User Manual for

IIIV m r M L M.Q E I

A comprehensive tool of QTN-by-environment interaction (QEI) detection based on compressed variance component mixed model in genome-wide association study

(version 1.0)

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

(soyzhang@mail.hzau.edu.cn)

Last updated on August, 2024

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 or appropriate. However, results are not guaranteed for any specific datasets. We strongly recommend that users validate the IIIVmrMLM.QEI results with other software packages.

Download website:
https://github.com/YuanmingZhang65/IIIVmrMLM.QEI.

Citation:

1 Zhang YW, Li M, Han XL, Zhang YM. IIIVmrMLM.QEI: A all-in-one tool for detecting more QTN-by-environment interactions in genome-wide association studies. **Computational and Structural Biotechnology Journal**, in submission

2 Li M, Zhang YW, Xiang Y, Liu MH, Zhang YM. IIIVmrMLM: the R and C++ tools associated with 3VmrMLM, a comprehensive GWAS method for dissecting quantitative traits. **Mol Plant** 2022; 15(8): 1251-1253

3 Li M, Zhang YW, Zhang ZC, Xiang Y, Liu MH, Zhou YH, Zuo JF, Zhang HQ, Chen Y, Zhang YM. A compressed variance component mixed model for detecting QTNs, and QTN-by-environment and QTN-by-QTN interactions in genome-wide association studies. **Molecular Plant** 2022, 15(4): 630-650

This work was supported by the National Natural Science Foundation of China (32270673, 32070557, and 31871242), and Huazhong Agricultural University Scientific & Technological Self-innovation Foundation (Program No. 2014RC020).

INTRODUCTION

1.1 Why IIIVmrMLM.QEI?

IIIVmrMLM.QEI is a comprehensive R tool for QTN-by-environment interaction (QEI) detection based on compressed variance-component mixed linear model. In IIIVmrMLM.QEI, multi-environment joint analysis via 3VmrMLM can be used to directly identify QTNs and QEIs for complex and omics traits, while a series of indirect indicators are used as phenotypes to indirectly identify QEIs using 3VmrMLM-random or 3VmrMLM-fixed (two single-environment analysis methods), including trait difference, regression intercept and coefficient, range, variance, standard deviation (SD), coefficient of variation (CV) and environmental factor (EF).

1.2 Getting started

The software package IIIVmrMLM.QEI runs only in R environment and can be freely downloaded from https://github.com/YuanmingZhang65/IIIVmrMLM.QEI and obtained from the maintainer, Dr Yuan-Ming Zhang at College of Plant Science and Technology, Huazhong Agricultural University (soyzhang@mail.hzau.edu.cn).

Note: Users need to install Rtools https://cran.r-project.org/bin/windows/Rtools/, while selecting the option of "Add rtools to system PATH" in Fig 1. The purpose is to ensure that the results can be written to the computer.

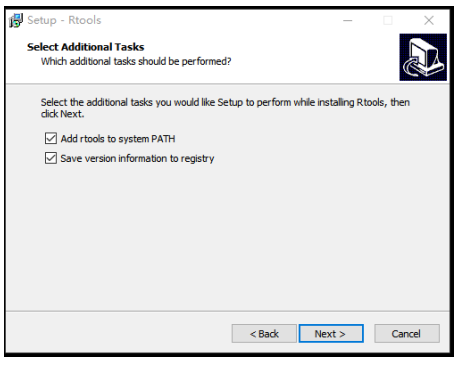

Figure 1. Install Rtools

1.2.1 Step-by-step installation

1.2.1.1 Install the add-on packages

install.packages(c("lars","RcppEigen","Rcpp","doParallel","data.table","MASS","openxlsx","BEDMatrix","bigmemory","stringr","biglasso","progress","ncvreg","coin","sampling","sbl"))

1.2.1.2 Install IIIVmrMLM.QEI

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

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

1.2.2 Run IIIVmrMLM.QEI

Once the software IIIVmrMLM.QEI is installed, users may run it using two commands:

library("IIIVmrMLM.QEI")

IIIVmrMLM.QEI(***)

If users re-use IIIVmrMLM.QEI, users also use the above two commands.

2. Function

2.1 Function IIIVmrMLM.QEI

2.1.1 Parameter settings

Parameter	Meaning	File format	Note
fileGen	File path & name of marker genotypic file in your computer, i.e., fileGen="D:/Users/Genotype"	PLINK binary filesets, including: **Genotype.bed**, **Genotype.bim**, and **Genotype.fam**	
filePhe	File path & name 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	File path & name in your computer, i.e., fileKin="D:/Users/Kinship.csv" or fileKin=NULL	*.csv; *.txt (Kinship matrix. **Row** & **Column**: individuals)	Table 2
filePS	File path & name 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)]	Table 3~5
PopStrType	Three types of population structures: *Q* (*Q* matrix), PCA (principal components), EvolPopStr (evolutionary population structure)		
fileCov	File path & name of covariate file in your computer,i.e., fileCov="D:/Users/Covariate.csv" or fileCov=NULL	*.csv; *.txt (Covariate. **Row**: individual; **Column**: covariate 1, 2, ---, *k* (No. of covariate)) Cate: categorical variable; Con: continuous variable	Table 6
method	One direct method and two indirect methods are available. Users may select one to three methods. For example, method=c("fixed","random")		
indicator	Seven indirect indicators may be used as phenotypes to indirectly identify QEIs using 3VmrMLM-random or 3VmrMLM-fixed, for example, Indicator=c("difference","RI","RC","EF","range","variance","SD","CV"), if "EF" are selected, environmental factor should be included in the column name of phenotype file, for example, if environmental factor are 10°C, 16°C, and 22°C, the column name should be written as "trait1 /10"," trait1/16" and " trait1/20".		
trait	Traits analyzed from number 1 to number 2, i.e., **trait=1:3** indicates that users analyze the first to third traits.		
n.en	Users need to add an environment number setting, i.e. **n.en=c(2,2,3)** (The numbers of environments are 2, 2, and 3, respectively, for the traits 1, 2, and 3)		
SearchRadius	Range of decollinearity, i.e., **SearchRadius=30**: only one most potentially associated QTN was selected within 30 kb.		
DrawPlot	This parameter is for all the six methods, including FALSE and TRUE. **DrawPlot=FALSE** indicates no figure output; **DrawPlot=TRUE** indicates the output of the Manhattan plot.		
Plotformat	This parameter is for all the figure files, including *.jpeg, *.png, *.tiff and *.pdf. **Plotformat="jpeg"** indicates the *.jpeg format of plot file.		
dir	Save path in your computer, i.e, "D:/Users/"		

2.1.2 Example

The codes

```
IIIVmrMLM.QEI(fileGen="D:/Users/Genotype",filePhe="D:/Users/Phenotype.csv",fileKin=NULL,filePS="D:/Users/PopStr.csv",PopStrType="Q",fileCov=NULL,method=c("fixed","random"),indicator=c("difference","RI","RC","range","variance","SD","CV","EF"),trait=1,n.en=c(3),SearchRadius=c(80),DrawPlot=FALSE,Plotformat="jpeg",dir="D:/Users/")
```

When users use the above codes to analyze real datasets in the software, all the data files can be found in the software. However, please note the positions (paths) of all the

dataset files, “80” (kb) in “SearchRadius=80” are for indicator, and please note the positions (paths) and names of all the data files when users analyze yourself datasets.

It should be noted that users must set "fileGen", "filePhe", "method", "indicator", "n.en", and "dir", and the other parameters can be default in function, including PopStrType="Q"; SearchRadius=20, DrawPlot=TRUE, Plotformat= "jpeg".

2.2.1 Data input format

Format for genotypic dataset “fileGen”

The file type of genotypes is “plink binary format” (.bed/.bim/.fam). The following provides a way to convert hapmap to .bed/.bim/.fam for users' reference.

First, use TASSEL to sort the hapmap file by running:

run_pipeline.pl -SortGenotypeFilePlugin -inputFile Genotype.hmp.txt -outputFile Genotype.sort.hmp.txt -fileType Hapmap

Then, use TASSEL to transform *.hmp to *.vcf by running:

run_pipeline.pl -fork1 -h Genotype.sort.hmp.txt -export -exportType VCF

Finally, use PLINK to transform *.vcf to plink binary filesets (.bed/.bim/.fam) by:

plink --vcf Genotype.vcf --make-bed --out Genotype

Format for the dataset “filePhe” (Table 1-1; Table 1-2) The **Phenotypic** file should be a file with ***.csv** or ***.txt** format. The first column lists individual ID, i.e., “B46”, and “<Phenotype>” should be showed in the first row. Among the other columns, each column lists all the observations for the trait, its trait name is showed in the first row, i.e., “trait1Env1”, and phenotypic values are in the corresponding rows of their individuals. Each trait has at least two columns, and each column is the phenotypes measured in an environment, "NA" indicates the missing or unknown phenotypes. In “trait1Env1”, users may change trait name of “trait1” and environmental name of “Env1” (Table 1-1). If environmental factors are included in the indirect indicator, such as temperature (10℃, 16℃, and 22℃), the column name of the phenotypic file should be written as trait under a environment, such as trait1/10, trait1/16, trait1/22, trait2/10, trait2/16, and trait2/22 (Table 1-2). Please note that **10** in “trait1/**10**” indicates **10℃** of environmental factor (temperature).

Table 1-1. The format of Phenotypic dataset

<Phenotype>	trait1Env1	trait1Env2	trait2Env1	trait2Env2	⋯
B46	42	43.02	44.32	42.11	⋯
B52	72.5	71.88	72.8	74.05	⋯
B57	41	41.7	41.42	42.13	⋯
B64	74.5	74.43	74.5	75.11	⋯
⋮	⋮	⋮	⋮		⋯

Table 1-2. The format of Phenotypic dataset included environmental factor

<Phenotype>	trait1/10	trait1/16	trait1/22	trait2/10	trait2/16	trait2/22	⋯
B46	42	43.02	44.32	42.11	40.23	45.51	⋯
B52	72.5	71.88	72.8	74.05	73.12	72.77	⋯
B57	41	41.7	41.42	42.13	43.51	41.44	⋯
B64	74.5	74.43	74.5	75.11	72.30	73.64	⋯
⋮	⋮	⋮	⋮				⋯

The format for dataset "fileKin" (Table 2) The Kinship file should be a file with ***.csv** or ***.txt** format. In the first column in Table 2, "263" is sample size (*n*), and "33-16", "Nov-38" and "A4226" 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 $n \times n$ matrix.

fileKin=NULL indicates that the Kinship matrix is calculated by the software IIIVmrMLM.QEI. Here only the above *n* individuals are used to calculate the Kinship matrix. fileKin="D:/Users/Kinship.csv" means that the K matrix with name Kinship.csv is uploaded from the folder "D:/Users". If the number and order of individuals in Kinship.csv are not consistent with those of the above *n* individuals, 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* individuals.

Table 2. The format of the Kinship dataset

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

***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 the number of the above common individuals and k is the number of sub-populations. In the first column, "<**PopStr**>" and "<**ID**>" should present in the first and second rows, respectively; "33-16", "Nov-38" and "A4226" are individual ID. In the 2nd to (k+1)-th columns, "Q_1" to "Q_k" indicate all the sub-populations. In the third row, "0.014", "0.972" and "0.014" are the posterior probabilities of the "33-16" individual in the 1st, 2nd and 3rd subpopulations, respectively. When the *Q* matrix is uploaded to the software, the software will automatically delete the column whose sum is the smallest.

Table 3. The format of the filePS dataset

<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
A188	0.013	0.982	0.005
A214N	0.762	0.017	0.221
A239	0.035	0.963	0.002
A272	0.019	0.122	0.859
⋮	⋮	⋮	⋮

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 the number of the common individuals and k is the number of principal components. In the first column, "<**PCA**>" and "<**ID**>" should present in the first and second rows, respectively; "33-16", "Nov-38" and "A4226" are individual ID. In the 2nd to (k+1)-th columns, "PC_1" to "PC_k" indicate the first to k-th principal components. In the second column, "0.306", ⋯, "0.216" are the scores of the first principal component for the 1st to 9-th individuals, respectively. Note that the software doesn't delete any principle components.

Table 4. The dataset format of principal components

<PCA>			
<ID>	PC1	PC2	PC3
33-16	0.306	0.029	0.226
Nov-38	-0.708	-2.071	1.413
A4226	-2.330	0.116	-0.824
A4722	1.059	0.470	-1.315
A188	-2.376	1.087	-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 the number of the common individuals. In the first column, "<**EvolPopStr**>" and "<**ID**>" should present in the first and second rows, respectively; "33-16", "Nov-38" and "A4226" are individual ID. In the second column, "EvolType" indicates the 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 dataset format of evolutionary population structure

<EvolPopStr>	
<ID>	EvolType
33-16	A
A4722	B
A188	A
A239	B
⋮	⋮

filePS=NULL indicates no inclusion of population structure in the genetic 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 aren't 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 dataset "fileCov" (Table 6) The "**Covariate**" dataset consists of the $(n+2) \times (k+1)$ matrix, where n is the number of the common individuals and k is the number of covariates. In the first column, "<**Covariate**>" and "<**ID**>" should present in the first and second rows, respectively. If covariate is categorical, it should be named as Cate_covariate*. If covariate is continuous, it should be named as Con_covariate* (Table 6).

fileCov=NULL indicates no inclusion of 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 datasets.

Table 6. The format of the 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
A188	B	C	213	374
⋮	⋮	⋮	⋮	⋮

2.2.2 Result

The **result** file (result-QEI_detection) includes three files:

1) *_K.csv (Kinship matrix calculated by IIIVmrMLM.QEI),
2) *_midresult.csv (middle results). 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 (position (bp) of markers on genome), and all the P-values of QEIs. These QEIs are identified by different approaches, e.g., pvalue.QE (RC_random) (P-value of QEI using 3VmrMLM-random-RC, in other words, regression coefficient is used as phenotypes to indirectly identify QEIs using 3VmrMLM-random), pvalue.QE (variance_fixed) (P-value of QEI using 3VmrMLM-random-var, in other words, variance is used as phenotypes to

indirectly identify QEIs using 3VmrMLM-fixed) (Table 7), and these QEIs are listed in one sheet one by one.

3) *_result.xlsx (final result, including significant and suggested QEIs detected by different indicators, if "difference", "RC", "RI", "range", "variance", "SD", "CV", "EF" are selected), and one Manhattan plots (if DrawPlot=TRUE) for all the QEIs.

Table 7 Midresult file of IIIVmrMLM.QEI

Marker	Chromosome	Position(bp)	pvalue.QEI (RI_random)	pvalue.QEI (variance_random)	⋯
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	⋯
⋮	⋮	⋮	⋮	⋮	⋯

In this file, all the columns are named as Trait ID, Trait_indicator_method, Marker (marker name), Chromosome, Position (position (bp) of markers on genome), LOD (QE) (LOD scores for QEIs), add (additive-by-environment interaction related effect), dom (dominant-by-environment interaction related effect), variance (the variance of each QEI), r^2 (%) (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 (the significance (SIG) of QTNs is 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) (Table 8)

Table 8 The final result file of IIIVmrMLM.QEI

Trait ID	Trait_indicator_method	Marker	Chromosome	Position (bp)	LOD (QE)	add	dom	variance	r^2(%)	P-value	significance
1	trait1_RC_fixed	PZB01647.1	1	231039372	11.0446	1.1628	5.8057	1.2459	2.8372	9.03E-12	SIG
1	trait1_RC_fixed	PZA02812.34	1	267615649	12.721	-1.3274	4.7147	1.4823	3.3754	1.90E-13	SIG
1	trait1_RC_fixed	PZA03073.28	3	168443662	14.6457	-1.1235	4.7367	1.8954	4.3162	2.26E-15	SIG
1	trait1_SD_random	PZA02957.4	1	281818425	18.9611	1.6327		1.3099	2.9828	9.25E-21	SIG
1	trait1_SD_random	PZA01122.1	4	12618115	19.6559	-1.7468	4.5194	1.6788	3.8229	2.21E-20	SIG
1	trait1_CV_random	PZA03305.5	1	286642725	4.7749	-0.311	5.8961	0.658	1.4984	1.68E-05	SIG
1	trait1_CV_random	PZA00176.8	2	10533421	10.299	1.187	0.065	1.1969	2.7257	5.03E-11	SIG
1	trait1_CV_random	PZB01642.1	5	12337501	13.8005	1.3689		0.5059	1.1521	1.56E-15	SIG

*_QEI_Manhattan plot: In the Manhattan plot, the median of -$\log_{10}$(P) values, calculated from all the methods selected by users, for each marker are used to draw the Manhattan plots. These selected approaches may be multi-environment joint analysis, and fourteen indirect methods, in which seven indirect indicators are used as phenotypes to indirectly identify QEIs using 3VmrMLM under random/fixed models. In the Manhattan plot, these dots are indicated by light colors. All the QEIs commonly identified by multiple approaches are indicated by the pink dots that are shown above dotted vertical lines, while all the QEIs identified by one single approach are indicated by the blue color dots that are also shown above dotted vertical lines (Fig 2).

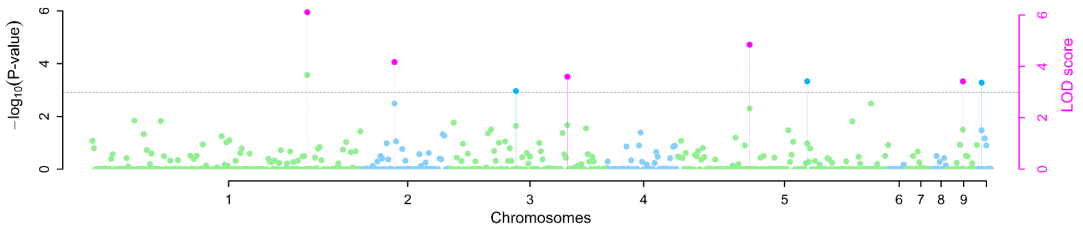

Figure 2. Manhattan plot

3. References

1. **Zhang YW, Li M, Han XL, Zhang YM**. IIIVmrMLM.QEI: A all-in-one tool for detecting more QTN-by-environment interactions in genome-wide association studies. **Computational and Structural Biotechnology Journal**, in submission.
2. **Li, M., Zhang, Y.W., Xiang, Y., Liu, M.H., Zhang, Y.M.** (2022a) IIIVmrMLM: The R and C++ tools associated with 3VmrMLM, a comprehensive GWAS method for dissecting quantitative traits. **Molecular Plant**, 15(8): 1251-1253.
3. **Li, M., Zhang, Y.W., Zhang, Z.C., Xiang, Y., Liu, M.H., Zhou, Y.H., *et al.*** (2022b) A compressed variance component mixed model for detecting QTNs and QTN-by-environment and QTN-by-QTN interactions in genome-wide association studies. **Molecular Plant**, **15**, 630–650.