Using GWAF package to conduct genome-wide association/interaction analysis and rare variant analysis with family data

Ming-Huei Chen and Qiong Yang
Departments of Neurology and Biostatistics
Boston University
Contact: qyang@bu.edu

Table of Contents

Overview

This package (GWAF) was designed mainly to analyze a batch of genotyped/imputed SNPs against a continuous or dichotomous phenotype measured on subjects of families for genetic association/interaction. The number of SNPs that can be analyzed at once depends on the memory capacity of your system. For genome-wide association (GWA) analysis and genome-wide interaction (GWI) analysis, if the memory is not enough to analyze all SNPs together, one can split the dataset by columns into several datasets, and analyze each of them sequentially using functions in this package. In addition, GWAF also provides functions for making genome-wide p-values plot, QQ plot and scripts for GWA/GWI analysis.

Methods

Both linear mixed effects (LME) model and generalized estimating equations (GEE) are used in this package to analyze continuous traits [1]. In LME modeling, a random intercept, person specific random effects correlated according to degree of relatedness (i.e. kinship coefficient) within a family, is used to account for within family correlation. For analyzing imputed genotype data, a more efficient and faster version of LME modeling is also implemented, which estimates polygenic variation only once when analyzing a batch of imputed SNPs (so the sample size is the same for each SNP). Logistic regressions via GEE is used in this package to analyze dichotomous traits, treating each pedigree (i.e. individuals with the same family id) as a cluster with independent working correlation structure used in the robust variance estimator. In addition, a generalized linear mixed effects model (GLMM) with logistic link and a normal random intercept for each cluster is also used to analyze dichotomous traits with observed genotypes, treating each pedigree as a cluster. Our package called functions lmekin(), geese(), and lmer() functions from packages coxme, geepack and lme4, respectively for aforementioned methods. LME and GLMM are recommended to be used for analyzing rare variants with continuous traits and dichotomous traits [2,3], respectively. And our package is a wrapper that enables users to analyze more than one SNP and automatically summarizing the results in an informative and convenient output..

Required Files

Before performing analyses with this package, following files have to be created.

1. **Pedigree file**: A file containing all the families is required. The column names should be exactly the same (case sensitive) as in following example based on **comma delimited format**. Missing father (fa) or mother (mo) ids should be 0. Individuals who are unrelated to anyone can be included as family of size 1.

```
famid,id,fa,mo,sex
1,10,0,0,1
```

```
1,11,0,0,2
1,12,10,11,1
1,13,10,11,1
3,32,0,0,1
3,33,0,0,2
3,334,32,33,1
3,335,32,33,2
10,50,0,0,1
11,60,0,0,2
```

2. **Kinship coefficient matrix for LME:** For LME, one should use kinship2 package to create the kinship coefficient matrix as an R object and save it on disk, as shown below. Then in the LME analysis, path to this R object file is supplied to the 'kinmat' argument in the LME analysis function (lmepack.batch).
   - Sample R code for create kinship coefficient matrix (must be named kmat in R as shown below)

```
library(GWAF)
library(kinship2)
kmat<-makekinship(ped$famid,ped$id,ped$fa,ped$mo)
## using twice the kinship coefficient
kmat<-kmat*2
## save the kinship matrix to a file
save(kmat,file="fhs_unrel_comb.kinship.Rdata")
```

   - In LME analyses, supply the path to the kinship coefficient matrix file to "kinmat" argument.

```
lmepack.batch(phenfile, genfile, pedfile, phen,
kinmat="fhs_unrel_comb.kinship.Rdata", …)
```

3. Format of phenotype and genotype files

Phenotype file contains unique individual id, phenotype and covariates. The header should contain "id", followed by other variable names. Use empty space for missing values. **Dichotomous phenotype must be coded as 0, 1 with 1 being affected**. Covariates values must be coded numerically (dichotomous covariate can have any two numeric values). Following is an example of the phenotype file based on **comma delimited format**:

```
id,phen1,phen2,covar1,covar2
10,100,1,1,0.2
112,,0,1,0.3
312,130,1,2,0.4
513,125,0,,0.5
```

Genotype file contains unique individual id and genotype data. The header should contain "id", followed by SNP names. **For genotyped SNPs, genotype should be coded as 0, 1, 2 representing the copies of the coded allele**. **While for imputed SNPs, genotype (allele dosage) is continuous and ranges from 0 to 2.** Use empty space for missing genotypes. SNP names should not contain special characters such as "-","/", etc. But "." and "_" are allowed. For example (based on **comma delimited format**):

```
id,SNP.1,SNP_2
10,0,1
11,,
12,1,2
13,2,0
```

Examples

Here are example function calls for analyzing a single phenotype against all genotyped SNPs in genotype file. Suppose `"phenfile.csv"`, `"genfile.csv"`, `"pedfile.csv"`, are directory paths to **comma delimited** phenotype, genotype and pedigree files respectively; "phen1" and "phen2" are the name of the continuous and binary phenotype to be analyzed, respectively. Kinship coefficient matrix file is "`fhs_unrel_comb.kinship.Rdata`". Replace geepack.lgst.batch() with glmm.lgst.batch() to use GLMM to analyze dichotomous traits with observed genotype data. Please note for analyzing imputed genotypes, use lmepack.batch.imputed()/lmeVpack.batch.imputed() and geepack.lgst.batch.imputed() functions, and the model argument is not available in these three functions.

LME:

```
library(GWAF)
lmepack.batch(phenfile="phenfile.csv",
genfile="genfile.csv", pedfile="pedfile.csv",
phen="phen1", model="a",
kinmat="fhs_unrel_comb.kinship.Rdata",
covars=c("covar1","covar2"), outfile="lme.result.csv")
#covars argument can be omitted if no covariates need to be adjusted
```

LME interaction:

```
library(GWAF)
lmepack.int.batch(phenfile="phenfile.csv",
genfile="genfile.csv", pedfile="pedfile.csv",
phen="phen1", outfile='lme.int.csv',
covars=c('age','sex'),
kinmat='fhs_unrel_comb.kinship.Rdata', cov.int='age',
sub='N', phen='phen1', sep.ped=',', sep.phe=',',
sep.gen=',', col.names=F)
```

GEE – binary trait:

```
library(GWAF)
geepack.lgst.batch(phenfile="phenfile.csv",
genfile="genfile.csv", pedfile="pedfile.csv",
phen="phen2", model="a", covars=c("covar1","covar2"),
outfile="gee.result.csv")
#covars argument can be omitted if no covariates need to be adjusted
```

GEE interaction – binary trait:

```
library(GWAF)
geepack.lgst.batch(phenfile="phenfile.csv",
genfile="genfile.csv", pedfile="pedfile.csv",
phen="phen2", covars=c("covar1","covar2"),
cov.int="covar1", sub="Y", outfile="gee.int.result.csv")
```

GEE – continuous trait:

```
library(GWAF)
geepack.quant.batch(phenfile="phenfile.csv",
genfile="genfile.csv", pedfile="pedfile.csv",
phen="phen1", model="a", covars=c("covar1","covar2"),
outfile="gee.quant.result.csv")
#covars argument can be omitted if no covariates need to be adjusted
```

GEE interaction – continuous trait:

```
library(GWAF)
geepack.quant.int.batch(phenfile="phenfile.csv",
genfile="genfile.csv", pedfile="pedfile.csv",
phen="phen1", outfile='gee.quant.int.csv',
covars=c('age','sex'), cov.int='age', sub='Y',
phen='phen1', sep.ped=',', sep.phe=',', sep.gen=',',
col.names=F)
```

Important: These functions are designed to analyze a single phenotype against all the SNP genotypes in a genotype file in a single call. To analyze multiple phenotypes, multiple calls of the functions are needed.

References

[1]. Chen MH and Yang Q. (2010) GWAF: an R package for genome-wide association analyses with family data. *Bioinformatics* 26(4):580-581.

[2]. Chen MH, Liu X, Wei F, Larson MG, Fox CS, Vasan RS and Yang Q. (2011) A comparison of strategies for analyzing dichotomous outcomes in genome-wide association studies with general pedigrees. *Genetic Epidemiology* 35:650-657.

[3]. Chen MH and Yang Q. GWAF for genome-wide interaction analyses and rare variant analyses with family data. In preparation.

Output

Output information: Output from a function call is saved to the file specified in *outfile* argument in each function. Table 1 describes the output columns for LME/GEE analyses for GWA analysis on a continuous trait with genotyped SNP. Table 2 describes the output columns for GEE/GLMM analyses for GWA analysis on a dichotomous trait with genotyped SNPs. Table 3 describes the output columns for LME/GEE analyses for GWA analysis on a continuous trait with imputed SNPs. Table 4 describes the output columns for GEE analyses for GWA analysis on a dichotomous trait with imputed SNPs. Table 5 describes the output columns for LME/GEE analyses for genome-wide interaction analyses on a continuous trait with genotyped/imputed SNP. Table 6 describes the output columns for GEE analyses for genome-wide interaction analyses on a dichotomous trait with genotyped/imputed SNPs.

Table 1: Output columns from LME/GEE analysis for GWA analysis on a continuous trait with genotyped SNPs (Genotype should be coded as 0, 1, 2 representing the copies of the coded allele)

Column	**Description**
phen	Phenotype Name
snp	SNP name
n0	number of subjects with non-missing phenotype and genotype 0
n1	number of subjects with non-missing phenotype and genotype 1
n2	number of subjects with non-missing phenotype and genotype 2
h2q§	% total phenotypic variance explained by the SNP, **not available in GEE**
Output fields for additive, dominant or recessive model	
beta	additive model: beta coefficient per 1 copy increment of coded allele; recessive model: beta coefficient for genotype 2 vs. all other genotypes; dominant model: beta coefficient for 1 and 2 combined vs. genotype 0;
chisq	Chi-square statistic for testing beta equal to zero
df	degrees of freedom for the chi-square statistic
model	model used in the analysis
pval	p-value of the chi-square statistic
Output fields for general model	
beta10	beta coefficient for genotype 1 vs. 0. If the dominant model is used in the analysis, this is the beta coefficient for genotype 1 and 2 combined vs. genotype 0.
beta20	beta coefficient for genotype 2 vs. 0
beta21	beta coefficient for genotypes 2 vs. 1
se10	standard error of beta10
se20	standard error of beta20
se21	standard error of beta21
chisq	Chi-square statistic for testing global hypothesis that both beta10 and beta20 equal zero
df	degrees of freedom of the chi-square statistic
model	model used in the analysis
pval	p-value of the chi-square statistic

§ $h_q^2 = max\left(0, \frac{\sigma_{G.null}^2 + \sigma_{e.null}^2 - \sigma_{G.full}^2 - \sigma_{e.full}^2}{Var(y)}\right)$, where *Var(y)* is the total phenotypic variance, $\sigma_{G.null}^2, \sigma_{e.null}^2$ are the polygenic variance and error variance when modeling without the tested SNP, and $\sigma_{G.full}^2, \sigma_{e.full}^2$ are the polygenic variance and error variance when modeling with the SNP.

Table 2. Output columns from GEE/GLMM analyses for GWA analysis on a dichotomous trait with genotyped SNPs (Genotype should be coded as 0, 1, 2 representing the copies of the coded allele)

Column	Description
phen	Phenotype Name
snp	SNP name
n0	number of subjects with non-missing phenotype and genotype 0
n1	number of subjects with non-missing phenotype and genotype 1
n2	number of subjects with non-missing phenotype and genotype 2
nd0	number of diseased subjects with genotype 0
nd1	number of diseased subjects with genotype 1
nd2	number of diseased subjects with genotype 2
miss.0	rate of missing genotypes among non-diseased subjects
miss.1	rate of missing genotypes among diseased subjects
miss.diff.p	P-value of test of differential missingness between unaffected and affected subjects
Output fields when additive, dominant or recessive model specified in control file	
beta	additive model: beta coefficient per 1 copy increment of coded allele; recessive model: beta coefficient for genotype 2 vs. genotypes 0 and 1 combined dominant model: beta coefficient for genotype 1 and 2 combined vs. genotype 0
se	standard error of beta
chisq	Chi-square statistic for testing beta equal to zero
df	degrees of freedom of the chi-square statistic
model	model used in the analysis
remark	warning or additional information for the analysis
pval	p-value of the chi-square statistic
Output fields for general model	
beta10	beta coefficient for genotype 1 vs. 0. If the dominant model is used in the analysis, this is the beta coefficient for genotype 1 and 2 combined vs. genotype 0.
beta20	beta coefficient of genotype with 2 copies of coded allele vs. that with 0 copy
beta21	beta coefficient of genotype with 2 copies of coded allele vs. that with 1 copy
se10	standard error of beta10
se20	standard error of beta20
se21	standard error of beta21
chisq	Chi-square statistic for testing at least one of the beta10 and beta20 not zero
df	degrees of freedom of the chi-square statistic
model*	model used in the analysis
remark†	warning or additional information for the analysis
pval	p-value of the chi-square statistic

*** When 0/low genotype counts occur, general model may be replaced by dominant model in analysis.**

† Remark column contains warning or additional information. Here is a detailed explanation of the meaning of each remark.

Remark	**Reason**
“not converged”	The GEE analysis did not converge. So results are not reliable and should be discarded.
“logistic reg”	Logistic regression assuming independent observations is performed, when the number of pedigrees with 2 or more individuals is less than 10 or there are zero genotype counts in any cell of snp by phenotype (3 by 2) table.
“exp count<5”	At least one expected count is less than 5 in 2xN table, N =number of genotype categories for general model, and N=2 for other models. The test results may have a higher false positive rate.
"not converged & exp count<5"	See above
“logistic reg& exp count<5”	See above
"collinearity"	If there are any covariates highly correlated with a snp (abs(correlation)>0.99999999),no analysis is performed.

Table 3: Output columns from LME/GEE analysis for GWA analysis on a continuous trait with imputed SNPs

Column	**Description**
phen	Phenotype Name
snp	SNP name
N	number of subjects with non-missing phenotype and genotype
AF	Imputed allele frequency of coded allele
h2q	% total phenotypic variance explained by the SNP (NOT available for GEE)
beta	beta coefficient per 1 copy increment of coded allele
se	standard error of beta
pval	p-value of the chi-square statistic

Table 4. Output columns from GEE analyses for GWA analysis on a dichotomous trait with imputed SNPs

Column	**Description**
phen	Phenotype Name
snp	SNP name
N	number of subjects with non-missing phenotype and genotype
Nd	number of subjects with non-missing phenotype and genotype in affected sample
AF	Imputed allele frequency of coded allele
AFd	Imputed allele frequency of coded allele in affected sample
beta	beta coefficient per 1 copy increment of coded allele;
se	standard error of beta
remark	warning or additional information for the analysis, please see 1st table on page 8
pval	p-value of the chi-square statistic

Table 5: Output columns from LME/GEE analysis for GWI analysis on a continuous trait with genotyped SNPs (Genotype should be coded as 0, 1, 2 representing the copies of the coded allele)

Column	Description
phen	Phenotype Name
snp	SNP name
covar_int	the covariate for interaction
n	sample size used in analysis
AF	allele frequency of the coded allele
model	genetic model used in analysis, **additive model only**
beta_snp	regression coefficient of SNP covariate
se_snp	standard error of beta_snp
pval_snp	p-value of testing SNP covariate
beta_int	regression coefficient of the interaction term
se_int	standard error of beta_int
pval_int	p-value of testing the interaction term
Additional output fields when stratified analyses are requested for a binary covariate for interaction	
beta_snp_cov0	regression coefficient of SNP covariate in stratified analysis using the subset where covar_int level is 0
se_snp_cov0	standard error of beta_snp_cov0
pval_snp_cov0	p-value of testing SNP covariate in stratified analysis using the subset where covar_int level is 0
beta_snp_cov1	regression coefficient of SNP covariate in stratified analysis using the subset where covar_int level is 0
se_snp_cov1	standard error of beta_snp_cov0
pval_snp_cov1	p-value of testing SNP covariate in stratified analysis using the subset where covar_int level is 0
Additional output fields when stratified analyses are not requested	
cov_beta_snp_beta_int	Covariance between beta_snp and beta_int

Table 6. Output columns from GEE analyses for GWI analysis on a dichotomous trait with genotyped SNPs (Genotype should be coded as 0, 1, 2 representing the copies of the coded allele)

Column	Description
phen	Phenotype Name
snp	SNP name
covar_int	the covariate for interaction
n	sample size used in analysis
AF	allele frequency of the coded allele
nd	number of subjects with phenotype and genotype in affected sample
AFd	allele frequency of the coded allele in affected sample
model	genetic model used in analysis, **additive model only**
beta_snp	regression coefficient of SNP covariate

se_snp	standard error of beta_snp
pval_snp	p-value of testing SNP covariate
beta_int	regression coefficient of the interaction term
se_int	standard error of beta_int
pval_int	p-value of testing the interaction term
remark	warning or additional information for the analysis, , please see 1st table on page 8
Additional output fields when stratified analyses are requested for a binary covariate for interaction	
beta_snp_cov0	regression coefficient of SNP covariate in stratified analysis using the subset where covar_int level is 0
se_snp_cov0	standard error of beta_snp_cov0
pval_snp_cov0	p-value of testing SNP covariate in stratified analysis using the subset where covar_int level is 0
beta_snp_cov1	regression coefficient of SNP covariate in stratified analysis using the subset where covar_int level is 0
se_snp_cov1	standard error of beta_snp_cov0
pval_snp_cov1	p-value of testing SNP covariate in stratified analysis using the subset where covar_int level is 0
Additional output fields when stratified analyses are not requested	
cov_beta_snp_beta_int	Covariance between beta_snp and beta_int

Acknowledgements

The authors thank Drs. Josée Dupuis, Kathryn L. Lunetta, L. Adrienne Cupples, Martin G. Larson, Anita L. DeStefano, and Jemma B. Wilk for their helpful comments on this package. The authors also thank Dr. Jinghua Zhao for his help with the kinship package, and Alisa N. Manning, Denver J. Lybarger and Andi Broka for their assistance.