User Manual for

FastBiCmrMLM

Fast Binary-trait **C**ompressed variance component
multi-locus **r**andom-SNP-effect **M**ixed **L**ogistic **M**odel tools
for genome-wide association study
(**version 0.0.1**)

Wang Jing-Tian, and Zhang Yuan-Ming
(soyzhang@mail.hzau.edu.cn)

Last updated June, 2024

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 FastBiCmrMLM results with those from other software packages, i.e., mrMLM, IIIVmrMLM, GMMAT, SAIGE and fastGWA-GLMM.

Download website:

https://github.com/YuanmingZhang65/FastBiCmrMLM

Citation:

Wang, J.T., Chang, X.Y., Zhao, Q. and Zhang, Y.M. FastBiCmrMLM: a fast and powerful compressed variance component mixed logistic model for big genomic case-control genome-wide association study. ***Brief Bioinform***, 2024, Accepted

This study was supported by the National Natural Science Foundation of China (32270673; 32070557).

This study made use of the datasets generated by the Wellcome Trust Case Control Consortium. A full list of the investigators who contributed to the generation of the data is available from www.wtccc.org.uk. Funding for WTCCC1 project was provided by the Wellcome Trust under awards 076113, 085475, and 090355.

1 Introduction

1.1 Why FastBiCmrMLM?

FastBiCmrMLM (**F**ast **Bi**nary-trait **C**ompressed variance component **m**ulti-locus **r**andom-SNP-effect **M**ixed **L**ogistic **M**odel) is an R package designed for case-control genome-wide association study (GWAS), which consider additive and dominant effects and their polygenic backgrounds in a compressed variance component mixed logistic model.

The current version, FastBiCmrMLM v0.0.1, features four algorithms:
1) FastBiCmrMLM: Primary algorithm for FastBiCmrMLM, which is used to analyze small samples with the number of individuals less than 1,000.
2) FastBiCmrMLM-Time: A fast version of FastBiCmrMLM, which is used to analyze the samples with the number of individuals from 1,000 to 20,000. FastBiCmrMLM-Time took about seven minutes for one disease of WTCCC1 datasets.
3) FastBiCmrMLM-RAM: Optimized for memory efficiency, which is used to save memory when sample size is large (≥ 20,000). In the analysis of biobank-scale simulation datasets (500,000 individuals and one million SNP markers), FastBiCmrMLM-RAM took about 13.5 hours on the server with 10 CPUs and the memory usage of 141 Gb for one sample.
4) FastBiCmrMLM-Hap: Haplotypes for each bin are constructed by adjacent linkage disequilibrium markers and used to associate the trait of interest using FastBiCmrMLM.

Algoirthms 1) to 3) can be used to detect structure variant.

FastBiCmrMLM v0.0.1 works only on Linux system.

1.2 Getting started

FastBiCmrMLM v0.0.1 is a package that runs both in the R environment and in the Linux environment, which can be freely downloaded from https://github.com/YuanmingZhang65/FastBiCmrMLM, or request from the maintainer, Dr Yuan-Ming Zhang at Crop Information Center, College of Plane

Science and Technology, Huazhong Agricultural University (soyzhang@hotmail.com; soyzhang@mail.hzau.edu.cn).

1.2.1 Installation in R environment

Within R environment on Linux system, the FastBiCmrMLM software can be installed by the following R codes:

First install the dependency packages by:

install.packages(c("data.table","doParallel","BEDMatrix","MASS","RcppArmadillo","RcppParallel","BH","bigmemory","dplyr","glmnet","SPAtest"))

Then install the FastBiCmrMLM package from local files by:

install.packages("E:/FastBiCmrMLM_0.0.1.zip",repos=NULL,type="win.binary")

1.2.2 Installation in Linux environment

Within Linux environment, the FastBiCmrMLM software can be installed by the following bash codes:

First, create a new conda environment by:

conda create -n "FastBiCmrMLM" r-essentials r-base=4.3

Second, activate the new environment by:

conda activate FastBiCmrMLM

Then, install the dependency packages by:

conda install -c conda-forge r-data.table r-doParallel r-BEDMatrix r-MASS r-RcppArmadillo r-RcppParallel r-BH r-bigmemory r-dplyr r-SPAtest r-glmnet

Finally, decompress FastBiCmrMLM.zip and install it with R code by:

unzip '/user/FastBiCmrMLM_Linux.zip' -d '/user/'

R

install.packages(paste0("/user/FastBiCmrMLM_Linux"), repos = NULL)

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

1.2.3 Run FastBiCmrMLM

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

library("FastBiCmrMLM")

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

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

2 Function FastBiCmrMLM

2.1 Parameter settings

Parameter	Meaning	File format	Note
fileGen	Name & path of genotypic file in your device, e.g., fileGen="D:/Users/Genotype".	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**
fileGRM	Name & path of individual genetic relationship file in your device, e.g., fileGRM="D:/Users/GRM.csv" or fileGRM=NULL. When "method=FastBiCmrMLM-RAM", fileGRM is set to 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)	**Table 3-5**
PopStrType	The types of population structure include "structure" (*Q matrix*), "PCA" (*principal components or covariates*), and "Evol" (*evolutionary population structure*).		
fileOut	Save path of the result in your device, e.g., "D:/Users/".		
method	Setting the use of four algorithms including "FastBiCmrMLM", "FastBiCmrMLM-Time", "FastBiCmrMLM-RAM" and "FastBiCmrMLM-Hap".		
trait	Traits analyzed from number 1 to number 2, e.g., trait=1:3 indicates that users analyze the first to third traits.		
SearchRadius	A physical distance of sliding window for removing some potential candidate variants with the collinearity (|Pearson correlation| ≥ 0.70) of the most significant one. Default value is SearchRadius=1.0e+6. Users can obtain more potential associated variants by setting a small value of SearchRadius. When causal variants are clustered on a chromosomal region, SearchRadius=0 is available.		
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.		
lasso_alpha	A compression coefficient for lasso, which is set between 0 and 1. Default value is lasso_alpha=1. When a loose compression is necessary, e.g., causal variants are clustered on a chromosomal region, users may set a small value of lasso_alpha, such as lasso_alpha=0.		
LODThreshold	A critical LOD score, which is larger than 0, (default LODThreshold=3), is used to select suggested variants.		
CutToBinary	A phenotypic threshold to cut the phenotype from categorical or continuous into binary one, e.g., CutToBinary=3 indicate that the individuals with phenotype ≤ 3 are set to 0 (control) while others are set to 1 (case). Default value is CutToBinary=NULL in scenarios of the input of binary traits.		
SubforGRM	A parameter indicates whether using a subset of variants to construct GRM (TRUE) or not (FALSE). Only used when "method=FastBiCmrMLM-Time" and "method=FastBiCmrMLM-RAM". Default value is SubforGRM=FALSE.		

SubNum	A parameter (>0) indicates the number of variants sampled[1] to calculate GRM. Only used when SubforGRM=TRUE. Default value is SubNum=1.0e+4.
GammaNum	A parameter (>0): the number of variants used only in "method=FastBiCmrMLM-Time" and "method=FastBiCmrMLM-RAM" to estimate Grammar ratio[2]. Default value is GammaNum=100.
ChunkNum	A parameter: the number of variants allocated to memory for each time in the step of single-variant scanning. Default value is ChunkNum=5e+4.
fastSPA	A parameter indicates whether using SPAtest[3] or not. Default value is fastSPA=FALSE (not using).
c	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.
NumHap	A parameter for setting the number of haplotypes for each bin genotype. Default value is NumHap=3.
genRegionFile	A file direction of function elements of genes. The first column is the name of function elements, the second column is the name of chromosome, the third and fourth columns are the left and right boundaries.

2.2 Parameter settings

The running codes for FastBiCmrMLM, FastBiCmrMLM-Time, and FastBiCmrMLM-RAM are as follows:

FastBiCmrMLM(fileGen="D:/Users/Genotype",filePhe="D:/Users/Phenotype.csv",fileGRM=NULL,filePS="D:/Users/PopStr.csv",PopStrType="PCA",fileOut="D:/Users/",method="FastBiCmrMLM-Time",trait=1,SearchRadius=1e+6,svpal=1e-5,LODThreshold=3,CutToBinary=NULL,SubforGRM=FALSE,SubNum=1e+4,GammaNum=100,ChunkNum=5e+4,fastSPA=FALSE)

The running code for FastBiCmrMLM-Hap algorithms:

FastBiCmrMLM(fileGen="D:/Users/Genotype",filePhe="D:/Users/Phenotype.csv",fileGRM=NULL,filePS="D:/Users/PopStr.csv",PopStrType="PCA",fileOut="D:/Users/",method="FastBiCmrMLM-Hap",trait=1,LODThreshold=3,CutToBinary=NULL,SubforGRM=FALSE,SubNum=1e+4,c=0.7,NumHap=3,genRegionFile=NULL)

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

Please note that the parameter *Xmx10g* specifies a maximum memory allocation of 10G for this step, and users can set it reasonably according to their own device specifications 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).

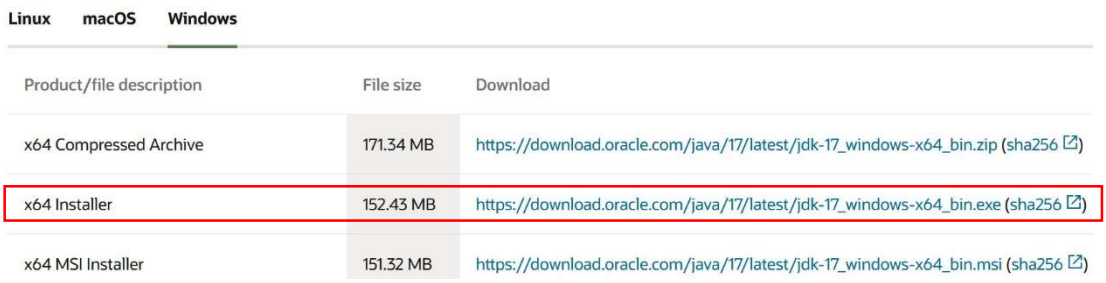
Java can be downloaded from https://www.oracle.com/java/technologies/downloads/#jdk17-windows

TASSEL can be downloaded from https://www.maizegenetics.net/tassel

PLINK can be downloaded from https://www.cog-genomics.org/plink/1.9/

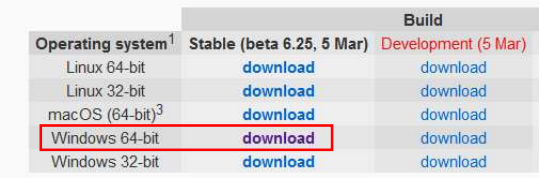
The downloaded files of Java, TASSEL, and PLINK are as follows.

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/

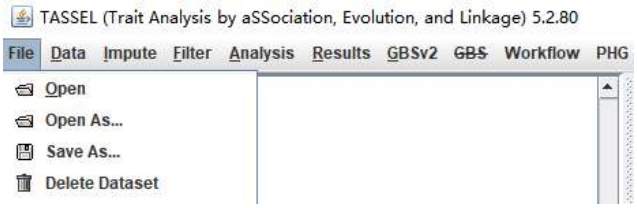
./run_pipeline -SortGenotypeFilePlugin -inputFile E:\FastBiCmrMLM\Genotype.hmp.txt
-outputFile E:\FastBiCmrMLM\Genotype.sort.hmp.txt -fileType Hapmap

./run_pipeline -fork1 -h E:\FastBiCmrMLM\Genotype.sort.hmp.txt -Xmx10g
-export E:\FastBiCmrMLM\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:\FastBiCmrMLM\Genotype.vcf --make-bed --out E:\FastBiCmrMLM\Genotype

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

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: phenotypic values of complex binary traits. The phenotypes missed: "NA".

Table 1. The format of phenotypic dataset

<Phenotype>	trait1	trait2	trait3	…
Ind46	0	1	0	…
Ind52	1	1	1	…
Ind57	1	1	NA	…
Ind64	0	0	0	…
Ind68	1	0	1	…
⋮	⋮	⋮	⋮	…

Format for genotypic dataset "fileGRM" (Table 2)

The "fileGRM" 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 GRM matrix is calculated by the “FastBiCmrMLM” software. When fileKin="D:/Users/GRM.csv", the GRM matrix with name GRM.csv is uploaded from the folder "D:/Users". If the number and order of individuals in the "GRM.csv" file do not match those in the phenotypic files, our software will attempt to match them.

Table 2. The format of GRM dataset

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	…
⋮	⋮	⋮	⋮	⋮	…

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
⋮	⋮

2.2.1 Result

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

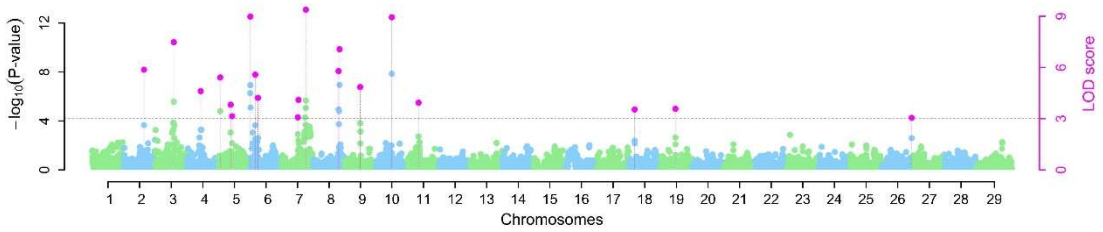
*_intermediate.csv: This file contains the results of genome-wide single-variants scanning in the first step. In this file, all the columns are named as “Marker” (variants name), “Chromosome”, “Position (bp)” (variants position (bp) on the genome), and “pvalue.Q” (the *P*-value for main-effect variants).

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
⋮	⋮	⋮	⋮

*_result.csv: The final results for significant and suggested variants. In this file, all the columns are named as “Marker” (marker name); “Chromosome”; “Position (bp)” (markers position (bp) on the genome); “effect type” (the type of effect for variants); “additive” (additive effect); “dominance” (dominance effect); variance (the variance of each variants); “r2(%)” (the proportion of total liability variance explained by each variants); “LOD” (LOD score); “*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); “tau” (the polygenic variance); and “inflation factor” (an indicator of goodness of the model used for association analysis[4]).

Marker	Chr	Position (bp)	effect type	additive	dominance	variance	r2(%)	LOD	P-value	significance	tau	inflation factor
null_16096	1	16097	ad	0.2968	1.4572	2.2651	13.095	63.616	2.43E-64	SIG	0.0113	1.1352
null_18665	1	18666	d	0	0.8884	0.7893	4.5628	45.8941	7.00E-48	SIG		
null_20132	1	20133	ad	0.4534	1.3063	2.1452	12.401	85.8537	1.41E-86	SIG		
null_24800	1	24801	a	1.157	0	1.3386	7.739	160.6026	7.42E-163	SIG		
null_30786	1	30787	d	0	1.1496	1.3216	7.6403	76.1055	3.36E-78	SIG		
null_30863	1	30864	d	0	0.3053	0.0932	0.5389	7.4661	4.53E-09	SIG		
null_41856	1	41857	ad	0.2594	0.4666	0.3322	1.9204	12.8412	1.44E-13	SIG		
null_41898	1	41899	a	0.2321	0	0.0539	0.3114	7.3386	6.13E-09	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 FastBiCmrMLM, 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 FastBiCmrMLM. Users can set different LOD threshold by setting LODThreshold (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.



3 Reference

1. Zhou, W. et al. Efficiently controlling for case-control imbalance and sample relatedness in large-scale genetic association studies. *Nat. Genet.* **50**, 1335–1341 (2018).
2. Svishcheva, G. R., Axenovich, T. I., Belonogova, N. M., van Duijn, C. M. & Aulchenko, Y. S. Rapid variance components-based method for whole-genome association analysis. *Nat. Genet.* **44**, 1166–1170 (2012).
3. Dey, R., Schmidt, E. M., Abecasis, G. R. & Lee, S. A fast and accurate algorithm to test for binary phenotypes and its application to PheWAS. *Am. J. Hum. Genet.* **101**, 37–49

(2017).
4. Tsepilov, Y. A. et al. Development and application of genomic control methods for genome-wide association studies using non-additive models. *PLoS One* **8**, e81431 (2013).
5. Wang Jing-Tian, Chang Xiao-Yu, Zhao Qiong, and Zhang Yuan-Ming. FastBiCmrMLM: a fast and powerful compressed variance component mixed logistic model for big genomic case-control genome-wide association study. *Brief. Bioinform.* Accepted (2024).