User Manual for

Fast3VmrMLM

Fast **3** **V**ariance components **m**ulti-locus **r**andom-SNP-effect **M**ixed **L**inear **M**odel tools for genome-wide association study

(**version 1.0**)

Wang Jing-Tian and Zhang Yuan-Ming

(soyzhang@mail.hzau.edu.cn; 450628680@qq.com)

Last updated July, 2025

Disclaimer: The software has undergone comprehensive testing by Yuan-Ming Zhang's Lab at the College of Plant Science and Technology, Huazhong Agricultural University. The results obtained from the software are generally reliable, correct, and appropriate. However, it is important to note that these results are not guaranteed for specific datasets. We strongly recommend that users integrate Fast3VmrMLM results with those from other software packages, i.e., mrMLM, 3VmrMLM and q3VmrMLM.

Download website:

https://github.com/YuanmingZhang65/Fast3VmrMLM

Citation:

Wang J#, Chen Y#, Shu G#, Zhao M#, Zheng A, Chang X, Li G, Wang Y* and Zhang YM*. Fast3VmrMLM: A fast algorithm that integrates genome-wide scanning with machine learning to accelerate gene mining and breeding by design for polygenic traits in large-scale GWAS datasets. Plant Communications 2025; 6: 101385.

This study was supported by the National Natural Science Foundation of China (32470657; 32270673).

1 Introduction

1.1 Why Fast3VmrMLM?

Fast3VmrMLM (**F**ast **3** **V**ariance components **m**ulti-locus **r**andom-SNP-effect **M**ixed **L**inear **M**odel) is an R package designed for fast and big data genome-wide association studies (GWAS), which consider additive and dominant effects and control for their polygenic backgrounds in a compressed variance component mixed linear model.

The current version, Fast3VmrMLM v1.0, features three modules:

1) Fast3VmrMLM: The genome-wide scanning plus machine learning framework was developed and integrated with advanced computational techniques to identify QTNs for complex traits using fast, efficient and large-scale GWAS algorithm, Fast3VmrMLM.
2) Fast3VmrMLM-Hap: Haplotypes of bins constructed by adjacent linkage disequilibrium markers are used to associate the trait of interest with the markers using the Fast3VmrMLM-Hap module.
3) Fast3VmrMLM-mQTL: To identify mQTLs, the Fast3VmrMLM-mQTL module can be used to associate the trait of interest with molecular variations, where there are two types of formats for the input file of the genotypes of molecular variations, including genes, lncRNA haplotypes and SVs. One is the "*.vcf" format with multi-allele marker, and another is the PLINK binary genotype format with the start and end positions on the genome for molecular markers (see Table 6).

Fast3VmrMLM v1.0 works only on Linux system.

1.2 Getting started

Fast3VmrMLM v1.0 (https://github.com/YuanmingZhang65/Fast3VmrMLM) is a package that runs in the R environment on the Linux system, which can be freely downloaded from the above website or requested from the maintainer, Dr Yuan-Ming Zhang at College of Plane Science and Technology, Huazhong Agricultural University (450625680@qq.com; soyzhang@hotmail.com).

1.2 Installation

Within the Linux system, the Fast3VmrMLM software can be installed by the following four steps:

1) To download and install Anaconda at https://repo.anaconda.com/archive/, in which the recommended installer file is Anaconda3-2021.05-Linux-x86_64.sh.

2) To create a new conda environment named Fast3VmrMLM by the following code:

```
conda create -n "Fast3VmrMLM" r-essentials r-base=4.3
```

3) To install the dependency packages.

- Install mamba for fast installation by:

  ```
  conda install -c conda-forge mamba
  ```

- Install dependency packages by following codes in Linux environment (or by *install.packages* in R environment):

  ```
  mamba install -c conda-forge r-Rcpp
  mamba install -c conda-forge r-RcppArmadillo
  mamba install -c conda-forge r-RcppParallel
  mamba install -c conda-forge r-data.table
  mamba install -c conda-forge r-MASS
  mamba install -c conda-forge r-openxlsx
  mamba install -c conda-forge r-BH
  ```

  (The above "*mamba*" can be replaced by "*conda*" if it doesn't work.)

- Install dependency packages "boost" by the following code in Linux environment:

  ```
  mamba install -c conda-forge boost
  ```

  or

  ```
  conda install -c conda-forge boost
  ```

4) To decompress Fast3VmrMLM.zip with the following bash code and install it with R code by:

```
unzip '/user/Fast3VmrMLM_Linux.zip' -d '/user/'
R
install.packages("/home/user/Fast3VmrMLM", repos = NULL)
```

The above installation may take some time, please be patient and wait.

If you have any questions about installation or anything else, please don't hesitate to contact Yuan-Ming Zhang (soyzhang@mail.hzau.edu.cn; 450625680@qq.com) or Jing-Tian Wang (1390032282@qq.com).

User Manual file Users can decompress the Fast3VmrMLM package and find the User Manual file (name: Instruction.pdf) in the folder of "…/ Fast3VmrMLM/inst".

Example datasets Users can decompress the Fast3VmrMLM package and find the example datasets in the folder of "…/Fast3VmrMLM/extdata".

1.2.3 Run Fast3VmrMLM

Once the software Fast3VmrMLM is installed, users may run it using two commands:

library("Fast3VmrMLM")

Fast3VmrMLM(***) (please see the example of § 2.2)

Before using Fast3VmrMLM in Linux, make sure to activate the conda environment of Fast3VmrMLM: *conda activate Fast3VmrMLM*. Users need to run *library("Fast3VmrMLM")* every time before utilizing the Fast3VmrMLM software package.

2 Function Fast3VmrMLM

2.1 Parameter settings

Parameter	Meaning	File format	Note
fileGen	Name & path of genotypic file in your device, e.g., fileGen="D:/Users/Genotype". **All file paths should exclude "~".**	PLINK binary files: **Genotype.bed+Genotype.bim+Genotype.fam**	
filePhe	Name & path of trait phenotypic file in your device, e.g., filePhe="D:/Users/Phenotype.csv".	*.csv (Phenotypic values. Row: individual; Column: traits)	**Table 1**
fileKin	Name & path of individual kinship file in your device, e.g., fileGRM="D:/Users/GRM.csv" or fileGRM=NULL.	*.csv (GRM. Row & Column: individuals)	**Table 2**
filePS	Name & path of covariates file in your device, e.g., filePS="D:/Users/covariates.csv" or filePS=NULL.	*.csv (Population structure. Row: individual; Column: sub-populations or covariates, e.g., sex and age)	**Tables 3-5**
filePedigree	filePedigree="D:/Users/Pedigree.csv" when using the NC II genetic mating design population, and filePedigree=NULL (the default value) for all others.	*.csv (Pedigree of NCII population consists of 3 individual-ID columns: the female & male parents, and their F_1 offspring)	**Table 7**
PopStrType	The types of population structure include "PopStrType=Q" (*Q matrix or ancestry coefficient matrix calculated by Admixture software*), "PopStrType=PC" (*principal components or covariates*), and "PopStrType=Evol" (*evolutionary population structure*).		
fileOut	Save path of the result in your device, e.g., "D:/Users/".		
genoType	Setting the algorithm as "SNP" (SNP-based Fast3VmrMLM algorithm), "Hap" (Haplotype-based Fast3VmrMLM-Hap algorithm) and "molecular" (molecular-based Fast3VmrMLM-mQTL algorithm)		
population	population="NCII" when using the NCII design, and population="nature" (the default value) for all others		
trait	Traits analyzed from number 1 to number 2, e.g., trait=1:3 indicates that users analyze the first to third traits.		
svrad	A physical distance of sliding window for removing potential candidate variants with the collinearity of the most significant one. Default value is svrad=2.0e+4 (bp). Users can obtain more potential associated variants by setting a small value of SearchRadius.		
svpal	A critical *P*-value (default svpal=1.0e-5) to select all the potentially associated variants in genome-wide single-variant scanning. The size of svpal may be changed based on sample size, such as from 1.0e-5 to 1.0e-2.		
svmlod	A critical LOD score, which is larger than 0, (default svmlod=3), is used to select suggested variants.		
SampleMarkersforGRM	A parameter indicates whether using a subset of variants to construct kinship (TRUE) or not (FALSE). Default value is SampleMarkersforGRM=FALSE.		
SampleMarkersforGRMNum	A parameter (>0) indicates the number of variants sampled to calculate kinship. Only used when SampleMarkersforGRM=TRUE. Default value is SampleMarkersforGRMNum=1.0e+4.		
c_threshold	A parameter for evaluating linkage disequilibrium of adjacent variants when constructing bin-based haplotypes, ranging from 0 to 1. Default value is c=0.7.		

numofHaplotypes	A parameter for setting the number of haplotypes for each bin genotype in the Fast3VmrMLM-Hap module. Default value is numofHaplotypes=3.
scainngParallel	A parameter indicates whether using parallel computing in genome-wide scanning. Default value is scainngParallel=FALSE.
nThreads	A parameter indicates the number of cores used in parallel computing. Default value is nThreads=20.
DrawPlot	A parameter indicates whether drawing and outputting the Manhattan plot based on the GWAS results. Default value is DrawPlot=FALSE.
Plotformat	A parameter indicates the format of the Manhattan plot. Default value is Plotformat=*.tiff.
MGinputClass	Setting the input file format "bed" (PLINK binary file) and "vcf" (.vcf format file), when variableType="molecular" in the Fast3VmrMLM-mQTL module. Default value is inputClass="vcf".
filegeneRegion	Setting the input file that indicates the genome intervals of each molecular marker when inputClass="bed" in the Fast3VmrMLM-mQTL module.

2.2 Running codes

The running code for Fast3VmrMLM is as follows:

Fast3VmrMLM(fileGen="/home/Genotype",filePhe="/home/phenotype.csv",fileKin=NULL,filePS="/home/PopStr.csv",PopStrType="PC",fileOut="/home/",genoType="SNP",trait=1,svrad=2e+4,svpal=1e-5,svmlod=3,scainngParallel=TRUE,nThreads=20,DrawPlot=FALSE,Plotformat=".tiff")*

The running code of Fast3VmrMLM when using NCII design is as follows:

Fast3VmrMLM(fileGen="/home/Genotype",filePhe="/home/phenotype.csv",fileKin=NULL,filePS="/home/PopStr.csv",filePedigree="/home/Pedigree.csv",PopStrType="PC",population="NCII",fileOut="/home/",genoType="SNP",trait=1,svrad=2e+4,svpal=1e-5,svmlod=3,scainngParallel=TRUE,nThreads=20,DrawPlot=FALSE,Plotformat=".tiff")*

Note that when Fast3VmrMLM was used for NCII design population, the genotype file (*fileGen*) should only consist of the genotypes of parents, and the genotypes of offsprings can be inferred from their parents by Fast3VmrMLM according to the pedigree (*filePedigree*).

The running code for Fast3VmrMLM-Hap algorithms:

Fast3VmrMLM(fileGen="/home/Genotype",filePhe="/home/phenotype.csv",fileKin=NULL,filePS="/home/PopStr.csv",PopStrType="PC",fileOut="/home/",genoType="Hap",trait=1,svrad=2e+4,svpal=1e-5,svmlod=3,c_threshold=0.7,numofHaplotypes=3)

The running code for Fast3VmrMLM-mQTL algorithms:

Fast3VmrMLM(fileGen="/home/Genotype",filePhe="/home/phenotype.csv",fileKin=NULL,filePS="/home/PopStr.csv",PopStrType="PC",fileOut="/home/",genoType="molecular",trait=1,svrad=2e+4,svpal=1e-5,svmlod=3,MGinputClass="vcf")

Users must set "fileGen", "filePhe", "trait", and "fileOut", while the other parameters may be default in function Fast3VmrMLM (see § 2.1).

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), which can be found the introduction from PLINK v1.9 (https://www.cog-genomics.org/plink2/.

Format for genotypic dataset "filePhe" (Table 1)

The file type of phenotypes for complex trait is *.csv, following the format outlined in Table 1. The first row in the first column: "<Phenotype>"; the second to nth rows in the first column: individual IDs or names, such as Ind46. The first row in other columns: trait names, such as "trait1", and the second to nth rows in other columns: values of phenotypic traits. The phenotypes missed: "NA".

Table 1. The format of phenotypic dataset

<Phenotype>	**trait1**	**trait2**	**trait3**	**…**
Ind46	91.03	88.32	87.67	…
Ind52	103.11	103.10	98.36	…
Ind57	92.07	116.30	NA	…
Ind64	128.5	101.20	101.02	…
Ind68	95.84	91.74	94.50	…
⋮	⋮	⋮	⋮	…

Format for genotypic dataset "fileKin" (Table 2)

Table 2. The format of knship dataset

Ind_1	Ind_2	Ind_3	Ind_4	Ind_5	Ind_6
1	0.700361011	0.599277978	0.675090253	0.620938628	…
0.700361011	1	0.620938628	0.666064982	0.653429603	…
0.599277978	0.620938628	1	0.561371841	0.5433213	…

0.675090253	0.666064982	0.561371841	1	0.615523466	…
0.620938628	0.653429603	0.5433213	0.615523466	1	…
⋮	⋮	⋮	⋮	⋮	…

The “fileKin” should be a file with *.csv format. All the kinship coefficients are listed as an n × n matrix. Both rows and columns represent individuals that arranged in the order of the *.fam file. The parameter “fileKin=NULL” indicates that the kinship matrix is calculated by the “Fast3VmrMLM” software. When fileKin="D:/Users/kinship.csv", the kinship matrix with name kinship.csv is uploaded from the folder "D:/Users". If the number and order of individuals in the "kinship.csv" file do not match those in the phenotypic files, our software will attempt to match them.

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

The Q matrix dataset in Table 3 consists of a (n+1) × (k+1) matrix, where n is sample size (the number of the above common individuals), and k is the number of sub-populations. The first column is “<Structure>” and individual IDs or names. In the 2nd to (k+1)-th columns, “Q1” to “Qk” indicate sub-populations. In the second 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.

Table 3. The Q matrix format of dataset filePS

<Structure>	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 or covariates for dataset “filePS” (Table 4)

The principal components or covariates dataset in Table 4 consists of a (n+1) × k matrix, where n is sample size (the number of the above common individuals), and k is the number of principal components and/or covariates. The first column is “<PCA>” and individual IDs or names. The 2nd to k-th columns indicate principal components and/or covariates.

Table 4. The principal components or covariates format of dataset filePS

<PCA>	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	…
⋮	⋮	⋮	⋮	

The evolutionary population structure for dataset "filePS" (Table 5)

The evolutionary population structure dataset in Table 5 consists of a (n+1) × 2 matrix, where n is sample size and the number of categories for variables is the number of sub-populations. The first column is "<EvolPopStr>" and individual IDs or names. The 2nd column indicates the evolutionary sub-population. Other population structure described by character type of variables are supported in this format.

Table 5. The evolutionary population structure format of dataset filePS

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

The format of the chromosome intervals of each molecular marker for the parameter filegeneRegion in § 2.1 (Table 6)

When genoType="molecular" and MGinputClass="bed", the genome intervals of each molecular marker should be set by users via the file in Table 6. The first column is its ID, the second column is its chromosome, the third and the fourth columns are the left and right physical positions, respectively.

Table 6. The chromosome intervals of each molecular marker

Zm00001eb000010	1	34617	40204
Zm00001eb000020	1	41214	46762
⋮	⋮	⋮	⋮

The format of the pedigree file in § 2.1 (Table 7)

When population="NCII", the filePedigree should set the following file of Table 7. There are three individual-ID columns: the female and male parents, and their F_1 offspring.

Table 7. The pedigree file

LT16	LT131	LT16_LT131
LT07	LT131	LT07_LT131
⋮	⋮	⋮

2.2.1 Result

The results include three files: *_intermediate.csv (intermediate results), *_result.csv (final results) and a Manhattan plot file.

*_intermediate.csv: This file contains the results of genome-wide single-variants scanning in the first step. In this file, all the columns are named as "MarkerID" (variants name), "CHR" (Chromosome), "POS" (variants position (bp) on the genome), and "pval" (the *P*-value for main-effect variants).

MarkerID	CHR	POS	pval
PZB00859.1	1	157104	0.292043111
PZA01271.1	1	1947984	0.185246808
PZA03613.2	1	2914066	0.99208603
PZA03613.1	1	2914171	0.999987108
⋮	⋮	⋮	⋮

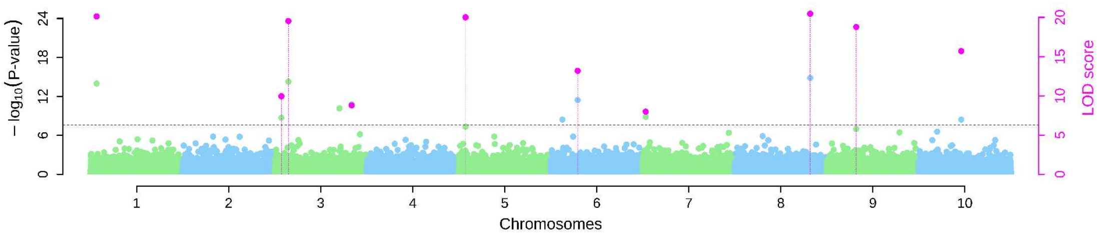
*_result.csv: The final result file for the Fast3VmrMLM module is as follows. All the columns are named as "Marker" (marker name); "Chromosome"; "Position (bp)" (markers position (bp) on the genome); "add" (additive effect); "dom" (dominance effect); "LOD" (LOD score); variance (the variance of each variants); "r2(%)" (the proportion of total liability variance explained by each variants); "*P*-value" (calculated from LOD score using χ^2 distribution); "significance" (significant (SIG) variants are based on Bonferroni correction, that is, critical *P*-value is 0.05/*m*, where *m* is the number of tests or variants, while suggested (SUG) variants are based on LOD ≥ 3.0, default). "A allele" is the nucleotide base with a dummy variable of 1.0.

Trait ID	Trait name	Marker	Chromo	Position (bp)	LOD	add	dom	variance	r^2(%)	P-value	signific	A allele
1	phe_1	null_653	1	654	31.5242	-0.4295	5.6711	7.9444	5.0209	3.00e-32	SIG	C
1	phe_1	null_20728	3	20729	9.9457	1.6284	2.7827	4.3464	2.7469	1.13e-10	SIG	G
1	phe_1	null_52928	6	52929	13.1912	3.5793	-0.163	6.4797	4.0952	6.45e-14	SIG	A
1	phe_1	null_78191	8	78192	65.6504	6.5965	-1.579	5.054	3.1942	2.25e-66	SIG	T

The final result file for the Fast3VmrMLM-Hap module is as follows. The “Marker_L” and “Marker_R” mean the markers on the left and right boundaries, respectively, of chromosome region (bin) determined by linkage disequilibrium, while “Position_L (bp)” and “Position_R (bp)” are for the left and right genomic positions, respectively. “Hap_1” to “Hap_numHap” are the haplotypic effects of the 1st to numHap-th haplotypes, where “numHap” is No. of haplotypes.

Trait	Trait	Marker_L	Marker_R	Chr	Position_L (bp)	Position_R (bp)	LOD	Hap_1	Hap_2	Hap_3	numHap	variance	r^2(%)	P-value	significance
1	phe_1	rs16857692	rs4668204	2	170672831	170705429	5.787	0.018	-0.870	0.018	3	0.2704	0.1268	1.63e-06	SUG
1	phe_1	rs10439267	rs6435173	2	203631692	203721357	125.5293	-2.213	-2.595	-2.213	3	12.6745	5.9419	2.99e-126	SIG
1	phe_1	rs1978865	rs7651321	3	61127327	61256881	82.8513	-1.924	-1.616	-1.924	3	6.88	3.2254	1.42e-83	SIG

*_Manhattan plot: Y-axis on the left-side reports -log10 *P*-values of variants, which are obtained from single-variant genome-wide scanning for all the variants in the first step of Fast3VmrMLM, while Y-axis on the right-side reports LOD scores, which are obtained from likelihood ratio test for suggested and significant variants, with the suggested threshold of LOD = 3.0 (dashed line), in the second step of Fast3VmrMLM. Users can set different LOD threshold by setting svmlod (see § 2.1). 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.



We recommend that all the significantly and suggested associated markers with the traits of interest are listed in a supplemental table, while all the all the significant and suggested QTNs with known and candidate genes are marked in the Manhattan plot.

3 Reference

1 Wang J, Chen Y, Shu G, Zhao M, Zheng A, Chang X, Li G, Wang Y, Zhang YM. Fast3VmrMLM: A fast algorithm that integrates genome-wide scanning with machine learning to accelerate gene mining and breeding by design for polygenic traits in large-scale GWAS datasets. Plant Communications 2025; 6: 101385.

4 Issues and solutions

1 **Issue:** There were “libstdc++.so.6: version `CXXABI_1.3.15' not found” reported after *library(Fast3VmrMLM)*.

Solution: Firstly, confirm that the R version is 4.3 and try to run Fast3VmrMLM again. If it doesn’t work, update the version of libstdc++ by the following code:

```
cd /home/user/anaconda3/envs/Fast3VmrMLM/lib
rm libstdc++.so
rm libstdc++.so.6
ln -s libstdc++.so.6.0.34 libstdc++.so
ln -s libstdc++.so.6.0.34 libstdc++.so.6
```