phen (total)
2		1	6	91.2693	5.4657	3.5896 *stm544tgagl	wPt-3728	29.8739	8.7301	123.7812	342.1955
3		1	11	49.4869	-12.1099	22.9343 cdo669b	csME1	146.6502	42.8557		
4		1	13	140.0322	-5.5446	6.1393 psr426	P40M54-4	30.7425	8.9839		
5		1	5	64.1152	-3.3388	2.6002 mwg950	psr596	11.1477	3.2577		
6											
7											
8											
9											
10											
11											
12											
13											

Figure 3.5.9. The results of in the **GCIM**

### 3.5.4 Draw plot

If the running is ended, user may visualize the result by selecting **plot** button in the Graphic menu. Before clicking **Save** button in the **Genome-wide composite interval mapping (GCIM) figure** window, you can set parameter according to your demand. The parameter settings in Figure 3.5.10 are for **general resolution**.

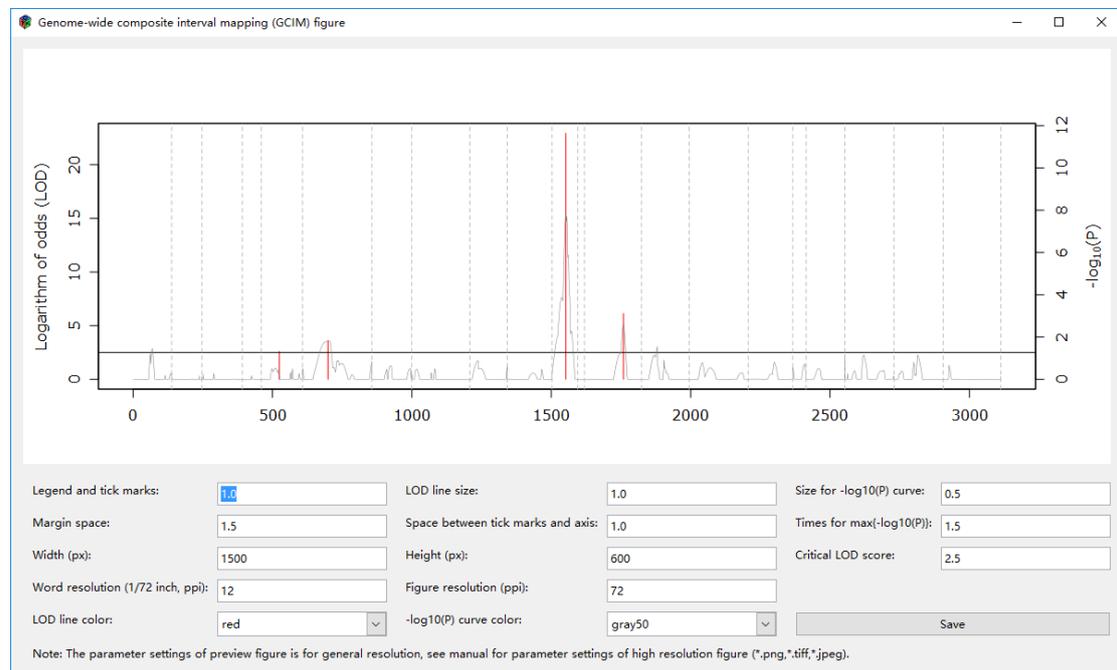


Figure 3.5.10. Genome-wide composite interval mapping (**GCIM**) figure using example data [Using example dataset in this software](#), the preview GCIM plots with the general resolution parameter settings were shown in **Figure 3.5.10**. [Using real dataset of the Triticale DH population in Wurschum et al. \(2014\) Genetics](#), the preview GCIM plot

with the general resolution parameter settings was shown in **Figure 3.5.11**.

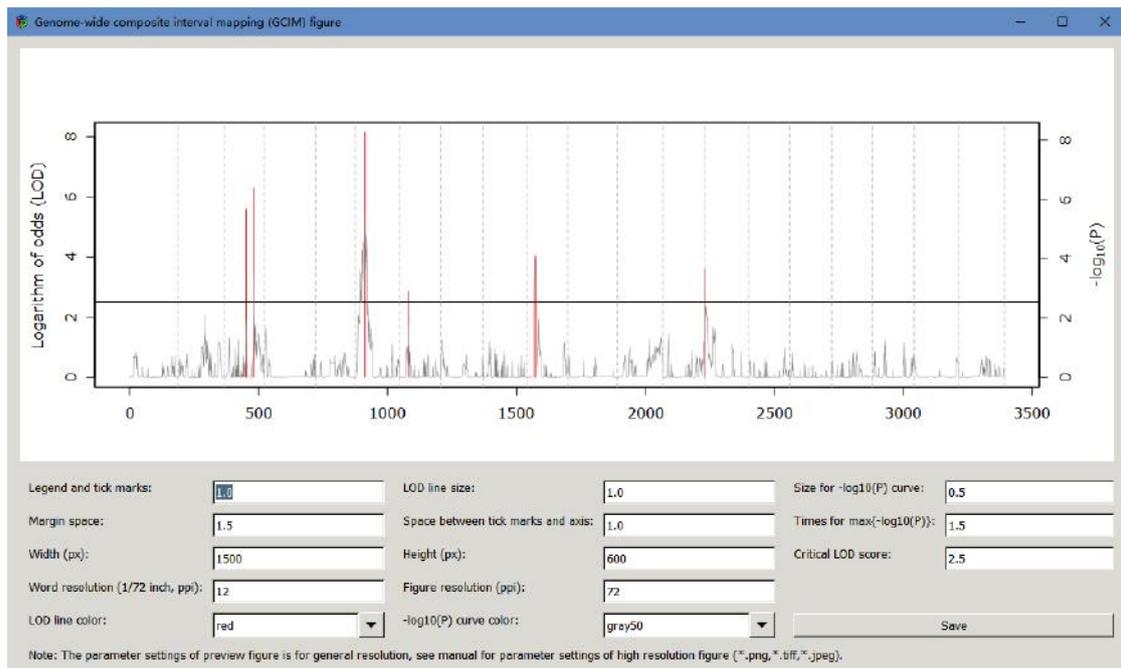


Figure 3.5.11. Genome-wide composite interval mapping (**GCIM**) figure using real data

If you want to obtain **high resolution** figure, we recommend the parameter settings shown in Figure 3.5.12. It is emphasized that preview figure may not seem same with **saved figure**. To give priority to save figure in application.

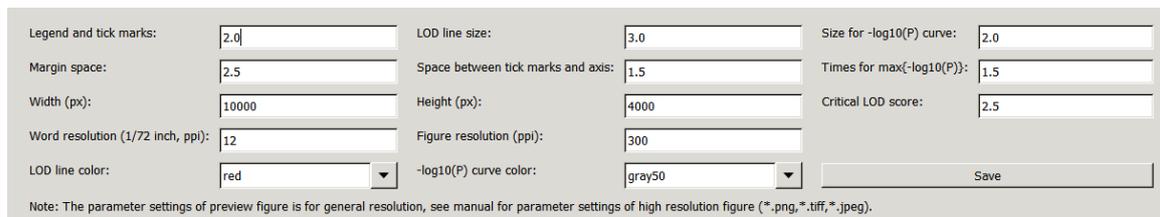


Figure 3.5.12. Parameter settings for high resolution figure of **GCIM** module

## 4 References

1. Zhang Y-M, Mao Y, Xie C, Smith H, Luo L, Xu S\*. 2005. Mapping quantitative trait loci using naturally occurring genetic variance among commercial inbred lines of maize (*Zea mays* L.). *Genetics* 169: 2267-2275
2. 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.
3. Wang Shi-Bo, Wen Yang-Jun, Ren Wen-Long, Ni Yuan-Li, Zhang Jin, Feng Jian-Ying, Zhang Yuan-Ming\*. Mapping small-effect and linked quantitative trait loci for complex traits in backcross or DH populations via

a multi-locus GWAS methodology. *Scientific Reports* 2016, 6: 29951.

4. 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, DOI: 10.1093/bib/bbw145
5. 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, DOI: 10.1371/journal.pcbi.1005357.
6. Zhang Jin<sup>#</sup>, Feng Jian-Ying<sup>#</sup>, 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, Accepted