BIOINFORMATICS APPLICATIONS NOTE Vol. 26 no. 2 2010, pages 270–272
doi:10.1093/bioinformatics/btp641

Genetics and population analysis

MAVEN: a tool for visualization and functional analysis of genome-wide association results

Kanchana Narayanan and Jing Li*

Department of Electrical Engineering and Computer Science, Case Western Reserve University Cleveland, OH, USA

Received on June 5, 2009; revised on November 11, 2009; accepted on November 12, 2009

Advance Access publication November 17, 2009

Associate Editor: Alex Bateman

ABSTRACT

Summary: We describe the features and implementation of a web application tool named MAVEN—for Management, Analysis, Visualization and rEsults shariNg of genome-wide association data using cutting edge technologies. Main capabilities include user data uploading and management, queries using a variety of criteria, visualization of results, interactive selections and seamless integration of users' data with databases at the National Center for Biotechnology Information (NCBI) for functional annotations of single nucleotide polymorphisms (SNPs) and genes.

Availability: http://cbc.case.edu/maven

Contact: jingli@case.edu

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

1 INTRODUCTION

Recently, genome-wide association studies (GWAS) have shown to be powerful in investigating the effect of inherited genetic variations on risks of complex diseases (McCarthy *et al.*, 2008). Within the last 2 years, scientists have successfully replicated genetic risks of many complex diseases such as cancers, heart diseases and diabetes using GWAS. As of this writing, there have been ∼325 published GWAS studies with ∼1500 SNPs that have been identified being associated with various diseases/conditions (Hindorff *et al.*, 2009). Though GWAS have shown some initial success, they also bring tremendous challenges to the community, not only in computations (i.e. hundreds of thousands of SNPs) and statistical analyses (e.g. multiple testing) but also in data management, access control, results visualization and sharing, integration with existing public resources (e.g. dbSNP, pubMed databases at NCBI) and interpretation of results. In this article, we present a web application tool (Fig. 1) named MAVEN, for Management, Analysis, Visualization and rEsults shariNg of GWA data using cutting edge technologies. The current implementation of MAVEN (version 1.1) allows users to directly upload their own GWAS results to be stored in a relational database, and to search their data using powerful and user-friendly filtering capabilities. MAVEN can conveniently assist genetic epidemiologists to visualize their GWAS results, to share results within their own group or with colleagues across the world. MAVEN also integrates information about SNPs and genes directly from NCBI databases and provides direct links to detailed information of each selected SNP and/or gene, which will allow researchers to easily narrow down interesting candidate disease genes for further functional analysis. None of the existing tools, including those recently developed ones (Aulchenko *et al.*, 2007; Gonzalez *et al.*, 2007; Hemminger *et al.*, 2006; Purcell *et al.*, 2007; Sole *et al.*, 2006) that are specific for GWAS, have fully considered or incorporated all these features. Almost all of these earlier tools focus primarily on data analysis such as quality control and cleaning, and single SNP- or haplotype-based statistical tests, which were greatly needed at earlier stages. However, they provide limited, if any, functionalities on result sharing, visualization and integrations, which are also greatly needed at late stages of GWAS. An exception is TAMAL (Hemminger *et al.*, 2006), which provides some query functions about SNPs given a list of genes. However, it does not integrate users' data with online information. While

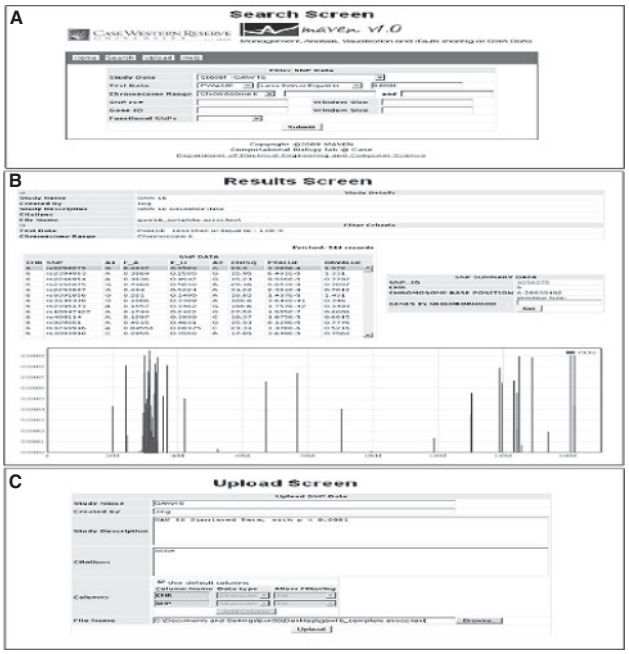

Fig. 1. Screen shots of MAVEN. (**A**) MAVEN allows users to filter data using various criteria. (**B**) Results are shown in a tabular format as well as in graphical display. (**C**) Users can define the structure of their data and upload data to the system.

*To whom correspondence should be addressed.

preparing the manuscript, we noticed a very recent tool named GWAS GUI (Chen *et al.*, 2009), which also aims to visualize GWA results. This provides independent evidence for the need of such tools. In comparison to GWAS GUI, MAVEN has the advantage that all annotations of SNPs and genes have been either stored in local databases or can be accessed directly from NCBI through links provided by MAVEN. In contrast, users themselves have to upload annotation information into GWAS GUI, which is difficult and tedious for many researchers in the epidemiology community. Another major difference is that MAVEN is provided as a web server tool, which makes sharing of results among colleagues from different locations extremely easy. In addition, MAVEN offers powerful filtering capabilities using different criteria, interactive selections, visualization of results and seamless integration with databases from NCBI.

2 FEATURES

Data Managements: for the current implementation, MAVEN accepts and stores GWAS *results* based on *single-locus* analysis methods, not the raw data, which makes sharing less problematic in terms of privacy. All data are stored and managed using a relational database management system. The system allows users to upload their own results into MAVEN by preparing their analysis results in a tabular text format [e.g. output file of the program pLink (Purcell *et al.*, 2007)]. Users can also submit other information such as descriptions of the study and references to their publications. The database is study oriented and also maintains some tables that are common to all studies. For example, SNP information and gene information have been downloaded from NCBI databases beforehand. They are identified by NCBI build number and will be updated as necessary. We have compiled all SNPs from many different SNP-chip platforms of the two major vendors (i.e. Illumina and Affymetrix). Users can also upload SNPs of their customized chips. For gene annotations, all three assemblies (reference, HuRef and Celera) have been considered. The reference assembly serves as the main coordinate for SNP/gene physical positions.

Filtering Capability: the tool offers several types of filtering capabilities for users to retrieve interesting SNPs and gene regions that are specific to their studies. Users can query SNPs according to (i) significance (e.g. P-value or χ^2 statistic) using a threshold value; (ii) chromosomes and physical positions with specified window sizes; (iii) identification number (rs#); (iv) functional roles (e.g. synonymous, non-synonymous, etc.); (v) specific gene regions; (vi) KEGG pathways, as well as combinations of all of the above criteria. MAVEN will visualize all SNPs that satisfy all the conditions.

Data Visualization: search results are displayed in two different formats: a *tabular format* and a *graphical format*. The result table displays the Chromosome, SNP rs# and other columns from the study dataset provided by users. Users can sort the records according to the values of a selected column or save the results into a file. The graphical display plots the value of each SNP as a line chart using the corresponding filtering field (e.g. P-value) as the Y-axis and chromosomes and/or base pair positions as the X-axis. Both the table and the chart support interactive and synchronized selections. When a user moves the mouse over the chart, summary statistics will be displayed. Users can also show SNP results in UCSC Genome Browsers as a customized track.

Functional Annotation of SNPs: when a user selects a particular SNP from the result table, or from the display graph, a request is sent to the backend which will access dbSNP database through NCBI eUtils to obtain up-to-date functional annotations of the SNP. MAVEN will output summary information about the selected SNP, which includes a link to its NCBI record, its chromosome position, gene name and a link to the gene if the SNP is within a gene region, its functional roles (synonymous, missense or nonsense, etc.), pathway information (links to KEGG) as well as disease information (links to OMIM) about the gene. Users can obtain more detailed information through these hyperlinks. Not all significant SNPs will be within gene regions. To better understand the roles of those SNPs, MAVEN provides a functionality to allow users to retrieve all genes from NCBI for a given window around a SNP. It will return all genes within that region with hyperlinks for detailed information.

Performance on Real Examples: we have performed extensive tests of MAVEN using a real GWA dataset with 500 K SNPs and a simulated data with around one million SNPs. Results using our local network have shown that MAVEN is efficient to handle such datasets. The most time-consuming operation is the data uploading and database creation step, which can be finished in a few minutes. This is not an issue because it is a one-time operation. All other operations can be finished almost instantly, partially due to the pagination mechanism implemented by MAVEN. For each query, MAVEN will first obtain the total number of SNPs that satisfy the criteria. At maximum, it only displays the first 500 SNP records and users can navigate through different pages for other SNPs. The limit usually does not cause any problem because users normally examine a handful of significant or interesting SNPs at a time.

Multiple phenotypes and/or multiple statistics: the current version of MAVEN is not restricted to a single phenotype or a single statistics. In the uploading stage, users can define any number of columns to be searchable. For example, users can have several P-values for a single phenotype, each based on a different test statistic or several P-values for different phenotypes.

3 IMPLEMENTATION

This web application has been developed using JAVA-J2EE technologies and follows the Model-View-Controller framework, which is a proven and convenient way to generate organized, modular applications that cleanly separate logic, style and data. Detailed information about implementation is provided as online Supplementary Materials.

ACKNOWLEDGEMENTS

We thank Yong Guo and Daniel Haig for their help.

Funding: NIH/NLM (LM008991); NIH/NCRR (RR03655); NSF (CRI0551603).

Conflict of Interest: none declared.

REFERENCES

Aulchenko,Y.S. *et al.* (2007) GenABEL: an R library for genome-wide association analysis. *Bioinformatics*, **23**, 1294–1296.

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.

Gonzalez,J.R. *et al.* (2007) SNPassoc: an R package to perform whole genome association studies. *Bioinformatics*, **23**, 644–645.

Hemminger,B.M. *et al.* (2006) TAMAL: an integrated approach to choosing SNPs for genetic studies of human complex traits. *Bioinformatics*, **22**, 626–627.

Hindorff,L. *et al.* (2009) A catalog of published genome-wide association studies. Available at http://www.genome.gov/26525384/ (last accessed date May 27, 2009).

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.

Sole,X. *et al.* (2006) SNPStats: a web tool for the analysis of association studies. *Bioinformatics*, **22**, 1928–1929.