BIOINFORMATICS APPLICATIONS NOTE Vol. 26 no. 4 2010, pages 580–581
doi:10.1093/bioinformatics/btp710

Genetics and population analysis

GWAF: an R package for genome-wide association analyses with family data

Ming-Huei Chen[1,2] and Qiong Yang[2,3,*]

[1]Department of Neurology, Boston University School of Medicine, Boston, MA 02118, [2]The National Heart, Lung, Blood Institute's Framingham Heart Study, Framingham, MA 01702 and [3]Department of Biostatistics, Boston University School of Public Health, Boston, MA 02118, USA

Received on September 28, 2009; revised on December 18, 2009; accepted on December 21, 2009

Advance Access publication December 29, 2009

Associate Editor: Jeffrey Barrett

ABSTRACT

Summary: GWAF, Genome-Wide Association analyses with Family, is an R package designed for GWAF. It implements association tests between a batch of genotyped or imputed single nucleotide polymorphisms (SNPs) and a binary or continuous trait with user specified genetic model, and generates informative results from the analyses. In addition, GWAF provides functions to visualize results. We evaluated GWAF using a simulated continuous trait and a binary trait dichotomized from the simulated continuous trait with real genotype data from the Framingham Heart Study's SNP Health Association Resource project.

Availability: http://cran.r-project.org/web/packages/GWAF/

Contact: qyang@bu.edu

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 INTRODUCTION

The completion of the Human Genome Project and the advances of the International Hapmap consortium have made genome-wide association (GWA) scans possible. Many software and packages are designed only for GWA studies with unrelated individuals. For family data, the correlations among individuals in a pedigree may cause false positives due to unexplained familial correlation. For this purpose, GWA analyses with family data (GWAF) package utilizes functions in existing R packages to properly model the residual correlations within families in the test of genotype–phenotype association. GWAF is a wrapper that enables users to analyze a batch of single nucleotide polymorphisms (SNPs) under user specified genetic model (additive, recessive, dominant or general) and covariates using existing functions, and that automatically summarizes the results in an informative and convenient format. The genotypes can be observed or imputed.

GWAF was developed from the functions that have been empirically tested through simulations and used in many GWAs publications from the SNP Health Association Resource (SHARe) project of Framingham Heart Study (FHS), http://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs000007.v6.p3. With the availability of SHARe data to general scientific community, this package would be a useful resource for investigators to analyze SHARe data or other GWA studies consisted of related individuals.

*To whom correspondence should be addressed.

2 IMPLEMENTATION

2.1 Input files

GWAF requires three input files: (i) pedigree file; (ii) genotype file; and (iii) phenotype file. The pedigree file shall contain pedigree information of family ID, individual ID, father ID, mother ID and sex. Each genotype file shall contain individual ID and the genotype data of an arbitrary number of SNPs. The phenotype file shall contain individual ID, phenotype of interest and covariate data. Please see package documentation (http://cran.r-project.org/web/packages/GWAF/) or Supplementary Material for more details.

2.2 Continuous traits

To account for the within pedigree familial correlation, GWAF uses a linear mixed effects model (LME) implemented in lmekin function in kinship package (http://cran.r-project.org/web/packages/kinship/) to test association between a continuous trait and each SNP in the genotype file under user specified genetic model. The only difference between lmekin() functions from kinship and GWAF is that under general model, GWAF uses the Wald chi-square test that provides a global test of genotype effects, while kinship only has *t*-test for each comparison of two genotypes. The within pedigree correlation matrix is modeled using kinship coefficient matrix (Abecasis *et al*., 2001) in GWAF, which can be easily generated with the kinship package. For genotyped SNPs, the user specified genetic model can be additive, dominant, recessive or general model. Except that general model uses a two degrees of freedom Wald chi-square test, the others use a one degree of freedom Wald chi-square test. While for imputed SNPs, the user specified genetic model option is not available and GWAF will analyze the imputed genotype data without recoding. Finally, GWAF also provides an estimate of the proportion of phenotype variance explained by the tested SNP, which is computed by

$$h_q^2 = \max\left(0, \frac{\sigma^2_{G.null} + \sigma^2_{e.null} - \sigma^2_{G.full} - \sigma^2_{e.full}}{Var(y)}\right)$$

where Var(y) is the total phenotypic variance, $\sigma^2_{G.null}$ and $\sigma^2_{e.null}$ are the polygenic variance and error variance when modeling without the tested SNP, $\sigma^2_{G.full}$ and $\sigma^2_{e.full}$ are the polygenic variance and error variance when modeling with the tested SNP.

2.3 Dichotomous traits

For a dichotomous trait, GWAF uses logistic regression via generalized estimating equations (GEE, Liang and Zeger, 1986) implemented in gee()

function in gee package (http://cran.r-project.org/web/packages/gee/) to test association between the phenotype of interest and each SNP in a genotype file with user specified genetic model. In GEE analysis, GWAF uses independence working correlation matrix with each family being a cluster in the robust variance estimate for the genotype effects. Similar to continuous traits, GWAF uses Wald chi-square test for the main effect. Again, except that general model uses a two degrees of freedom Wald chi-square test, the others use a one degree of freedom Wald chi-square test. In addition, Fisher's exact test is carried out to test whether differential missingness exists between affected and unaffected sample which can be used to judge potential genotyping quality discrepancy between cases and controls.

The gee() function in the gee package can encounter convergence or hanging (unlimited looping) issues mainly due to SNPs with 0/low counts in the disease-genotype contingency table that frequently happens to low minor allele frequency (MAF) SNPs. In such cases, GWAF may employ logistic regression instead of GEE and provide a remark. Users should pay caution to the remarks for the top SNPs or are suggested to filter out SNPs with low MAFs.

2.4 Output file

Please see the manual of our functions at http://cran.r-project.org/web/packages/GWAF/ for details. Sample outputs were also provided in Supplementary Tables S1–S6.

2.5 Genome-wide *P*-values plot and quantile–quantile *P*-values plot

After GWA analyses are completed, to briefly visualize and examine the results, GWAF provides GWplot() function for genome-wide plot of $-\log_{10}(P\text{-values})$ versus genomic position and qq() function for quantile–quantile (QQ) plot of observed $-\log_{10}(P\text{-values})$ versus expected $-\log_{10}(P\text{-values})$ in bitmap format. GWplot() requires the *P*-value, the chromosome number and the physical position of each SNP. The *P*-values are negatively logarithm transformed with base 10 in both plots. In GWplot(), the users are allowed to specify two *P*-value cut-offs, one indicates genome-wide significance with the default of $5E-8$ and the other indicates suggestive genome-wide significance with the default of $4E-7$. SNPs with genome-wide significance are presented in red, while SNPs between the two cut-offs are plotted in blue. The qq() function makes the QQ plot of *P*-values against a uniform (0,1) distribution. The genomic control parameter λ (Devlin and Roeder, 1999) that indicates systematic inflation in GWA results for one degree of freedom chi-square statistics corresponding to the *P*-values is also presented in the QQ plot.

3 EXAMPLE

We applied GWAF to a simulated continuous trait and a binary trait with real 550K genotype data from Framingham Heart Study's SHARe project. Phenotypes were simulated on 8481 individuals genotyped with call rate >97% from 1494 real pedigrees. The continuous traits were randomly generated following multivariate normal distribution, with a quantitative trait locus (QTL), a polygenic and a residual variance component using the program SOLAR (Almasy and Blangero 1998). SNP rs1570092 on chromosome 1 of good genotype quality was selected to be the single QTL explaining 1% of total phenotypic variance (=1) and the polygenic heritability was set to be 0.3. Additive genetic model was used in simulation and association analyses.

To create binary traits, we dichotomized the simulated continuous traits by assuming 10% population prevalence and an additive genetic model with genotype relative risk of 1.3. The additive genetic model was also used in the association analyses.

The results of LME and GEE are presented in the genome-wide *P*-values plots and QQ plots (Supplementary Fig. S1). Both genome-wide *P*-values plots show that the genome-wide significant SNPs were all close to the QTL, rs1570092. The genomic control factors (λ) are 1 and 1.02 for LME and GEE, respectively, showing no global inflation of false positives. Note, λ is computed as the empirical median divided by its expectation under the χ^2_1 distribution. Thus λ does not reflect how much deviation the tail has from the 45° line, as in our example, a notable deviation in the tail is observed with λ close to 1. For GEE results, 21 SNPs with MAF <0.01 were excluded. Both LME and GEE identified rs1570092 as the most significant SNP [LME *P*-value = $1.64E-22$ (explained 1.28% phenotype variation); GEE *P*-value = $6.52E-10$, Odds Ratio = 1.427]. For a batch of 1000 genotyped SNPs, GEE takes ~4.5 min and LME takes ~1.8 h to complete the analyses using a Linux cluster with 2× Dual-Core AMD Opteron(tm) Processor 2218 HE and total 12 GB RAM, running Rocks 4.3 Linux Cluster Distribution from San Diego Supercomputer Center. LME (uses 405 MB of RAM) takes longer time because of its model complexity and the estimation of the proportion of phenotypic variance explained by the tested SNP that requires two analyses (under null and full models) for each SNP. For analyzing imputed genotypes, GWAF takes less time because of no recoding.

4 FUTURE WORK

We are in the process of expanding GWAF to perform gene-environmental interaction and better handling rare variants.

ACKNOWLEDGEMENTS

The authors thank Dr Josée Dupuis, Dr Kathryn L. Lunetta, Dr L. Adrienne Cupples, Dr Martin G. Larson, Dr Anita L. DeStefano and Dr Jemma B. Wilk for their helpful comments on the 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. This research was conducted in part using data and resources from the Framingham Heart Study of the National Heart Lung and Blood Institute of the National Institutes of Health and Boston University School of Medicine.

Funding: National Heart, Lung and Blood Institute's Framingham Heart Study (contract no. N01-HC-25195) and its contract with Affymetrix, Inc for genotyping services (contract no. N02-HL-6-4278). A portion of this research utilized the Linux Cluster for Genetic Analysis (LinGA-II) funded by the Robert Dawson Evans Endowment of the Department of Medicine at Boston University School of Medicine and Boston Medical Center.

Conflict of Interest: none declared.

REFERENCES

Abecasis,G.R. *et al*. (2001) Association analysis in a variance components framework. *Genet. Epidemiol*., **21**(Suppl. 1), S341–S346.

Almasy,L. and Blangero,J. (1998) Multipoint quantitative-trait linkage analysis in general pedigrees. *Am. J. Hum. Genet.*, **62**, 1198–1211.

Devlin,B. and Roeder,K. (1999) Genomic control for association studies. *Biometrics*, **55**, 997–1004.

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