User Manual for

q3VmrMLM

quantile-based 3VmrMLM tools for multi-locus GWAS

(version 1.0)

Wen-Xian Sun, Xiao-Yu Chang, Yuan-Ming Zhang

(soyzhang@mail.hzau.edu.cn)

Last updated on October, 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, and appropriate. However, results are not guaranteed for any specific datasets. We strongly recommend that users integrate the q3VmrMLM results with those from other software packages, i.e., 3VmrMLM, mrMLM, GEMMA, EMMAX, and PLINK.

Download website: https://github.com/YuanmingZhang65/q3VmrMLM

The work was supported by the National Natural Science Foundation of China (32070557 and 32270673).

1. Introduction

1.1 Why q3VmrMLM?

q3VmrMLM (quantile-based 3 Variance-component multi-locus random-SNP-effect Mixed Linear Model) program is an R package for multi-locus genome-wide association studies (GWAS) that identifies QTNs and QTN-by-environment interactions (QEIs) for quantitative and multi-omics traits.

1.2 LDAK-Thin Model

The q3VmrMLM software utilizes the LDAK-Thin Model (Speed et al., 2020) from the LDAK software (https://dougspeed.com/) to select the specific markers from the original genotype file. The following command is used for this selection process:

ldak6 --thin thin --bfile fileGen --window-prune 0.98 --window-kb 100

Once the relevant markers are selected, the q3VmrMLM software proceeds to compute the LD-adjusted Kinship matrix and conducts the GWAS analysis. The entire process is seamlessly integrated into the q3VmrMLM software, which automatically performs the necessary steps to ensure an efficient and user-friendly experience. Users can rely on the software to handle all operations without requiring additional input.

1.3 Getting started

The q3VmrMLM software runs only in the R software environment and can be obtained from the maintainer, Dr Yuan-Ming Zhang at College of Plant Science and Technology, Huazhong Agricultural University (soyzhang@mail.hzau.edu.cn).

1.3.1 One-Click installation

Under Linux system, the user needs the following command line to create the environment to install q3VmrMLM and dependent packages

```
conda create -n r43
conda activate r43
conda install -c conda-forge r-base=4.3.1
```

Within the R environment, the q3VmrMLM and dependent packages can be installed as follows.

Please install the dependency packages first

```
install.packages(c("quantreg","RcppEigen","Rcpp","doParallel","data.table","MASS","openxlsx","BEDMatrix","bigmemory","stringr","progress","matrixStats","collapse","bigstatsr","Matrix","nnet","SKAT","doFuture","gaston","devtools"))
```

Package MORST and bivas need to be installed from github.

devtools::install_github("mxcai/bivas")

devtools::install_github("yaowuliu/MORST")

Then install the q3VmrMLM package from local files as described below

unzip /home/ Users name /q3VmrMLM.zip

conda activate r43

R

install.packages("/home/ Users name /q3VmrMLM", repos = NULL)

Note that the software path is "/home/Users name/".

User Manual file Users can decompress the q3VmrMLM package and find the user manual file (name: Instruction.pdf) in the folder of "…/q3VmrMLM/inst".

1.3.2 Run q3VmrMLM from the command line

./q3VmrMLM -dir /home/SWX/1111 -fileGen plink -filePhe phenotype_all_env.csv -method Single_env

2. Function q3VmrMLM

2.1 Parameter settings

Parameter	Meaning	File format	Note
fileGen	Name of genotypic file in your computer, e.g., -fileGen Genotype	PLINK binary files: **Genotype.bed+Genotype. bim+Genotype.fam**	
filePhe	Name of trait phenotypic file in your computer, e.g., -filePhe Phenotype.csv	*.csv; *.txt (Phenotypic values. **Row**: individual; **Column**: traits)	Table 1
fileKin	Name of individual kinship file in your computer, e.g., -fileKin Kinship.csv	*.csv; *.txt (Kinship matrix. **Row** & **Column**: individuals)	Table 2
filePS	Name of population structure file in your computer, e.g., -filePS PopStr.csv	*.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), e.g., -PopStrType PCA(Q, EvolPopStr)		
fileCov	Name of covariate file in your computer, e.g., -fileCov Covariate.csv	*.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 e.g., -method Single_env(Multi_env)		
trait	Traits analyzed from number 1 to number 3, e.g., **-trait "c(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**, e.g., **-trait "c(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.)		

<table>
<tr><td>SearchRadius</td><td>In the setup of decollinearity parameter SearchRadius, only one parameter should be set in QTN and QEI detection, the value is -SearchRadius a, indicating the fact that only one potentially associated QTN or QEI is selected within a kb. The value of SearchRadius depends on species, traits and marker density. The value may be set as zero in low marker density and Monte Carlo simulation studies and 200 kb in high marker density, such as 1439 rice hybrid dataset (Huang et al., 2015). This value depends on several factors like LD distances of species, traits, marker density and so on. Users may set several values to find the best ones in real data analysis and the purpose is to reduce the effect of collinearity among selected markers on the results. Incorporating information on traits, marker density and other relevant factors improves the reliability and interpretability of GWAS results.
<table>
<tr><th>Species</th><th>LD distances (kb)</th><th>References</th></tr>
<tr><td>rice</td><td>≈100~500</td><td>Huang et al., 2010</td></tr>
<tr><td>wheat</td><td>≈4200</td><td>Pang et al., 2020</td></tr>
<tr><td>cultivated soybean</td><td>≈250</td><td>Lam et al., 2010</td></tr>
<tr><td>maize</td><td>≈1~5</td><td>Yan et al., 2009</td></tr>
<tr><td>Arabidopsis</td><td>≈250</td><td>Nordborg et al., 2002</td></tr>
<tr><td>barley</td><td>≈100-200</td><td>Caldwell et al., 2006</td></tr>
<tr><td>cotton</td><td>≈220</td><td>Ma et al., 2021</td></tr>
</table></td></tr>
<tr><td>svpal</td><td>In the setup of critical P-value parameter svpal, one critical P-value (default -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. The P-value should be set between 0 and 1.</td></tr>
<tr><td>DrawPlot</td><td>-DrawPlot FALSE: no figure output; -DrawPlot TRUE: the output of Manhattan plot. The Default value is TRUE.</td></tr>
<tr><td>Plotformat</td><td>The format for figures storage, including *.jpeg, *.png, *.tiff, and *.pdf. Default value is –Plotformat "pdf", means *.pdf format.</td></tr>
<tr><td>dir</td><td>Save path of the results in your computer, e.g., -dir /home/Users/.</td></tr>
</table>

2.2 Single_env: The detection of QTNs for quantitative traits

An example code for Monte Carlo simulation datasets

The user can run the q3VmrMLM process by simply typing the following command line at the command prompt. Where 'fileGen', 'filePhe', 'method', 'trait', 'n_en', 'dir', fileKin; filePS; PopStrType; fileCov; SearchRadius; svpal; DrawPlot; Plotformat; is a list of user specified parameters.

cd /home/username/XXX

Note XXX indicates the directory (folder) for the q3VmrMLM.sh

chmod +x q3VmrMLM

./q3VmrMLM -dir /home/username/ -fileGen genotype -filePhe phenotype.csv -method Single_env -trait “c(1)” -n_en “c(1)” -fileKin NULL -filePS NULL -PopStrType NULL -fileCov NULL-SearchRadius 0 -svpal 0.01 -DrawPlot TRUE -Plotformat "pdf"

- dir Paths to the input and output files
- fileGen Name of trait genotypic file
- filePhe Name of trait phenotypic file

- method Methods for analysing associations, including Single_env and Multi_env
-trait Analysis of the ith phenotype

Users must set "fileGen", "filePhe", "method", "trait", and "dir", while the other parameters may be the default in function q3VmrMLM, including fileKin=NULL; filePS=NULL; PopStrType="Q"; fileCov=NULL; SearchRadius=0; svpal=0.01; DrawPlot=TRUE; Plotformat="pdf"

In real data analysis, "SearchRadius 0" should be changed to "SearchRadius 200". Note that 200 may not be the best, and users may try to find the best ones.

2.2.1 Data input format

Format for genotypic dataset "fileGen"

The genotype file type 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 user reference.

Under linux system, please install the Linux versions of the TASSEL and PLINK software first, and then follow 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 the 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/

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:\q3VmrMLM\Genotype.hmp.txt -outputFile E:\q3VmrMLM\Genotype.sort.hmp.txt -fileType Hapmap

./run_pipeline -fork1 -h E:\q3VmrMLM\Genotype.sort.hmp.txt -Xmx10g -export E:\q3VmrMLM\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:\q3VmrMLM\Genotype.vcf --make-bed --out E:\q3VmrMLM\Genotype. Note that, in code cd path, path is the location where PLINK is decompressed.

Format for phenotypic dataset "filePhe" (Table 1)

The type of phenotype file for a 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 missing phenotypes: "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 of Table 2, "262"

is the sample size (n), and “33-16”, “A4226”, and “A4722” are the individual IDs. Note that “*n*” is the number of individuals in common between the phenotypic and genotypic datasets. All the kinship coefficients are listed as an *n* × *n* matrix.

Table 2. The format of 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 will be calculated by the “q3VmrMLM” software. Only the marker information of the above *n* individuals is used to calculate kinship matrix.

fileKin="/home/User name/Kinship.csv" means that the Kinship matrix named Kinship.csv will be uploaded from the folder "D:/Users". If the number and order of individuals in Kinship.csv do not match those of the above *n* individuals in the genotypic and phenotypic files, our software can adjust the K matrix so that the number and order of the transferred K matrix match those of the above *n* common individuals.

Q matrix format for dataset “filePS” (Table 3)

The Q matrix dataset in Table 3 consists of an (*n*+2) × (*k*+1) matrix, where *n* is the sample size (the number of the common individuals above), and *k* is the number of subpopulations. 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, “Q_1” 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 were 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 an (*n*+2) × (*k*+1) matrix, where *n* is

the sample size (the number of the common individuals), and *k* is the number of principal components. In the first column, "<PCA>" and "<ID>" must appear 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, "PC_1" to "PC_k" indicate the first to *k*th principal components. In the second column, "0.306" is the score of the first principal component of the 1st individual. Note that the software does not delete any principal components.

Table 4. The format of the principal components in 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 an (*n*+2) × 2 matrix, where *n* is the sample size (the number of the common individuals). In the first column, "<EvolPopStr>" and "<ID>" must appear 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 format of the evolutionary population in the filePS dataset

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

filePS=NULL indicates that there is no population structure. filePS="/home/Users name/PopStr.csv" means that the population structure dataset named PopStr.csv is uploaded from the folder "D:/Users" folder. If the number and order of individuals in PopStr.csv do not match those in the above common individuals, our software can adjust the population structure matrix so that the number and order of the new matrix match 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 the sample size (the number of the common individuals), and k is the number of covariates. In the first column, “<Covariate>” and “<ID>” must appear in the first and second rows, respectively. If the covariate is categorical, the names are Cate_covariate*. If the 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 means that there is no covariate. fileCov="D:/Users/covariate.csv" means that the covariates named covariate.csv are uploaded from the “D:/Users” folder. If the number and order of individuals in the uploaded file do not match those in the above common individuals, our software will 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) contains three files: *_K.csv (Kinship matrix computed by q3VmrMLM), *_midresult.csv (intermediate results), and *_result.xlsx (final results), and one Manhattan plot (if DrawPlot=TRUE).

*_midresult.csv: This is the result 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 (marker position (bp) on the genome), and pvalue.Q (the P-value of QTN).

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 (marker 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^2distribution), and significance (significant (SIG) QTNs are based on Bonferroni correction, i.e., 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	r^2(%)	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: The left *y*-axis reports -$\log_{10}$(P-values), which are obtained from genome-wide single-marker scanning for all the markers in the first step of q3VmrMLM, while the right *y*-axis reports LOD scores obtained from likelihood ratio test for significant and suggested QTNs, with the thresholds of 0.05/*m* and LOD = 3.0 (dashed line), respectively, in the second step of q3VmrMLM. These LOD scores are shown as points with straight lines. When LOD scores ≥ 20, the obtained LOD scores are transformed as LOD' = 20 + (LOD – 20)/100 to make the Manhattan plot more beautiful.

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

An example code for Monte Carlo simulation data

The user can run the q3VmrMLM process by simply typing the following command line at the command prompt. Where 'fileGen', 'filePhe', 'method', 'trait', 'n.en', 'dir', fileKin; filePS; PopStrType; fileCov; SearchRadius; svpal; DrawPlot; Plotformat; is a list of user specified parameters,

./q3VmrMLM -dir /home/username/ -fileGen genotype -filePhe phenotype.csv -method c('Multi_env') -trait "c(1:2)" -n_en "c(2,2)" -fileKin NULL -filePS NULL -PopStrType NULL -fileCov NULL-SearchRadius 0 -svpal 0.01 -DrawPlot TRUE -Plotformat "pdf"

In real data analysis, "SearchRadius=0" should be changed to "SearchRadius=200". Note that 200 may not be the best, and users may try to find the best ones.

Users must set "fileGen", "filePhe", "method" (or module), "trait", "n.en", and "dir", while the other parameters may be default in function q3VmrMLM, 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 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 file filePhe contains the phenotypic values of three traits, and the number of environments for each trait is 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>	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 first row in the first column: "<Phenotype>", while the second to *n*th rows in the first column are 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 phenotype file is organized by trait, each trait has at least two columns, and each column is the phenotypes measured in an environment. The missing phenotypes are represented by "NA".

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

2.3.2 Result

The result file (result-QEI_detection) contains three files: *_K.csv (kinship matrix computed by q3VmrMLM), *_midresult.csv (intermediate results), and *_result.xlsx (final result, including two sheets, significant/suggested QTNs (result.Q) and significant / suggested QEIs (result.QEI)), and a Manhattan plot (if drawPlot=TRUE), for QTN and QEI.

*_midresult.csv: This is the result of a single marker scan on the genome in the first step. In this file, all the columns are named as Marker (marker name), Chromosome, Position

(marker position (bp) on the genome), p-value (Q+E) (the global p-value for QTNs and QEIs).

Marker	Chromosome	Position (bp)	p-value (Q+E)
PZB00859.1	1	157104	0.812972071
PZA01271.1	1	1947984	0.993594668
PZA03613.2	1	2914066	0.99306619
PZA03613.1	1	2914171	0.99997328
PZA03614.2	1	2915078	0.971495059
⋮	⋮	⋮	⋮

result.Q: The results are for significant / suggested QTNs. In this sheet, all the columns are named as Trait ID, Trait Name, Marker (marker name), Chromosome, Position (marker position (bp) on the genome), LOD (Q) (LOD score for 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 QTN detection using χ^2 distribution), and significance (significant (SIG) QTNs are based on Bonferroni correction, i.e., critical P-value is 0.05/*m*, where *m* is the number of 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 result for significant / suggested QEIs. In this sheet, all the columns are named as Trait ID, Trait Name, Marker (marker name), Chromosome, Position (marker 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), r^2 (%) 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, i.e., critical P-value is 0.05/*m*, where *m* is the number of 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

*_QTN _Manhattan plot: Manhattan plot for QTNs. The left *y*-axis reports -$\log_{10}$(P-values) of the global P-values for QTNs and QEIs obtained from genome-wide single-marker scanning for all the markers in the first step of q3VmrMLM, while the right *y*-axis reports LOD scores, which are obtained from the likelihood ratio test for significant / suggested QTNs, with the thresholds of 0.05/*m* and LOD = 3.0 (dashed line), respectively, in the second step of q3VmrMLM. These LOD scores are shown as points with straight lines.

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

*_QEI_Manhattan plot: Manhattan plot for QEIs. The left *y*-axis reports -$\log_{10}$(P-values) of the global P-values for QTNs and QEIs obtained from genome-wide single-marker scan for all the markers in the first step of q3VmrMLM, while the right *y*-axis reports LOD scores obtained from the likelihood ratio test for significant / suggested QEIs, with the thresholds of 0.05/*m* and LOD = 3.0 (dashed line), in the second step of q3VmrMLM. These LOD scores are shown as points with straight lines.

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

References

1 Caldwell, K. S., Russell, J., Langridge, P., and Powell, W. (2006). Extreme population-dependent linkage disequilibrium detected in an inbreeding plant species, Hordeum vulgare. Genetics. 172(1):557-567.

2 Huang, X., Wei, X., Sang, T., Zhao, Q., Feng, Q., Zhao, Y., Li, C., Zhu, C., Lu, T., Zhang, Z., Li, M., Fan, D., Guo, Y., Wang, A., Wang, L., Deng, L., Li, W., Lu, Y., Weng, Q., Liu, K., ··· Han, B. (2010). Genome-wide association studies of 14 agronomic traits in rice landraces. Nat Genet. 42(11):961-967.

3 Lam, H. M., Xu, X., Liu, X., Chen, W., Yang, G., Wong, F. L., Li, M. W., He, W., Qin, N., Wang, B., Li, J., Jian, M., Wang, J., Shao, G., Wang, J., Sun, S. S., and Zhang, G. (2010). Resequencing of 31 wild and cultivated soybean genomes identifies patterns of genetic diversity and selection. Nat Genet. 42(12):1053-1059.

4 Li, M., Zhang, Y. W., Xiang, Y., Liu, M. H., and Zhang, Y. M. (2022). 3VmrMLM: The R

and C++ tools associated with 3VmrMLM, a comprehensive GWAS method for dissecting quantitative traits. Mol Plant. 15(8):1251–1253.

5 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.

6 Ma, Y., Min, L., Wang, J., Li, Y., Wu, Y., Hu, Q., Ding, Y., Wang, M., Liang, Y., Gong, Z., Xie, S., Su, X., Wang, C., Zhao, Y., Fang, Q., Li, Y., Chi, H., Chen, M., Khan, A. H., Lindsey, K., ⋯ Zhang, X. (2021). A combination of genome-wide and transcriptome-wide association studies reveals genetic elements leading to male sterility during high temperature stress in cotton. New Phytol. 231(1):165-181.

7 Nordborg, M., Borevitz, J. O., Bergelson, J., Berry, C. C., Chory, J., Hagenblad, J., Kreitman, M., Maloof, J. N., Noyes, T., Oefner, P. J., Stahl, E. A., and Weigel, D. (2002). The extent of linkage disequilibrium in Arabidopsis thaliana. Nat Genet. 30(2):190-193.

8 Pang, Y., Liu, C., Wang, D., St Amand, P., Bernardo, A., Li, W., He, F., Li, L., Wang, L., Yuan, X., Dong, L., Su, Y., Zhang, H., Zhao, M., Liang, Y., Jia, H., Shen, X., Lu, Y., Jiang, H., Wu, Y., … Liu, S. (2020). High-Resolution Genome-wide Association Study Identifies Genomic Regions and Candidate Genes for Important Agronomic Traits in Wheat. Mol Plant. 13(9):1311-1327.

9 Speed, D., Holmes, J., and Balding, D. J. (2020). Evaluating and improving heritability models using summary statistics. Nat. Genet. 52:458–462.

10 Yan, J., Shah, T., Warburton, M. L., Buckler, E. S., McMullen, M. D., & Crouch, J. (2009). Genetic characterization and linkage disequilibrium estimation of a global maize collection using SNP markers. PLoS One. 4(12):e8451.

11 Sun, W.X., Chang, X.Y., Chen, Y., Zhao, Q., Zhang, Y.M. The integration of quantile regression with 3VmrMLM identifies more QTNs and QTNs-by-environment interactions using SNP and haplotype-based markers. Plant Communications, Provisionally Accepted.