BIOINFORMATICS APPLICATIONS NOTE Vol. 26 no. 19 2010, pages 2474–2476
doi:10.1093/bioinformatics/btq452

Genetics and population analysis

Advance Access publication August 10, 2010

Genevar: a database and Java application for the analysis and

Tsun-Po Yang[1], Claude Beazley[1], Stephen B. Montgomery[1,2], Antigone S. Dimas[1,2,3], Maria Gutierrez-Arcelus[2], Barbara E. Stranger[1,4], Panos Deloukas[1] and Emmanouil T. Dermitzakis[1,2,*]

[1]Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1HH, UK, [2]Department of Genetic Medicine and Development, University of Geneva Medical School, Geneva CH-1211, Switzerland, [3]Wellcome Trust Centre for Human Genetics, Oxford OX3 7BN, UK and [4]Division of Genetics, Department of Medicine, Harvard Medical School, Brigham and Women's Hospital, Boston, MA 02115, USA

Associate Editor: Martin Bishop

ABSTRACT

Summary: Genevar (GENe Expression VARiation) is a database and Java tool designed to integrate multiple datasets, and provides analysis and visualization of associations between sequence variation and gene expression. Genevar allows researchers to investigate expression quantitative trait loci (eQTL) associations within a gene locus of interest in real time. The database and application can be installed on a standard computer in database mode and, in addition, on a server to share discoveries among affiliations or the broader community over the Internet via web services protocols.

Availability: http://www.sanger.ac.uk/resources/software/genevar

Contact: emmanouil.dermitzakis@unige.ch

Received on May 12, 2010; revised on July 22, 2010; accepted on August 4, 2010

1 INTRODUCTION

Expression quantitative trait loci (eQTL) mapping, where gene expression profiling is treated as a phenotypic trait in genome-wide association studies (GWAS), has successfully been employed to uncover genetic variants that influence expression variation in recent studies (Dixon *et al.*, 2007; Stranger *et al.*, 2007a). Single-nucleotide polymorphism (SNP)–gene associations from eQTL analysis can be investigated in populations (Stranger *et al.*, 2007b) or among tissue types (Dimas *et al.*, 2009; Heinzen *et al.*, 2008). In addition to genome-wide eQTL identification, combinations of eQTLs and lead SNPs identified by GWAS have been provided to interrogate the mechanisms underlying disease susceptibility at specific loci (Grundberg *et al.*, 2009; Nica *et al.*, 2010; Zeller *et al.*, 2010). However, an analytical and visualization tool, together with a structured repository for multiple datasets, is still needed to facilitate the investigation of loci of interest and to share data publicly and among collaborators.

Here, we present Genevar, a database and Java tool designed to provide: (i) data warehousing; (ii) real-time computation of correlation significance; (iii) visualization of mapping results in a user-friendly interface; and (iv) an added web services platform that is implemented as a bridge between the server and multiple users. Genevar allows published data to be visually accessible in a secure fashion, without the need for users to download raw data. Through interactive analysis pipelines, researchers are able to rapidly investigate, for instance, *cis*-acting eQTLs at the locus of interest.

Complementing already available standalone tools (Chen *et al.*, 2009; Ge *et al.*, 2008), a database-centric architecture enables Genevar to perform complex queries on-the-fly and does not have a high memory requirement for prior reading in large-scale datasets. Furthermore, exploiting the convenience of web-based (Wang *et al.*, 2003; Zou *et al.*, 2007) and web-launch (Mueller *et al.*, 2005) tools, a Java interface was developed that connects to both database and web services. The main advantage of this system design is that users can switch between public services and local data on the same interface. Default services at the Sanger Institute currently contain gene expression profiling and genotypic data from the following two datasets: lymphoblastoid cell lines from eight HapMap3 populations (824 individuals, unpublished data); and three cell types derived from umbilical cords of 75 Geneva GenCord individuals (Dimas *et al.*, 2009).

2 FEATURES

Genevar has two main functionalities in *cis*-eQTL analysis: (i) identifying eQTLs in genes of interest, and (ii) observing SNP–gene associations surrounding SNPs of interest (Fig. 1). Additional features include SNP–probe association plots and external links to three major genome browsers. Either *cis*- or *trans*-eQTLs can be plotted in the SNP–probe association plot module. Mapping results are listed in tree nodes in a structural manner, and information can be saved as PNG diagrams or exported as tab-delimited lists for further use in presentations or publications.

Genevar is compatible with PLINK (Purcell *et al.*, 2007) genotype data formats and any tab-delimited expression/genotyping file in our format. After uploading datasets onto the database, Genevar presents expression profiling data and individual genotypes in two cataloged management panels. Once a group of datasets is selected in the follow-up analysis pipelines, the software automatically prompts available expression–genotype pairs for the user to choose from.

*To whom correspondence should be addressed.

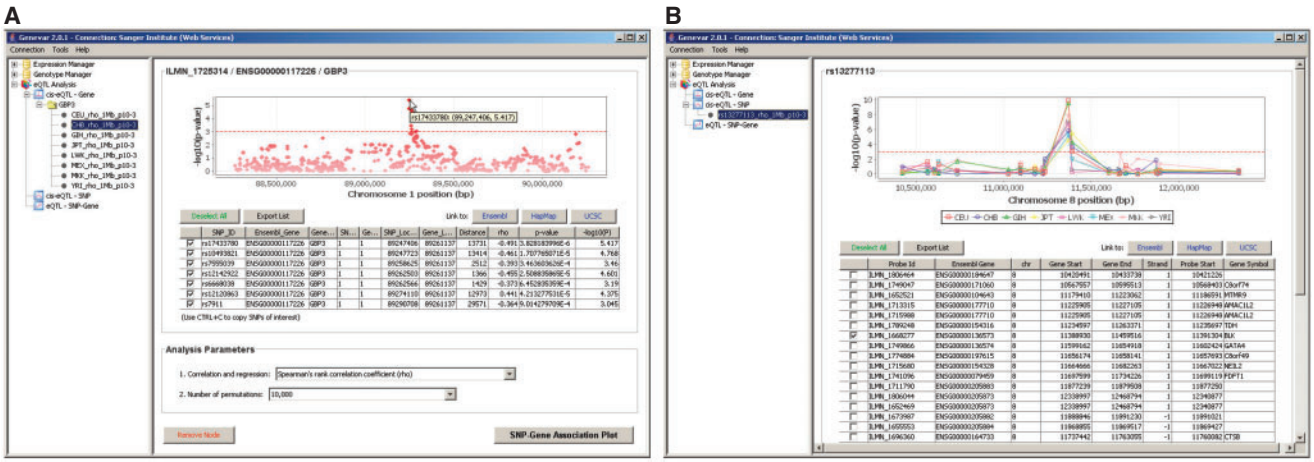

Fig. 1. Results of Genevar: a scatter plot represents observed eQTLs in a 2 Mb window centering the *GBP3* locus in HapMap3 CHB (**A**), and a line chart illustrates observed SNP–gene associations in a 2 Mb region surrounding rs13277113 SNP in eight HapMap3 populations (**B**).

Spearman's rank correlation coefficient is performed to estimate the strength of relationship between alleles and gene expression intensities, linear regression is also used to model the relationship between the two variables. To test the significance of the relationship, a *t*-statistic is employed with $n-2$ degrees of freedom for both correlation and regression analysis (Stranger *et al.*, 2007b). The software allows the user to adjust the window size centering on the gene/SNP of interest (e.g. 2 Mb) and user-defined *P*-value threshold (e.g. $P<0.001$) for the featured *cis*-eQTL analysis. Alternatively, non-parametric permutation *P*-values are also provided in the subsequent association plot module to further evaluate the significance of nominal *P*-values. In order to construct a distribution of the test statistic, under the null hypothesis of no SNP–probe associations, expression intensities are randomly re-assigned to individuals' genotypes, then correlation coefficient and statistical significance are re-computed for the relabeled traits, and this procedure is repeated 10 000 times (Stranger *et al.*, 2005).

We recommend users to launch Genevar via Java Web Start from our homepage for the most up-to-date version. After launching, Genevar is initially in web services mode connecting to the Sanger Institute. The user can then make another services connection to affiliated institutes, or switch to database mode connecting directly to user's local database. Genevar can be run completely offline in database mode as there is no communication between the Java interface and Sanger server.

Future work will include modified visualization for displaying next-generation sequence data, e.g. RNA-Seq (Montgomery *et al.*, 2010); and implementation of methylation modules to interrogate epigenomic data.

3 IMPLEMENTATION

This approach to relational database design is an attempt to systematically decompose traditional flat files, which are one record per line and have no structural relationships between the records, into grouped dimension tables and to reduce data redundancy. A normalized and structured repository is suitable to warehouse all kinds of data format regardless of the file size and field numbers. Most importantly, the advantage of using database indexing on expression and genotype fact tables highly stabilize retrieval performance with the subsequent but reasonable cost of slower uploads and increased disk space. The only limitation when the datasets grew would be the storage space as this is a trade-off for query speed.

To maximize the potential of Genevar as a platform shared among affiliations, Genevar has been extended to interact with web services protocols to enhance data security; the database schema will be deployed behind and protected by the firewall, whereas only a secure frontend webpage acting as a middle layer will be accessible to the user over the Internet.

Genevar uses Hibernate library (http://www.hibernate.org) to map object-oriented models onto MySQL relational database tables (http://www.mysql.com) in the back-end, and acquires Apache CXF framework (http://cxf.apache.org) to wrap up database queries and business logics into middle-layer services. Finally, a Tomcat server (http://tomcat.apache.org) is used to provide services in the front-end. For a standalone database-mode Genevar, only a MySQL database is required to be installed on user's local machine. Association results are visualized in genomic views by JFreeChart library (http://www.jfree.org/jfreechart/). A gene-centered scatter plot represents observed SNP–gene associations around genes of interest, and a SNP-centered line chart illustrates observed eQTLs surrounding SNPs of interest (Fig. 1).

Tested on a 1.6 GHz Pentium Centrino laptop with 1 GB of RAM, Genevar was able to upload a 75×23k expression dataset onto the database and built up indexes in 1 min; another 23 min were required for the 75×400k genotype file. Once it is uploaded, Genevar can fetch per SNP–probe pairs from these 75 individuals in <0.0257 s from the database, and calculates Spearman's *rho*s and nominal *P*-values for 486 SNP–probe pairs in 3 s.

ACKNOWLEDGEMENTS

We thank Guillaume Smits and Johan Rung (EMBL-EBI) for their suggestions on improving the functionalities. We also thank Richard

Jeffs, James Smith, Paul Bevan (Sanger Webteam) and Andrew Bryant (Database Team) for helpful support on this project.

Funding: Wellcome Trust and Louis-Jeantet Foundation.

Conflict of Interest: none declared.

REFERENCES

Chen,W. *et al*. (2009) GWAS GUI: graphical browser for the results of whole-genome association studies with high-dimensional phenotypes. *Bioinformatics*, **25**, 284–285.

Dimas,A.S. *et al*. (2009) Common regulatory variation impacts gene expression in a cell type-dependent manner. *Science*, **325**, 1246–1250.

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

Ge,D. *et al*. (2008) WGAViewer: software for genomic annotation of whole genome association studies. *Genome Res.*, **18**, 640–643.

Grundberg,E. *et al*. (2009) Population genomics in a disease targeted primary cell model. *Genome Res.*, **19**, 1942–1952.

Heinzen,E.L. *et al*. (2008) Tissue-specific genetic control of splicing: implications for the study of complex traits. *PLoS Biol.*, **6**, 2869–2879.

Montgomery,S.B. *et al*. (2010) Transcriptome genetics using second generation sequencing in a Caucasian population. *Nature*, **464**, 773–777.

Mueller,M. *et al*. (2005) eQTL Explorer: integrated mining of combined genetic linkage and expression experiments. *Bioinformatics*, **22**, 509–511.

Nica,A.C. *et al*. (2010) Candidate causal regulatory effects by integration of expression QTLs with complex trait genetic associations. *PLoS Genet.*, **6**, e1000895.

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.

Stranger,B.E. *et al*. (2005) Genome-wide associations of gene expression variation in humans. *PLoS Genet.*, **1**, e78.

Stranger,B.E. *et al*. (2007a) Relative impact of nucleotide and copy number variation on gene expression phenotypes. *Science*, **315**, 848–853.

Stranger,B.E. *et al*. (2007b) Population genomics of human gene expression. *Nat. Genet.*, **38**, 1217–1224.

Wang,J. *et al*. (2003) WebQTL: web-based complex trait analysis. *Neuroinformatics*, **1**, 299–308.

Zeller,T. *et al*. (2010) Genetics and beyond–the transcriptome of human monocytes and disease susceptibility. *PLoS One*, **5**, e10693.

Zou,W. *et al*. (2007) eQTL Viewer: visualizing how sequence variation affects genome-wide transcription. *BMC Bioinformatics*, **8**, 7–11.