Bioinformatics, 32(15), 2016, 2392–2393
doi: 10.1093/bioinformatics/btw160
Advance Access Publication Date: 24 March 2016
Applications Note

OXFORD

Genetics and population analysis

FREGAT: an R package for region-based association analysis

Nadezhda M. Belonogova[1,*,†], Gulnara R. Svishcheva[1,2,†] and Tatiana I. Axenovich[1,3]

[1]Institute of Cytology and Genetics, Siberian Branch of the Russian Academy of Sciences, Novosibirsk, [2]Vavilov Institute of General Genetics, the Russian Academy of Sciences, Moscow, Russia and [3]Novosibirsk State University, Novosibirsk, Russia

*To whom correspondence should be addressed.
†The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors.
Associate Editor: Oliver Stegle

Received on November 9, 2015; revised on March 2, 2016; accepted on March 20, 2016

Abstract

Summary: Several approaches to the region-based association analysis of quantitative traits have recently been developed and successively applied. However, no software package has been developed that implements all of these approaches for either independent or structured samples. Here we introduce *FREGAT* (Family REGional Association Tests), an R package that can handle family and population samples and implements a wide range of region-based association methods including burden tests, functional linear models, and kernel machine-based regression. *FREGAT* can be used in genome/exome-wide region-based association studies of quantitative traits and candidate gene analysis. *FREGAT* offers many useful options to empower its users and increase the effectiveness and applicability of region-based association analysis.
Availability and Implementation: https://cran.r-project.org/web/packages/FREGAT/index.html
Supplementary Information: Supplementary data are available at *Bioinformatics* Online.
Contact: belon@bionet.nsc.ru

1 Introduction

The development of new and effective whole-exome and whole-genome resequencing technologies demands the establishment of powerful and computationally efficient statistical methods to test the associations between rare variants and complex traits. The methods developed for the analysis of common variants are underpowered because of the small number of observations for any given variant (Eichler *et al.*, 2010). The statistical power of the association analysis of rare variants is expected to increase when genetic variants in a region of interest are tested simultaneously rather than separately (Eichler *et al.*, 2010).

Several approaches to the region-based association analysis of quantitative traits have been developed and successively applied in practice. The simplest approach uses various methods for collapsing rare variants within a region of interest (burden tests). In this case, a set of rare variants in a region is described by a single genetic variable that is then tested for association (Dering *et al.*, 2011). Other approaches wherein genotype effects are estimated as fixed effects are based on the multiple regression model, especially on its functional data analysis-based version (functional linear models, FLMs) (Fan *et al.*, 2013, 2014). An alternative approach treats the regional association as a random effect and employs kernel machine-based regression (i.e. sequence kernel association tests, SKAT and SKAT-O) (Kwee *et al.*, 2008; Lee *et al.*, 2012; Wu *et al.*, 2011).

Each of these methods has their own advantages and limitations. All have been adopted for pedigree samples (Belonogova *et al.*, 2013; Chen *et al.*, 2013; Feng *et al.*, 2015; Oualkacha *et al.*, 2013; Schifano *et al.*, 2012; Svishcheva *et al.*, 2014, 2015). However, no software package has yet been developed that implements all of these popular approaches for handling population or family samples. We developed the Family REGional Association Tests (*FREGAT*) package, an R package that can handle population and family samples and implements a wide range of region-based association methods including

burden tests, FLMs, and kernel machine-based regression. *FREGAT* can be used for the genome/exome-wide association studies of quantitative traits and offers many useful options to increase the effectiveness and applicability of region-based association analysis.

2 Implementation

In *FREGAT*, the inheritance of a quantitative trait in the sample of n genetically related individuals is described by the linear mixed model as $\boldsymbol{y} = \boldsymbol{X\alpha} + \boldsymbol{r} + \boldsymbol{g} + \boldsymbol{\varepsilon}$, where $\boldsymbol{y}$ denotes the $n \times 1$ vector of phenotypes; $\boldsymbol{X}$ represents the $n \times p$ matrix of p covariates; $\boldsymbol{\alpha}$ is the $p \times 1$ vector of the regression coefficients of the covariates; $\boldsymbol{g}$ and $\boldsymbol{\varepsilon}$ are the $n \times 1$ vectors of random polygenic and environmental effects, respectively; $\boldsymbol{r}$ is the $n \times 1$ vector of effects of the analyzed region modeled as either random or fixed variables depending on the chosen method. For family samples, vector $\boldsymbol{g}$ is assumed to be distributed as $N(0, 2\sigma_g^2\boldsymbol{K})$, where $\boldsymbol{K}$ is the $n \times n$ kinship matrix, and σ_g^2 is the variance component that models within-family correlations; $\boldsymbol{\varepsilon} \sim N(0, \sigma_e^2\boldsymbol{I}_n)$, where $\boldsymbol{I}_n$ is the $n \times n$ identity matrix and σ_e^2 is the variance component of random errors. All *FREGAT* methods assume that parameters $\boldsymbol{\alpha}$, σ_g^2 and σ_e^2 are estimated once for each trait under the null model ($\boldsymbol{r} = 0$). This step is similar for all methods (Fig. 1). Within the package, the null model estimates obtained in the first run can be saved and reused in all subsequent runs for the same trait. This feature increases the analysis speed because the parameter estimation under the null model can be computationally intensive, particularly in large samples. If the kinship matrix is not set, the program does not include within-family correlations in the linear mixed model and individuals are treated as unrelated.

Comparison with native implementations demonstrates that calculations are correct for all *FREGAT* methods (Supplementary Figures S1 and S2). Run times are comparable with those of the fastest existing programs, with SKAT-O and FLM methods in *FREGAT* being the fastest among available implementations (Supplementary Table S1). The run time of fitting the null model has cubic dependence on sample size, while the run times of the association tests exhibit quadratic or slower growth with the sample size and region size (Supplementary Figures S3 and S4). All tests are fast enough to perform a whole-exome region-based association study of a large structured sample in a reasonable time even on a single processor. Parallel calculations are enabled by *foreach* and *doParallel* R packages (Calaway *et al.*, 2015a,b) and increase the speed of large analyses on multi-core machines. *FREGAT* can accept genotypic data stored as *snp.data* class to enhance memory usage (Aulchenko *et al.*, 2007), but such data will not be processed as raw data. For VCF and PLINK binary input files, functions from *seqminer* (Zhan and Liu, 2015) and *snpStats* (Clayton, 2015) R packages are used to read them gene-wise. The run time of the null model step is significantly improved compared with the *FFBSKAT* package (Svishcheva *et al.*, 2014).

All *FREGAT* methods support covariates, sliding window analysis and non-additive models. *FREGAT* can be used to analyze a single candidate gene, an entire exome and the sliding windows of the desired size in a genome-wide region-based association study.

For each method, *FREGAT* offers an extensive list of options for an informative and flexible analysis. Kernel-based tests allows the assignment of different kernel functions (e.g. linear, quadratic and identity-by-state (IBS)-based) and estimation of P values by using either Davies' or Kuonen's methods. It also provides an option to estimate the proportion of variance explained by region (see details in Supplementary Materials). Arbitrary weight functions can be used in burden tests and kernel machine-based regression. FLMs within *FREGAT* consider all known analytical issues of this relatively novel approach to simplify calculations and improve the interpretability of the results.

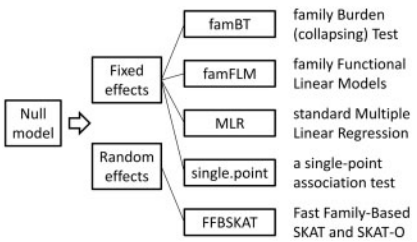

Fig. 1. Methods of region-based association analysis implemented in *FREGAT* package

Acknowledgement

We thank Dr Anatoly Kirichenko for technical support.

Funding

This work was supported by the Russian Foundation for Basic Research (13-04-00272, 14-04-00126, 16-04-00360), and Federal Agency of Scientific Organizations (VI.53.2.2, 0324-2015-0003).

Conflict of Interest: none declared.

References

Aulchenko,Y.S. *et al.* (2007) GenABEL: an R library for genome-wide association analysis. *Bioinformatics*, **23**, 1294–1296.

Belonogova,N.M. *et al.* (2013) Region-based association analysis of human quantitative traits in related individuals. *PLoS One*, **8**, e65395.

Calaway,R. *et al.* (2015a) Foreach: Foreach looping construct for R. R package version 1.4.3.

Calaway, R. *et al.* (2015b) doParallel: Foreach parallel adaptor for the 'parallel' package. R package version 1.0.10.

Chen,H. *et al.* (2013) Sequence kernel association test for quantitative traits in family samples. *Genet. Epidemiol.*, **37**, 196–204.

Clayton,D. (2015) snpStats: SnpMatrix and XSnpMatrix classes and methods. R package version 1.20.0.

Dering,C. *et al.* (2011) Statistical analysis of rare sequence variants: an overview of collapsing methods. *Genet. Epidemiol*, **35** (**suppl 1**), S12–S17.

Eichler,E.E. *et al.* (2010) Missing heritability and strategies for finding the underlying causes of complex disease. *Nat. Rev. Genet.*, **11**, 446–450.

Fan,R. *et al.* (2014) Generalized functional linear models for gene-based case-control association studies. *Genet. Epidemiol.*, **38**, 622–637.

Fan,R. *et al.* (2013) Functional linear models for association analysis of quantitative traits. *Genet. Epidemiol.*, **37**, 726–742.

Feng,S. *et al.* (2015) Methods for association analysis and meta-analysis of rare variants in families. *Genet. Epidemiol.*, **39**, 227–238.

Kwee,L.C. *et al.* (2008) A powerful and flexible multilocus association test for quantitative traits. *Am. J. Hum. Genet.*, **82**, 386–397.

Lee,S. *et al.* (2012) Optimal tests for rare variant effects in sequencing association studies. *Biostatistics*, **13**, 762–775.

Oualkacha,K. *et al.* (2013) Adjusted sequence kernel association test for rare variants controlling for cryptic and family relatedness. *Genet. Epidemiol.*, **37**, 366–376.

Schifano,E.D. *et al.* (2012) SNP set association analysis for familial data. *Genet. Epidemiol.*, **36**, 797–810.

Svishcheva,G.R. *et al.* (2014) FFBSKAT: fast family-based sequence kernel association test. *PLoS One*, **9**, e99407.

Svishcheva,G.R. *et al.* (2015) Region-based association test for familial data under functional linear models. *PLoS One*, **10**, e0128999.

Wu,M.C. *et al.* (2011) Rare-variant association testing for sequencing data with the sequence kernel association test. *Am. J. Hum. Genet.*, **89**, 82–93.

Zhan,X. and Liu,D.J. (2015) SEQMINER: an R-package to facilitate the functional interpretation of sequence-based associations. *Genet. Epidemiol.*, **39**, 619–623.