User Manual for mrMLM

1. Introduction

mrMLM, a R package, aims to provide a user-friendly interface to conduct genome-wide association study (GWAS) via a multi-locus random-SNP-effect mixed linear model (mrMLM) methodology and to visualize the results. It works on the platforms of Windows, Linux and MacOS. The GUI is based on available add-on package RGtk2, via the aid of another package gWidgetsRGtk2. The visualization of results is based on package qqman, such as Manhattan and QQ plots.

2. Installation

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 complete the installation of Gtk2 libraries.

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).

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/).

2.2 Install R

Download R from CRAN (https://cran.r-project.org/) and install it by running the file.

2.3 Install the R 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/). Install them in order, as some depend on others. Then install them directly within R environment using the below command:

install.packages(pkgs=c("RGtk2","cairoDevice","gWidgets","gWidgetsRGtk2","RGtk2Extras","qqman"))

2.4 Install mrMLM

The mrMLM package is freely available at the CRAN (https://cran.r-project.org/web/packages/mrMLM/index.html or soyzhang@mail.hzau.edu.cn or soyzhang@hotmail.com), you can download or request this R software. Then install mrMLM software directly within R environment using the below command:

install.packages(pkgs="mrMLM")

3. Running

The procedure is described below. Launch mrMLM within R environment by command: *library(mrMLM)*, after that the GUI of mrMLM turned out like the following dialog.

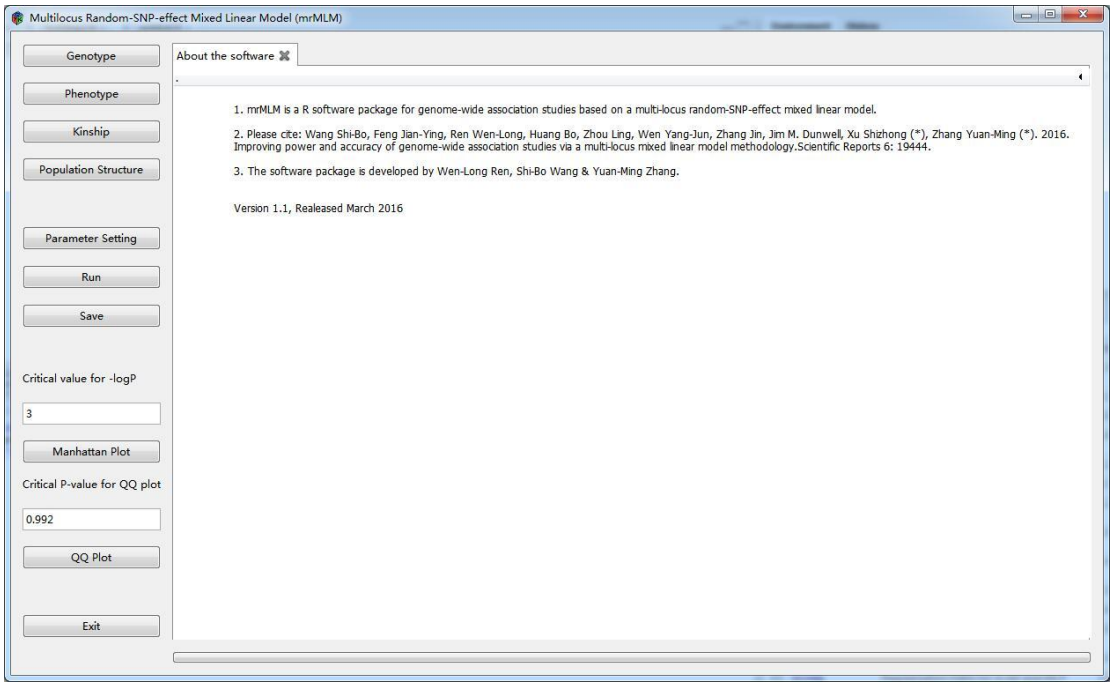


Figure 1. A snapshot of the GUI of mrMLM

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

3.1 Input Files

Use the buttons of **Genotype**, **Phenotype**, **Kinship** and **Population Structure** to import the datasets of genotypes, trait observations (phenotypes), kinship matrix and population structure matrix, respectively. If one file is imported successfully, there will be one tabbed page added to the notebook. Meanwhile, the **Kinship** has two options: one is to import directly the known kinship matrix and another is to compute the kinship matrix using the mrMLM software. And the **Population Structure** has two options as well. If population structure has no effect on GWAS, the population structure matrix may be not included in the mixed linear model of the GWAS. If not, you should incorporate the population structure matrix into your GWAS.

Note: About the **input files format** in details, please see **Direction 1** in the end of the manual.

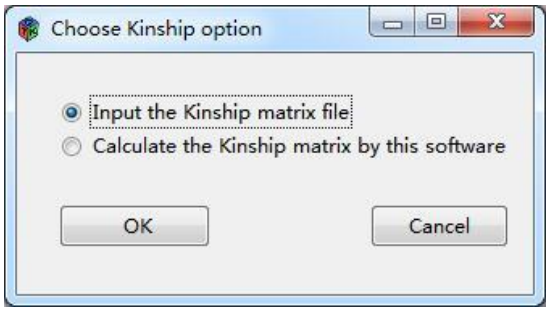


Figure 2. A snapshot of the **Kinship** dialog

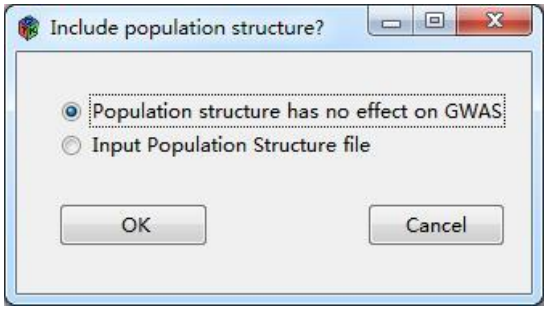


Figure 3. A snapshot of the **population structure** dialog

3.2 Run Program

Use the **Parameter Setting** button to set parameters before run the program. If not, the parameters will keep the default settings of related parameters. "Critical P-value in rMLM" is used to determine the number of SNPs in the rMLM. If the critical P-value is changed from 0.01 to 0.05, the number of SNPs will increase in the rMLM. "Search radius of candidate gene (kb)" means to choose the only one marker with least P-value and to abandon the other markers in the search radius. Use the **Run** button to execute the software. If the program starts, there will be a progress bar showing the running status and "**Please be patient...**" words in the bottom of the interface. If the work have finished, there will be "**All done.**" instead of "**Please be patient...**".

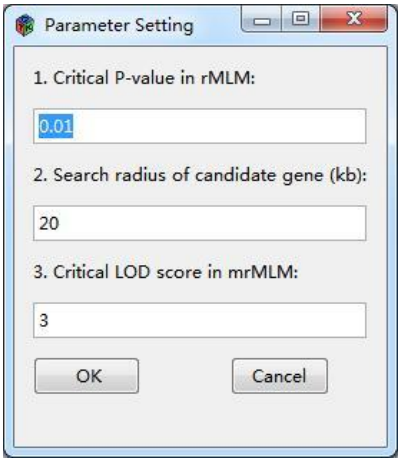


Figure 4. A snapshot of **Parameter Setting** dialog

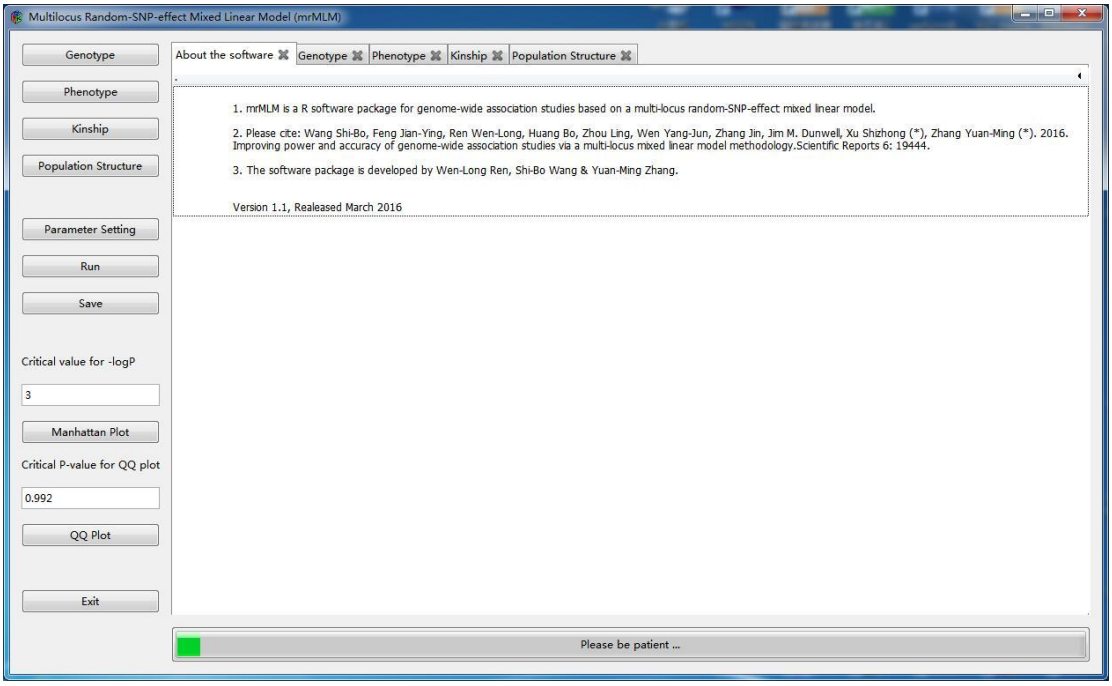
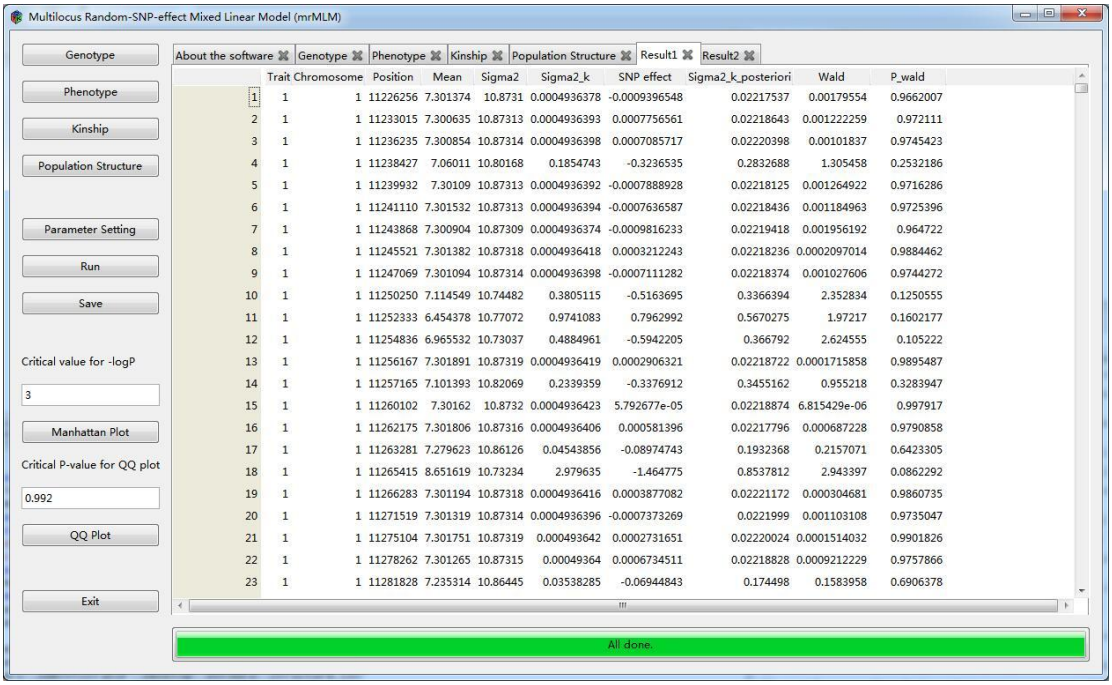


Figure 5. A snapshot of a **running** program



	Trait	Chromosome	Position	Mean	Sigma2	Sigma2_k	SNP effect	Sigma2_k_posteriori	Wald	P_wald
1	1	1	11226256	7.301374	10.8731	0.0004936378	-0.0009396548	0.02217537	0.00179554	0.9662007
2	1	1	11233015	7.300635	10.87313	0.0004936393	0.0007756561	0.02218643	0.001222259	0.972111
3	1	1	11236235	7.300854	10.87314	0.0004936398	0.0007085717	0.02220398	0.00101837	0.9745423
4	1	1	11238427	7.06011	10.80168	0.1854743	-0.3236535	0.2832688	1.305458	0.2532186
5	1	1	11239932	7.30109	10.87313	0.0004936392	-0.0007888928	0.02218125	0.001264922	0.9716286
6	1	1	11241110	7.301532	10.87313	0.0004936394	-0.0007636587	0.02218436	0.001184963	0.9725396
7	1	1	11243868	7.300904	10.87309	0.0004936374	-0.0009816233	0.02219418	0.001956192	0.964722
8	1	1	11245521	7.301382	10.87318	0.0004936418	0.0003212243	0.02218236	0.0002097014	0.9884462
9	1	1	11247069	7.301094	10.87314	0.0004936398	-0.0007111282	0.02218374	0.001027606	0.9744272
10	1	1	11250250	7.114549	10.74482	0.3805115	-0.5163695	0.3366394	2.352834	0.1250555
11	1	1	11252333	6.454378	10.77072	0.9741083	0.7962992	0.5670275	1.97217	0.1602177
12	1	1	11254836	6.965532	10.73037	0.4884961	-0.5942205	0.366792	2.624555	0.105222
13	1	1	11256167	7.301891	10.87319	0.0004936419	0.0002906321	0.02218722	0.0001715858	0.9895487
14	1	1	11257165	7.101393	10.82069	0.2339359	-0.3376912	0.3455162	0.955218	0.3283947
15	1	1	11260102	7.30162	10.8732	0.0004936423	5.792677e-05	0.02218874	6.815429e-06	0.997917
16	1	1	11262175	7.301806	10.87316	0.0004936406	0.000581396	0.02217796	0.000687228	0.9790858
17	1	1	11263281	7.279623	10.86126	0.04543856	-0.08974743	0.1932368	0.2157071	0.6423305
18	1	1	11265415	8.651619	10.73234	2.979635	-1.464775	0.8537812	2.943397	0.0862292
19	1	1	11266283	7.301194	10.87318	0.0004936416	0.0003877082	0.02221172	0.000304681	0.9860735
20	1	1	11271519	7.301319	10.87314	0.0004936396	-0.0007373269	0.0221999	0.001103108	0.9735047
21	1	1	11275104	7.301751	10.87319	0.000493642	0.0002731651	0.02220024	0.0001514032	0.9901826
22	1	1	11278262	7.301265	10.87315	0.00049364	0.0006734511	0.02218828	0.0009212229	0.9757866
23	1	1	11281828	7.235314	10.86445	0.03538285	-0.06944843	0.174498	0.1583958	0.6906378

Figure 6. A snapshot of finished program (**Results in the rMLM**) (**Result1**)

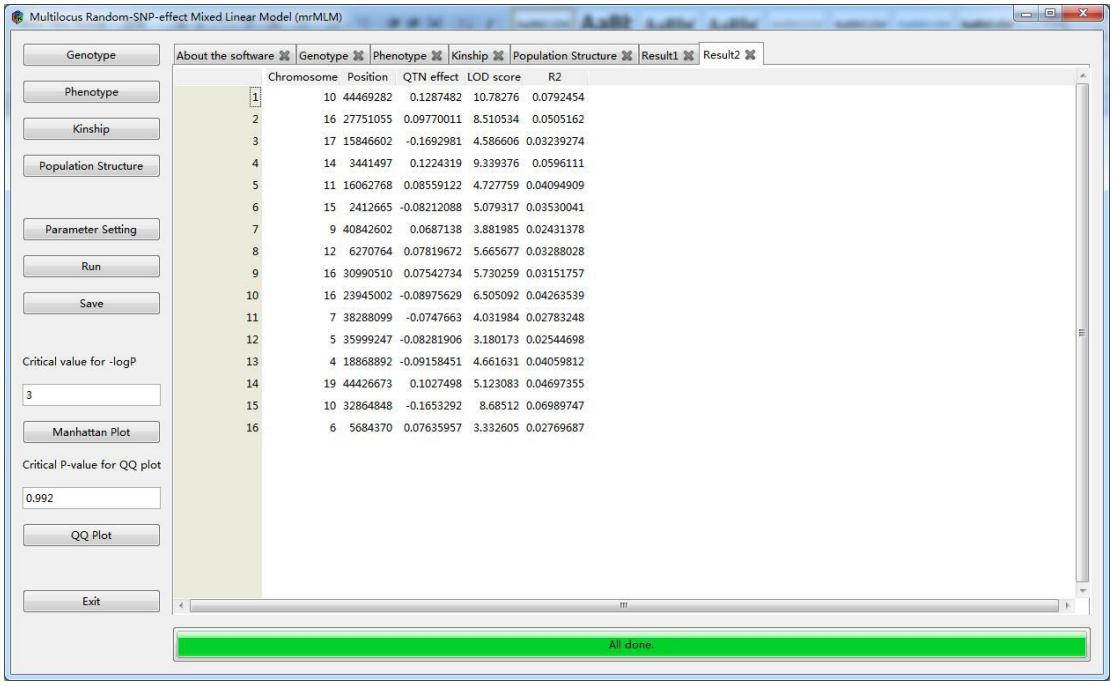


Figure 7. A snapshot of finished program (**Results in the mrMLM**) (**Result2**)

3.3 Save Results

Use **Save** button to save the results as ***.csv** format file. There will have an option to save **Results in the rMLM** (**Result1**) or **Results in the mrMLM** (**Result2**). If click **OK** button (after **Save** button), a dialog is used to choose the pathway and the file name of the saving results.

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

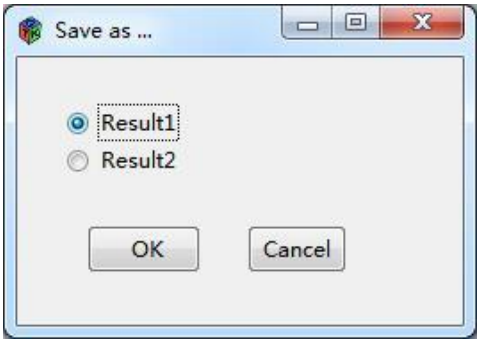


Figure 8. A snapshot of **Save** choice dialog

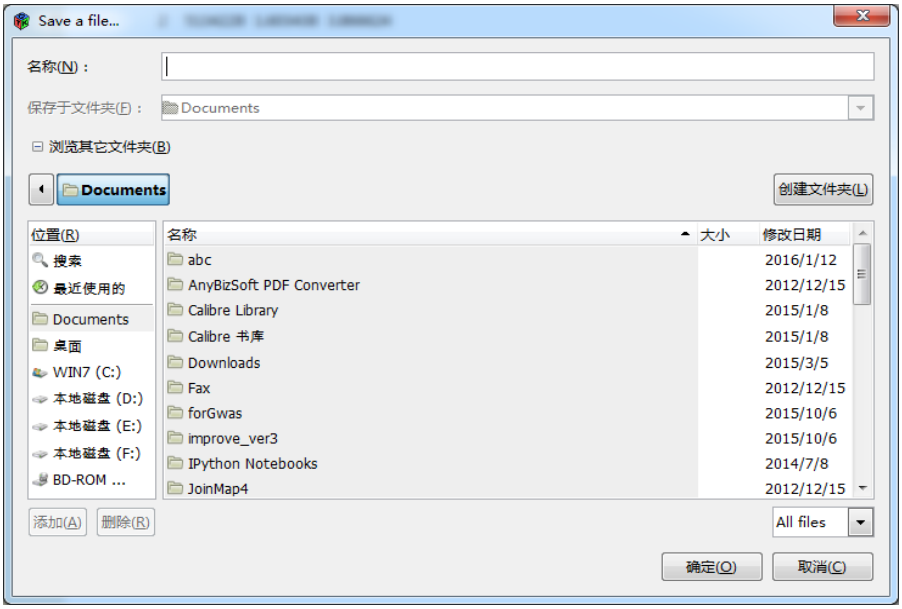


Figure 9. A snapshot of **Save** dialog

Warning: It is better not to include other languages except English in the pathway and file name. Otherwise, there may be something wrong.

4. Visualization of Results

If the program have finished, you can have the visualization of the results. Before use the **Manhattan Plot** button, please set the critical value for $-\log_{10}(P)$, which is defaulted the value of 3. Before use the **QQ Plot** button, please set the critical P-value for QQ plot, which is defaulted the value of 0.992. This is because the P-values are a mixture of a χ^2 distribution with one degree of freedom and a point mass at one.

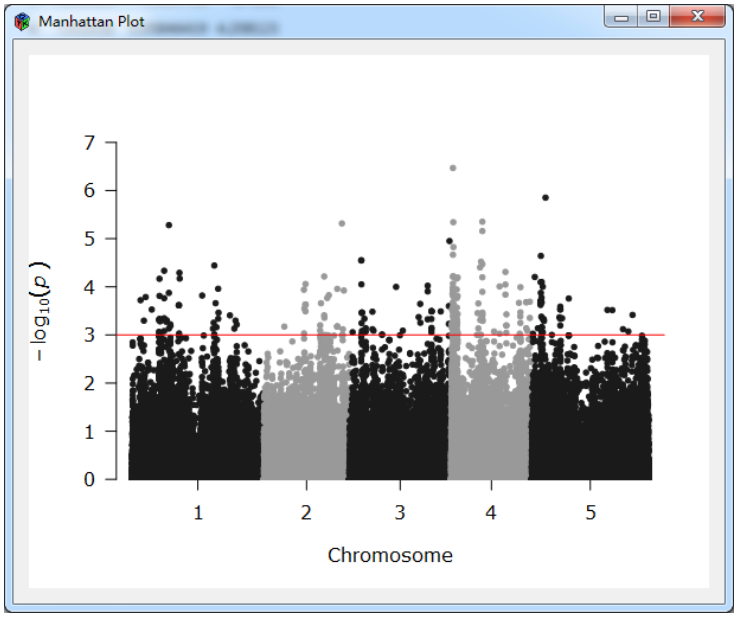


Figure 10. A snapshot of **Manhattan Plot**

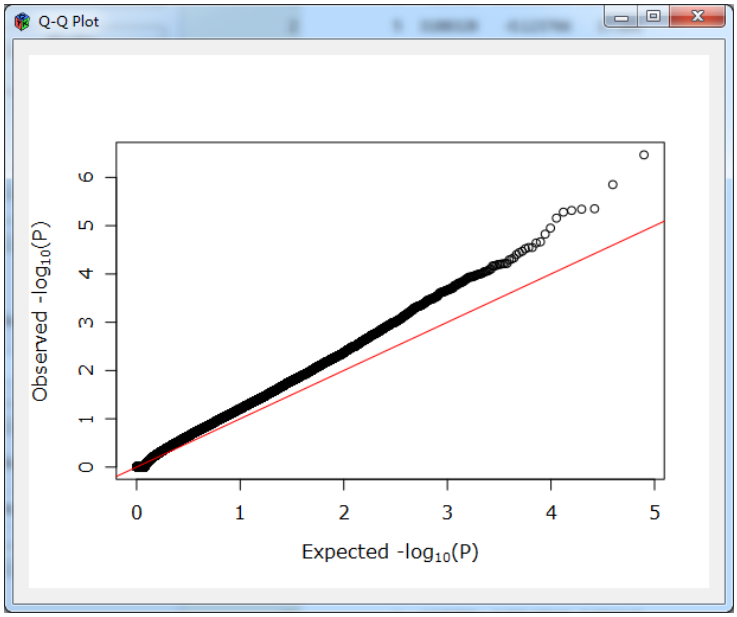


Figure 11. A snapshot of **QQ Plot**

Directions

Direction 1: Explanation of input files in details

D1.1 The Genotype file:

The **Genotype** file should be a ***.csv** format file. The first column stands for chromosomes, the second column stands for the marker's position (bp) in the chromosomes and the remaining columns stand for samples. The row stands for marker locus and the column is for each individual. The marker values should be numeric values, and the characters such as A, T, C, G are not accepted. The file does not need the row and column names.

	A	B	C	D	E	F	G	H
1	1	7166	1	1	1	1	1	1
2	1	31727	1	1	1	1	1	1
3	1	31734	1	1	1	1	1	1
4	1	32926	1	1	-1	-1	-1	-1
5	1	83093	1	1	-1	-1	-1	-1
6	1	93692	1	1	1	1	1	1
7	1	121762	1	1	1	1	1	1
8	1	121775	1	1	1	1	1	1
9	1	128283	1	1	1	1	1	1
10	1	175580	1	1	1	1	1	1
11	1	195402	1	1	-1	-1	1	1
12	1	195406	1	1	1	1	1	1
13	1	212039	1	1	-1	-1	1	1
14	1	226616	1	1	1	1	1	1
15	1	230076	1	1	-1	-1	-1	-1

Figure D1.1. A snapshot of part of **Genotype** file

D1.2 The Phenotype file:

The Phenotype File should be a ***.csv** format file. The rows stand for samples. The column stand for trait. Now the phenotype file can include only one trait once. The file did not need row names and column names.

	A
1	11.678
2	10.146
3	11.325
4	12.101
5	12.336
6	12.687
7	12.912
8	12.101
9	13.058
10	11.181
11	11.366
12	10.279
13	11.471
14	13.295
15	12.721

Figure D1.2. A snapshot of part of **Phenotype** file

D1.3 The Kinship file

The Kinship file should be a ***.csv** format file. The number of rows (or columns) equals to the number of individuals in the sample. And the kinship matrix is a symmetric matrix. The file does not need the row and column names.

	A	B	C	D	E	F	G	H
1	1	0.432	0.282	0.264	0.48	0.314	0.248	0.244
2	0.432	1	0.266	0.252	0.364	0.33	0.252	0.244
3	0.282	0.266	1	0.662	0.266	0.224	0.602	0.606
4	0.264	0.252	0.662	1	0.332	0.178	0.496	0.488
5	0.48	0.364	0.266	0.332	1	0.306	0.32	0.316
6	0.314	0.33	0.224	0.178	0.306	1	0.154	0.15
7	0.248	0.252	0.602	0.496	0.32	0.154	1	0.952
8	0.244	0.244	0.606	0.488	0.316	0.15	0.952	1

Figure D1.3. A snapshot of part of the **Kinship** file

D1.4 The Population Structure file

The **Population Structure** file should be a ***.csv** format file. The population structure matrix may be calculated from the **Structure** software. If population structure matrix has ***k*** columns, please input its arbitrary **(*k* − 1)** columns. Each row stands for each individual from 1 to *n* (sample size). The file does not need the row and column names.

	A	B
1	0.016	0.976
2	0.01	0.921
3	0.001	0.998
4	0.006	0.993
5	0.281	0.714
6	0.247	0.425
7	0.004	0.995
8	0.002	0.996
9	0.004	0.995
10	0.485	0.402
11	0.322	0.676
12	0.997	0.002
13	0.98	0.001
14	0.924	0.066
15	0.519	0.336

Figure D1.4. A snapshot of part of the **Population Structure** file

Direction 2: Explanation of Result1 and Result2 in details

D2.1 Explanation of Result1 in details

The **Result1** table with ten columns shows the results from the rMLM (random-SNP-effect mixed linear model) method. The corresponding column names are as follows: trait code (such as 1), chromosome, marker's position (bp) in the chromosome, population mean value (Mean), residual variance (σ^2, Sigma2), priori variance of the *k*th SNP effect (ϕ_k^2, Sigma2_*k*), SNP effect (γ_k, Effect), posteriori variance of SNP effect ($\mathrm{var}(\gamma_k)$, Sigma2_*k*_posteriori), Wald test statistic value, and the P-value of Wald test, respectively.

	A	B	C	D	E	F	G	H	I	J
1	Trait	Chromosome	Position	Mean	Sigma2	Sigma2_k	SNP effect	Sigma2_k_posteriori	Wald	P_wald
2	1	1	7166	3.950667	0.066528	0.006193	-0.05462971	0.055971464	0.952631	0.329051
3	1	1	31727	3.898178	0.066507	0.002284	-0.03394233	0.033625871	1.018911	0.312777
4	1	1	31734	3.898178	0.066507	0.002284	-0.03394233	0.033625871	1.018911	0.312777
5	1	1	32926	3.898684	0.066751	3.03E-06	4.55E-05	0.00173725	0.000684	0.979128
6	1	1	83093	3.898672	0.06675	3.03E-06	7.62E-05	0.001737648	0.001922	0.965029
7	1	1	93692	3.898689	0.066751	3.03E-06	-3.53E-05	0.001740023	0.000411	0.983821
8	1	1	121762	3.898688	0.066751	3.03E-06	5.74E-06	0.001740777	1.09E-05	0.99737
9	1	1	121775	3.898688	0.066751	3.03E-06	5.74E-06	0.001740777	1.09E-05	0.99737
10	1	1	128283	3.898696	0.066751	3.03E-06	-8.65E-06	0.001740574	2.47E-05	0.996033
11	1	1	175580	3.898683	0.066751	3.03E-06	1.52E-05	0.001740737	7.63E-05	0.993031
12	1	1	195402	3.898643	0.066751	3.03E-06	5.96E-05	0.00173767	0.001177	0.972629
13	1	1	195406	3.898674	0.066751	3.03E-06	2.73E-05	0.001740386	0.000246	0.987496
14	1	1	212039	3.891241	0.06667	0.000307	0.008785774	0.015120777	0.337607	0.561214
15	1	1	226616	3.898679	0.066751	3.03E-06	1.84E-05	0.001740656	0.000112	0.991561

Figure D2.1. A snapshot of Results in the rMLM (**Result1**)

D2.2 Explanation of Result2 in details

The Result2 table with five columns shows the final results of the mrMLM method.

The corresponding column names are as follows: chromosome, marker's position (bp) in the

chromosome, QTN effect, LOD score, and the proportion of phenotypic variance explained by the putative QTN, respectively.

	A	B	C	D	E
1	Chromosome	Position	QTN effect	LOD score	R2
2	10	44469282	0.128748206	10.78275608	0.079245
3	16	27751055	0.097700109	8.510534059	0.050516
4	17	15846602	-0.16929809	4.586605736	0.032393
5	14	3441497	0.122431926	9.339376295	0.059611
6	11	16062768	0.085591218	4.727758577	0.040949
7	15	2412665	-0.08212088	5.079317195	0.0353
8	9	40842602	0.068713798	3.881985266	0.024314
9	12	6270764	0.078196725	5.665677397	0.03288
10	16	30990510	0.075427343	5.730258539	0.031518
11	16	23945002	-0.08975629	6.505092387	0.042635
12	7	38288099	-0.0747663	4.031984169	0.027832
13	5	35999247	-0.08281906	3.180173055	0.025447
14	4	18868892	-0.09158451	4.661630749	0.040598
15	19	44426673	0.102749777	5.12308303	0.046974

Figure D2.2. A snapshot of Results in the mrMLM (**Result2**)