

MAGEE: Mixed Model Association Test for  
GEne-Environment Interaction  
Version 1.3.1

Xinyu Wang  
Human Genetics Center  
Dept. of Biostatistics and Data Science  
School of Public Health  
The University of Texas Health Science Center at Houston  
Email: Xinyu.Wang@uth.tmc.edu

Han Chen  
Human Genetics Center  
Dept. of Epidemiology, Human Genetics and Environmental Sciences  
School of Public Health  
Center for Precision Health  
School of Biomedical Informatics  
The University of Texas Health Science Center at Houston  
Email: Han.Chen.2@uth.tmc.edu

Duy Pham  
Human Genetics Center  
Dept. of Epidemiology, Human Genetics and Environmental Sciences  
School of Public Health  
The University of Texas Health Science Center at Houston  
Email: duy.t.pham@uth.tmc.edu

Kenneth Westerman  
Department of Medicine  
Clinical and Translational Epidemiology Unit  
Mongan Institute  
Massachusetts General Hospital  
Email: KEWESTERMAN@mgh.harvard.edu

Cong Pan  
Human Genetics Center  
Dept. of Epidemiology, Human Genetics and Environmental Sciences  
School of Public Health  
The University of Texas Health Science Center at Houston  
Email: cong.pan@uth.tmc.edu

October 12, 2023

# Contents

<b>1</b>	<b>Introduction</b>	<b>5</b>
<b>2</b>	<b>The model</b>	<b>5</b>
2.1	The full model . . . . .	5
2.2	GEI tests . . . . .	5
2.2.1	Interaction variance component test (IV) . . . . .	5
2.2.2	Interaction hybrid test using Fisher's method (IF) . . . . .	5
2.3	Joint tests . . . . .	6
2.3.1	Joint variance component test (JV) . . . . .	6
2.3.2	Joint hybrid test using Fisher's method (JF) . . . . .	6
2.3.3	Joint hybrid test using double Fisher's procedures (JD) . . . . .	6
<b>3</b>	<b>Getting started</b>	<b>6</b>
3.1	Downloading <i>MAGEE</i> . . . . .	6
3.2	Installing <i>MAGEE</i> . . . . .	6
<b>4</b>	<b>Input</b>	<b>7</b>
4.1	Object . . . . .	7
4.2	Genotypes . . . . .	7
4.3	Group definition file . . . . .	8
<b>5</b>	<b>Running <i>MAGEE</i></b>	<b>8</b>
5.1	Fitting GLMM . . . . .	8
5.2	Single variant tests . . . . .	8
5.2.1	Pooled analysis . . . . .	9
5.2.2	meta-analysis . . . . .	9
5.3	Variant set tests . . . . .	9
5.3.1	Pooled analysis . . . . .	9
5.3.2	meta-analysis . . . . .	11
<b>6</b>	<b>Output</b>	<b>12</b>
<b>7</b>	<b>Advanced options</b>	<b>14</b>
7.1	Missing genotypes . . . . .	14
7.2	Parallel computing . . . . .	15
7.3	Variant filters . . . . .	15
7.4	Internal minor allele frequency weights . . . . .	16
7.5	Allele flipping . . . . .	16
7.6	$P$ values of weighted sum of chi-squares . . . . .	16
7.7	Other options . . . . .	16
<b>8</b>	<b>Version</b>	<b>17</b>
8.1	Version 0.1.1 (February 25, 2020) . . . . .	17
8.2	Version 1.0.0 (May 1, 2021) . . . . .	17
8.3	Version 1.0.1 (November 13, 2021) . . . . .	17
8.4	Version 1.0.2 (January 27, 2022) . . . . .	17
8.5	Version 1.1.0 (March 24, 2022) . . . . .	17

8.6	Version 1.1.1 (April 12, 2022) . . . . .	17
8.7	Version 1.2.0 (June 2, 2022) . . . . .	17
8.8	Version 1.3.0 (April 18, 2023) . . . . .	18
8.9	Version 1.3.1 (October 12, 2023) . . . . .	18
<b>9</b>	<b>Contact</b>	<b>18</b>
<b>10</b>	<b>Acknowledgments</b>	<b>18</b>

# 1 Introduction

*MAGEE* is an R package for gene-environment interaction (GEI) tests and joint tests (testing the marginal genetic effects and GEI effects simultaneously) for genome-wide association studies (GWAS) and large-scale sequencing studies.<sup>1</sup> Based on the generalized linear mixed models (GLMMs),<sup>2</sup> the tests within the *MAGEE* framework are highly efficient.

For GWAS, *MAGEE* performs single-variant tests for GEI and joint effects. For rare variant analysis, *MAGEE* performs group tests based on user-defined variant sets. The group-based tests include two GEI tests and three joint tests: interaction variance component test (IV), interaction hybrid test using Fisher’s method (IF), joint variance component test (JV), joint hybrid test using Fisher’s method (JF), and joint hybrid test using double Fisher’s procedures (JD). Before running *MAGEE* for analyzing the data across the whole genome, a global null model that only accounts for covariates (not including any genetic main effects) is fitted. The model should be fitted using the R package GMMAT.<sup>3</sup>

## 2 The model

### 2.1 The full model

The full model of *MAGEE* is:

$$g(\mu_i) = \mathbf{X}_i\boldsymbol{\alpha} + \mathbf{G}_i\boldsymbol{\beta} + \mathbf{K}_i\boldsymbol{\gamma} + r_i,$$

where  $g(\cdot)$  is the link function of  $\mu_i$ , and  $\mu_i$  is the conditional mean of the phenotype for individual  $i$  given covariates  $\mathbf{X}_i$ , genotypes  $\mathbf{G}_i$  and a random intercept  $r_i$ .  $\mathbf{X}_i$  is a row vector of  $p$  covariates including an intercept,  $\mathbf{G}_i$  is a row vector of  $q$  variants, and  $\mathbf{K}_i$  is a row vector of  $m \times q$  pairwise GEI terms for  $m$  environmental factors (which are a subset of the  $p$  covariates in  $\mathbf{X}_i$ ) and  $q$  variants. Accordingly,  $\boldsymbol{\alpha}$  is a  $p \times 1$  vector for the covariate effects,  $\boldsymbol{\beta}$  is a  $q \times 1$  vector for the genetic main effects, and  $\boldsymbol{\gamma}$  is the  $mq \times 1$  vector for GEI effects. Assuming the sample size is  $N$ , the length  $N$  vector for the random intercept  $\mathbf{r} \sim N(0, \sum_{l=1}^L \lambda_l \boldsymbol{\Psi}_l)$ , where  $\lambda_l$  are the variance component parameters for  $L$  random effects, and  $\boldsymbol{\Psi}_l$  are  $N \times N$  known relatedness matrices.

### 2.2 GEI tests

#### 2.2.1 Interaction variance component test (IV)

IV test assumes  $\boldsymbol{\gamma} \sim N(0, \tau \mathbf{W}_K^2)$ , where  $\mathbf{W}_K$  is an  $mq \times mq$  predefined diagonal weight matrix for GEI. The weight matrix can be arbitrarily defined by the users, using either functional annotation scores<sup>4-6</sup> or a function of the minor allele frequency (MAF).<sup>7</sup> Testing for GEI effects  $H_0 : \boldsymbol{\gamma} = 0$  is then equivalent to testing the variance component parameter  $H_0 : \tau = 0$  versus  $H_1 : \tau > 0$ .

#### 2.2.2 Interaction hybrid test using Fisher’s method (IF)

IF test is a hybrid test that combines a burden-type test<sup>8</sup> and an adjusted variance component test,<sup>7</sup> which are asymptotically independent. When the true mean of interaction

effects  $\gamma$  is not close to 0, IF test is supposed to achieve superior power than the IV test. IF test assumes  $\gamma \sim N(\mathbf{W}_K \mathbf{1}_{mq} \gamma_0, \tau \mathbf{W}_K^2)$ , where  $\mathbf{1}_{mq}$  is a vector of 1's with length  $mq$ , and testing for GEI effects  $H_0 : \gamma = 0$  is equivalent to testing  $H_0 : \gamma_0 = \tau = 0$  versus  $H_1 : \gamma_0 \neq 0$  or  $\tau > 0$ .

## 2.3 Joint tests

### 2.3.1 Joint variance component test (JV)

JV test is a variance component joint analysis for genetic main effects and GEI effects simultaneously. JV test assumes  $\beta \sim N(0, \theta \mathbf{W}_G^2)$  and  $\gamma \sim N(0, \tau \mathbf{W}_K^2)$ , where  $\mathbf{W}_G$  is a  $q \times q$  predefined diagonal weight matrix for genetic effects. Testing for  $H_0 : \beta = \gamma = 0$  is equivalent to testing for  $H_0 : \theta = \tau = 0$  versus  $H_1 : \theta > 0$  or  $\tau > 0$ .

### 2.3.2 Joint hybrid test using Fisher's method (JF)

JF test combines burden and variance component test and jointly analyze the genetic main effects and GEI effects. JF test assumes  $\beta \sim N(\mathbf{W}_G \mathbf{1}_q \beta_0, \theta \mathbf{W}_G^2)$  and  $\gamma \sim N(\mathbf{W}_K \mathbf{1}_{mq} \gamma_0, \tau \mathbf{W}_K^2)$ , and test for  $H_0 : \beta_0 = \theta = \gamma_0 = \tau = 0$  versus  $H_1 : \beta_0 \neq 0$  or  $\theta > 0$  or  $\gamma_0 \neq 0$  or  $\tau > 0$ . The JF test statistic combines the  $P$  value for each parameter at once through Fisher's method,<sup>9</sup> which follows a Chi-square distribution with 8 degrees of freedom.

### 2.3.3 Joint hybrid test using double Fisher's procedures (JD)

JD test is also a hybrid joint analysis method for genetic main effects and GEI effects. JD test has the same assumption for  $\beta$  and  $\gamma$  as JF test, but it combines the  $P$  values for the 4 parameters following an alternative strategy. Instead of combining the 4  $P$  values at once, JD test combines the  $P$  value for genetic main effect (test for  $\beta_0 = \theta = 0$ ), and then combine this  $P$  value with the IF test  $P$  value (test for  $\gamma_0 = \tau = 0$ ) to get the joint test  $P$  value. All the combination procedures use Fisher's method. The JF test statistic follows a Chi-square distribution with 4 degrees of freedom.

Note: The main effect variance component test (MV) in *MAGEE* is the same as SKAT for related samples.<sup>10</sup> The main effect hybrid test using Fisher's method test (MF) in *MAGEE* is the same as the efficient hybrid test *SMMAT-E*<sup>11</sup> in the GMMAT package.

## 3 Getting started

### 3.1 Downloading *MAGEE*

*MAGEE* is an open source project and is freely available for download at <https://github.com/xwang21/MAGEE>. It can also be found as a regular R package and downloaded from CRAN (<https://CRAN.R-project.org/package=MAGEE>).

### 3.2 Installing *MAGEE*

*MAGEE* links to R packages Rcpp and RcppArmadillo, and also imports R packages Rcpp, CompQuadForm, foreach, parallel, Matrix, methods, GMMAT, data.table. *MAGEE* requires Bioconductor packages SeqArray and SeqVarTools to work with genotype files in

the GDS format. In addition, *GMMAT* requires *testthat* to run code checks during development, and *doMC* to run parallel computing in **glmm.gei** and **MAGEE** for genotype files in the GDS format (however, *doMC* is not available on Windows and these functions will switch to a single thread). These dependencies should be installed before installing *MAGEE*.

For optimal computational performance, it is recommended to use an R version configured with the Intel Math Kernel Library (or other fast BLAS/LAPACK libraries). See the instructions on building R with Intel MKL (<https://software.intel.com/en-us/articles/using-intel-mkl-with-r>).

Here is an example for installing *MAGEE* and all its dependencies in an R session (assuming none of the R packages other than the default has been installed):

```
> ## try http:// if https:// URLs are not supported
> ## remove "doMC" below if you are running Windows
> install.packages(c("devtools", "RcppArmadillo", "CompQuadForm", "doMC",
+                  "foreach", "Matrix", "GMMAT", "BiocManager", "testthat", "data.table"),
+                  repos = "https://cran.r-project.org/")
> BiocManager::install(c("SeqArray", "SeqVarTools"))
> devtools::install_github("https://github.com/large-scale-gxe-methods/MAGEE")
```

## 4 Input

*MAGEE* requires an object from fitting the null model using the **glmm.kin** function from the *GMMAT* package, and a genotype file in a GDS or BGEN format. For rare variant analysis, a user-defined group definition file is also required. Specified formats of these files are described as follows.

### 4.1 Object

*MAGEE* can perform analysis of gene by multiple environmental factors on multiple traits. To fit the null model, the phenotype and covariates (include the environmental factors of interest) should be saved in a data frame. If the samples are related, the relatedness should be known positive semidefinite matrices  $\mathbf{V}_k$  as an R matrix (in the case of a single matrix) or an R list (in the case of multiple matrices). Refer to the *GMMAT* user manual (<https://cran.r-project.org/web/packages/GMMAT/vignettes/GMMAT.pdf>) to learn the method of fitting the null model. The class of the object should be either "glmmkin" or "glmmkin.multi".

### 4.2 Genotypes

*MAGEE* can take genotype files either in the GDS format or in any version of the BGEN format. Genotypes in Variant Call Format (VCF) and PLINK binary PED format can be converted to the GDS format using *seqVCF2GDS* and *seqBED2GDS* functions from the *SeqArray* package:

```
> SeqArray::seqVCF2GDS("VCF_file_name", "GDS_file_name")
> SeqArray::seqBED2GDS("BED_file_name", "FAM_file_name", "BIM_file_name",
+                      "GDS_file_name")
```

### 4.3 Group definition file

For rare variant analysis, a user-defined group definition file with no header and 6 columns (variant set id, variant chromosome, variant position, variant reference allele, variant alternate allele, weight) is also required. For example, here we show the first 6 rows of the example group definition file "SetID.withweights.txt":

Set1	1	1	T	A	1
Set1	1	2	A	C	4
Set1	1	3	C	A	3
Set1	1	4	G	A	6
Set1	1	5	A	G	9
Set1	1	6	C	A	9

Note that each variant in the group definition file is matched by chromosome, position, reference allele and alternate allele with variants from the GDS file. One genetic variant can be included in different groups with possibly different weights. If no external weights are needed in the analysis, simply replace the 6th column by all 1's.

## 5 Running *MAGEE*

If *MAGEE* has been successfully installed, you can load it in an R session using

```
> library(MAGEE)
```

There are 2 functions in *MAGEE*: for single variant GEI and joint analysis, use **glmm.gei**; for rare variant set-based GEI and joint analysis, use **MAGEE**; Details about how to use these functions, their arguments and returned values can be found in the R help document of *MAGEE*. For example, to learn more about **MAGEE** in an R session you can type

```
> ?MAGEE
```

### 5.1 Fitting GLMM

Both **MAGEE** and **glmm.gei** requires a "glmmkin" or "glmmkin.multi" class object that contains a fitted GLMM null model. The object can be obtained from the **glmmkin** function from the R package GMMAT. For more examples and details about the **glmmkin** function, see the GMMAT manual (<https://cran.r-project.org/web/packages/GMMAT/vignettes/GMMAT.pdf>). Below is an example of fitting a GLMM using the **glmmkin** function from GMMAT:

```
> library(GMMAT)
> GRM.file <- system.file("extdata", "GRM.txt.bz2", package = "MAGEE")
> GRM <- as.matrix(read.table(GRM.file, check.names = FALSE))
> model0 <- glmmkin(disease ~ age + sex, data = pheno, kins = GRM,
+                   id = "id", family = binomial(link = "logit"))
```

### 5.2 Single variant tests

Here is a simple example of single variant score tests using **glmm.gei**:



### 5.2.1 Pooled analysis

```
> infile <- system.file("extdata", "geno.gds", package = "MAGEE")
> gds_outfile <- tempfile()
> glmm.gei(model0, interaction='sex', geno.file = infile,
+          outfile = gds_outfile)
```

The first argument in **glmm.gei** is the returned **glmmkin** class object from fitting the null model. The argument "interaction" can be either a character vector indicating one or multiple environmental factors, or a numerical vector indicating the column numbers for the environmental factors in the covariate matrix. The argument "geno.file" is the name (and path if not in the current working directory) of the genotype file, and the argument "outfile" is the name of the output file.

Alternatively, if your genotype information is saved as a BGEN file "geno.bgen" and includes a BGEN sample file "geno.sample", you can use:

```
> infile <- system.file("extdata", "geno.bgen", package = "MAGEE")
> gds_outfile <- tempfile()
> glmm.gei(model0, interaction='sex', geno.file = infile,
+          outfile = gds_outfile)
```

The function **glmm.gei** returns no value for GDS and BGEN genotype files.

### 5.2.2 meta-analysis

```
> infile1 <- system.file("extdata", "meta1.txt", package = "MAGEE")
> infile2 <- system.file("extdata", "meta2.txt", package = "MAGEE")
> infile3 <- system.file("extdata", "meta3.txt", package = "MAGEE")
> infile4 <- system.file("extdata", "meta4.txt", package = "MAGEE")
> infile5 <- system.file("extdata", "meta5.txt", package = "MAGEE")
> outfile <- tempfile()
> glmm.gei.meta(files = c(infile1, infile2, infile3, infile4, infile5),
+               interaction="sex", outfile = outfile)
```

In this example, the first argument in **glmm.gei.meta** is tab or space delimited plain text files (or compressed files that can be recognized by the R function `read.table`) with at least the following columns: SNPID, CHR, POS, Non\_Effect\_Allele, Effect\_Allele, N\_Sample, AF, Beta\_Marginal, SE\_Beta\_Marginal, P\_Value\_Marginal, Beta\_G, Beta\_G\_sex, SE\_Beta\_G, SE\_Beta\_G\_sex, Cov\_Beta\_G\_G-sex, P\_Value\_Interaction, P\_Value\_Joint. Generally, if each study performs score tests using genotypes in PLINK binary PED format or GDS format, the score test output from **glmm.score** can be directly used as input files. The argument "interaction" can be either a character vector indicating one or multiple environmental factors, or a numerical vector indicating the column numbers for the environmental factors in the covariate matrix. The argument "outfile" is the name of the output file.

## 5.3 Variant set tests

### 5.3.1 Pooled analysis

Variant set tests in a single study (or a pooled analysis of multiple studies) can be performed using the function **MAGEE**. In addition to an object returned from the function

**glmmkin**, a group definition file with no header and 6 columns (variant set id, variant chromosome, variant position, variant reference allele, variant alternate allele, weight) is also required, as described in **section 4.3**. An example of running **MAGEE**:

```
> geno.file <- system.file("extdata", "geno.gds", package = "MAGEE")
> group.file <- system.file("extdata", "SetID.withweights.txt",
+                             package = "MAGEE")
> MAGEE(model0, interaction='sex', geno.file, group.file,
+         group.file.sep = "\t", tests=c("JV", "JF", "JD"))
```

The first argument in **MAGEE** is the returned **glmmkin** class object from fitting the null model. The argument "interaction" can be either a character vector indicating one or multiple environmental factors, or a numerical vector indicating the column numbers for the environmental factors in the covariate matrix. The argument "geno.file" is the name (and path if not in the current working directory) of the genotype file, and the argument "group.file" is the name of the group definition file. The users can choose one or more test types as "IV", "IF", "JV", "JF", and "JD" in the "tests" argument. Note that the JV test also returns the  $P$  value from MV and IV tests, and the JF and JD tests also return the  $P$  value from MF and IF tests. Therefore, the above example gives the test results for all the seven tests. The **MAGEE** function returns a data.frame object for both GDS and BGEN genotype file inputs. Below are examples for the first 5 rows of the example output:

group	n.variants	miss.min	miss.mean	miss.max
Set1	20	0	0.000875	0.0175
Set2	20	0	0.000000	0.0000
Set3	20	0	0.000000	0.0000
Set4	20	0	0.000000	0.0000
Set5	20	0	0.000000	0.0000
...				
freq.min	freq.mean	freq.max	freq.strata.min	freq.strata.max
0.5000	0.8150402	0.99125	0.47	0.9950
0.6400	0.8795625	0.99125	0.63	0.9950
0.5675	0.8385000	0.98875	0.56	0.9950
0.5075	0.7450625	0.98375	0.50	0.9900
0.5050	0.7266250	0.98375	0.49	0.9900
...				
MV.pval	MF.pval	IV.pval	IF.pval	
0.1161530	0.1888730	0.2309887	0.2999593	
0.8984427	0.9611505	0.7955216	0.7048124	
0.4849650	0.5054350	0.6238591	0.2223911	
0.3678975	0.1128065	0.3670468	0.1513834	
0.1360848	0.3095582	0.6059774	0.7587549	
...				
JV.pval	JF.pval	JD.pval		
0.1239074	0.20058700	0.2192965		
0.9547726	0.94709001	0.9412548		
0.6642510	0.34011753	0.3580810		
0.4054061	0.07679027	0.0865809		
0.2882450	0.57320809	0.5751443		

The first column contains the group name (group) followed by the number of variants in the group in the second column (n.variants). The results are included in the next 15 columns: the minimum, mean, and maximum average missing genotype rate for all variants in the group (miss.min/miss.mean/miss.max), the minimum, mean, and maximum allele frequency for all variants in the group (freq.min/freq.mean/freq.max), the minimum and maximum allele frequency for all variants in the group after stratification (freq.strata.min/freq.strata.max), and  $P$  values for the MV test (MV.pval), MF test (MF.pval), IV test (IV.pval), IF test (IF.pval), JV test (JV.pval), JF test (JF.pval), and JD test (JD.pval).

### 5.3.2 meta-analysis

```
> geno.file <- system.file("extdata", "geno.gds", package = "MAGEE")
> group.file <- system.file("extdata", "SetID.withweights.txt",
+                             package = "MAGEE")
> meta.files.prefix <- tempfile()
> MAGEE.meta(meta.files.prefix = meta.files.prefix,
+             group.file=group.file,
+             tests=c("JV", "JF", "JD"))
```

The first argument in **MAGEE.meta** is a vector of intermediate files' prefix with length equal to the number of studies. The argument "group.file" is the name of the group definition file. The users can choose one or more test types as "IV", "IF", "JV", "JF", and "JD" in the "tests" argument. The **MAGEE.meta** function returns a data.frame object for both GDS and BGEN genotype file inputs. Below are examples for the first 5 rows of the example output:

group	n.variants				
Set1	20				
Set2	20				
Set3	20				
Set4	20				
Set5	20				
...					
MV.pval	MF.pval	IV.pval	IF.pval		
0.1161530	0.1888730	0.2309887	0.2999593		
0.8984427	0.9611505	0.7955216	0.7048124		
0.4849650	0.5054350	0.6238591	0.2223911		
0.3678975	0.1128065	0.3670468	0.1513834		
0.1360848	0.3095582	0.6059774	0.7587549		
...					
JV.pval	JF.pval	JD.pval			
0.1239074	0.20058700	0.2192965			
0.9547726	0.94709001	0.9412548			
0.6642510	0.34011753	0.3580810			
0.4054061	0.07679027	0.0865809			
0.2882450	0.57320809	0.5751443			

The first column contains the group name (group) followed by the number of variants in the group in the second column (n.variants). The results are included in the next 7

columns:  $P$  values for the MV test (MV.pval), MF test (MF.pval), IV test (IV.pval), IF test (IF.pval), JV test (JV.pval), JF test (JF.pval), and JD test (JD.pval).

## 6 Output

The single variant test function **glmm.gei** generates a tab-delimited plain text output file. Here we show the header and the first five rows of the example output for each genotype file input.

If you use a GDS genotype file "geno.gds", here are the header and the first 5 rows of the example output "outfile.txt" using the default settings from **glmm.gei**:

SNPID	CHR	POS	Non_Effect_Allele	Effect_Allele	N_Sample	AF	N_sex_0
SNP1	1	1	T	A	393	0.9745547	197
SNP2	1	2	A	C	400	0.5000000	200
SNP3	1	3	C	A	400	0.7925000	200
SNP4	1	4	G	A	400	0.7012500	200
SNP5	1	5	A	G	400	0.5937500	200
...							
AF_sex_0	N_sex_1	AF_sex_1	Beta_Marginal	SE_Beta_Marginal	Beta_G-sex		
0.9720812	196	0.9770408	-0.43565484	0.4684802	0.5017660		
0.4700000	200	0.5300000	0.07576315	0.1469115	0.1162287		
0.7825000	200	0.8025000	0.01743008	0.1807686	0.4599819		
0.6850000	200	0.7175000	0.07688790	0.1571101	0.3479766		
0.6150000	200	0.5725000	-0.09464890	0.1537993	-0.2899459		
...							
SE_Beta_G-sex	P_Value_Marginal	P_Value_Interaction	P_Value_Joint				
0.9287819	0.3524062	0.5890309	0.5608414				
0.2913865	0.6060598	0.6899805	0.8085364				
0.3563337	0.9231854	0.1967474	0.4326500				
0.3081733	0.6245666	0.2588309	0.4689541				
0.3032559	0.5382872	0.3390169	0.5239105				

The first 5 columns are extracted from the GDS file: SNP ("annotation/id"), CHR ("chromosome"), POS ("position"), reference and alternate alleles ("allele"). Results are included in 13 columns for the ALT allele: the sample size N\_Sample (with non-missing genotypes), the allele frequency (AF), the number of non-missing samples (N\_sex\_0, and N\_sex\_1) and allele frequency of the effect allele (AF\_sex\_0 and AF\_sex\_1) for each combination of strata for all of the categorical exposure or interaction covariate, the coefficient estimate for the marginal genetic effect (Beta\_Marginal), the standard error (SE) of the marginal genetic effect (SE\_Beta\_Marginal), the coefficient estimate for the interaction term sex (Beta\_G-sex), the model-based SE associated with any GxE term (SE\_Beta\_G-sex), the marginal effect score test  $P$  value P\_Value\_Marginal, the gene-environment interaction test  $P$  value P\_Value\_Interaction, and the joint test  $P$  value P\_Value\_Joint.

If you use a BGEN genotype file "geno.bgen", here are the header and the first 5 rows of the example output "outfile.txt" using the default settings from **glmm.gei**:

SNPID	RSID	CHR	POS	Non_Effect_Allele	Effect_Allele	N_Sample	AF
SNP1	SNP1	1	1	T	A	393	0.974555
SNP2	SNP2	1	2	A	C	400	0.500000

SNP3	SNP3	1	3	C	A	400	0.792500
SNP4	SNP4	1	4	G	A	400	0.701250
SNP5	SNP5	1	5	A	G	400	0.593750
...							
N_sex_0	AF_sex_0	N_sex_1	AF_sex_1	Beta_Marginal	SE_Beta_Marginal		
197	0.972081	196	0.977041	-0.4356550	0.468480		
200	0.470000	200	0.530000	0.0757631	0.146912		
200	0.782500	200	0.802500	0.0174301	0.180769		
200	0.685000	200	0.717500	0.0768879	0.157110		
200	0.615000	200	0.572500	-0.0946489	0.153799		
...							
Beta_G-sex	SE_Beta_G-sex	P_Value_Marginal	P_Value_Interaction	P_Value_Joint			
0.501766	0.928782	0.352406	0.589031	0.560841			
0.116229	0.291386	0.606060	0.689980	0.808536			
0.459982	0.356334	0.923185	0.196747	0.432650			
0.347977	0.308173	0.624567	0.258831	0.468954			
-0.289946	0.303256	0.538287	0.339017	0.523911			

The first 6 columns are copied from the BGEN file: the SNP, RSID, chromosome CHR, physical position POS, and reference and alternate alleles ("allele"). Results are included in 13 columns for the second allele in the BGEN file: the sample size N\_Sample (with non-missing genotypes), the allele frequency (AF), the number of non-missing samples (N\_sex\_0, N\_sex\_1) and allele frequency of the effect allele (AF\_sex\_0, AF\_sex\_1) for each combination of strata for all of the categorical exposure or interaction covariate, the coefficient estimate for the marginal genetic effect (Beta\_Marginal), the SE of the marginal genetic effect (SE\_Beta\_Marginal), the coefficient estimate for the interaction term sex (Beta\_G-sex), the model-based SE associated with GxE term (SE\_Beta\_G-sex), the marginal effect score test *P* value P\_Value\_Marginal, the gene-environment interaction test *P* value P\_Value\_Interaction, and the joint test *P* value P\_Value\_Joint.

For both GDS and BGEN file formats, if the argument `meta.output = TRUE`, **glmm.gei** will output additional columns containing the coefficients and variance-covariance of the interaction terms.

The meta-analysis function **glmm.gei.meta** generates a tab-delimited plain text output file. Here are the header and the first 5 rows of the example output from the meta-analysis:

SNPID	CHR	POS	Non_Effect_Allele	Effect_Allele	N_Samples	AF
1:6:T:G	1	6	T	G	10000	0.6681618
1:9:A:A	1	9	A	A	10000	0.8820916
1:4:A:A	1	4	A	A	10000	0.8959009
1:3:A:G	1	3	A	G	20000	0.4858367
1:7:T:G	1	7	T	G	20000	0.3754258
...						
Beta_Marginal	SE_Beta_Marginal	P_Value_Marginal	Beta_G	Beta_G_sex		
0.32866874	0.9899530	7.398859e-01	0.30315698	0.506471492		
0.43239601	0.1320372	1.057351e-03	1.04968727	1.153496562		
0.56527936	0.7841878	4.710037e-01	0.29184287	1.083371475		
0.05649831	0.5634919	9.201342e-01	0.48271148	-0.073251696		
0.14983613	0.3414406	6.607810e-01	-0.17989467	-0.230739884		
...						

SE_Beta_G	SE_Beta_G_sex	Cov_Beta_G_G_sex	P_Value_Interaction	P_Value_Joint
0.1459865	0.9128646	0.2895442852	0.5790208	0.949407885
0.1495806	0.7204614	0.2154526653	0.1093653	1.000000000
1.1220222	1.2256722	0.1288226309	0.3767503	0.665978825
0.4875232	0.9367640	0.1250295669	0.8823533	0.573433101
0.5301991	0.8262347	0.0973737560	0.7874718	0.923653779
...				

The first 3 columns are set by the function **glmm.gei.meta** to denote SNP name and alleles. The rest of the columns are: reference and alternate alleles ("allele"), the sample size N\_Sample (with non-missing genotypes), the allele frequency (AF), the coefficient estimate for the marginal genetic effect (Beta\_Marginal), the SE of the marginal genetic effect (SE\_Beta\_Marginal), the coefficient estimate for the interaction term sex (Beta\_G-sex), the model-based SE associated with GxE term (SE\_Beta\_G-sex), the marginal effect score test *P* value P\_Value\_Marginal, the gene-environment interaction test *P* value P\_Value\_Interaction, and the joint test *P* value P\_Value\_Joint.

In variant set tests **MAGEE**, if "meta.file.prefix" is specified, space-delimited intermediate files for single variant scores and binary intermediate files for covariance matrices will be generated. Here are the header and the first 5 rows of the example intermediate file "MAGEE.meta.score.1":

group	chr	pos	ref	alt	N	missrate	altfreq
Set1	1	1	T	A	393	0.0175	0.974554707379135
Set1	1	2	A	C	400	0	0.5
Set1	1	3	C	A	400	0	0.7925
Set1	1	4	G	A	400	0	0.70125
Set1	1	5	A	G	400	0	0.59375
...							
G.SCOR							K.SCOR.1
-1.98499773963038							0.758459905129889
3.51031642023436							1.41604692383531
0.533400376147224							3.62151052066005
3.11494101140768							3.61126108351086
-4.00135050078827							-3.03952592152864
...							

The first 5 columns are copied from the group definition file, indicating the variant set (group) id, variant chromosome, variant position, variant reference allele, variant alternate allele, respectively. Results are included in 6 columns: the sample size N (with non-missing genotypes), the genotype missing rate missrate, the alt allele frequency altfreq, the score statistic SCORE of alt allele, the score of the first environmental factor.

## 7 Advanced options

### 7.1 Missing genotypes

It is recommended to perform genotype quality control prior to analysis to impute missing genotypes or filter out SNPs with high missing rates. However, *MAGEE* does allow

missing genotypes, and imputes to the mean value by default (`missing.method = "impute2mean"`) in both **glmm.gei** and **MAGEE**. Alternatively, in **glmm.gei** missing genotypes can be omitted from the analysis using

```
missing.method = "omit"
```

In variant set tests using **MAGEE**, instead of imputing missing genotypes to the mean value, you can impute missing genotypes to 0 (homozygous reference allele) using

```
missing.method = "impute2zero"
```

## 7.2 Parallel computing

Parallel computing can be enabled in **glmm.gei** and **MAGEE** using the argument `"ncores"` to specify how many cores you would like to use on a computing node. By default `"ncores"` is 1, meaning that these functions will run in a single thread.

For **glmm.gei**, if you enable parallel computing, multiple temporary files will be placed in the directory. For example, if your `"ncores = 12"` and you specify `"glmm.gei.gds.testoutfile.txt"` as your output file name, then 12 files `"glmm.gei.gds.testoutfile.txt_tmp.1"`, `"glmm.gei.gds.testoutfile.txt_tmp.2"`, ..., `"glmm.gei.gds.testoutfile.txt_tmp.12"` will be generated from each thread to store the results. The results from each temporary file will then be combined into a single file with the output file name `"glmm.gei.gds.testoutfile.txt"` as the file name when all threads have completed.

If your R is configured with Intel MKL and you would like to enable parallel computing, it is recommended that you set the environmental variable `"MKL_NUM_THREADS = 1"` before running R to avoid hanging. Alternatively, you can do this at the beginning of your R script by using

```
> Sys.setenv(MKL_NUM_THREADS = 1)
```

For Mac OS users using R configured with OpenBLAS, the R package `RhpcBLASctl` may help set the number of threads used by OpenBLAS to 1. The following lines of code can be used at the beginning of your R script:

```
> #install.packages("RhpcBLASctl")
> library(RhpcBLASctl)
> blas_set_num_threads(1)
```

## 7.3 Variant filters

Variants can be filtered in **glmm.gei** and **MAGEE** based on minor allele frequency (MAF) and missing rate filters. The argument `"MAF.range"` specifies the minimum and maximum MAFs for a variant to be included in the analysis. By default the minimum MAF is  $1 \times 10^{-7}$  and the maximum MAF is 0.5, meaning that only monomorphic markers in the sample will be excluded (if your sample size is no more than 5 million). The argument `"miss.cutoff"` specifies the maximum missing rate for a variant to be included in the analysis. By default it is set to 1, meaning that no variants will be removed due to high genotype missing rates.

## 7.4 Internal minor allele frequency weights

Internal weights are calculated based on the minor allele frequency (NOT the effect allele frequency, therefore, variants with effect allele frequencies 0.01 and 0.99 have the same weights) as a beta probability density function. Internal weights are multiplied by the external weights given in the last column of the group definition file. To turn off internal weights, use

```
MAF.weights.beta = c(1, 1)
```

to assign flat weights, as a beta distribution with parameters 1 and 1 is a uniform distribution on the interval between 0 and 1.

## 7.5 Allele flipping

In variant set tests **MAGEE**, by default the alt allele is used as the coding allele and variants in each variant set are matched strictly on chromosome, position, reference and alternate alleles.

The argument "auto.flip" allows automatic allele flipping if a specified variant is not found in the genotype file, but a variant at the same chromosome and position with reference allele matching the alternate allele in the group definition file "group.file", and alternate allele matching the reference allele in the group definition file "group.file", to be included in the analysis. Please use with caution for whole genome sequence data, as both ref/alt and alt/ref variants at the same position are not uncommon, and they are likely two different variants, rather than allele flipping.

The argument "use.minor.allele" allows using the minor allele instead of the alt allele as the coding allele in variant set tests.

## 7.6 *P* values of weighted sum of chi-squares

In variant set tests **MAGEE**, you can use 3 methods in the "method" argument to compute *P* values of weighted sum of chi-square distributions: "davies",<sup>12</sup> "kuonen"<sup>13</sup> and "liu".<sup>14</sup> By default "davies" is used, if it returns an error message in the calculation, or a *P* value greater than 1, or less than  $1 \times 10^{-5}$ , "kuonen" method will be used. If "kuonen" method fails to compute the *P* value, "liu" method will be used.

## 7.7 Other options

By default, genotypes are centered to the mean before the analysis in single variant tests **glmm.gei**. You can turn this feature off by specifying

```
center = FALSE
```

to use raw genotypes.

In **glmm.gei**, by default 100 SNPs are tested in a batch. You can change it using the "nperbatch" argument, but the computational time can increase substantially if it is either too small or too large, depending on the performance of your computing system.

In the variant set tests **MAGEE**, by default the group definition file "group.file" should be tab delimited, but you can change it using the "group.file.sep" argument.

There is a "Garbage.Collection" argument (default FALSE), if turned on, **MAGEE** will call the function **gc** for each variant set tested. It helps save memory footprint, but the computation speed might be slower.



## 8 Version

### 8.1 Version 0.1.1 (February 25, 2020)

Initial public release of *MAGEE*.

### 8.2 Version 1.0.0 (May 1, 2021)

1. Support BGEN file format in both **glmm.gei** and **MAGEE** functions.
2. Allow adjustment for interaction covariates in both **glmm.gei** and **MAGEE** functions.
3. Include a `meta.output` argument for **glmm.gei** to output additional summary statistics for the interaction terms.

### 8.3 Version 1.0.1 (November 13, 2021)

1. Supported multiple phenotype analysis in **MAGEE**.
2. Supported longitudinal data analysis in **glmm.gei** and **MAGEE**.
3. Updated automatic tests for **glmm.gei** and **MAGEE**.

### 8.4 Version 1.0.2 (January 27, 2022)

1. Fixed a `dgesdd` bug from **MASS::ginv** in **MAGEE**.
2. Fixed a minor bug on the interaction term in **MAGEE.prep** and **MAGEE.lowmem**.

### 8.5 Version 1.1.0 (March 24, 2022)

1. Edited the names of output headers in **glmm.gei**.
2. Added new output headers in **glmm.gei**.
3. Fixed bugs on longitudinal data analysis in **glmm.gei**.
4. Fixed bugs on interaction covariates in **glmm.gei**.
5. Updated automatic tests for **glmm.gei**.

### 8.6 Version 1.1.1 (April 12, 2022)

1. Fixed bugs on inverse of singular matrix **glmm.gei**.

### 8.7 Version 1.2.0 (June 2, 2022)

1. Added meta-analysis functions **glmm.gei.meta** and **MAGEE.meta**.

## 8.8 Version 1.3.0 (April 18, 2023)

1. Check for system copies of `zstd` and `libdeflate` libraries.
2. Fixed a minor bug in reading the ID column for bgen sample files in **MAGEE** and **glmm.gei**.
3. Fixed a minor bug in calling the internal function **fix.dgesdd** in **glmm.gei**.
4. Replaced **read.table** by the more efficient function **data.table::fread**.
5. Implemented a new argument "AF.strata.range" to filter variants based on their environmental exposure stratum-specific coding allele frequencies in **MAGEE**.
6. Removed **context** in `testthat` tests.
7. Replaced **print** and **cat** in the code by **message** and **warning**.
8. Minor changes in the man directory, per CRAN policy.

## 8.9 Version 1.3.1 (October 12, 2023)

1. Bioconductor packages **SeqArray** and **SeqVarTools** moved to Suggests.

## 9 Contact

Please refer to the R help document of *MAGEE* for specific questions about each function. For comments, suggestions, bug reports and questions, please contact Han Chen (Han.Chen.2@uth.tmc.edu). For bug reports, please include an example to reproduce the problem without having to access your confidential data.

## 10 Acknowledgments

This work was supported by National Institutes of Health (NIH) grants R00 HL130593 and R01 HL145025.

## References

- [1] Wang, X., Lim, E., Liu, C, Sung, Y. J., Rao, D. C., Morrison, A. C., Boerwinkle, E., Manning, A. K., and Chen, H. Efficient gene-environment interaction tests for large biobank-scale sequencing studies. *Genetic Epidemiology* **44**, **8**, 908–923 (2020) .
- [2] Breslow, N. E. and Clayton, D. G. Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association* **88**, 9–25 (1993).
- [3] Chen, H., Wang, C., Conomos, M. P., Stilp, A. M., Li, Z., Sofer, T., Szpiro, A. A., Chen, W., Brehm, J. M., Celedón, J. C., Redline, S., Papanicolaou, G. J., Thornton, T. A., Laurie, C. C., Rice, K. and Lin, X. Control for Population Structure and Relatedness for Binary Traits in Genetic Association Studies via Logistic Mixed Models. *The American Journal of Human Genetics* **98**, 653–666 (2016).

- [4] Kircher, M., Witten, D. M., Jain, P., O’Roak, B., Cooper, G. M., and Shendure, J. A general framework for estimating the relative pathogenicity of human genetic variants. *Nature Genetics* **46(3)**, 310-315 (2014).
- [5] Rentzsch, P., Witten, D., Cooper, G. M., Shendure, J., and Kircher, M. CADD: Predicting the deleteriousness of variants throughout the human genome. *Nucleic Acids Research* **47**, D886-D894 (2019).
- [6] Rogers, M. F., Shihab, H. A., Mort, M., Cooper, D. N., Gaunt, T. R., and Campbell, C. FATHMM-XF: Accurate prediction of pathogenic point mutations via extended features. *Computer Applications in the Biosciences; Bioinformatics* **34(3)**, 511-513 (2018).
- [7] Wu, M. C., Lee, S., Cai, T., Li, Y., Boehnke, M. and Lin, X. Rare-variant association testing for sequencing data with the sequence kernel association test. *The American Journal of Human Genetics* **89**, 82–93 (2011).
- [8] Li, B. and Leal, S. M. Methods for detecting associations with rare variants for common diseases: Application to analysis of sequence data. *The American Journal of Human Genetics* **83**, 311–321 (2008).
- [9] Fisher, R. A. Statistical methods for research workers. *Journal of Comparative Pathology and Therapeutics* **41**, 261-262 (1928).
- [10] Chen, H., Meigs, J. B., and Dupuis, J. Sequence kernel association test for quantitative traits in family samples. *Genetic Epidemiology* **37(2)**, 196 (2013).
- [11] Chen, H., Huffman, J. E., Brody, J. A., Wang, C., Lee, S., Li, Z., Gogarten, S. M., Sofer, T., Bielak, L. F., Bis, J. C., *et al.* Efficient variant set mixed model association tests for continuous and binary traits in large-scale whole-genome sequencing studies. *The American Journal of Human Genetics* **104**, 260–274 (2019).
- [12] Davies, R. B. Algorithm AS 155: The Distribution of a Linear Combination of  $\chi^2$  Random Variables. *Journal of the Royal Statistical Society. Series C (Applied Statistics)* **29**, 323–333 (1980).
- [13] Kuonen, D. Saddlepoint Approximations for Distributions of Quadratic Forms in Normal Variables. *Biometrika* **86**, 929–935 (1999).
- [14] Liu, H., Tang, Y. and Zhang, H. H. A new chi-square approximation to the distribution of non-negative definite quadratic forms in non-central normal variables. *Computational Statistics & Data Analysis* **53**, 853–856 (2009).