BIOINFORMATICS APPLICATIONS NOTE Vol. 27 no. 13 2011, pages 1871–1872
doi:10.1093/bioinformatics/btr301

Genome analysis

Advance Access publication May 17, 2011

Genome-wide association studies pipeline (GWASpi): a desktop application for genome-wide SNP analysis and management

Fernando Muñiz-Fernandez[1,2], Angel Carreño-Torres[1,2], Carlos Morcillo-Suarez[1,2,3] and Arcadi Navarro[1,2,3,4,5,*]

[1]Institut de Biología Evolutiva, Universitat Pompeu Fabra, Biomedical Research Park (PRBB), Barcelona, [2]Population Genomics Node (GNV8) National Institute for Bioinformatics (INB), Barcelona, [3]National Genotyping Centre (CeGen), Barcelona, [4]Institucio Catalana de Recerca i Estudis Avançats, ICREA, Barcelona and [5]Departament de Ciències Experimentals i de la Salut, Universitat Pompeu Fabra, Dr Aiguader 88, 08003 Barcelona, Spain

Associate Editor: Dmitrij Frishman

ABSTRACT

Motivation: Genome-wide association studies (GWAS) based on single nucleotide polymorphism (SNP) arrays are the most widely used approach to detect loci associated to human traits. Due to the complexity of the methods and software packages available, each with its particular format requiring intricate management workflows, the analysis of GWAS usually confronts scientists with steep learning curves. Indeed, the wide variety of tools makes the parsing and manipulation of data the most time consuming and error prone part of a study. To help resolve these issues, we present GWASpi, a user-friendly, multiplatform, desktop-able application for the management and analysis of GWAS data, with a novel approach on database technologies to leverage the most out of commonly available desktop hardware. GWASpi aims to be a start-to-finish GWAS management application, from raw data to results, containing the most common analysis tools. As a result, GWASpi is easy to use and reduces in up to two orders of magnitude the time needed to perform the fundamental steps of a GWAS.

Availability: Freely available on the web at http://www.gwaspi.org. Implemented in Java, Apache-Derby and NetCDF-3, with all major operating systems supported.

Contact: gwaspi@upf.edu; arcadi.navarro@upf.edu

Received on January 3, 2011; revised on April 26, 2011; accepted on May 9, 2011

1 INTRODUCTION

Genome-wide association studies (GWAS) have come to be the choice method to detect loci associated with human hereditary traits, especially diseases (McCarthy *et al.*, 2008). The number of studies published yearly based on these arrays, has constantly increased from 3 in 2003 to 384 in 2010 (Yu *et al.*, 2008). In parallel, reference databases such as HapMap (International and Consortium, 2003), HGDP (Cavalli-sforza, 2005) and the 1000 Genomes Project (The 1000 Genomes Project Consortium, 2010) are being released together with an ever-growing range of analytical methods and software packages.

The most common freely available tools in the field, however, leave it to the user to tackle the jungle of formats and the bulk of raw data generated by GWAS. Also, learning how to apply the methods commonly used in GWASs takes significant time, extending an already lengthy and arduous data gathering phase. Thus, the steep learning curve and the burden of manipulation of the raw data still makes the access to GWAS a costly endeavour for researchers or institutions without Bioinformatics personnel. To contribute to solve this problem and make GWAS an achievable effort for smaller teams, as well as for the sake of speeding up raw data management in a consistent, self-contained way for the general researcher community, we have developed the GWAS Pipeline (GWASpi).

GWASpi has not been designed to replace other well-established applications (Purcell *et al.*, 2007; Browning and Browning, 2007) insofar as it does not offer the whole breadth of statistical analysis methods used in GWAS. Instead, the current version of GWASpi focuses in case–control studies and offers a solution for teams in need of concise and quickly obtained results as well as compact, comprehensive and agile dataset management to complement existing analysis packages. Some basic quality-control features are also integrated in GWASpi.

2 APPROACH

Genotype data consists of dense, static information, in the sense that it is formed by large numbers of genotypes (up to billions) that may need to be accessed many times but do not change once they have been ascertained. Subsets of data may be required some times for separate analysis, but they can be generated easily and quickly as new, independent datasets that, again, will be static. The NetCDF-3 database technology developed by the University Corporation for Atmospheric Research (UCAR, http://www2.ucar.edu) and originally intended for meteorological data storage has proven to be well suited to this kind of data. On top of this database, Java's LinkedHashMap objects are used to retrieve slabs of data in an ordered list fashion that is searchable by index.

3 METHODS

The efficient storage of genotypes and the posterior retrieval of specific subsets of this data is a highly desirable feature for the types of analysis that are being performed on these datasets. Generally, this type of functionality is achieved through the use of Relational Databases (RDB) with the help of SQL language. Nevertheless, RDB technologies have poor scaling properties for the type of data that composes genome-wide studies. GWASpi makes

*To whom correspondence should be addressed.

use of an embedded JavaDB SQL database for its internal management as well as for storage of data related to samples (such as their disease status or geographic origin). However, a crucial feature of GWASpi is that all genotype data and analysis results, which constitute the bulk of information in a GWAS, are stored in the Array-oriented Scientific Data Format (http://www.unidata.ucar.edu/software/netcdf/). Among the existing implementations of this technology, NetCDF-3 was chosen over others for the availability of a pure Java implementation of its API, which ensures cross-platform portability in conjunction with the availability of many high-performance libraries and tools (JfreeChart, Apache Derby). Plots and charts are handled by the JFreeChart library.

3.1 Common usage

From the point of view of the user, GWASpi is a cross-platform Java application with integrated tools and features that make the processing of genotype data more agile and speedy. All the features included in GWASpi are accessible through a self-teaching, user-friendly, graphical interface, while the heaviest functionalities can also be launched via a command-line interface. The basic usage of GWASpi starts by uploading raw genotype data exported from a proprietary genotyping platform or standard formats such as Affymetrix GW6.0, Illumina, PLINK, HapMap, etc. Once the data has been imported into GWASpi's database, the standard quality controls are performed, such as computing the number of missing genotypes per sample and per SNP, controlling for allelic mismatch and measuring individual heterozygosity. Next, a genotype frequency count is executed, which automatically includes a Hardy–Weinberg quality control. Finally, allelic, genotypic and trend association tests are performed, which generate tables, reports and their corresponding Manhattan and QQ-plots. A navigator is available to view the detail of a chromosome area around a given marker, providing a first glimpse of the results of the GWAS. These tables and plots also link each marker to a selection of reference databases such as Ensembl, NCBI's dbSNP and the UCSC genome browser among others.

3.2 Performance

The general application performance parameters benchmark as follows:

RAM usage: GWASpi, as a Java application, gets assigned a fixed size of the system's RAM and will never request more than this number. The application has been designed to work on common desktop or laptop hardware, typically using 2 GB of RAM per 10^6 SNPs with the number of samples only being limited by available disk space.

Disk usage: the disk usage increase is proportional to the number of genotypes, which is equal to the number of SNPs times the sample size. A genotype matrix 2×10^6 SNPs and 1000 samples large occupies about 4 GB on disk.

Speed: depending on the format and the number of genotypes, the loading of a dataset into the GWASpi can take as little as a few minutes (using GWASpi's own format) up to hours. As a guideline, a study with 900.000 SNPs and 1000 samples in Affymetrix GW 6.0 format takes a total of about 4 h for data load, quality control, Hardy–Weinberg and association tests as well as for producing the corresponding reports and charts.

4 CONCLUSION

The integration of key pre-processing, quality control, data management, analysis and reporting steps within a single application is a distinctive advantage we present in GWASpi. With this application, we want to offer a solid and scalable platform, even if the current version focuses in basic case–control designs. New quality-control and analysis methods as well as new formats are scheduled to be added into GWASpi. High-quality charts and reports sorted by relevance are generated automatically, delivering the results visually and in tables. Reference databases and chart navigation tools are integrated in the application providing an agile starting point for continuing research. For the user, GWASpi simplifies drastically the cumbersome workflows so common in GWAS pipelines and the many intermediate data files to be generated, checked and re-processed. The learning curve for GWASpi's usage is very accessible as all available operations are documented in an online contextual help page.

ACKNOWLEDGEMENTS

We are indebted to Francesc Calafell and Hafid Laayouni for their valuable discussion. Thanks are given to UCAR's NetCDF team for providing a method to suit GW matrices.

Funding: Spanish National Institute for Bioinformatics (www.inab.org); MICINN PSE (PSS-010000-2009-1 to A.N.).

Conflict of Interest: none declared.

REFERENCES

Browning,S.R. and Browning,B.L. (2007) Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. *Am. J. Hum. Genet.*, **81**, 1084–1097.

Cavalli-sforza,L.L. (2005) The Human Genome Diversity Project: past, present and future. *Genetics*, **6**, 3–10.

International HapMap Consortium (2003) The International HapMap Project. *Nature*, **426**, 789–796.

McCarthy,M.I. *et al.* (2008) Genome-wide association studies for complex traits: consensus, uncertainty and challenges. *Nat. Rev. Genet.*, **9**, 356–369.

Purcell,S. *et al.* (2007) PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.*, **81**, 559–575.

The 1000 Genomes Project Consortium (2010) A map of human genome variation from population-scale sequencing. *Nature,* **467**, 1061–1073.

Yu,W. *et al.* (2008) HuGE Watch: tracking trends and patterns of published studies of genetic association and human genome epidemiology in near-real time. *Eur. J. Hum. Genet. EJHG*, **16**, 1155–1158.