Using GWAF package to conduct association 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 (Genome-Wide Association analyses with Family data) was designed mainly to analyze a batch of genotyped/imputed SNPs against a continuous or dichotomous phenotype measured on subjects of families for association. The number of SNPs that can be analyzed at once depends on the memory capacity of your system. For genome-wide association studies (GWAS), 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 GWAS.

Methods

Linear mixed effects model (LME) is used in this package to analyze continuous traits, with person specific random effects correlated according to degree of

relatedness (i.e. kinship coefficient) within a family to account for within family correlation. Logistic regressions via generalized estimating equations (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. These methods were implemented in the lmekin() and gee() functions in packages kinship and gee, 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 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 (lme.batch).

- Sample R code for create kinship coefficient matrix (must be named kmat in R as shown below)

```
library(kinship)
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.

```
lme.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” is the name of the phenotype to be analyzed. Kinship coefficient matrix file is `“fhs_unrel_comb.kinship.Rdata”`. Please note for analyzing imputed genotypes, use lme.batch.imputed() and gee.lgst.batch.imputed() functions, and the model argument is not available in these two functions.

LME:

```
library(GWAF)
lme.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 adjsuted
```

GEE:

```
library(GWAF)
gee.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 adjsuted
```

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.

Output

Output information: Output from a function call is saved to the file specified in *outfile* argument in each function. Tables 1 and 2 describe the output columns for LME and GEE analyses with genotyped SNPs, respectively. Tables 3 and 4 describe the output columns for LME and GEE analyses with imputed SNPs, respectively.

Table 1: Output columns from LME analysis 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
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

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

Table 2. Output columns from GEE analyses 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 analysis 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
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 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
pval	p-value of the chi-square statistic

† Remark column contains warning or additional information. Here is a detailed explanation of the meaning of each remark. The genotype counts are computed based on rounded imputed genotypes.

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.

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.