BIOINFORMATICS APPLICATIONS NOTE Vol. 30 no. 21 2014, pages 3134–3135
doi:10.1093/bioinformatics/btu435

Genetics and population analysis

Advance Access publication July 16, 2014

lrgpr: interactive linear mixed model analysis of genome-wide association studies with composite hypothesis testing and regression diagnostics in R

Gabriel E. Hoffman[1,2,3,*], Jason G. Mezey[3,4] and Eric E. Schadt[1,2]

[1]Department of Genetics and Genomic Sciences, [2]Icahn Institute for Genomics and Multiscale Biology, Icahn School of Medicine at Mount Sinai, New York, NY, USA, [3]Department of Biological Statistics and Computational Biology, Cornell University, Ithaca, USA and [4]Department of Genetic Medicine, Weill Cornell Medical College, New York, NY, USA

Associate Editor: Jeffrey Barrett

ABSTRACT

Summary: The linear mixed model is the state-of-the-art method to account for the confounding effects of kinship and population structure in genome-wide association studies (GWAS). Current implementations test the effect of one or more genetic markers while including prespecified covariates such as sex. Here we develop an efficient implementation of the linear mixed model that allows composite hypothesis tests to consider genotype interactions with variables such as other genotypes, environment, sex or ancestry. Our R package, lrgpr, allows interactive model fitting and examination of regression diagnostics to facilitate exploratory data analysis in the context of the linear mixed model. By leveraging parallel and out-of-core computing for datasets too large to fit in main memory, lrgpr is applicable to large GWAS datasets and next-generation sequencing data.

Availability and implementation: lrgpr is an R package available from lrgpr.r-forge.r-project.org

Contact: gabriel.hoffman@mssm.edu

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

Received on March 3, 2014; revised on June 12, 2014; accepted on July 2, 2014

1 INTRODUCTION

Genetic confounding owing to kinship and population structure is a common cause of inflation in genome-wide association studies (GWAS) test statistics and can lead to a substantial increase in false-positive results (Price *et al.*, 2010). Linear mixed models have been widely adopted to correct for genetic confounding in GWAS analysis (Kang *et al.*, 2010; Listgarten *et al.*, 2013; Long *et al.*, 2013; Svishcheva *et al.*, 2012; Zhou and Stephens, 2012), and the low-rank linear mixed model has advantages in terms of power and computational efficiency (Lippert *et al.*, 2011; Listgarten *et al.*, 2012). Yet, current software uses a 'one size fits all' paradigm where the analyst selects covariates and a genetic similarity metric, and the program performs a standard analysis on each genetic marker. As GWAS datasets have become larger and more complex, there is great potential for a custom analysis to identify biologically relevant associations not found by a standard analysis.

*To whom correspondence should be addressed.

2 METHODS

We have developed an efficient and user-friendly R package, lrgpr, that facilitates custom exploratory data analysis of GWAS datasets by combining the well-established linear mixed model with novel statistical, diagnostic and interactive functionality. The package's main function, `lrgpr()`, is designed with much of the same functionality as the standard `lm()` function for linear regression and takes advantage of R's interactive paradigm for exploratory data analysis (R Core Team, 2013). This function allows visualization of diagnostic plots that are essential for complex datasets to ensure that the regression model is not badly misspecified (Fox, 2008). Combining interactive analysis with model diagnostics allows the analyst to examine the relevance of additional covariates or non-linear effects of covariates on the phenotype. The `lrgpr()` function also provides composite hypothesis testing using a Wald statistic in the context of the linear mixed model to allow tests of epistasis as well as genotype interactions with other variables such as environment, sex or ancestry, where these variables are fit as fixed effects. The function is able to fit the linear mixed model using either a full- or low-rank genetic similarity matrix (Zhou and Stephens, 2012; Lippert *et al.*, 2011) and can learn the appropriate rank by cross-validation (Listgarten *et al.*, 2012) or model selection criteria (Hoffman, 2013).

For genome-wide analysis, the function `lrgprApply()` allows time and memory-efficient fitting of the linear mixed model for millions of genetic markers and can apply a composite hypothesis test on a large scale. This function applies a fast approximation, which reuses estimates from the null model (Lippert *et al.*, 2011), but has an option to use an exact method to reestimate variance components for each marker (Zhou and Stephens, 2012). Moreover, `lrgprApply()` can efficiently remove markers in the region being tested from the random effect to increase power (Listgarten *et al.*, 2012). The complementary function `glmApply()` fits fixed-effect linear and logistic models. These functions allow analysis of large datasets in parallel on multicore computers. They are designed to take advantage of the bigmemory package (Kane *et al.*, 2013) for out-of-core computing to efficiently process datasets that cannot fit into main memory.

3 FEATURES

The lrgpr package provides the following:

- Seamless interactive R interface to arbitrarily large datasets through `big.matrix` from the bigmemory package.
- Scalable fixed-effect linear or logistic regression for millions of hypothesis tests using `glmApply()`
- Fitting a full- or low-rank linear mixed model with `lrgpr()`

Table 1. Speed comparison of lrgpr with widely used programs for two simulated datasets

samples, markers	plink + 10 PC's[1]	EMMAX[2]	GEMMA[3]	FaST-LMM[4]	GRAMMAR-gamma[5]	mmscore[5]	GWFGLS[6]	lrgpr[7]
5 K, 500 K	103 m 1 s	45 m 49 s	221 m 11 s	25 m 22 s	6 m 3 s	147 m 53 s	206 m	17 m 47 s
10 K, 1 M	207 m 10 s	332 m 34s	1542 m 8 s	NA[8]	NA[9]	NA[9]	3287 m 6 s	198 m 39 s

Analysis was run with default settings on an 8 core Intel® Xeon® E5-2687W @ 3.10 GHz with 64 Gb RAM using R 3.1.0 compiled with the Intel® Math Kernel Library. Overhead for file conversion is not included.
[1]v1.07 [2]Multithreaded version from 2/10/2012 [3]v0.92 [4]v2.06.20130802 [5]GenABEL v1.8.0 [6]MixABEL v0.0.9.1 with DatABEL v0.1.6 [7]v0.1.0 [8]Requires more than 64Gb of memory [9]Dataset exceeds hardware-independent size limit that GenABEL can load.

- Data-adaptive construction of the genetic similarity matrix for the linear mixed model with `criterion.lrgpr()` and `cv.lrgpr()`
- Scalable linear mixed model regression for millions of hypothesis tests using `lrgprApply()`
- Ability to define arbitrary interaction models and perform composite hypothesis tests with `glmApply()`, `lrgpr()` and `lrgprApply()`

4 APPLICATION

The main contribution of the lrgpr software is its flexibility and integration into the R environment while being scalable to large datasets. This framework facilitates integration of existing analyses in R and rapid prototyping of novel methods. To illustrate its efficiency and scalability, we applied lrgpr to two simulated datasets of 5000 samples with 500 000 markers and 10 000 samples with 1 million markers. Although lrgpr is more flexible than other software, we ran the full-rank linear mixed model reusing variance component estimates from the null model to make a fair comparison between methods. All programs were run with default parameters using the same genetic similarity matrix. The runtimes required to fit a linear mixed model on this dataset are shown for lrgpr and six widely used programs, in addition to plink, which fits a fixed-effects linear model (Table 1). [Running GEMMA and FaST-LMM with the same grid search for estimating variance components as lrgpr uses increases the runtime (Supplementary Table S1).] The runtimes indicate that lrgpr, despite its flexible and user-friendly interface, is competitive with existing software.

5 DISCUSSION

As most analysis of GWAS datasets have been performed under the 'one size fits all' paradigm, there is great potential for a custom exploratory reanalysis to examine novel aspects of existing datasets to further elucidate the molecular mechanisms of complex traits. Moreover, Yang *et al.* (2014) emphasizes that the optimal analysis depends on the population stratification, kinship, sample size, genetic architecture, disease prevalence and study design of each dataset. The lrgpr package allows an analyst to apply multiple variations of existing methods to customize an analysis based on the empirical properties of a specific dataset.

This description of the lrgpr software is necessarily brief, and we provide a detailed tutorial illustrating the functionality of the software on the package's Web site.

ACKNOWLEDGEMENTS

The authors thank Roman Kosoy, Sushila Shenoy, Sarah Brooks, Cris Van Hout and Monica Ramstetter for feedback on the software, and Michael Kane for help with the bigmemory R package.

Funding: This work was supported by a fellowship from the Cornell Center for Comparative and Population Genomics, NSF grants IOS1026555 and DEB0922432 (Cornell University), and NIH grants R01AG046170 and R01MH095034, and a grant from The Leona M. and Harry B. Helmsley Charitable Trust (Icahn School of Medicine at Mount Sinai). This work was supported in part through the computational resources and staff expertise provided by Cornell University and the Department of Scientific Computing at the Icahn School of Medicine at Mount Sinai.

Conflict of interest: none declared.

REFERENCES

Fox,J. (2008) *Applied Regression Analysis and Generalized Linear Models*. 2nd edn. Sage, London.
Hoffman,G.E. (2013) Correcting for population structure and kinship using the linear mixed model: theory and extensions. *PLoS One*, **8**, e75707.
Kane,M.J. *et al.* (2013) Scalable strategies for computing with massive data. *J. Stat. Softw.*, **55**.
Kang,H.M. *et al.* (2010) Variance component model to account for sample structure in genome-wide association studies. *Nat. Genet.*, **42**, 348–354.
Lippert,C. *et al.* (2011) FaST linear mixed models for genome-wide association studies. *Nat. Methods*, **8**, 833–835.
Listgarten,J. *et al.* (2012) Improved linear mixed models for genome-wide association studies. *Nat. Methods*, **9**, 525–526.
Listgarten,J. *et al.* (2013) A powerful and efficient set test for genetic markers that handles confounders. *Bioinformatics*, **29**, 1526–1533.
Long,Q. *et al.* (2013) JAWAMix5: an out-of-core HDF5-based java implementation of whole-genome association studies using mixed models. *Bioinformatics*, **29**, 1220–1222.
Price,A.L. *et al.* (2010) New approaches to population stratification in genome-wide association studies. *Nat. Rev. Genet.*, **11**, 459–463.
R Core Team. (2013) R: a language and environment for statistical computing. R Foundation for Statistical Computing. Vienna, Austria.
Svishcheva,G.R. *et al.* (2012) Rapid variance components-based method for whole-genome association analysis. *Nat. Genet.*, **44**, 1166–1170.
Yang,J. *et al.* (2014) Advantages and pitfalls in the application of mixed-model association methods. *Nat. Genet.*, **46**, 100–106.
Zhou,X. and Stephens,M. (2012) Genome-wide efficient mixed-model analysis for association studies. *Nat. Genet.*, **44**, 821–824.