User Manual for

lll V m r M L M

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

(**version 1.0**)

Mei Li, Ya-Wen Zhang, Yuan-Ming Zhang

(soyzhang@mail.hzau.edu.cn)

Last updated on May, 2022

Disclaimer: While extensive testing has been performed by Yuan-Ming Zhang's Lab at College of Plant Science and Technology, Huazhong Agricultural University, the results are, in general, reliable, correct, and appropriate. However, results are not guaranteed for any specific datasets. We strongly recommend that users integrate the IIIVmrMLM results with those from other software packages, i.e., mrMLM, GEMMA, EMMAX, and PLINK.

Citation:

Li, M., Zhang, Y.W., Zhang, Z.C., Xiang, Y., Liu, M.H., Zhou, Y.H., Zuo, J.F., Zhang, H.Q., Chen, Y., and Zhang, Y.M. (2022). A compressed variance component mixed model for detecting QTNs and QTN-by-environment and QTN-by-QTN interactions in genome-wide association studies. Mol. Plant **15**:630–650.

The work was supported by the National Natural Science Foundation of China (32070557 and 31871242), and Huazhong Agricultural University Scientific & Technological Self-Innovation Foundation (2014RC020).

1. Introduction

1.1 Why IIIVmrMLM?

IIIVmrMLM (3 Variance-component multi-locus random-SNP-effect Mixed Linear Model) program is an R package of multi-locus genome-wide association studies (GWAS), which identifies main-effect QTNs and QTN-by-environment and QTN-by-QTN interactions (QEIs and QQIs) for complex and multi-omics traits.

1.2 Getting started

Before users install our IIIVmrMLM package, users should first install the newest versions of R, Rtools, and Rstudio software.

The software package IIIVmrMLM runs only in the R software environment and can be requested from the maintainer, Dr Yuan-Ming Zhang at College of Plant Science and Technology, Huazhong Agricultural University (soyzhang@mail.hzau.edu.cn).

1.2.1 One-Click installation

Within R environment, the IIIVmrMLM software can be installed as below.
First install the dependency packages:
install.packages(c("lars","RcppEigen","Rcpp","doParallel","data.table","MASS","openxlsx","BEDMatrix","bigmemory","stringr","biglasso","progress","ncvreg","coin","sampling","sbl"))

Then install the IIIVmrMLM package from local files.

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

1.2.2 Run IIIVmrMLM

Once the software IIIVmrMLM is installed, users may run it using two commands:
library("IIIVmrMLM")
IIIVmrMLM(***) (***: please see the example of § 2.2, 2.3 and 2.4)

Users need to run **library("IIIVmrMLM")** every time before users use this software package.

2. Function IIIVmrMLM

2.1 Parameter settings

Parameter	Meaning	File format	Note
fileGen	Name & path of genotypic file in your computer, i.e., fileGen="D:/Users/Genotype".	PLINK binary files: **Genotype.bed+Genotype.bim+Genotype.fam**	
filePhe	Name & path of trait phenotypic file in your computer, i.e., filePhe="D:/Users/Phenotype.csv".	*.csv; *.txt (Phenotypic values. **Row**: individual; **Column**: traits)	Table 1
fileKin	Name & path of individual kinship file in your computer, i.e., fileKin="D:/Users/Kinship.csv" or fileKin=NULL.	*.csv; *.txt (Kinship matrix. **Row** & **Column**: individuals)	Table 2
filePS	Name & path of population structure file in your computer, i.e., filePS="D:/Users/PopStr.csv" or filePS=NULL.	*.csv; *.txt [Population structure. **Row**: individual; **Column**: sub-populations 1, 2, ···, *k* (No. of sub-populations)]	Tables 3~5
PopStrType	The types of population structure include *Q* (Q matrix), PCA (principal components), and EvolPopStr (evolutionary population structure).		
fileCov	Name & path of covariate file in your computer, i.e., fileCov="D:/Users/Covariate.csv" or fileCov=NULL.	*.csv; *.txt (**Row**: individual; **Column**: covariates 1, 2, ···, *k* (no. of covariates)) **Cate**: categorical variable; **Con**: continuous variable	Table 6
method	Three GWAS methods: single- and multi-environment analysis, and epistasis detection, i.e., method="Single_env", method="Multi_env", or method="Epistasis".		
trait	Traits analyzed from number 1 to number 2, i.e., **trait=1:3** indicates that users analyze the first to third traits.		
	If method="Multi_env", users need to add a parameter **n.en** to indicate the number of environments for each trait in the **filePhe**, i.e., **trait=1:2** (Analyzing the first to second traits)**; n.en=c(2,2,3)** (The filePhe file contains the phenotypic values of three traits, and the environmental numbers of each trait are 2,2,and 3, respectively.)		
SearchRadius	In the setup of decollinearity parameter **SearchRadius**, only one parameter should be set in QTN and QEI detection, i.e., **SearchRadius=20**, indicating the fact that only one potentially associated QTN or QEI is selected within 20 kb, while two parameters should be set in QQI detection, i.e., **SearchRadius=c(10,20)**, the first parameter "10" (kb) is set for main-effect QTNs and the second one "20" (kb) is set for QQIs.		
svpal	In the setup of critical P-value parameter **svpal**, the critical P-value 0.01 (**svpal=0.01**) is set to select all the potentially associated QTNs and QEIs in genome-wide single-marker scanning in QTN and QEI detection, while two critical P-values 0.01 and 0.01 are set to select main-effect QTNs and QQIs, respectively, in QQI detection, i.e., **svpal=c(0.01,0.01)**, the first parameter is for main-effect QTNs and the second one is for QQIs.		
sblgwas_t	Only needed **in QQI detection**, a number between [-3,0] to control sparseness of the "sblgwas" function, default value is -1.		
DrawPlot	**DrawPlot=FALSE:** no figure output; **DrawPlot=TRUE:** the output of Manhattan plot.		
Plotformat	The format for figures storage, including *.jpeg, *.png, *.tiff, and *.pdf. i.e, **Plotformat="pdf"** means *.pdf format.		
Chr_name_com	The common part of all chromosome names, deleted when drawing the Manhattan plot, i.e, "chr".		
dir	Save path of the result in your computer, i.e, "D:/Users".		

2.2 Single_env: The detection of main-effect QTNs for complex traits

IIIVmrMLM(fileGen="D:/Users/Genotype",filePhe="D:/Users/Phenotype.csv",
fileKin=NULL,filePS="D:/Users/PopStr.csv",PopStrType="Q",fileCov=NULL,
method="Single_env",trait=c(1:3),SearchRadius=20,svpal=0.01,
DrawPlot=TRUE,Plotformat="pdf",Chr_name_com=NULL,dir="D:/Users/")

Users must set "fileGen", "filePhe", "method", "trait", and "dir", while the other parameters may be default in function *IIIVmrMLM*, including *fileKin=NULL; filePS=NULL; PopStrType="Q"; fileCov=NULL; SearchRadius=20; svpal=0.01; DrawPlot=TRUE; Plotformat="pdf"; Chr_name_com=NULL*.

2.2.1 Data input format

Format for genotypic dataset "fileGen"

The file type of genotypes is "plink binary format" (Genotype.bed + Genotype.bim + Genotype.fam). The following provides a way to convert hapmap to plink binary files (*.bed + *.bim + *.fam) for users reference .

Under linux system, please install the Linux versions of the TASSEL and PLINK software first, and then conduct the four steps below:

1. First, set a path to the location where original and converted datasets are stored, e.g., running:

   cd /home/data

2. Then, use TASSEL to sort the hapmap file by running:

   /home/tassel-5-standalone/run_pipeline.pl -SortGenotypeFilePlugin -inputFile Genotype.hmp.txt -outputFile Genotype.sort.hmp.txt -fileType Hapmap

3. Next, use TASSEL to transform *.hmp.txt to *.vcf by running:

   /home/tassel-5-standalone/run_pipeline.pl -fork1 -h Genotype.sort.hmp.txt -Xmx10g -export -exportType VCF

   Note that *-Xmx10g* indicates that a maximum of 10G of memory is allocated to this step, and users can allocate it reasonably according to their own device conditions and the size of Hapmap data.

4. Finally, use PLINK to transform *.vcf to plink binary files (*.bed + *.bim + *.fam) by running:

   /home/PLINK/plink --vcf Genotype.vcf --make-bed --out Genotype

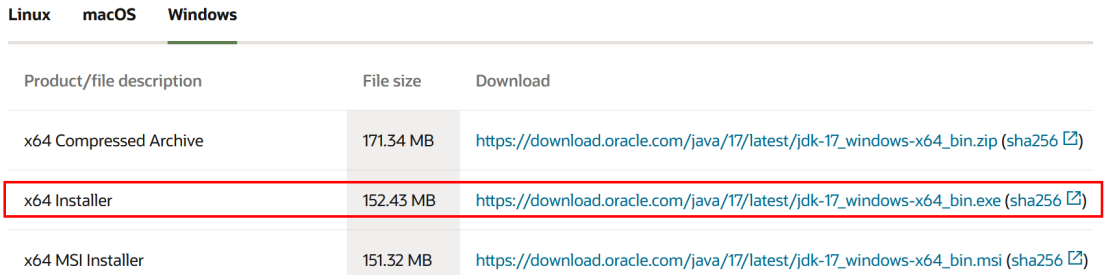
Under windows system, please install the Windows versions of the Java, TASSEL, and

PLINK software first (The use of TASSEL requires Java runtime environment). Note that Java and TASSEL should be installed, while PLINK should be downloaded, decompressed, and used (it is unnecessary to install).

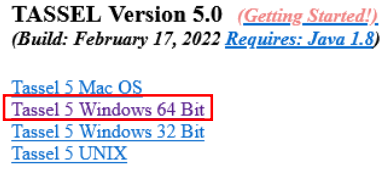
Java can be downloaded from
https://www.oracle.com/java/technologies/downloads/#jdk17-windows
TASSEL can be downloaded from https://www.maizegenetics.net/tassel
PLINK can be downloaded from https://www.cog-genomics.org/plink/1.9/

The downloaded files of Java, TASSEL, and PLINK are as follows.

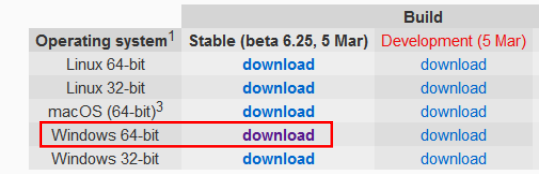
Java:



TASSEL:



PLINK:



And then conduct the two steps below:

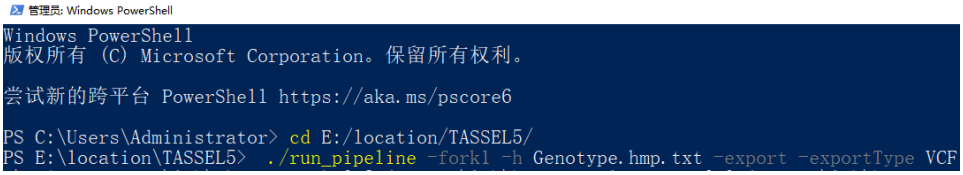
1. First, use TASSEL to transform *.hmp.txt to *.vcf. Open Windows PowerShell on your computer and run the following three codes (almost the same as the above codes in Linux system):

   cd E:/location/TASSEL5/

   ./run_pipeline -SortGenotypeFilePlugin -inputFile E:\IIIVmrMLM\Genotype.hmp.txt -outputFile E:\IIIVmrMLM\Genotype.sort.hmp.txt -fileType Hapmap

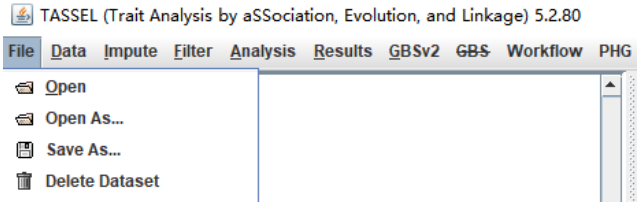
   ./run_pipeline -fork1 -h E:\IIIVmrMLM\Genotype.sort.hmp.txt -Xmx10g -export E:\IIIVmrMLM\Genotype -exportType VCF

   Note that *-Xmx10g* indicates that a maximum of 10G of memory is allocated to this step, and users can allocate it reasonably according to their own device conditions and the size of Hapmap data.



If the data file is small, the above three codes may be implemented via the interface version of TASSEL to convert Hapmap file to VCF file using the below operations:

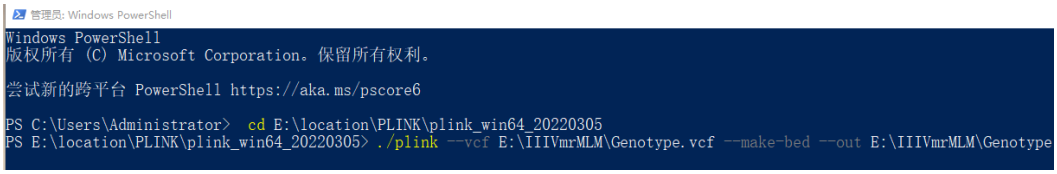
File—> Open As—>Format: Hapmap (Sort Positions)—> File—>Save As—>Format: VCF



2. Then, use PLINK to transform *.vcf to plink binary filesets (*.bed + *.bim + *.fam). Open Windows PowerShell on your computer and run the following two codes:

   cd E:\location\PLINK\plink_win64_20220305

   ./plink --vcf E:\IIIVmrMLM\Genotype.vcf --make-bed --out E:\IIIVmrMLM\Genotype



Note that, in code *cd path*, path is the location where PLINK is decompressed.

Format for phenotypic dataset "filePhe" (Table 1)

The file type of phenotypes for complex trait is *.csv or *.txt, as shown below. The first row in the first column: "<Phenotype>"; the second to *n*th rows in the first column: individual IDs or names, such as B46. The first row in other columns: trait names, such as "trait1", and the second to *n*th rows in other columns: phenotypic values of complex traits. The phenotypes missed: "NA".

Table 1. The format of phenotypic dataset

<Phenotype>	trait1	trait2	trait3	…
B46	42	43.02	44.32	…
B52	72.5	71.88	72.8	…
B57	41	41.7	41.42	…
B64	74.5	74.43	74.5	…
B68	65	66.4	65.33	…
⋮	⋮	⋮	⋮	…

The format for kinship dataset "fileKin" (Table 2)

The "fileKin" should be a file with *.csv or *.txt format. In the first column in Table 2, "262" is sample size (*n*), and "33-16", "A4226", and "A4722" are individual ID. Note that "*n*" is the number of common individuals between the phenotypic and genotypic datasets. All the kinship coefficients are listed as an *n* × *n* matrix.

Table 2. The format of the Kinship dataset

262						
33-16	1	0.700361011	0.599277978	0.675090253	0.620938628	…
A4226	0.700361011	1	0.620938628	0.666064982	0.653429603	…
A4722	0.599277978	0.620938628	1	0.561371841	0.5433213	…
A188	0.675090253	0.666064982	0.561371841	1	0.615523466	…
A214N	0.620938628	0.653429603	0.5433213	0.615523466	1	…
⋮	⋮	⋮	⋮	⋮	⋮	…

fileKin=NULL indicates that the Kinship matrix is calculated by the "IIIVmrMLM" software. Here only the marker information of the above *n* individuals is used to calculate Kinship matrix.

fileKin="D:/Users/Kinship.csv" means that the K matrix with name Kinship.csv is uploaded from the folder "D:/Users". If the number and order of individuals in Kinship.csv are not consistent with those of the above *n* individuals in genotypic and phenotypic files, our software may match the K matrix in order that the number and order of the transferred K matrix are consistent with those in the above *n* common individuals.

***Q* matrix format for dataset "filePS" (Table 3)**

The Q matrix dataset in Table 3 consists of a (*n*+2) × (*k*+1) matrix, where *n* is sample size (the number of the above common individuals), and *k* is the number of sub-populations. In the first column, "<PopStr>" and "<ID>" must present in the first and second rows, respectively; "33-16", "Nov-38" and "A4226" are individual IDs or names. In the 2nd to (*k*+1)-th columns, "Q1" to "Q*k*" indicate sub-populations. In the third row, "0.014", "0.972" and "0.014" are posterior probabilities that the individual "33-16" is belong to the 1st, 2nd, and 3rd subpopulations, respectively. When the Q matrix was uploaded to the software, the software would automatically delete the column in which the column sum is the smallest if their sums are all equal to one.

Table 3. The Q matrix format of dataset filePS

<PopStr>			
<ID>	Q1	Q2	Q3
33-16	0.014	0.972	0.014
Nov-38	0.003	0.993	0.004
A4226	0.071	0.917	0.012
A4722	0.035	0.854	0.111
⋮	⋮	⋮	⋮

Principal components format for dataset "filePS" (Table 4)

The principal component dataset in Table 4 consists of a (n+2) × (k+1) matrix, where n is sample size (the number of the common individuals), and k is the number of principal components. In the first column, "<PCA>" and "<ID>" must present in the first and second rows, respectively; "33-16", "Nov-38", and "A4226" are individual IDs or names. In the 2nd to (k+1)-th columns, "PC1" to "PCk" indicate the first to kth principal components. In the second column, "0.306" is the score of the first principal component for the 1st individuals. Note that any principle components are not deleted in the software.

Table 4. The principal components format of the filePS dataset

<PCA>			
<ID>	PC1	PC2	PC3
33-16	0.306	0.029	0.226
Nov-38	-0.708	-0.271	1.413
A4226	-2.330	0.116	-0.824
A4722	1.059	0.470	-0.135
⋮	⋮	⋮	⋮

Evolutionary population structure format for dataset "filePS" (Table 5)

The evolutionary population structure dataset in Table 5 consists of a (n+2) × 2 matrix, where n is sample size (the number of the common individuals). In the first column, "<EvolPopStr>" and "<ID>" must present in the first and second rows, respectively; "33-16", "Nov-38" and "A4226" are individual IDs or names. In the second column, "EvolType": evolutionary type, i.e., the evolutionary types for individuals "33-16" and "A4722" are "A" and "B", respectively, such as wild (A), landrace (B), and bred (C) soybeans.

Table 5. The evolutionary population format of the filePS dataset

<EvolPopStr>	
<ID>	EvolType
33-16	A
Nov-38	B
A4226	A
A4722	B
⋮	⋮

filePS=NULL indicates no population structure in the model. filePS="D:/Users/PopStr.csv" means that population structure dataset with name PopStr.csv is uploaded from the folder "D:/Users". If the number and order of individuals in PopStr.csv are not consistent with those of the above common individuals, our software may match the population structure matrix in order that the number and order of new matrix are consistent with those in the above common individuals.

The format for covariate dataset "fileCov" (Table 6)

The "**Covariate**" dataset consists of the (n+2) × (k+1) matrix, where n is sample size (the number of the common individuals), and k is the number of covariates. In the first column, "**<Covariate>**" and "**<ID>**" must present in the first and second rows, respectively. If covariate is categorical, the names are Cate_covariate*. If covariate is continuous, the names are Con_covariate* (Table 6).

Table 6. The format of fileCov dataset

<Covariate>				
<ID>	Cate_covariate1	Cate_covariate2	Con_covariate1	Con_covariate2
33-16	A	C	349.5	374
Nov-38	B	C	205	452
A4226	A	D	300	374
A4722	A	D	190	452
⋮	⋮	⋮	⋮	

fileCov=NULL indicates no covariates in the genetic model. fileCov="D:/Users/covariate.csv" means that the covariates with name covariate.csv are uploaded from the folder "D:/Users". If the number and order of individuals in the uploaded file are not consistent with those in the above common individuals, our software need to change the number and order of individuals in order to match the above genotypic and phenotypic datasets.

2.2.2 Result

The result file (result-main_QTN_detection) includes three files: *_K.csv (Kinship matrix calculated by IIIVmrMLM), *_midresult.csv (intermediate results), and *_result.xlsx (final results), and one Manhattan plot (if DrawPlot=TRUE).

*_midresult.csv: This is the results of single marker scanning on the genome in the first step. In this file, all the columns are named as Marker (marker name), Chromosome, Position (markers position (bp) on the genome), and pvalue.Q (the P-value for main-effect QTNs).

Marker	Chromosome	Position (bp)	pvalue.Q
PZB00859.1	1	157104	0.292043111
PZA01271.1	1	1947984	0.185246808
PZA03613.2	1	2914066	0.99208603
PZA03613.1	1	2914171	0.999987108
PZA03614.2	1	2915078	0.976018023
⋮	⋮	⋮	⋮

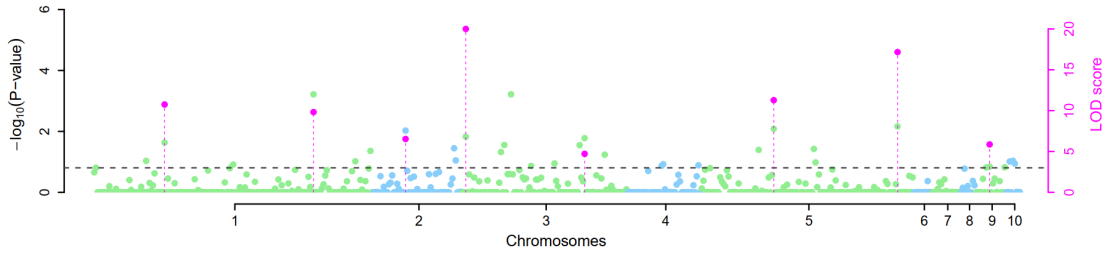
*_result.xlsx: The final results for significant and suggested QTNs. In this file, all the columns are named as Trait ID, Trait name, Marker (marker name), Chromosome, Position (markers position (bp) on the genome), LOD (LOD score), add (additive effect), dom (dominant effect), variance (the variance of each QTN), r^2 (%) (the proportion of total phenotypic variance explained by each QTN), P-value (calculated from LOD score using χ^2 distribution), and significance (significant (SIG) QTNs are based on Bonferroni correction, that is, critical P-value is 0.05/*m*, where *m* is the number of tests or markers, while suggested (SUG) QTNs are based on LOD ≥ 3.0, default).

Trait ID	Trait name	Marker	Chromosome	Position (bp)	LOD	add	dom	variance	r2(%)	P-value	significance
1	trait1	PZA03214.3	1	245136244	11.7017	7.0691		26.4678	6.3725	2.12416E-13	SIG
1	trait1	PZA03188.4	1	280719882	10.1462	-7.1265	0.896	34.9545	8.4157	7.14861E-11	SIG
1	trait1	PZA03559.1	2	15810363	7.872	5.5235	17.383	31.2681	7.5282	1.34367E-08	SIG
1	trait1	PZB01892.1	3	161573186	20.8753	-9.9062		16.0224	3.8576	1.07536E-22	SIG
1	trait1	PZA03647.3	3	185318086	4.4864	4.3915		14.9657	3.6032	5.48594E-06	SIG
1	trait1	PZA00112.5	5	13664679	13.5465	-7.7296		24.641	5.9327	2.8295E-15	SIG
1	trait1	PZA03042.5	5	64413280	16.8506	-8.392	-38.9679	18.2141	4.3853	1.4125E-17	SIG
1	trait1	PZB00379.1	9	26661626	5.1546	-4.2882	-27.6411	21.0427	5.0663	7.00752E-06	SIG

*_Manhattan plot: Y-axis on the left side reports -$\log_{10}$ P-values, which are obtained from single-marker genome-wide scanning for all the markers in the first step of IIIVmrMLM,

while Y-axis on the right side reports LOD scores, which are obtained from likelihood ratio test for significant and suggested QTNs, with the threshold of LOD = 3.0 (dashed line), in the second step of IIIVmrMLM. These LOD scores are shown in points with straight lines.

If LOD score ≥ 20, the LOD scores obtained are transformed as LOD' = 20 + (LOD – 20)/100 in order that the Manhattan plot is more beautiful.



2.3 Multi_env: Detection of QTN-by-environment interactions for complex traits

IIIVmrMLM(fileGen="D:/Users/Genotype",
filePhe="D:/Users/Phenotype_multi_env.csv",
fileKin=NULL,filePS="D:/Users/PopStr.csv",PopStrType="Q",fileCov=NULL,
method="Multi_env",trait=1:2,n.en=c(2,2),SearchRadius=20, svpal=0.01,
DrawPlot=TRUE,Plotformat="pdf",Chr_name_com=NULL,dir="D:/Users/")

Users must set "fileGen", "filePhe", "method", "trait", "n.en", and "dir", while the other parameters may be default in function *IIIVmrMLM*, including *fileKin=NULL; filePS=NULL; PopStrType="Q"; fileCov=NULL; SearchRadius=20; svpal=0.01; DrawPlot=TRUE; Plotformat="pdf"; Chr_name_com=NULL*.

Compared to the detection of main-effect QTNs for complex traits in single environment (Single_env), there are three main changes in QEI detection (Multi_env).
1) The phenotype file should be arranged according to traits, the number of environments for each trait is greater than or equal to 2;
2) method="Multi_env";
3) Add a vector n.en to represent the number of environments for each trait in the *filePhe*. For example, n.en=c(2,2,3) (The *filePhe* file contains the phenotypic values of three traits, and the environmental numbers of each trait are 2, 2, and 3, respectively).

2.3.1 Data input format

Format for the dataset "filePhe"

The type of phenotypic file for complex trait is *.csv or *.txt, as shown below.

<Phenotype>	trait1Env1	trait1Env2	trait2Env1	trait2Env2	...
B46	38.04	34.45	38.71	35.72	...
B52	38.64	40.85	43.04	34.97	...
B57	41.54	33.82	45.10	33.23	...
B64	40.82	35.20	39.14	30.93	...
B68	33.40	33.99	38.04	39.24	...
⋮	⋮	⋮	⋮	⋮	...

The first row in the first column: "<Phenotype>", while the second to *n*th rows in the first column: individual names (or IDs), such as B46. The first rows from the second column: trait names, such as "**trait1Env1**", while the other rows from the second column: phenotypic values of complex traits. The phenotypic file is arranged by traits, each trait has at least two columns, and each column is the phenotypes measured in an environment. The missed phenotypes are represented by "NA".

Format for datasets "fileGen", "fileKin", "filePS", "fileCov" are same as those in main-effect QTN detection (Single_env).

2.3.2 Result

The result file (result-QEI_detection) includes three files: *_K.csv (Kinship matrix calculated by IIIVmrMLM), *_midresult.csv (intermediate results), and *_result.xlsx (final result, including two sheets, significant and suggested main-effect QTNs (result.Q) and significant and suggested QEIs (result.QEI)), and two Manhattan plots (if DrawPlot=TRUE), one for Main-effect QTN and one for QEI.

*_midresult.csv: This is the results of single marker scanning on the genome in the first step. In this file, all the columns are named as Marker (marker name), Chromosome, Position (markers position (bp) on the genome), pvalue.Q (the P-value for main-effect QTNs), and pvalue.QE (the P-value for QEIs).

Marker	Chromosome	Position (bp)	pvalue.Q	pvalue.QE
PZB00859.1	1	157104	0.812972071	0.928177513
PZA01271.1	1	1947984	0.993594668	0.988087592
PZA03613.2	1	2914066	0.99306619	0.993440946
PZA03613.1	1	2914171	0.99997328	0.999970187
PZA03614.2	1	2915078	0.971495059	0.983226371
⋮	⋮	⋮	⋮	

result.Q: The results are for significant and suggested main-effect QTNs. In this sheet, all

the columns are named as Trait ID, Trait name, Marker (marker name), Chromosome, Position (markers position (bp) on the genome), LOD (Q) (LOD score for main-effect QTNs), add (additive effect), dom (dominant effect), variance (the variance of each QTN), r^2 (%) (the proportion of total phenotypic variance explained by each QTN), P-value (calculated from LOD score in main-effect QTN detection using χ^2 distribution), and significance (significant (SIG) QTNs are based on Bonferroni correction, that is, critical P-value is 0.05/*m*, where *m* is the number of tests or markers, while suggested (SUG) QTNs are based on LOD ≥ 3.0, default).

Trait ID	Trait name	Marker	Chromosome	Position (bp)	LOD (Q)	add	dom	variance	r2(%)	P-value	significance
1	trait1	PZB01647.1	1	231039372	11.0446	1.1628	5.8057	1.2459	2.8372	9.03249E-12	SIG
1	trait1	PZA02812.34	1	267615649	12.721	-1.3274	4.7147	1.4823	3.3754	1.9032E-13	SIG
1	trait1	PZA02957.4	1	281818425	18.9611	1.6327		1.3099	2.9828	9.25024E-21	SIG
1	trait1	PZA03305.5	1	286642725	4.7749	-0.311	5.8961	0.658	1.4984	1.67971E-05	SIG
1	trait1	PZA00176.8	2	10533421	10.299	1.187	0.065	1.1969	2.7257	5.02727E-11	SIG
1	trait1	PZA03073.28	3	168443662	14.6457	-1.1235	4.7367	1.8954	4.3162	2.26368E-15	SIG
1	trait1	PZA01122.1	4	12618115	19.6559	-1.7468	4.5194	1.6788	3.8229	2.21205E-20	SIG
1	trait1	PZB01642.1	5	12337501	13.8005	1.3689		0.5059	1.1521	1.56227E-15	SIG

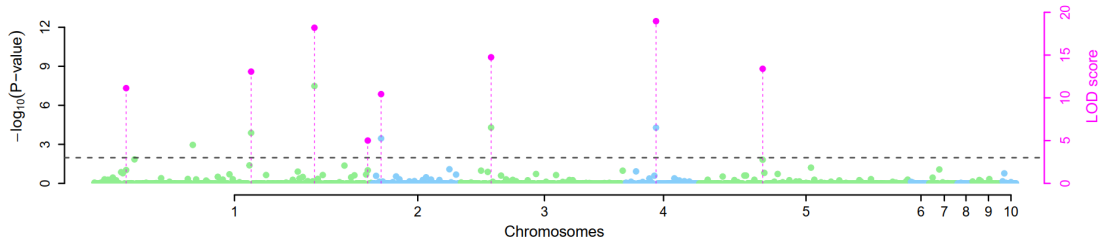
result.QEI: This is the results for significant and suggested QEIs. In this sheet, all the columns are named as Trait ID, Trait name, Marker (marker name), Chromosome, Position (markers position (bp) on the genome), LOD (QE) (LOD score for QEIs). add*env*k* (additive effect in environment *k*), dom*env*k* (dominant effect in environment *k*), variance (the variance of each QEI), r2 (%) (r2 (%) is the proportion of total phenotypic variance explained by each QEI), P-value (calculated from LOD score in QEI detection using χ^2 distribution), and significance (significant (SIG) QEIs are based on Bonferroni correction, that is, critical P-value is 0.05/*m*, where *m* is the number of tests or markers, while suggested (SUG) QEIs are based on LOD ≥ 3.0, default).

Trait ID	Trait name	Marker	Chromosome	Position (bp)	LOD (QE)	add*env1	dom*env1	add*env2	dom*env2	variance	r2(%)	P-value	significance
1	trait1	tb1.15	1	264847721	9.7984	-1.152		1.152		1.3272	3.0223	1.85152E-11	SIG
1	trait1	PZA03191.3	3	185290309	10.0227	-1.1749		1.1749		1.3804	3.1435	1.09283E-11	SIG
1	trait1	PZA00281.1	5	9965510	12.9872	-1.3908	-0.356	1.3908	0.356	1.9268	4.3877	1.0311E-13	SIG
1	trait1	PZB00869.2	5	32366232	4.7824	-0.7892	-0.5069	0.7892	0.5069	0.6214	1.4151	1.65109E-05	SIG
1	trait1	PZA03042.1	5	64413079	5.7313	-0.814	-3.6012	0.814	3.6012	0.8164	1.8591	1.85763E-06	SIG

*_Q_Manhattan plot: Manhattan plot for Main-effect QTNs. Y-axis on the left side reports -$\log_{10}$ P-values of main-effect QTNs, which are obtained from single-marker genome-wide scanning for all the markers in the first step of IIIVmrMLM, while Y-axis on the right side reports LOD scores, which are obtained from likelihood ratio test for significant and

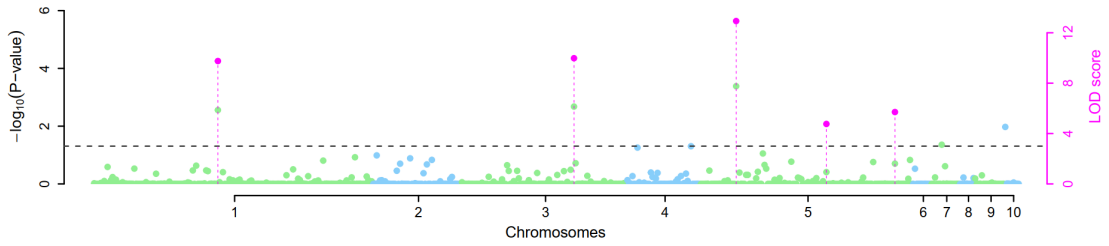
suggested QTNs, with the threshold of LOD = 3.0 (dashed line), in the second step of IIIVmrMLM. These LOD scores are shown in points with straight lines.

If LOD score ≥ 20, the LOD scores obtained are transformed as LOD' = 20 + (LOD – 20)/100 in order that the Manhattan plot is more beautiful.



*_QEI_Manhattan plot: Manhattan plot for QEIs. Y-axis on the left side reports -$\log_{10}$ P-values of QEIs, which are obtained from single-marker genome-wide scanning for all the markers in the first step of IIIVmrMLM, while Y-axis on the right side reports LOD scores, which are obtained from likelihood ratio test for significant and suggested QEIs, with the threshold of LOD = 3.0 (dashed line), in the second step of IIIVmrMLM. These LOD scores are shown in points with straight lines.

If LOD score ≥ 20, the LOD scores obtained are transformed as LOD' = 20 + (LOD – 20)/100 in order that the Manhattan plot is more beautiful.



2.4 Epistasis: Detection of epistatic QTNs for complex traits

IIIVmrMLM(fileGen="D:/Users/Genotype",filePhe="D:/Users/Phenotype.csv",
fileKin=NULL,filePS="D:/Users/PopStr.csv",PopStrType="Q",fileCov=NULL,
method="Epistasis",trait=c(1:3),SearchRadius=c(0,1),svpal=c(0.01,0.01),
sblgwas_t=-1,DrawPlot=TRUE,Plotformat="pdf",Chr_name_com=NULL,
dir="D:/Users/")

Users must set "fileGen", "filePhe", "method", "trait", and "dir", while the other parameters may be default in function *IIIVmrMLM*, including *fileKin=NULL; filePS=NULL; PopStrType="Q"; fileCov=NULL; SearchRadius=c(10,20); svpal=c(0.01,0.01); sblgwas_t=-1; DrawPlot=TRUE; Plotformat="pdf"; Chr_name_com=NULL*.

In epistasis detection *SearchRadius* and *svpal* are two-dimensional vectors, the first parameter is set for main-effect QTNs and the second one is set for QQIs. *sblgwas_t* is a number between [-3,0] to control sparseness of the "sblgwas" function, default value is -1.

Format for datasets "fileGen", "fileKin", "filePS", "fileCov" are same as those in main-effect QTN detection (Single_env).

The result file (result-Epi_detection) includes two files: *_K.csv (Kinship matrix calculated by IIIVmrMLM) and *_result.xlsx (including two sheets, significant and suggested QTNs (result.Q) and significant and suggested QTN-by-QTN interactions (QQIs) (result.QQI)), and two Manhattan plots (if DrawPlot=TRUE), one for Main-effect QTN and one for QQI.

result.Q: The results are for significant and suggested QTNs. In this sheet, all the columns are named as Trait ID, Trait name, Marker (marker name), Chromosome, Position (markers position (bp) on the genome), LOD (Q) (LOD score for main-effect QTN), add (additive effect), dom (dominant effect), variance (the variance of each QTN), r^2 (%) (the proportion of total phenotypic variance explained by each QTN), P-value (calculated from LOD score of main-effect QTNs in QQI detection using χ^2 distribution), and significance (significant (SIG) QTNs are based on Bonferroni correction, that is, critical P-value is 0.05/*m*, where *m* is the number of tests or markers, while suggested (SUG) QTNs are based on LOD ≥ 3.0**,** default).

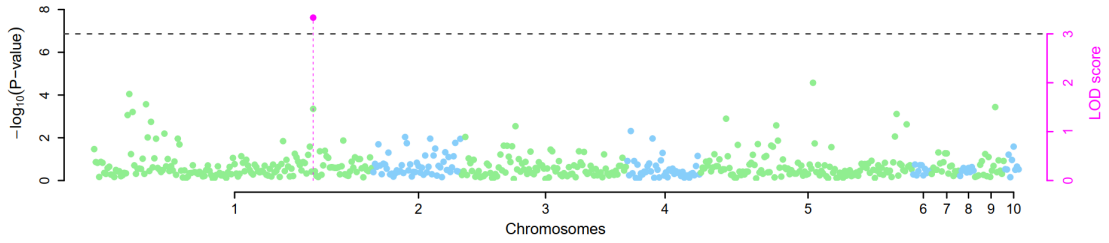
Trait ID	Trait name	Marker	Chromosome	Position (bp)	LOD (Q)	add	dom	variance	r2(%)	P-value	significance
1	trait1	PZA03188.4	1	280719882	3.3317	-5.8445	0.5376	23.8219	5.7354	0.00046606	SUG

result.QQI: This is the results of significant and suggested QQIs. In this sheet, all the columns are named as Trait ID, Trait name, Marker_*i* (name of marker *i* in an interaction pair), Chr_*i* (chromosome of marker *i* in an interaction pair), and Posi_*i* (position of marker *i* in an interaction pair, bp) (*i*=1,2), LOD (LOD score), aa.effect (additive-additive effect), ad.effect (additive-dominant effect), da.effect (dominant-additive effect), dd.effect (dominant-dominant effect), variance (the variance of each QQI), r2 (%) (the proportion of total phenotypic variance explained by each QQI), P-value (calculated from LOD score of epistasis in QQI detection using χ2 distribution), and significance (significant (SIG) QQIs are based on Bonferroni correction, that is, critical P-value is 0.05/*m*, where *m* is the number of tests or markers, while suggested (SUG) QQIs are based on LOD ≥ 3.0**,** default).

Trait ID	Trait name	Marker_1	Chr_1	Posi_1	Marker_2	Chr_2	Posi_2	LOD	aa.effect	ad.effect	da.effect	dd.effect	variance	r2(%)	P-value	significance
1	trait1	PZA03301.2	1	240574247	PZA03665.2	1	241430615	5.2653	-3.6773				12.9649	3.1215	8.47516E-07	SUG
1	trait1	PZA03214.1	1	245136387	PZA03336.3	2	11193471	3.5036	3.7019				12.986	3.1265	5.90067E-05	SUG
1	trait1	PZB01427.5	1	270364633	PZA02204.1	1	278690966	3.2263	-2.1569	0.0736			4.4934	1.0819	0.000594087	SUG
1	trait1	PZB00119.1	1	286269951	PZA00449.2	5	37511797	4.1247	-5.19	-1.0276			20.532	4.9434	7.50614E-05	SUG
1	trait1	PZA00108.4	2	13779970	PZB02017.3	2	20958616	3.8997	-3.9178		-0.7314		14.5959	3.5141	0.000126029	SUG
1	trait1	PZA03559.1	2	15810363	PZB01482.1	7	3671683	5.8034	5.3545		0.7908		27.1068	6.5263	1.57329E-06	SUG
1	trait1	PZD00027.1	3	169757113	PZA00281.13	5	9965534	3.5998	3.5533	0.3447			8.3321	2.0061	0.000251375	SUG
1	trait1	PZB01919.1	3	178235128	PZA03107.1	4	2851075	3.3939	-3.7499				13.9399	3.3562	7.70701E-05	SUG
1	trait1	PZB02080.1	4	4980568	PZA00188.1	7	2998280	3.7489	-4.9426				18.2097	4.3842	3.25358E-05	SUG
1	trait1	PZA00112.5	5	13664679	PZB01983.2	5	23437363	8.8132	-8.2499				35.1406	8.4606	1.88252E-10	SIG
1	trait1	PZA02792.16	5	21771297	PZA00710.1	5	61492543	3.7439	-4.5708		-0.1197		20.4298	4.9188	0.000180398	SUG

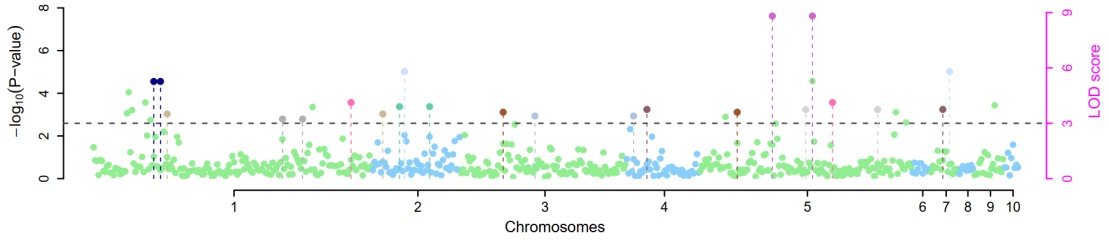
*_Q_Manhattan plot: Manhattan plot for Main-effect QTNs. Y-axis on the left side reports -$\log_{10}$ P-values of main-effect QTNs, which are obtained from single-marker genome-wide scanning for all the markers in the first step of IIIVmrMLM, while Y-axis on the right side reports LOD scores, which are obtained from likelihood ratio test for significant and suggested QTNs, with the threshold of LOD = 3.0 (dashed line), in the second step of IIIVmrMLM. These LOD scores are shown in points with straight lines.

If LOD score ≥ 20, the LOD scores obtained are transformed as LOD' = 20 + (LOD – 20)/100 in order that the Manhattan plot is more beautiful.



*_QQI_Manhattan plot: Manhattan plot for QQIs. Y-axis on the left side reports -$\log_{10}$ P-values of main-effect QTNs, which are obtained from single-marker genome-wide scanning for all the markers in the first step of IIIVmrMLM, while Y-axis on the right side reports LOD scores, which are obtained from likelihood ratio test for significant and suggested QQIs (a pair of QQI have the same color), with the threshold of LOD = 3.0 (dashed line), in the second step of IIIVmrMLM. These LOD scores are shown in points with straight lines.

If LOD score ≥ 20, the LOD scores obtained are transformed as LOD' = 20 + (LOD – 20)/100 in order that the Manhattan plot is more beautiful.



This software doesn't provide quantile-quantile plots. The IIIVmrMLM includes three steps. In the first step, single-marker genome-wide scanning is conducted, and its purpose is to obtain potentially associated markers. In the second step, all the selected markers are placed into a multi-locus genetic model, all the effects are estimated by empirical Bayes, and all the non-zero effects are further identified by likelihood ratio test for significant and suggested QTNs. Thus, QQ plot from the P-values in the scanning is important for existing methods rather than IIIVmrMLM. In the third step, around all the significant and suggested loci, previously reported and candidate genes may be mined using multi-omics data and bioinformatics analyses. If there are known or candidate genes around a suggested locus, the suggested locus is reliable.

3. Reference

Li, M., Zhang, Y.W., Zhang, Z.C., Xiang, Y., Liu, M.H., Zhou, Y.H., Zuo, J.F., Zhang, H.Q., Chen, Y., and Zhang, Y.M. (2022). A compressed variance component mixed model for detecting QTNs and QTN-by-environment and QTN-by-QTN interactions in genome-wide association studies. Mol. Plant **15**:630–650.