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

Gabriel E. Hoffman 1; Jason G. Mezey 2,3† and Eric E. Schadt 1‡

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

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 in order to consider genotype interactions with variables such as other genotypes, environment, sex, or ancestry. Our R package, Irgpr, allows interactive model fitting and examination of regression diagnostics in order 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, Irgpr is applicable to very large GWAS datasets and next-generation sequencing data.

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

Contact: gabriel.hoffman@mssm.edu

1 INTRODUCTION

Genetic confounding due 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 positives (Price et al., 2010). Linear mixed models have been widely adopted to correct for genetic confounding in GWAS analysis (Svishcheva et al., 2012; Zhou and Stephens, 2012; Kang et al., 2010; Listgarten et al., 2013; Long et al., 2013), and the low rank linear mixed model has advantages in terms of power and computational efficiency (Listgarten et al., 2012; Lippert et al., 2011). 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.

2 METHODS

We have developed an efficient and user-friendly R package, Irgpr, 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 nonlinear 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 very fast approximation which reuses estimates from the null model (Lippert et al., 2011), but has an option to use an exact method to re-estimate variance components for each marker (Zhou and Stephens, 2012). Moreover, lrgprApply() can efficiently remove markers in the region being tested from the random effect in order 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 multi-core computers. They are designed to take advantage of the bigmemory package (Kane et al., 2013) for out-of-core computing in order to efficiently process datasets that cannot fit into main memory.

3 FEATURES

The Irgpr package provides:

- Seamless, interactive R interface to arbitrarily large datasets through big.matrix from the bigmemory package.

* to whom correspondence should be addressed
† indicates co-senior authors

© The Author (2014). Published by Oxford University Press. All rights reserved. For Permissions, please email: journals.permissions@oup.com
Table 1. Speed comparison of lrgpr with widely used programs for two simulated datasets. Analysis was run with default settings on an 8 core Intel® Xeon® ES-2687W @ 3.10GHz with 64 GB RAM using R 3.1.0 compiled with the Intel® Math Kernel Library. Overhead for file conversion is not included.

<table>
<thead>
<tr>
<th>samples, markers</th>
<th>plink + 10 PC’s¹</th>
<th>EMMAX²</th>
<th>GEMMA³</th>
<th>FaST-LMM⁴</th>
<th>GRAMMAR-gamma⁵</th>
<th>mmscore⁶</th>
<th>GWFGLS⁷</th>
<th>lrgpr⁸</th>
</tr>
</thead>
<tbody>
<tr>
<td>5K, 500K</td>
<td>103m 1s</td>
<td>45m 49s</td>
<td>221m 11s</td>
<td>25m 22s</td>
<td>6m 3s</td>
<td>147m 53s</td>
<td>206m</td>
<td>17m 47s</td>
</tr>
<tr>
<td>10K, 1M</td>
<td>207m 10s</td>
<td>332m 34s</td>
<td>1542m 8s</td>
<td>NA⁹</td>
<td>NA⁹</td>
<td>NA⁹</td>
<td>3287m 6s</td>
<td>198m 39s</td>
</tr>
</tbody>
</table>

¹v1.07 ²Multithreaded version from 2/10/2012 ³v0.92 ⁴v2.06.20130802 ⁵GenABEL v1.8.0 ⁶MixABEL v0.0.9.1 with DatABEL v0.1.6 ⁷v1.0.9.⁸Requires more than 64Gb of memory ⁹Dataset exceeds hardware-independent size limit that GenABEL can load

- 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()
- 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. In order to illustrate its efficiency and scalability, we applied lrgpr to two simulated datasets of 5,000 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 in order 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 Lrgr and six widely used programs, in addition to plink, which fits a fixed-effects linear model (Table 1). (We note that running GEMMA and FaST-LMM with the same grid search for estimating variance components as lrgpr uses increases the runtime (Supplementary Table 1).) The runtimes indicate that lrgpr, despite its flexible and user-friendly interface, is very 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 re-analysis 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 in order 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 website.

ACKNOWLEDGEMENT

We 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 NSF grants IOS1026555 and DEB0922432. This work was supported in part through the computational resources and staff expertise provided by the Department of Scientific Computing at the Icahn School of Medicine at Mount Sinai.

REFERENCES


