Bioinformatics, 35(24), 2019, 5346–5348
doi: 10.1093/bioinformatics/btz567
Advance Access Publication Date: 22 July 2019
Applications Note

Genetics and population analysis

Genetic association testing using the GENESIS R/Bioconductor package

Stephanie M. Gogarten [1,*], Tamar Sofer [2,3], Han Chen [4,5], Chaoyu Yu[1], Jennifer A. Brody[6], Timothy A. Thornton[1], Kenneth M. Rice[1] and Matthew P. Conomos [1,*]

[1]Department of Biostatistics, University of Washington, Seattle, WA 98195, USA, [2]Division of Sleep and Circadian Disorders, Department of Medicine, Brigham and Women's Hospital, Boston, MA, USA, [3]Department of Biostatistics, Harvard T. H. Chan School of Public Health, Boston, MA, USA, [4]Human Genetics Center, Department of Epidemiology, Human Genetics and Environmental Sciences, School of Public Health, The University of Texas Health Science Center at Houston, Houston, TX 77030, USA, [5]Center for Precision Health, School of Public Health and School of Biomedical Informatics, The University of Texas Health Science Center at Houston, Houston, TX 77030, USA and [6]Cardiovascular Health Research Unit, Department of Medicine, University of Washington, Seattle, WA 98101, USA

*To whom correspondence should be addressed.

Associate Editor: Alfonso Valencia

Received on February 14, 2019; revised on May 24, 2019; editorial decision on July 15, 2019; accepted on July 16, 2019

Abstract

Summary: The Genomic Data Storage (GDS) format provides efficient storage and retrieval of genotypes measured by microarrays and sequencing. We developed GENESIS to perform various single- and aggregate-variant association tests using genotype data stored in GDS format. GENESIS implements highly flexible mixed models, allowing for different link functions, multiple variance components and phenotypic heteroskedasticity. GENESIS integrates cohesively with other R/Bioconductor packages to build a complete genomic analysis workflow entirely within the R environment.

Availability and implementation: https://bioconductor.org/packages/GENESIS; vignettes included.

Contact: sdmorris@uw.edu or mconomos@uw.edu

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

1 Introduction

The Genomic Data Storage (GDS) format provides efficient storage and retrieval of genotype and annotation data for bi-allelic and multi-allelic genomic variants called from microarrays, sequencing or genotype imputation. A suite of tools for genomic analysis that utilizes the computational efficiency afforded by GDS format has been built through a collection of R/Bioconductor packages (Gogarten *et al.*, 2012; Zheng *et al.*, 2012, 2017). These packages are used to convert genotype data stored in other formats (e.g. VCF, PLINK) to GDS; perform sample- and variant-level quality control; efficiently compute kinship coefficients and ancestry principal components (PCs) and perform various other genomic analyses.

Here, we present GENESIS, an R/Bioconductor package that performs mixed-model based single- and aggregate-variant association tests for quantitative, binary and count phenotypes, using genotype data stored in GDS format. With the inclusion of GENESIS, a comprehensive GDS-based genomic analysis pipeline, from data formatting to association testing, can be built entirely within the R environment.

To meet the memory limit of a given computer and allow for parallelization, GENESIS uses iterator classes defined in the R/Bioconductor packages GWASTools and SeqVarTools to read and manipulate data from 'blocks' of consecutive variants on the GDS file. The user defines an iterator object (GenotypeIterator for microarray or SeqVarIterator for sequencing) that provides a connection to the GDS file, along with the size of the block (number of variants) and potential filters to apply prior to reading the genotypes. To allow GENESIS association tests to work with both iterator types, we designed the software to separate the code that reads, or iterates, through large datasets from the code that performs computations and statistical tests on blocks of data. This also facilitates easier code maintenance, unit testing and implementation of new features.

2 Genetic association testing

We focus here on implementation of mixed models, which are the prevailing method for genetic association testing in typical population studies. (Linear and logistic regression, used for unrelated samples, are special cases of mixed models. These can easily be handled in GENESIS; see the Supplementary Material for an example and performance data.) Mixed model genetic association testing with GENESIS comprises three steps: (i) inferring population structure and relatedness, (ii) fitting the null model and (iii) testing variants for association, either individually or in aggregate.

2.1 Population structure and relatedness inference

GENESIS implements the PC-AiR (Conomos *et al.*, 2015) and PC-Relate (Conomos *et al.*, 2016a) methods, which together provide accurate population structure inference and kinship estimation. PC-AiR relies on efficient computation of ancestry PCs provided by the SNPRelate package.

2.2 Fitting a null model

Fitting mixed models on large samples is computationally expensive, requiring significant memory and CPU time. As fitting a mixed model for each variant genome-wide is prohibitively expensive, it is standard to fit one 'null model' under the null hypothesis of no genetic association, and subsequently use score tests to assess each variant's association.

GENESIS fits linear mixed models for quantitative phenotypes and generalized linear mixed models for binary and count phenotypes via the penalized quasi-likelihood (Breslow and Clayton, 1993) approach of Chen *et al.* (2016). Both use the average information REML procedure (Gilmour *et al.*, 1995). Sample dependence due to genetic similarity is accounted for by including a random effects term with covariance matrix proportional to a genetic relationship matrix (GRM) or kinship matrix (KM). When using a KM, ancestry representative vectors such as PCs should also be included as fixed effects in the null model to adjust for population structure.

Unlike much available software, GENESIS allows multiple variance components and does not restrict their form. In addition to the standard GRM/KM term, other variance components can be used to account for e.g. shared environments (Conomos *et al.*, 2016b), as well as phenotypic heteroskedasticity by study subgroup (Snijders and Berkhof, 2008).

2.3 Testing variants for association

GENESIS can test genetic variants for association either individually or in aggregate. Required output from the null model is stored efficiently to expedite computation of these tests. Score tests and approximations to Wald tests are provided for single variants. Available aggregate-variant tests include burden, SKAT (Chen *et al.*, 2013; Wu *et al.*, 2011), SKAT-O (Lee *et al.*, 2012), fastSKAT (Lumley *et al.*, 2018) and SMMAT (Chen *et al.*, 2019) methods. Aggregate units can be defined using a sliding window approach, or can be customized; e.g. genes or pathways. Variant weights can be specified either as a function of minor allele frequency via a beta distribution; e.g. Wu *et al.* (2011), or customized by the user; e.g. utilizing annotation features such as CADD scores (Kircher *et al.*, 2014).

3 Sparse GRM/KM for efficient computation

Even fitting only one null model per analysis, the computational burden in large samples may still be prohibitive. One reason is matrix inversion, which has long been known to be a hurdle (Thompson and Shaw, 1990). For estimating the variance components, it is efficient to treat each pedigree as its own cluster; i.e. to add up over the pedigrees (O'Connell, 2014). GENESIS implements such an analysis by using a sparse, block-diagonal GRM/KM; the R package Matrix (Bates and Maechler, 2018) is used for sparse matrix storage and linear algebra methods. With sparsity, the computational complexities of the null model and score tests are $O(K\tilde{N}^3)$ and $O(K\tilde{N}^2M)$, respectively, where K is the number of clusters, $\tilde{N}$ is the maximum cluster size and M is the number of variants tested. In practice, these are reduced to $O(\tilde{N}^3)$ and $O(\tilde{N}^2M)$ when one largest cluster dominates, or to $O(K)$ and $O(KM)$ when there are many small clusters (e.g. a study of trios).

A pedigree-based KM is sparse by nature, but pedigrees are often unavailable or incomplete. In contrast, an empirical GRM/KM estimated from genotype data captures all relatedness, but is dense, with no entries equal to 0. When computational burden is a concern, an empirical GRM/KM can be made sparse: we recommend grouping samples such that any pair with an estimated relatedness greater than a specified threshold is in the same cluster. All pairwise estimates within a cluster are kept, even if they are below the threshold. All pairwise estimates between clusters are set to 0, creating a sparse, block-diagonal matrix.

To illustrate the computational advantage of using sparse matrices, we analyzed a simulated heritable quantitative trait measured on 100 000 samples. We compared the computational performance of fitting the null model using a dense GRM/KM to that of using the same GRM/KM made sparse with varying cluster sizes (Supplementary Fig. S1). Compared to the dense GRM/KM, the analysis using the sparse GRM/KM with clusters of 1000 samples took 0.6% of the CPU time (24 min versus 67 h) and 2.0% of the memory (16 GB versus 820 GB).

To investigate the statistical impact of sparsity, we compared association P values of ∼24 M variants when using different empirical GRMs, KMs and PCs to account for structure in a heritable quantitative trait simulated on 2504 samples from 1000 Genomes (Supplementary Table S3 and Supplementary Fig. S2). The dense KM was made sparse at a 5th degree relatedness threshold (≈0.011) using the recommended algorithm; this sparse KM had 2236 clusters, of which 2080 were singletons and the largest had 23 members. The differences in P values when using this sparse KM rather than the dense KM were small; 99.9% of variants had differences in $-\log_{10}(p) < 0.25$, over 99.9999% had differences <0.5 and the maximum difference was only 0.59. Given a variant with a 'true' P value of 5.0×10^{-8}, a difference in $-\log_{10}(p) < 0.25$ would correspond to a reported P value $\in (8.9 \times 10^{-8}, 2.8 \times 10^{-8})$.

4 Discussion

GENESIS adds to an existing collection of R/Bioconductor packages to provide a cohesive, computationally efficient set of genomic analysis tools that utilize GDS format, all within the R environment. Workflows may utilize the extensive collection of genome annotations and associated packages that are part of Bioconductor for tasks such as assigning rare variants to genes in an aggregate association test. Working examples are provided in the GENESIS package vignettes.

Making an empirical GRM/KM sparse can be thought of as approximating low levels of relatedness as 'unrelated.' In the simulations presented here, we observed that the computational gain afforded by sparsity is significant in large samples, and the differences in calculated *P* values are unlikely to change the impact of results. We expect this to generally hold true, but the magnitude of the impact of this approximation will likely depend on the genetic architecture of the phenotype; we plan to explore this further in future work.

The structure of the GENESIS code allows straightforward addition of new methods and algorithms. Features currently in development include saddle point approximations to calibrate logistic mixed model *P* values when there is case–control imbalance (Dey *et al.*, 2017; Zhou *et al.*, 2018).

Funding

The work presented in this paper was supported by the NHLBI through the TOPMed Data Coordinating Center (3R01HL-120393-02S1) and NHLBI grant U01HL137162.

Conflict of Interest: none declared.

References

Bates,D. and Maechler,M. (2018) *Matrix: Sparse and Dense Matrix Classes and Methods*. R package version 1.2-14. https://CRAN.R-project.org/package=Matrix (April 2018, date last accessed).

Breslow,N.E. and Clayton,D.G. (1993) Approximate inference in generalized linear mixed models. *J. Am. Stat. Assoc.*, **88**, 9–25.

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

Chen,H. *et al.* (2016) Control for population structure and relatedness for binary traits in genetic association studies via logistic mixed models. *Am. J. Hum. Genet.*, **98**, 653–666.

Chen,H. *et al.* (2019) Efficient variant set mixed model association tests for continuous and binary traits in large-scale whole genome sequencing studies. *Am. J. Hum. Genet.*, **104**, 260–274.

Conomos,M.P. *et al.* (2015) Robust inference of population structure for ancestry prediction and correction of stratification in the presence of relatedness. *Genet. Epidemiol.*, **39**, 276–293.

Conomos,M.P. *et al.* (2016a) Genetic diversity and association studies in us Hispanic/Latino populations: applications in the Hispanic Community Health Study/Study of Latinos. *Am. J. Hum. Genet.*, **98**, 165–184.

Conomos,M.P. *et al.* (2016b) Model-free estimation of recent genetic relatedness. *Am. J. Hum. Genet.*, **98**, 127–148.

Dey,R. *et al.* (2017) A fast and accurate algorithm to test for binary phenotypes and its application to PheWAS. *Am. J. Hum. Genet.*, **101**, 37–49.

Gilmour,A.R. *et al.* (1995) Average information REML: an efficient algorithm for variance parameter estimation in linear mixed models. *Biometrics*, **51**, 1440–1450.

Gogarten,S.M. *et al.* (2012) GWASTools: an R/Bioconductor package for quality control and analysis of genome-wide association studies. *Bioinformatics*, **28**, 3329–3331.

Kircher,M. *et al.* (2014) A general framework for estimating the relative pathogenicity of human genetic variants. *Nat. Genet.*, **46**, 310.

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

Lumley,T. *et al.* (2018) FastSKAT: sequence kernel association tests for very large sets of markers. *Genet. Epidemiol.*, **42**, 516–527.

O'Connell,J. (2014) *MMAP User Guide*. University of Maryland, Baltimore, MA.

Snijders,T.A. and Berkhof,J. (2008) Diagnostic checks for multilevel models. In: *Handbook of Multilevel Analysis*. Springer, New York, NY, pp. 141–175.

Thompson,E. and Shaw,R. (1990) Pedigree analysis for quantitative traits: variance components without matrix inversion. *Biometrics*, **46**, 399–413.

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.

Zheng,X. *et al.* (2017) SeqArray—a storage-efficient high-performance data format for WGS variant calls. *Bioinformatics*, **33**, 2251–2257.

Zheng,X. *et al.* (2012) A high-performance computing toolset for relatedness and principal component analysis of SNP data. *Bioinformatics*, **28**, 3326–3328.

Zhou,W. *et al.* (2018) Efficiently controlling for case-control imbalance and sample relatedness in large-scale genetic association studies. *Nat. Genet.*, **50**, 1335.