BIOINFORMATICS APPLICATIONS NOTE Vol. 25 no. 2 2009, pages 284–285
doi:10.1093/bioinformatics/btn600

Genetics and population analysis

GWAS GUI: graphical browser for the results of whole-genome association studies with high-dimensional phenotypes

Wei Chen*, Liming Liang and Gonçalo R. Abecasis

Center for Statistical Genetics, Department of Biostatistics, University of Michigan, 1420 Washington Heights, Ann Arbor, MI 48109, USA

Received on August 6, 2008; revised on November 3, 2008; accepted on November 14, 2008

Advance Access publication November 20, 2008

Associate Editor: Martin Bishop

ABSTRACT

Summary: We describe an interactive package that provides graphical overviews of the results of whole-genome association studies in datasets with rich multi-dimensional phenotypic information, such as global surveys of gene expression. Windows, Linux and Mac binaries are available from our website.

Availability: http://www.sph.umich.edu/csg/weich/software.html

Contact: weich@umich.edu

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

1 INTRODUCTION

Recently, genome-wide association scans (GWAS) have been used to successfully dissect a variety of complex traits, ranging from discrete clinical outcomes such as asthma and diabetes (Moffatt *et al*., 2007; Scott *et al*., 2007; WTCCC, 2007) to continuous traits as diverse as height, weight, global gene expression and blood lipid levels (Dixon *et al*., 2007; Frayling *et al*., 2007; Sanna *et al*., 2008; Scuteri *et al*., 2007; Willer *et al*., 2008). The amount of information generated in these studies is staggering and interpreting their results requires efficient computational tools for data analysis and visualization. This challenge is most noticeable when high-dimensional data (such as microarray gene expression data or proteomics data) are analyzed. In this case, the results of whole genome association studies can include billions of data points (Cheung *et al.*, 2005; Dixon *et al*., 2007; Moffatt *et al*., 2007). Realizing the full benefits of these studies requires an efficient way to share data among collaborators and with other researchers, both before and after the data are published. Here, we present a tool that facilitates interactive browsing of the results from whole genome association studies. To illustrate the capabilities of our browser, we used it to create an interactive interface for the results of a recent genome-wide association study of global gene expression (Dixon *et al*., 2007). The objective of the Dixon *et al*. (2007) study was to build a database that would allow researchers to systematically examine potential effects of disease-associated variants on transcript expression and our interactive browser makes it easy for many researchers to explore the data.

A diverse set of statistical methods can be used to examine the association between phenotypes of interest and single nucleotide polymorphism (SNP) data. For example, χ^2 test statistics, *P*-values, effect size estimates and their standard errors, as well as SNP-specific heritability estimates are all commonly reported in GWAS studies. When there are tens of thousands of phenotypic outcomes and hundreds of thousands SNPs, the result set is usually very large, containing several million statistics and easily totaling several gigabytes. These datasets can be integrated into specialized local databases for further investigation, but it can be challenging for researchers without extensive database or programming skills to access results. Our GWAS GUI (Graphic User Interface) is intended to provide a convenient tool for interacting with arbitrary GWAS result sets and to facilitate searches and displays of GWAS results in graph or tabular form. We hope our tool will facilitate data sharing within collaborative groups and with the public at large.

2 FEATURES OF GWAS GUI BROWSER

Our GWAS GUI browser is an interactive package that facilitates rapid interactive browsing of whole-genome association study results. It is designed to handle thousands of phenotypes, and thus can handle very rich datasets, such as those where global surveys of gene expression are combined with genome-wide SNP data. The browser also allows users to interact with the results of simpler scans, such as scans that focus on a single discrete outcome or a small number of related traits. To evaluate the program, we have applied it to several large datasets, including a study evaluating the association between 408 273 SNPs and the levels of 54 675 transcripts representing 20 599 known genes and assessed in lymphoblastoid cell lines from approximately 400 children (Dixon *et al*., 2007). After this initial evaluation, we released an early version of the program, named the mRNA by SNP browser (MRBS), when the Dixon *et al*. (2007) paper was published. In addition to the visualization tool, the full GWAS GUI browser includes a data preparation tool that can be used to organize tabulated results into an indexed database for rapid browsing. There are two main browsing interfaces within our browser: (i) an interface that retrieves all results for a specific trait and (ii) an interface that retrieves all results in a specific genomic region. In either view, results can typically be retrieved almost instantaneously. In the 'trait-centric' view, the browser can tabulate and sort a summary of user provided association test results (e.g. effect size, standard error, heritability estimates, test statistics and *P*-value) and quickly generate plots that summarize the distribution of a user-specified test statistics along the genome. Alternatively, in the 'position-centric' view, the

*To whom correspondence should be addressed.

Fig. 1. An illustration of the GWAS GUI browser interface. This example demonstrates how to display the results for a specific region. Several large statistics have been highlighted with blue circles by selecting the corresponding rows. The top transcripts ordered by maximum statistic within the region are tabulated in the right panel.

browser can tabulate all significant association test statistics (using a user-defined threshold) in a target region and plot the results for multiple traits. Optionally, information such as the location of nearby genes can also be displayed (Fig. 1). For convenience, both interfaces allow the browser to link the results to external databases chosen by the user, such as the University of California Santa Cruz (UCSC) genome browser, where users can examine the genomic context of each result in detail. When the user requests a SNP that is not included in the current dataset, linkage disequilibrium (LD) and tag information from the International HapMap Consortium can be used to suggest a backup tag-SNP. Figure 1 is an illustration of the browser interface after searching for a specific SNP position using the 'position-centric' view. Four SNPs of interest have been highlighted by the user in the tabular view (bottom left) and are circled in the graphical view.

3 EXAMPLES OF APPLICATION

Allowing large groups of scientists to browse and interact with the results of large multi-dimensional GWAS can be extremely helpful. For example, prior to the publication of the Dixon *et al*. (2007) gene expression paper, we used an early version of our browser to share preliminary results with several colleagues. This led to the observation that SNPs in an intergenic region on chromosome 5p13 that were associated with Crohn's Disease were also associated with transcript levels of PTGER4 suggesting that PTGER4 may be the primary candidate gene for Crohn's disease on chromosome 5. The Crohn's-associated SNPs are >200 Kb away from the nearest annotated gene. The result is published and described in detail elsewhere (Libioulle *et al*., 2007). Since then, many others have browsed our results resulting in several potential links between SNPs, human disease and mRNA transcript levels.

The current version of the GWAS GUI browser program is not restricted to gene-expression data, but is intended as a general tool that provides graphical overviews of whole-genome association study results for arbitrary phenotypes. The extended program allows users to load their own data files, tests statistics and genomic annotation files into the browser in a standardized text format. Generally, the traits can be any outcomes of interest, such as case–control indicators, expression values and many other continuous or categorical measurements. Arbitrary meta-data about each trait can be tracked and displayed. We expect that the browser will be particularly helpful when multiple-related traits are studied. In this setting, the browser simplifies the initial comparison of signals for different-related traits in regions of interest.

4 IMPLEMENTATION

The GWAS GUI browser program was implemented in C++ using the Qt4 toolkit (open-source version 4.4 Trolltech Inc.). It has been tested on Windows, Linux and Mac workstations. The system requirements depend on the size of input datasets which can range from a dataset examining a single trait dataset and hundreds of thousands of genetic markers to large-scale genome-wide gene-expression datasets with tens of thousands of traits and markers. On a modern Windows Workstation, the initial indexing of a set of results generated by PLINK (Purcell *et al*., 2007), MERLIN (Chen *et al*., 2007) or another whole-genome analysis tools and including approximately 300 000 SNPs requires ~200 Mb of RAM and 5–10 min of computing time. After indexing, opening the same dataset and browsing the data should be nearly instantaneous and require only 60 Mb RAM.

Funding: National Human Genome Research Institute; National Heart Lung and Blood Institute.

Conflict of Interest: G.R.A. is a Pew Scholar of the Biomedical Sciences and is supported by the Pew Charitable Trusts.

REFERENCES

Chen,W.M. *et al*. (2007) Family-based association tests for genome-wide association scans. *Am. J. Hum. Genet.,* **81**, 913–926.

Cheung,V.G. *et al.* (2005) Mapping determinants of human gene expression by regional and genome-wide association. *Nature*, **437**, 1365–1369.

Dixon,A.L. *et al*. (2007) A genome-wide association study of global gene expression. *Nat. Genet*., **39**, 1202–1207.

Frayling,T.M. *et al*. (2007) A common variant in the FTO gene is associated with body mass index and predisposes to childhood and adult obesity. *Science,* **316**, 889–894.

Libioulle,C. *et al*. (2007) Novel Crohn disease locus identified by genome-wide association maps to a gene desert on 5p13.1 and modulates expression of PTGER4. *PLoS Genet.*, **3**, e58.

Moffatt,M.F. *et al*. (2007) Genetic variants regulating ORMDL3 expression contribute to the risk of childhood asthma. *Nature*, **448**, 470–473.

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

Sanna,S. *et al*. (2008) Common variants in the GDF5-UQCC region are associated with variation in human height. *Nat. Genet.,* **40**, 198–203.

Scott,L.J. *et al*. (2007) A genome-wide association study of type 2 diabetes in Finns detects multiple susceptibility variants. *Science*, **316**, 1341–1345.

Scuteri,A. *et al*. (2007) Genome-wide association scan shows genetic variants in the FTO gene are associated with obesity-related traits. *PLoS Genet.*, **3**, 1200–1210.

Wellcome Trust Case Control Consortium (2007) Genome-wide association study of 14 000 cases of seven common diseases and 3000 shared controls. *Nature*, **447**, 661–678.

Willer,C.J. *et al*. (2008) Newly identified loci that influence lipid concentrations and risk of coronary artery disease. *Nat. Genet.,* **40**, 161–169.