BIOINFORMATICS APPLICATIONS NOTE Vol. 28 no. 24 2012, pages 3329–3331
doi:10.1093/bioinformatics/bts610

Genetics and population analysis

Advance Access publication October 10, 2012

GWASTools: an R/Bioconductor package for quality control and analysis of genome-wide association studies

Stephanie M. Gogarten[1,*], Tushar Bhangale[1,2], Matthew P. Conomos[1], Cecelia A. Laurie[1], Caitlin P. McHugh[1], Ian Painter[3], Xiuwen Zheng[1], David R. Crosslin[1], David Levine[1], Thomas Lumley[1,4], Sarah C. Nelson[1], Kenneth Rice[1], Jess Shen[1], Rohit Swarnkar[1], Bruce S. Weir[1] and Cathy C. Laurie[1]

[1]Department of Biostatistics, University of Washington, Seattle, WA, [2]Department of Bioinformatics and Computational Biology, Genentech Inc, South San Francisco, CA, [3]Department of Health Services, University of Washington, Seattle, WA, USA and [4]Department of Statistics, University of Auckland, Auckland, New Zealand

Associate Editor: Jeffrey Berrett

ABSTRACT

Summary: GWASTools is an R/Bioconductor package for quality control and analysis of genome-wide association studies (GWAS). GWASTools brings the interactive capability and extensive statistical libraries of R to GWAS. Data are stored in NetCDF format to accommodate extremely large datasets that cannot fit within R's memory limits. The documentation includes instructions for converting data from multiple formats, including variants called from sequencing. GWASTools provides a convenient interface for linking genotypes and intensity data with sample and single nucleotide polymorphism annotation.

Availability and implementation: GWASTools is implemented in R and is available from Bioconductor (http://www.bioconductor.org). An extensive vignette detailing a recommended work flow is included.

Contact: sdmorris@uw.edu

Received on August 2, 2012; revised on October 3, 2012; accepted on October 6, 2012

1 INTRODUCTION

Despite the increasing popularity of next-generation sequencing, the relatively low cost of chip arrays and the recent introduction of exome chips continue to make genome-wide association studies (GWAS) an attractive option for many investigators. The success of GWAS depends in large part on rigorous quality control (QC) and quality assurance (QA) to ensure that false-positive signals are removed and that real signals are not masked by low-quality data (Laurie *et al.*, 2010). We present GWASTools, an R/Bioconductor package to facilitate QC/QA of large single nucleotide polymorphism (SNP) datasets.

As SNP arrays have increased in size (commonly >1 million SNPs), it is often impossible to store an entire dataset in R's virtual memory, even at maximum compression of one byte per genotype. Analysing floating-point values, such as allelic probe intensity, requires even more storage space. Existing R packages for GWAS, such as snpStats (Clayton and Leung, 2007), store genotype matrices in memory and do not include intensity data. We use the network common data form (NetCDF) format to store large genotype and intensity datasets on disk with efficient access to subsets of the data. For an example dataset of 5000 samples and 1 000 000 SNPs (which exceeds R's limit of 2^{31} elements in a single array), iterative read time for a NetCDF file is up to 22 times faster than a PED file.

The benefits of GWASTools include the interactive analysis provided by R's interface and the ability to include intensity data. Intensity data can be used to detect sex chromosome aneuploidies (which can be confused with sex mis-annotation), autosomal anomalies (which generate genotyping errors) and evaluation of clustering by genotype call. GWASTools supports the following format conversions: input—plain text, PLINK, Variant Call Format, imputed genotypes (IMPUTE2, BEAGLE, MaCH); output—PLINK, snpStats objects.

2 DATA STRUCTURES

GWASTools uses R's S4 object system to provide formally defined methods and classes for GWAS data and annotation.

Genotype and intensity data are stored in NetCDF files with dimensions (SNP, sample) and are accessed through the NcdfGenotypeReader and NcdfIntensityReader classes. The ncdf package is used by these classes as the base-level interface to the NetCDF files.

Both the SNP and sample dimensions in a NetCDF file have associated annotation. For SNPs, this usually includes chromosome, position, reference SNP ID (rsID) and alleles. For samples, annotation can include sex, phenotype, pedigree and genotyping batch. Often the same sample is genotyped more than once to use the concordance between the two genotyping instances (or 'scans') as a QC measure. Each SNP and scan is given a unique integer ID that serves as the primary key between the NetCDF files and the annotation. SNP and scan annotations are stored in instances of SnpAnnotationDataFrame and ScanAnnotationDataFrame, respectively, which extend the AnnotatedDataFrame class in the Biobase package. These classes store both annotation values and metadata (a description of each annotation variable). Some variables are required (snpID, chromosome and position in an

*To whom correspondence should be addressed.

SnpAnnotationDataFrame and scanID in a ScanAnnotation DataFrame), but any number of other variables of different types may also be included.

An important component of GWAS analysis is ensuring that annotation is mapped correctly to its corresponding genotype data. The classes GenotypeData and IntensityData each contain a NetCDF reader object, a SNP annotation object and a scan annotation object. On object creation, validity methods ensure that the primary keys of the NetCDF and annotation objects match. Most functions in GWASTools take GenotypeData and IntensityData objects as arguments, such that all necessary information (e.g. genotype, chromosome and sex) is contained in a single object. Common methods are provided to access data across multiple classes, including getSnpID, getScanID, getChromosome, getPosition, getSex and getGenotype.

In designing GWASTools, we took care to separate the application programming interface of the GenotypeData and IntensityData classes from the format in which the data are stored. This design allows for extension to other data formats without changing any of the functions that use GenotypeData and IntensityData objects. For example, GWASTools contains classes for storing SNP and scan annotation in SQLite format, and for storing genotype data in-memory as a matrix or on disk in the highly efficient genomic data structure (GDS) format used by SNPRelate (Zheng *et al*., 2012). A GenotypeData object can be constructed with these objects as well. In addition to the functionality already included in GWASTools, one could make use of the GenotypeData and IntensityData classes to apply the many statistical functions available in R to GWAS data.

3 AVAILABLE FUNCTIONALITY

In this section, we present some common tasks utilizing the GWASTools package.

3.1 Creating NetCDF files

NetCDF files can be created from data in plain text format, such as files for Illumina and Affymetrix arrays released by a genotyping centre or from PLINK PED/MAP files.

3.2 Statistics by SNP or scan

Allele frequency, missing call rate or heterozygosity can all be calculated by SNP and by scan, with the option to exclude certain SNPs or scans. For sex chromosomes, male and female subjects are considered separately. Mean intensity can be calculated by scan.

3.3 Batch quality

Missing call rate by scan can be compared with genotyping batch to look for outliers. The allele frequency of each batch compared with all other batches can also be an indicator of problematic batches.

3.4 Chromosome anomalies

The B allele frequency (BAF) and log R ratio (LRR) are useful metrics for detecting chromosome anomalies (Conlin *et al*., 2010; Peiffer *et al*., 2006). BAF and LRR values may be calculated from the genotype and intensity data if they are not included in the raw data files. GWASTools uses circular binary segmentation (Venkatraman and Olshen, 2007) to find change points in BAF and detects loss of heterozygosity by identifying homozygous runs with change in LRR. See Laurie *et al*. (2012) for an application of this method.

3.5 Relatedness

GWASTools contains functions for checking pedigrees for accuracy, as well as inferring pairwise relationships from and plotting kinship coefficients. The companion package SNPRelate is recommended for relatedness and population structure, and its native data format may be used in GWASTools.

3.6 SNP quality

SNP quality can be addressed by (i) checking the genotyping concordance between duplicate scans of the same sample; (ii) looking for Mendelian errors in families; and (iii) finding deviations from Hardy–Weinberg equilibrium in control subjects. Other useful filters involve removing SNPs with poor genotyping quality scores, high missing call rate, low minor allele frequency and sex differences in allele frequency and heterozygosity.

3.7 Association tests

Regression tests (on genotype calls and imputed dosages) and survival analysis can be performed.

3.8 Plotting

GWASTools contains many plotting functions, including genotype cluster plots, BAF/LRR plots with chromosome ideograms, quantile–quantile plots and Manhattan plots.

3.9 Summary and Performance

For a full description of recommended QC methods and GWASTools functionality, see Laurie *et al*. (2010) and the vignette 'GWAS Data Cleaning' supplied with GWASTools.

We compared the performance of GWASTools against PLINK on a system with a 2.66 Ghz Intel Core i5 processor with 16 GB RAM and running Mac OS X 10.6.8. The test dataset had 269 samples and 1 134 514 SNPs. (i) Missing call rate by SNP: GWASTools 72s, PLINK with PED format 352s, PLINK with binary format 153s. (ii) Mendelian errors: GWASTools 199s, PLINK with PED format 382s, PLINK with binary format 181s.

ACKNOWLEDGEMENTS

GWASTools was developed and tested using data from the Gene-Environment Association Studies Consortium (GENEVA, Cornelis *et al*., 2010, http://www.genevastudy.org). The authors thank the anonymous reviewers for the helpful comments that improved the article.

Funding: National Institutes of Health, GENEVA Coordinating Center (U01 HG 004446); GARNET Coordinating Center (U01 HG 005157).

Conflict of Interest: none declared.

REFERENCES

Clayton,D. and Leung,H.-T. (2007) An R package for analysis of whole-genome association studies. *Hum. Hered.*, **64**, 45–51.

Conlin,L.K. *et al.* (2010) Mechanisms of mosaicism, chimerism and uniparental disomy identified by single nucleotide polymorphism array analysis. *Hum. Mol. Genet.*, **19**, 1263–1275.

Cornelis,M.C. *et al.* (2010) The gene, environment association studies consortium (GENEVA): maximizing the knowledge obtained from GWAS by collaboration across studies of multiple conditions. *Genet. Epidemiol.*, **34**, 364–372.

Laurie,C.C. *et al.* (2010) Quality control and quality assurance in genotypic data for genome-wide association studies. *Genet. Epidemiol.*, **34**, 591–602.

Laurie,C.C. *et al.* (2012) Detectable clonal mosaicism from birth to old age and its relationship to cancer. *Nat. Genet.*, **44**, 642–650.

Peiffer,D.A. *et al.* (2006) High-resolution genomic profiling of chromosomal aberrations using Infinium whole-genome genotyping. *Genome Res.*, **16**, 1136–1148.

Venkatraman,E.S. and Olshen,A.B. (2007) A faster circular binary segmentation algorithm for the analysis of array CGH data. *Bioinformatics*, **23**, 657–663.

Zheng,X. *et al.* (2012) A high-performance computing toolset for relatedness and principal component analysis of SNP data. *Bioinformatics* [Epub ahead of print, doi:10.1093/bioinformatics/bts606, October 11, 2012].