

# Package ‘GWAF’

January 2, 2012

**Type** Package

**Title** Genome-Wide Association analyses with Family data

**Version** 1.2

**Date** 2009-11-25

**Author** Ming-Huei Chen <mhchen@bu.edu> and Qiong Yang <qyang@bu.edu>

**Maintainer** Ming-Huei Chen <mhchen@bu.edu>

**Depends** gee, kinship

**Description** Functions to test genetic associations between SNPs and a continuous/dichotomous trait using family data, and to make genome-wide p-value plot and QQ plot.

**License** GPL (>= 2)

**Repository** CRAN

**Date/Publication** 2009-11-26 08:12:25

## R topics documented:

GWAF-package . . . . .	2
auto . . . . .	2
gee.lgst . . . . .	4
gee.lgst.batch . . . . .	5
gee.lgst.batch.imputed . . . . .	7
gee.lgst.imputed . . . . .	9
GWplot . . . . .	10
lme.batch . . . . .	11
lme.batch.imputed . . . . .	13
lmekin . . . . .	14
qq . . . . .	15

<b>Index</b>	<b>16</b>
--------------	-----------

---

 GWAF-package

*Genome-Wide Association analyses with Family data*


---

### Description

GWAF package provides functions to fit Linear Mixed Effects (LME) model accounting for within pedigree correlation for testing genetic association between genotyped/imputed SNPs and a continuous trait and functions to fit Generalized Estimation Equation (GEE) model using each pedigree as a cluster with independent working correlation matrix for testing genetic association between SNPs and a dichotomous trait. In addition, GWAF package also provides functions for making genome-wide p-values plot and QQ plot containing genomic control parameter estimate and generating scripts for genome-wide association analysis.

### Details

Package: GWAF  
 Type: Package  
 Version: 1.2  
 Date: 2009-11-25  
 License: GPL (>= 2)

### Author(s)

Qiong Yang <qyang@bu.edu> and Ming-Huei Chen <mhchen@bu.edu>  
 Maintainer: Ming-Huei Chen <mhchen@bu.edu>

---

 auto

*function to generate scripts for genome-wide association analysis using GWAF*


---

### Description

Given a phenotype file, a pedigree file, a phenotype of interest, covariates, analysis of interest (can be 'lme', 'lme.imputed', 'gee' or 'gee.imputed'), a path or directory (genopath) that keeps genotype files, and other arguments, auto function generates one R script and one shell script that executes the R script for each genotype file, and one list file that executes all shell scripts in batch mode.

### Usage

```
auto(genopath, phenfile, pedfile, outfile, phen, covars = NULL, analysis, lib.loc, model = NULL,
kinmat = NULL, col.names = F, sep.ped = ",", sep.phe = ",", sep.gen = ",")
```

**Arguments**

genopath	a character string indicating the path or directory that keeps genotype files to be analyzed
phenfile	a character string naming the phenotype file for reading (see format requirement in details)
pedfile	a character string naming the pedigree file for reading (see format requirement in details)
outfile	a character string naming the result file for writing
phen	a character string for a phenotype name in phenfile
covars	a character vector for covariates in phenfile
analysis	a character string indicating the analysis of interest available in GWAF package, can be 'lme', 'lme.imputed', 'gee' or 'gee.imputed'
lib.loc	a character string indicating the location of GWAF package
model	a single character of 'a','d','g', or 'r', with 'a'=additive, 'd'=dominant, 'g'=general and 'r'=recessive models; Not appropriate/needed for analyzing imputed SNPs
kinmat	a character string naming the file where kinship coefficient matrix is kept; needed for LME analyses
col.names	a logical value indicating whether the output file should contain column names
sep.ped	the field separator character for pedigree file
sep.phen	the field separator character for phenotype file
sep.gen	the field separator character for genotype file

**Details**

auto function generates one R script and one shell script that executes the R script for each genotype file, and one list file that can execute all shell scripts in batch mode to analyze all genotype data in genopath. These scripts are named based on the phenotype of interest, the analysis of interest and the time these scripts are generated. After creating these scripts, auto function gives a message telling the user how to submit ALL the jobs (using ksh XXXX.lst). When a submitted job is completed, a log file indicating which genotype file was analyzed will be generated and the R script and the shell script will be removed. The number of log files should equal to the number of genotype files in genopath, if all jobs are completed. All the results will be written and appended to the user specified single output file. Different outfile should be assigned for different genopath to avoid over-writing.

**Value**

No value is returned. Instead, results are written to outfile.

**Author(s)**

Ming-Huei Chen <mhchen@bu.edu> and Qiong Yang <qyang@bu.edu>

---

gee.lgst                      *function for testing association between a dichotomous trait and a genotyped SNP in family data using GEE*

---

### Description

Fit logistic regression via GEE to test association between a dichotomous phenotype and one genotyped SNP in a genotype file with user specified genetic model. Each family is treated as a cluster, with independence working correlation matrix used in the robust variance estimator. This function is called in gee.lgst.batch function to apply association test to all SNPs in the genotype data.

### Usage

```
gee.lgst(snp, phen, test.dat, covar=NULL, model="a")
```

### Arguments

snp	genotype data of a SNP
phen	a character string for a phenotype name in phenfile
test.dat	the product of merging phenotype, genotype and pedigree data, should be ordered by "famid"
covar	a character vector for covariates in phenfile
model	a single character of 'a','d','g', or 'r', with 'a'=additive, 'd'=dominant, 'g'=general and 'r'=recessive models

### Details

The gee.lgst function tests association between a dichotomous trait and a SNP from a dataset that contains phenotype, genotype and pedigree data, where the dataset needs to be ordered by famid.

### Value

Please see output in gee.lgst.batch.

### Author(s)

Qiong Yang <qyang@bu.edu> and Ming-Huei Chen <mhchen@bu.edu>

### References

Liang, K.Y. and Zeger, S.L. (1986) Longitudinal data analysis using generalized linear models. *Biometrika*, **73** 13–22.

Zeger, S.L. and Liang, K.Y. (1986) Longitudinal data analysis for discrete and continuous outcomes. *Biometrics*, **42** 121–130.

Vincent J Carey. Ported to R by Thomas Lumley (versions 3.13 and 4.4) and Brian Ripley (version 4.13). gee: Generalized Estimation Equation solver. [4.13]. 2007. Ref Type: Computer Program, <http://cran.r-project.org/>

**See Also**

gee() function from package gee

---

gee.lgst.batch	<i>function to test genetic association between a dichotomous trait and a batch of genotyped SNPs in families using GEE</i>
----------------	---

---

**Description**

Fit logistic regression via GEE to test association between a dichotomous phenotype and all genotyped SNPs in a genotype file with user specified genetic model. Each pedigree is treated as a cluster, with independence working correlation matrix used in the robust variance estimator. This function applies the same trait-SNP association test to all SNPs in the genotype data. The trait-SNP association test is carried out by gee.lgst function where the the gee() function from package gee is used.

**Usage**

```
gee.lgst.batch(genfile, phenfile, pedfile, outfile, phen, covars=NULL,
model="a", col.names=T, sep.ped=",", sep.phe=",", sep.gen=",")
```

**Arguments**

genfile	a character string naming the genotype file for reading (see format requirement in details)
phenfile	a character string naming the phenotype file for reading (see format requirement in details)
pedfile	a character string naming the pedigree file for reading (see format requirement in details)
outfile	a character string naming the result file for writing
phen	a character string for a phenotype name in phenfile
covars	a character vector for covariates in phenfile
model	a single character of 'a','d','g', or 'r', with 'a'=additive, 'd'=dominant, 'g'=general and 'r'=recessive models
col.names	a logical value indicating whether the output file should contain column names
sep.ped	the field separator character for pedigree file
sep.phe	the field separator character for phenotype file
sep.gen	the field separator character for genotype file

## Details

The `gee.lgst.batch` function first reads in and merges phenotype-covariates, genotype and pedigree files, then tests the association of phen against all SNPs in `genfile`. `genfile` contains unique individual id and genotype data, with the column names being "id" and SNP names. For each genotyped SNP, the genotype data should be coded as 0, 1, 2 indicating the numbers of coded alleles. The SNP names in genotype file should not have any dash, '-' and other special characters (dots and underscores are OK). `phenfile` contains unique individual id, phenotype and covariates data, with the column names being "id" and phenotype and covariate names. `pedfile` contains pedigree information, with the column names being "famid", "id", "fa", "mo", "sex". In all files, missing value should be an empty space, except missing parental id in `pedfile`. Only phenotypes with two categories are analyzed. A phenotype should be coded as 0 and 1, with 1 denoting affected and 0 unaffected. SNPs with low genotype counts (especially minor allele homozygote) may be omitted or analyzed with dominant model or analyzed with logistic regression. The `gee.lgst.batch` function fits Generalized Estimation Equation (GEE) model using each pedigree as a cluster with `gee.lgst` function from GWAF package and 'gee' function from gee package.

## Value

No value is returned. Instead, results are written to `outfile`. When the genetic model is 'a', 'd' or 'r', the result includes the following columns. When the genetic model is 'g', beta and se are replaced with `beta10`, `beta20`, `beta21`, `se10`, `se20`, `se21`.

<code>phen</code>	phenotype name
<code>snp</code>	SNP name
<code>n0</code>	the number of individuals with 0 copy of minor alleles
<code>n1</code>	the number of individuals with 1 copy of minor alleles
<code>n2</code>	the number of individuals with 2 copies of minor alleles
<code>nd0</code>	the number of individuals with 0 copy of minor alleles in affected sample
<code>nd1</code>	the number of individuals with 1 copy of minor alleles in affected sample
<code>nd2</code>	the number of individuals with 2 copies of minor alleles in affected sample
<code>miss.0</code>	Genotype missing rate in unaffected sample
<code>miss.1</code>	Genotype missing rate in affected sample
<code>miss.diff.p</code>	P-value of differential missingness test between unaffected and affected samples
<code>beta</code>	regression coefficient of SNP covariate
<code>se</code>	standard error of beta
<code>chisq</code>	Chi-square statistic for testing beta not equal to zero
<code>df</code>	degree of freedom of the Chi-square statistic
<code>model</code>	model actually used in the analysis
<code>remark</code>	warning or additional information for the analysis, 'not converged' indicates the GEE analysis did not converge; 'logistic reg' indicates GEE model is replaced by logistic regression; 'exp count<5' indicates any expected count is less than 5 in phenotype-genotype table; 'not converged and exp count<5', 'logistic reg & exp count<5' are noted similarly; 'collinearity' indicates collinearity exists between SNP and some covariates

pval	p-value of the chi-square statistic
beta10	regression coefficient of genotype with 1 copy of minor allele vs. that with 0 copy
beta20	regression coefficient of genotype with 2 copy of minor allele vs. that with 0 copy
beta21	regression coefficient of genotype with 2 copy of minor allele vs. that with 1 copy
se10	standard error of beta10
se20	standard error of beta20
se21	standard error of beta21

**Author(s)**

Qiong Yang <qyang@bu.edu> and Ming-Huei Chen <mhchen@bu.edu>

---

gee.lgst.batch.imputed

*function to test genetic association between a dichotomous trait and a batch of imputed SNPs in families using GEE*

---

**Description**

Fit logistic regression via GEE to test association between a dichotomous phenotype and all imputed SNPs in a genotype file. Each family is treated as a cluster, with independence working correlation matrix used in the robust variance estimator. This function applies the same trait-SNP association test to all SNPs in the imputed genotype data. The trait-SNP association test is carried out by gee.lgst.imputed function where the the gee() function from package gee is used.

**Usage**

```
gee.lgst.batch.imputed(genfile, phenfile, pedfile, outfile, phen,
covars=NULL, col.names=T, sep.ped=",", sep.phe=",", sep.gen=",")
```

**Arguments**

genfile	a character string naming the genotype file for reading (see format requirement in details)
phenfile	a character string naming the phenotype file for reading (see format requirement in details)
pedfile	a character string naming the pedigree file for reading (see format requirement in details)
outfile	a character string naming the result file for writing
phen	a character string for a phenotype name in phenfile

<code>covars</code>	a character vector for covariates in phenfile
<code>col.names</code>	a logical value indicating whether the output file should contain column names
<code>sep.ped</code>	the field separator character for pedigree file
<code>sep.phe</code>	the field separator character for phenotype file
<code>sep.gen</code>	the field separator character for genotype file

### Details

Similar to the details for `gee.lgst.batch` but here the SNP data contains imputed genotypes (allele dosages) that are continuous and range from 0 to 2. In addition, the user specified genetic model argument is not available.

### Value

No value is returned. Instead, results are written to `outfile`.

<code>phen</code>	phenotype name
<code>snp</code>	SNP name
<code>N</code>	the number of individuals in analysis
<code>Nd</code>	the number of individuals in affected sample in analysis
<code>AF</code>	imputed allele frequency of coded allele
<code>AFd</code>	imputed allele frequency of coded allele in affected sample
<code>beta</code>	regression coefficient of SNP covariate
<code>se</code>	standard error of beta
<code>remark</code>	warning or additional information for the analysis, note that the genotype counts are based on rounded imputed genotypes; 'not converged' indicates the GEE analysis did not converge; 'logistic reg' indicates GEE model is replaced by logistic regression; 'exp count<5' indicates any expected count is less than 5 in phenotype-genotype table; 'not converged and exp count<5', 'logistic reg & exp count<5' are noted similarly; 'collinearity' indicates collinearity exists between SNP and some covariates
<code>pval</code>	p-value of the chi-square statistic

### Author(s)

Qiong Yang <qyang@bu.edu> and Ming-Huei Chen <mhchen@bu.edu>

---

gee.lgst.imputed	<i>function for testing association between a dichotomous trait and an imputed SNP in family data using GEE</i>
------------------	---

---

## Description

Fit logistic regression via GEE to test association between a dichotomous phenotype and one imputed SNP in a genotype file. Each family is treated as a cluster, with independence working correlation matrix used in the robust variance estimator. This function is called in `gee.lgst.batch.imputed` function to apply association test to all imputed SNPs in the genotype data.

## Usage

```
gee.lgst.imputed(snp, phen, test.dat, covar = NULL)
```

## Arguments

snp	imputed genotype data of a SNP
phen	a character string for a phenotype name in phenfile
test.dat	the product of merging phenotype, genotype and pedigree data, should be ordered by "famid"
covar	a character vector for covariates in phenfile

## Details

Similar to the details for `gee.lgst` function but here the SNP data contains imputed genotypes (allele dosages) that are continuous and range from 0 to 2. In addition, the user specified genetic model argument is not available.

## Value

Please see output in `gee.lgst.batch.imputed`.

## Author(s)

Qiong Yang <qyang@bu.edu> and Ming-Huei Chen <mhchen@bu.edu>

## References

Liang, K.Y. and Zeger, S.L. (1986) Longitudinal data analysis using generalized linear models. *Biometrika*, **73** 13–22.

Zeger, S.L. and Liang, K.Y. (1986) Longitudinal data analysis for discrete and continuous outcomes. *Biometrics*, **42** 121–130.

Vincent J Carey. Ported to R by Thomas Lumley (versions 3.13 and 4.4) and Brian Ripley (version 4.13). `gee`: Generalized Estimation Equation solver. [4.13]. 2007. Ref Type: Computer Program, <http://cran.r-project.org/>

**See Also**

gee() function from package gee

---

GWplot	<i>function for making genome-wide p-values plot</i>
--------	--

---

**Description**

GWplot function plots  $-\log_{10}$  p-value based on SNP's chromosomal position in bitmap format.

**Usage**

```
GWplot(data, pval, pos, chr, chr.plot=c(1:22,"X"), title.text="",
ylim=Inf, outfile, cutoff1=5e-08, cutoff2=4e-07)
```

**Arguments**

data	a dataframe that contains p-values, chromosome number and physical position of SNPs
pval	a character string correspond to the name of the p-value column
pos	a character string correspond to the name of column with SNP physical positions
chr	a character string correspond to the name of column with SNP chromosome number
chr.plot	the chromosomes of interest for GWplot; either 1:22 or c(1:22,"X"), default chr.plot=c(1:22,"X"), "X" for X chromosome
title.text	the title of the genome-wide p-value plot
ylim	the maximum of $-\log_{10}$ p-value to be plotted, useful when not want to plot extremely small p-values
outfile	the file name (xxxx.bmp) for output plot in bitmap format
cutoff1	genome-wide significance; default is 5E-8 ; p-values below this threshold will be highlighted in red
cutoff2	suggestive genome-wide significance; default is 4E-7; p-values below this threshold but above cutoff1 will be highlighted in blue

**Details**

When the dataset has 0 p-value, GWplot will generate pvalzero.csv that contain the results with 0 p-value and make the genome-wide p-values plot by replacing 0 p-value with 5E-324. P-values that reach genome-wide significance are displayed in red color; P-values that reach suggestive genome-wide significance but not genome-wide significance are displayed in blue color.

**Author(s)**

Qiong Yang <qyang@bu.edu> and Ming-Huei Chen <mhchen@bu.edu>

---

lme.batch	<i>function to test genetic association for a continuous trait in families using Linear Mixed Effects model</i>
-----------	---

---

### Description

Fit linear mixed effect model to test association between a continuous phenotype and all SNPs in a genotype file with user specified genetic model. The SNP genotype is treated as fixed effects, and a random effect correlated according to degree of relatedness within a family is also fitted. In each trait-SNP association test, the `lmekin()` function which is modified from the same named function in package `kinship` is used.

### Usage

```
lme.batch(phenfile, genfile, pedfile, phen, kinmat, model="a", covars=NULL,
          outfile, col.names=T, sep.ped=",", sep.phe=",", sep.gen=",")
```

### Arguments

genfile	a character string naming the genotype file for reading (see format requirement in details)
phenfile	a character string naming the phenotype file for reading (see format requirement in details)
pedfile	a character string naming the pedigree file for reading (see format requirement in details)
outfile	a character string naming the result file for writing
phen	a character string for a phenotype name in phenfile
covars	a character vector for covariates in phenfile
model	a single character of 'a','d','g', or 'r', with 'a'=additive, 'd'=dominant, 'g'=general and 'r'=recessive models
kinmat	a character string naming the file where kinship coefficient matrix is kept
col.names	a logical value indicating whether the output file should contain column names
sep.ped	the field separator character for pedigree file
sep.phe	the field separator character for phenotype file
sep.gen	the field separator character for genotype file

### Details

The `lme.batch` function first reads in and merges phenotype-covariates, genotype and pedigree files, then tests the association of phen against all SNPs in `genfile`. `genfile` contains unique individual id and genotype data, with the column names being "id" and SNP names. For each SNP, the genotype data should be coded as 0, 1, 2 indicating the numbers of less frequent alleles. The SNP name in genotype file should not have any dash, '-' and other special characters (dots and underscores are OK). `phenfile` contains unique individual id, phenotype and covariates data,

with the column names being "id" and phenotype and covariate names. pedfile contains pedigree information, with the column names being "famid","id","fa","mo","sex". In all files, missing value should be an empty space, except missing parental id in pedfile. SNPs with low genotype counts (especially minor allele homozygote) may be omitted or analyzed with dominant model. The 'lme.batch' function fits linear mixed effects (LME) model using with 'lme.batch' function from GWAF package and a modified 'lme.kin' function from kinship package.

### Value

No value is returned. Instead, results are written to outfile. When the genetic model is 'a', 'd' or 'r', the result includes the following columns. When the genetic model is 'g', beta and se are replaced with beta10, beta20, beta21, se10, se20, se21 .

phen	phenotype name
snp	SNP name
n0	the number of individuals with 0 copy of minor alleles
n1	the number of individuals with 1 copy of minor alleles
n2	the number of individuals with 2 copies of minor alleles
h2q	the portion of phenotypic variation explained by the SNP
beta	regression coefficient of SNP covariate
se	standard error of beta
chisq	Chi-square statistic for testing beta not equal to zero
df	degree of freedom of the Chi-square statistic
model	model actually used in the analysis
pval	p-value of the chi-square statistic
beta10	regression coefficient of genotype with 1 copy of minor allele vs. that with 0 copy
beta20	regression coefficient of genotype with 2 copy of minor allele vs. that with 0 copy
beta21	regression coefficient of genotype with 2 copy of minor allele vs. that with 1 copy
se10	standard error of beta10
se20	standard error of beta20
se21	standard error of beta21

### Author(s)

Qiong Yang <qyang@bu.edu> and Ming-Huei Chen <mhchen@bu.edu>

## References

kinship package: mixed-effects Cox models, sparse matrices, and modeling data from large pedigrees. Beth Atkinson (atkinson@mayo.edu) for pedigree functions. Terry Therneau (therneau@mayo.edu) for all other functions. 2007. Ref Type: Computer Program <http://cran.r-project.org/>.

Abecasis, G. R., Cardon, L. R., Cookson, W. O., Sham, P. C., & Cherny, S. S. Association analysis in a variance components framework. *Genet Epidemiol*, **21** Suppl 1, S341-S346 (2001).

---

lme.batch.imputed	<i>function to test associations between a continuous trait and a batch of imputed SNPs in families using Linear Mixed Effects model</i>
-------------------	--

---

## Description

Fit linear mixed effects model to test association between a continuous phenotype and all imputed SNPs in a genotype file. The SNP genotype is treated as fixed effects, and a random effect correlated according to degree of relatedness within a family is also fitted. In each trait-SNP association test, the `lmekin()` function which is modified from the same named function in package `kinship` is used.

## Usage

```
lme.batch.imputed(phenfile, genfile, pedfile, phen, kinmat, covars=NULL,
  outfile, col.names=T, sep.ped=",", sep.phe=",", sep.gen=",")
```

## Arguments

phenfile	a character string naming the phenotype file for reading (see format requirement in details)
genfile	a character string naming the genotype file for reading (see format requirement in details)
pedfile	a character string naming the pedigree file for reading (see format requirement in details)
phen	a character string for a phenotype name in phenfile
kinmat	a character string naming the file where kinship coefficient matrix is kept
covars	a character vector for covariates in phenfile
outfile	a character string naming the result file for writing
col.names	a logical value indicating whether the output file should contain column names
sep.ped	the field separator character for pedigree file
sep.phe	the field separator character for phenotype file
sep.gen	the field separator character for genotype file

**Details**

Similar to the details for `lme.batch` function but here the SNP data contains imputed genotypes (allele dosages) that are continuous and range from 0 to 2. In addition, the user specified genetic model argument is not available.

**Value**

No value is returned. Instead, results are written to `outfile`.

<code>phen</code>	phenotype name
<code>snp</code>	SNP name
<code>N</code>	the number of individuals in analysis
<code>AF</code>	imputed allele frequency of coded allele
<code>h2q</code>	the portion of phenotypic variation explained by the SNP
<code>beta</code>	regression coefficient of SNP covariate
<code>se</code>	standard error of beta
<code>pval</code>	p-value of the chi-square statistic

**Author(s)**

Qiong Yang <qyang@bu.edu> and Ming-Huei Chen <mhchen@bu.edu>

**References**

kinship package: mixed-effects Cox models, sparse matrices, and modeling data from large pedigrees. Beth Atkinson (atkinson@mayo.edu) for pedigree functions. Terry Therneau (therneau@mayo.edu) for all other functions. 2007. Ref Type: Computer Program <http://cran.r-project.org/>.

Abecasis, G. R., Cardon, L. R., Cookson, W. O., Sham, P. C., & Cherny, S. S. Association analysis in a variance components framework. *Genet Epidemiol*, **21** Suppl 1, S341-S346 (2001).

---

<code>Imekin</code>	<i>function for linear mixed effects modeling with a kinship coefficient matrix</i>
---------------------	---

---

**Description**

A similar function to `Imekin` from kinship package, but using Wald test, instead of t test. Please see <http://cran.r-project.org/web/packages/kinship/kinship.pdf> for more information.

**See Also**

`Imekin()` function from package kinship

---

`qq`*function to make Qantile-Qantile (QQ) plot for p-values*

---

**Description**

qq function makes the QQ plot of p-values against a uniform (0,1) distribution. The genomic control parameter for one degree freedom chi-square statistics corresponding to the p-values is also plotted.

**Usage**

```
qq(pvalue, outfile)
```

**Arguments**

pvalue	P-values of interest.
outfile	the file name (xxxx.bmp) for output QQ plot in bitmap format

**Author(s)**

Qiong Yang <qyang@bu.edu> and Ming-Huei Chen <mhchen@bu.edu>

# Index

auto, [2](#)

gee.lgst, [4](#)

gee.lgst.batch, [5](#)

gee.lgst.batch.imputed, [7](#)

gee.lgst.imputed, [9](#)

GWAF (GWAF-package), [2](#)

GWAF-package, [2](#)

GWplot, [10](#)

lme.batch, [11](#)

lme.batch.imputed, [13](#)

lmekin, [14](#)

qq, [15](#)