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
Kanchana.Narayanan@case.edu, JingLi@case.edu

***Abstract*—Genome-wide association studies (GWAS) provide a new and powerful approach to investigate the effect of inherited genetic variation on risks of complex diseases. With recent advances in genotyping technology, genome-wide association studies are now becoming a reality. Within the past two years, scientists have successfully replicated genetic risks of several complex diseases including cancers, obesity, and type 2 diabetes using GWAS. And more data from GWAS are expected in an accelerated rate. However, management, analysis, visualization, and interpretation of genome wide association data are particularly difficult, primarily because GWAS may consists of hundreds of thousands SNPs (single nucleotide polymorphisms) from thousands individuals. This paper describes the features and implementation of a web application tool named MAVEN for Management, Analysis, Visualization and rEsults shariNg of GWA data using cutting edge technologies. In addition, MAVEN seamlessly integrates users own data with databases at the National Center for Biotechnology Information (NCBI), which allows users to directly obtain functional annotations of SNPs and genes that are relevant to their own research interests. Therefore, MAVEN can effectively facilitate the functional analysis of GWAS.**

Keywords-Visualization; Genome-wide association studies; Functional analaysis

I. INTRODUCTION

Most common diseases such as neurodegenerative diseases (*e.g.*, Alzheimer's disease (AD) and Parkinson's disease), cardiovascular diseases, various cancers, diabetes, osteoporoses are complex diseases that involve multiple genes, their interactions, environmental factors, and gene-environment interactions. Two most commonly used approaches for disease gene mapping are linkage and association studies. Linkage studies examine the cosegregation patterns of genetic markers and the trait of interest within families, and establish linkage of disease genes and markers through recombination fractions. Association studies evaluate correlations between genetic polymorphisms and phenotypes at the population level with the aim to identify genetic loci that are in linkage disequilibrium with causal variants. Recently, genome-wide association studies (GWAS) have shown to be powerful in investigating the effect of inherited genetic variation on risks of complex diseases [1, 2]. Within the last two years, scientists have successfully replicated genetic risks of several complex diseases including cancers, heart diseases, and diabetes using GWAS [1, 2]. Though GWAS has shown some initial success, it also brings tremendous challenges to the community, not only in computation (i.e., hundreds of thousands SNPs) and statistical analysis (e.g., multiple testing), but also in data management, access control, results visualization and sharing, integration with existing public resources (e.g., dbGaP, dbSNP, pubMed databases at NCBI) and interpretation of results. None of existing statistical genetics tools, including those recently developed ones [3-6] that are specific for GWAS, have fully considered or incorporated all these features. In this article, we present a web application named MAVEN, for Management, Analysis, Visualization and rEsults shariNg of GWA data using cutting edge technologies. The current implementation of MAVEN (version 1.0) allows users to directly upload their own GWAS results which will be stored in relational databases. The data can then be searched using various filtering functionalities to select interesting SNPs according to their significances, chromosomal and physical positions, a particular SNP or a gene and their surrounding neighborhoods, as well as functional roles of SNPs (i.e., synonymous vs. non-synonymous). The search results will be presented in tabular as well as graphical formats. MAVEN also integrates information about SNPs and genes directly from National Center for Biotechnology Information (NCBI) databases and provides direct links to detailed

This research is supported by National Institutes of Health grant LM008991.

978-0-7695-3685-9/09 $25.00 © 2009 IEEE
DOI 10.1109/OCCBIO.2009.27

information with respect to each specific SNP/gene. MAVEN can conveniently assist genetic epidemiologists to visualize their GWAS results, to share the results within their own group or with colleagues across the world. Furthermore, by integrating information about SNPs and genes from NCBI databases, researchers now can easily narrow down interesting candidate disease genes for further functional analysis. The system can be directly accessed from our server at http://cbc.case.edu/MAVEN and source code will be freely available to non-commercial users upon requests.

II. FUNCTIONALITIES

MAVEN implements the following features:

Data Managements: For the current implementation, MAVEN only accepts and stores GWAS *results* based on *single-locus* analysis methods, not the raw data. Summary results in general are adequate for many users and at the same time, protection of data privacy can be enforced because no personal information will be available at any time. All data are stored and managed using a relational database management system (DBMS). The system allows users to upload their own results into MAVEN, with the hope that eventually, MAVEN will become a database that collects most existing published GWAS results. Users can prepare their analysis results in a tabular text format and submit them along with some description information about their studies, including references to their publications (Figure 1). To make it more convenient, users can directly use the output file of the program pLink[6], which is a popular tool for analyzing GWAS data. The database is study-oriented and also maintains some tables that are common to all studies. For example, SNP information (e.g., SNP IDs, physical positions) and gene information (e.g., gene names, IDs, physical positions) have been downloaded from NCBI databases beforehand. They are identified by NCBI build number and will be updated as necessary.

Filtering Capability: To better view results that are interesting to researchers, we have implemented a variety of filtering functions using different criteria (Figure 2):

- Study selection – Users can only view results from a particular study because different studies in general donot have a direct relationship.
- p-Value filter – Most users are interested in significant SNPs, which can be selected based on their *p*-values.
- Physical positions – Users can obtain SNPs in a candidate region by specifying a chromosome and their physical positions in basepairs.
- Search by a SNP – Users can provide the rs# of a SNP and a range parameter in basepairs to obtain all SNPs in that neighborhood.
- Search by a gene – Users can provide a gene ID or gene name and a range parameter in basepairs to obtain all SNPs in the neighborhood of a particular gene.
- Search by functional roles of SNPs – In many cases, one is more interested in SNPs that are in gene regions. MAVEN allows users to limit their search only in gene regions, or only search for synonymous or non-synonymous SNPs.

Users can specify one or more criteria and MAVEN will return all SNPs that satisfy all the conditions.

Figure 1. Interface for uploading genome-wide association results in MAVEN.

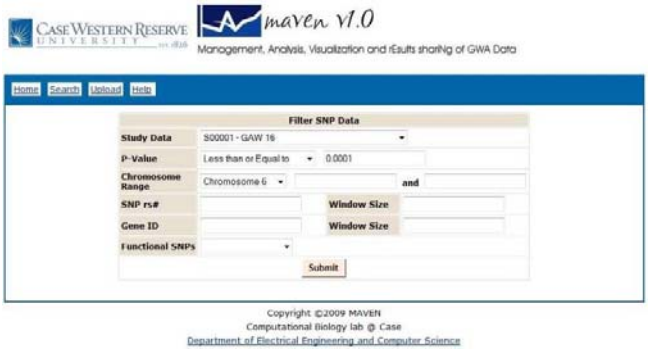

Figure 2. Users can search interesting SNPs by a variaty of critieria.

Data Visualization: MAVEN displays search results in two different formats: a tabular format and a graphical format. Currently, the result table displays the following

information for each SNP: SNP rs number, chromosome number and the physical position of the SNP on the chromosome in basepair, minor allele and its frequency in cases and controls, major allele, value of allelic χ^2-statistics and its associated *p*-value, and the value of odd-ratio. Users can also sort the records according to the values of a selected column. By default, the *graphical display* plots the significance of each SNP as a line chart using its -$\log_{10}$(*p*_value) on the Y-axis and the base pair position of the SNP on a chromosome as the X-axis. The chart is embedded as an image on the web page.

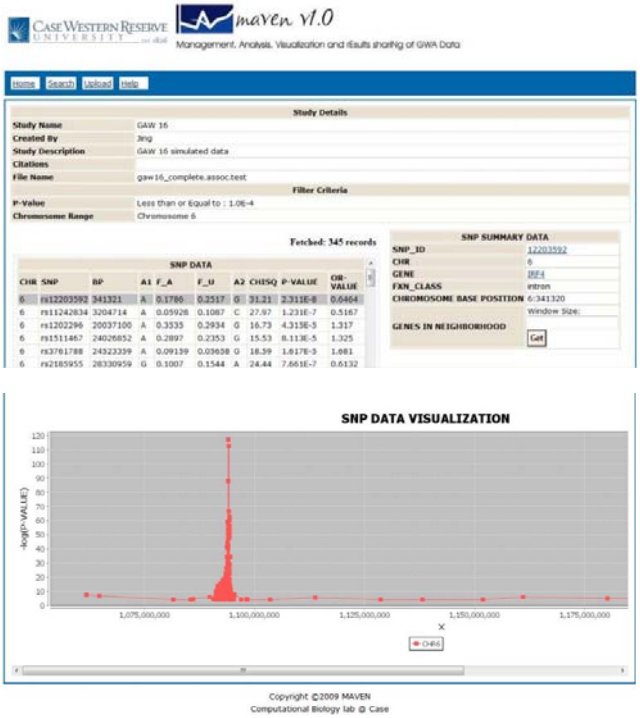

Figure 3. Search results are shown in a tabular format(above) and in a graphical format (bottom).

Functional Annotation of SNPs: When users click a particular row on the SNP Data Table, a request is sent to the backend which will access NCBI SNP database (dbSNP) to obtain some basic functional annotation information about the SNP, which includes: whether the SNP is in a gene region or not, whether the SNP is a synonymous SNP or a non-synonymous SNP if it is in a coding region. If users are willing to obtain more detailed information about the SNP or the gene, they can click the SNP/gene ID hyperlink, which will obtain all relevant results about the SNP/gene from NCBI databases. When a significant SNP is not within any gene regions, we provide a functionality to allow users to specify a neighborhood by a physical distance around that SNP. MAVEN will then display all gene IDs within that window with hyperlinks for detailed information.

SNP SUMMARY DATA	
SNP_ID	12203592
CHR	6
GENE	IRF4
FXN_CLASS	intron
CHROMOSOME BASE POSITION	6:341320
GENES IN NEIGHBORHOOD	Window Size: 200000 Get 56940, 642335, 642316, 3662, 727829, 727827, 55770

Figure 4. Functional annotation of a particular SNP and genes around its neighborhood. Detailed information can be obtained from NCBI databases through hyperlinks.

Efficiency Tests: The Upload functionality has been tested with file containing 500K SNPs. The upload time takes about 5mins. The results page implements paginations, displaying 500 records per page. Hence the search time is constant no matter how many records are returned.

III. System Architecture and Implementation Details

This web application has been developed using JAVA-J2EE technologies. The rapid ascension and acceptance of this technology can be attributed to its design and programming features and particularly the fact that it is portable to different hardware and software platforms. The architecture of this application mainly follows the Model-View-Controller (MVC) framework which is a proven and convenient way to generate organized, modular applications that cleanly separate logic, style, and data. More specifically, the front end for this application has been written using the Struts and Tiles Framework. In the Java world, Struts is one of the best-known and most cited open source embodiments of MVC. The project's core functionality and the view support have been improved, by incorporating the Tiles view component framework to strengthen support for component-based development, to increase reuse, and to enhance consistency. Tiles provides a means of compartmentalizing the front end so that common elements, shared over several screens, need to be coded only once and reused wherever needed. The application implementation has been divided into four layers (Figure 5); each layer has been assigned a particular

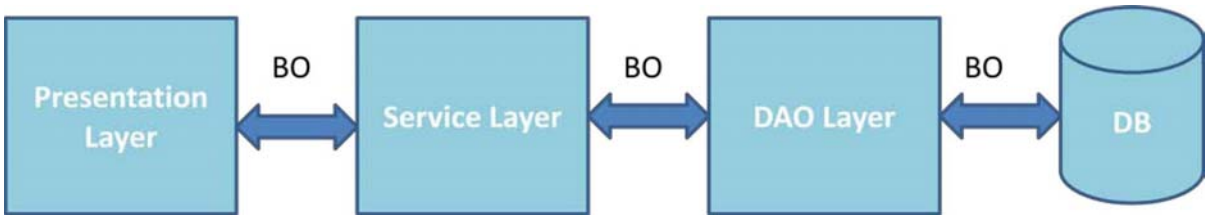

Figure 5. Implementation architecture.

responsibility. The actual data is stored in a Microsoft SQL Server database (*the Database Layer*). The result of each study is stored in one table. A separate table stores SNP information including SNP IDs and physical positions, which is common for all studies. The data in the database is accessed using JDBC access techniques through the *DAO (Data Access Object) Layer*. *The Service Layer* acts as a mediator between the presentation layer and the DAO layer. The philosophy behind introducing this layer is to segregate the business logic and the data access logic from the presentation layer. This would facilitate the flexibility to change the data access methods in the future without affecting the *presentation layer* in any manner. For instance, we might want to read data from a different database management system or directly from a flat file in the future, and the presentation layer would not have to be altered in any manner. The presentation layer mainly incorporated the User Interface part of the application. It mainly includes the Struts Action and Forms classes. The presentation layer is responsible for handling user requests from the browser and communicating with the service layer to obtain the information requested by the user. The Form class encapsulates the various BO's (Business Objects) which will be used to display the data on the JavaServer Pages. Data is transferred back and forth using these BO's. The presentation layer purely acts as receiver of user requests and does not perform any kind of validations or DB access. The user interface is handled by using Javascript. CSS stylesheets have been employed to style the various components displayed on the screen. JFreeChart**,** which is an open-source library, is used to generate the graphical display. AJAX (Asynchronous JavaScript and XML) technology allows parts of the web page to communicate independently with the server therefore the entire screen need not be refreshed when only a part of the data is being changed on the screen. The web server is implemented using Apache Tomcat.

IV. Discussion and Future Work

Though the current implementation of MAVEN provides many useful functions for visualizing results from GWAS, it can be enhanced in many ways. For example, it is common that for a particular study, multiple traits might be measured and statistical tests can be performed for all the traits. Though statistical results of multiple traits can be loaded to the database, our current implementation can only search SNPs based on statistics of one trait and provide graphical visualization of that statistics. One can partially solve the problem by creating multiple studies, each with a different trait, but cross-reference of different studies is not allowed in the current version. In addition, in certain circumstances, researchers might have some prior information about disease pathways and they might want to retrieve all SNPs within the coding regions of all genes in a particular pathway. Search SNPs through a pathway or direct integration with pathway databases is desirable. All such functionalities are currently being considered. Furthermore, we also plan to make the chart area of the web page to be interactive so that when a user clicks on a data point, a request would be sent to the server to fetch the summary data for that SNP.

References

[1] W. T. C. C. Consortium, "Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls," *Nature*, vol. 447, pp. 661-78, 2007.

[2] M. I. McCarthy, G. R. Abecasis, L. R. Cardon, D. B. Goldstein, J. Little, J. P. Ioannidis, and J. N. Hirschhorn, "Genome-wide association studies for complex traits: consensus, uncertainty and challenges," *Nat Rev Genet*, vol. 9, pp. 356-69, 2008.

[3] X. Sole, E. Guino, J. Valls, R. Iniesta, and V. Moreno, "SNPStats: a web tool for the analysis of association studies," *Bioinformatics*, vol. 22, pp. 1928-9, 2006.

[4] Y. S. Aulchenko, S. Ripke, A. Isaacs, and C. M. van Duijn, "GenABEL: an R library for genome-wide association analysis," *Bioinformatics*, vol. 23, pp. 1294-6, 2007.

[5] J. R. Gonzalez, L. Armengol, X. Sole, E. Guino, J. M. Mercader, X. Estivill, and V. Moreno, "SNPassoc: an R package to perform whole genome association studies," *Bioinformatics*, vol. 23, pp. 644-5, 2007.

[6] S. Purcell, B. Neale, K. Todd-Brown, L. Thomas, M. A. Ferreira, D. Bender, J. Maller, P. Sklar, P. I. de Bakker, M. J. Daly, and P. C. Sham, "PLINK: a tool set for whole-genome association and population-based linkage analyses," *Am J Hum Genet*, vol. 81, pp. 559-75, 2007.