BIOINFORMATICS ORIGINAL PAPER Vol. 27 no. 17 2011, pages 2422–2425
doi:10.1093/bioinformatics/btr389

Databases and ontologies Advance Access publication June 27, 2011

BRISK—research-oriented storage kit for biology-related data

Alan Tan, Ben Tripp and Denise Daley*
James Hogg iCAPTURE Center, Department of Medicine, University of British Columbia (UBC), Vancouver, BC, Canada V6Z1Y6

Associate Editor: Martin Bishop

ABSTRACT

Motivation: In genetic science, large-scale international research collaborations represent a growing trend. These collaborations have demanding and challenging database, storage, retrieval and communication needs. These studies typically involve demographic and clinical data, in addition to the results from numerous genomic studies (omics studies) such as gene expression, eQTL, genome-wide association and methylation studies, which present numerous challenges, thus the need for data integration platforms that can handle these complex data structures. Inefficient methods of data transfer and access control still plague research collaboration. As science becomes more and more collaborative in nature, the need for a system that adequately manages data sharing becomes paramount.

Results: Biology-Related Information Storage Kit (BRISK) is a package of several web-based data management tools that provide a cohesive data integration and management platform. It was specifically designed to provide the architecture necessary to promote collaboration and expedite data sharing between scientists.

Availability and Implementation: The software, documentation, Java source code and demo are available at http://genapha.icapture.ubc.ca/brisk/index.jsp. BRISK was developed in Java, and tested on an Apache Tomcat 6 server with a MySQL database.

Contact: denise.daley@hli.ubc.ca

Received on February 11, 2011; revised on June 18, 2011; accepted on June 20, 2011

1 INTRODUCTION

In biomedical studies, statistical power is important because some associations can be undetected if statistical power is low. A simple way to improve power is to increase the sample size; this drives scientists to collaborate and share their data. Several consortia have been created to facilitate collaborative scientific efforts on national and international scales. Large-scale collaborations generate substantial amounts of data that is useful for analysis but can be difficult to manage. Delegating access and managing information for such large sets of data can be problematic if not appropriately handled. Labs can adequately manage large amounts of data by using laboratory information management systems. However, delegating access to data can be a problem as some groups may want to limit access to unpublished data (Schadt *et al.*, 2010). Studies of scientific publications, from the past 20 years, have shown a continuous increase in collaboration between scientists (Sonnenwald, 2007). New tools need to be developed to account for the increasingly collaborative nature of science.

We were driven to develop Biology-related Information Storage Kit (BRISK) to facilitate collaboration and simplify data sharing for investigators in the AllerGen (The Allergy, Genes and Environment Network) consortium (http://www.allergen-nce.ca/ and http://genapha.icapture.ubc.ca/index.do). BRISK is an open-sourced software package that provides the infrastructure necessary for efficient communication between collaborators regarding data and sample storage, data formats and data retrieval. It has the ability to handle and organize large datasets, including data generated in longitudinal studies. BRISK has been used by investigators to retrieve information from high-throughput Genome-Wide Association Studies (GWAS) with hundreds of thousands of markers (>500 000), to smaller scale Illumina Golden Gate arrays with 1536 markers.

2 FUNCTIONALITY

BRISK is a package of three Java EE web applications that can be used independently, or as a package to provide a unifying system for managing samples and data. Communication between tools in BRISK is made possible by the underlying database. The BRISK software package has three main services; a Sample-based laboratory information Management System, Investigator Services and Web Information Services. Figure 1 provides an overview of BRISK's structure and the functionality of the three services. BRISK was developed for deployment on a server running Apache Tomcat 6.0 and a MySQL database, with a Linux OS; however, these are not strict requirements. The underlying code of the application is written in Java 1.6, and uses Hibernate 3.3.1 to facilitate the mapping of Java classes to database tables. The web interfaces are generated by JavaServer Pages backed by the Struts Application Framework.

BRISK was designed for clinical investigators and both wet (genotyping biology, microbiology) and dry laboratory (statistical and administrative) users. BRISK is intended to be installed by a database manager/developer and comes with deployment instructions, and both user and technical manuals for developers and permissions documentation which can be downloaded from the BRISK site (http://genapha.icapture.ubc.ca/brisk/documentation.do). In addition we have created an interactive demonstration (demo site) preloaded with practice data that will allow users to explore the utility of the BRISK package.

2.1 Sample-based Laboratory Information Management System

Sample-based Laboratory Information Management System (SLIMS) is a feature-rich laboratory information management

*To whom correspondence should be addressed.

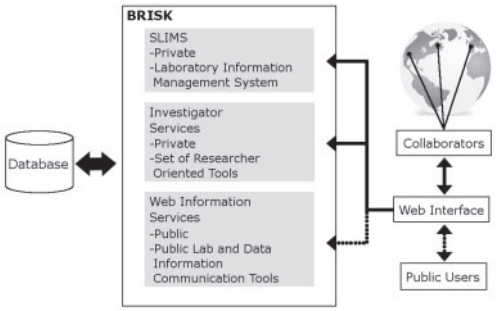

Fig. 1. A brief overview of BRISK's structure and functions.

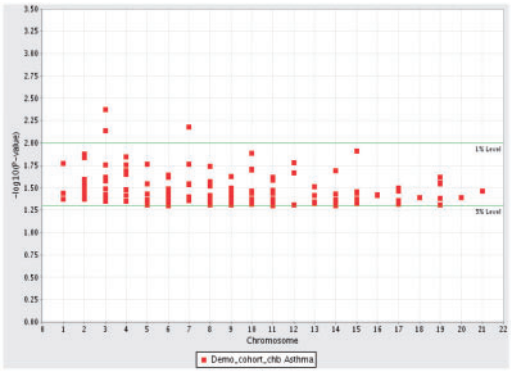

Fig. 2. Example of Manhattan plots for P-values ≤ 0.05.

system (Van Rossum *et al*., 2010) It is intended to be used internally by lab staff to manage sample, subject and container data. Since the last release version of SLIMS, four new key features—barcode support, remote sample entry, attachment of files, Illumina importer—have been implemented into SLIMS to make it an invaluable tool for sample management. The addition of barcode functionality, for samples, provides lab staff greater control over the management of physical samples. A new interface made specifically for remote-entry was developed. Samples can be tagged with barcodes and added into the database at the site of sample collection; this is crucial for the management of samples in multicenter studies. Users can attach datasheets and digitized documents onto each sample, providing a centralized location for investigators to obtain sample information. The attached files are stored as Binary Large Objects on the database, so there is no discrimination of file types when uploading. Database updates and insertions are made simpler through a new data importer. When genotyping data returns from a genotyping run at a centre that uses the Illumina platform, we can use the exported data file generated by Illumina BeadStudio to automatically populate the database with little effort from the user.

2.2 Investigator services

Intended to be used primarily by principal investigators, statisticians and data analysts, Investigator Services acts as a centralized location for retrieving data. It allows each user to be self sufficient, without the need to rely on another staff member to generate analysis files for them. After the database has been populated with data from a genome wide association study (GWAS), Investigator Services provides users with the ability to generate a variety of data formats, which can then be used as input files for the various software packages used in the statistical analysis. This has been accomplished with the use of a graphical user interface (GUI) that questions the user about the data format and content. Once the user has selected the dataset, to be analyzed they are asked if they want to filter data by removing duplicate subjects, twins and other relevant information. For family datasets, the user is queried on the data formats (trio's or case-control). The next screen asks the user what genes to include, the phenotype to be analyzed, covariates to be added and for continuous variables do they want to convert the data to a binary phenotype. The final step asks the user for the data format, current options include UNPHASED, binary PLINK, FBAT and SAGE file formats.

This allows each data analyst to work off a standardized set of data to prevent the use of outdated or modified data. An automated permissions request system is implemented to facilitate data requests and maintain integrity of consortium studies. Each user can be limited to the amount of data they have access to. If a user wants access to data they are not currently authorized to access, primary investigators can be notified to provide the user with the necessary permissions. The status of data requests can be monitored using a web-based user interface. Additionally, all data requests are date and time stamped to provide an electronic record of data requests and data releases. In our personal implementation of BRISK, Investigator Services is only available via an internal network link, this is to protect and maintain the security of confidential and sensitive information contained in the genotype and phenotype databases which could be used to identify individuals. Research Oriented tools (analysis files, raw genotyping QC results and protocols, questionnaire, environmental exposure information) refers to utilities that can interface directly with raw genotypes, or link genotypes, phenotypes, exposures or questionnaire responses together. Direct access to these databases is restricted to individuals that are on-site and within the secure network. Information requests from outside investigators are accepted by a public server, the request is then automatically forwarded to the internal network by the either database administrator or the local Principal Investigator in charge of the study.

2.3 Web Information Services

Web Information Services is the public arm of our services and was developed as an automatic data retrieval system for the massive amounts of data generated in GWAS. It has two main functionalities: it provides the lab with a place to describe their research and present their findings, and it allows the public to explore and interact with the data. The default Web Information Services template can be modified with ease to fit the theme of the lab, and the information that needs to be presented. By default, when data are imported into the database through SLIMS, data are put into the 'Private' dataset therefore only accessible by users with login credentials. However, data can be made accessible to the public by adding the data to the 'Public' dataset. Communication Tools includes a suite of utilities that allows users to graphically view, plot and interact with association results. Manhattan plots can be plotted to the user specifications, a GUI guides the user through the selection of variables to be plotted, see Figure 2 for an example plot of association results.

Basic analysis tools are available for users that are interested in taking a further look at the data; these tools include many of the features first developed for PATH (Zamar *et al*., 2009) including a suite of mapping tools that connect to external databanks such as the National Center for Biotechnology Information (NCBI), and Kyoto Encyclopedia of Genes and Genomes (KEGG) (Kanehisa, 2002; Kanehisa and Goto, 2000; Kanehisa *et al*., 2006, 2008) and the Genetic Association Database (Becker *et al*., 2004; http://geneticassociationdb.nih.gov). Custom mapping tools allow the user to build their own pathways and view pathways and test for gene–gene interactions. Web Information Services has been set-up using the Genapha website as a model (http://genapha.icapture.ubc.ca/index.do). Genapha was initially designed to host the research results and findings from a collaborative international genetics initiative to identify genes related to asthma and allergy phenotypes (Daley *et al*., 2009). During the development Genapha, we continued to add new tools and features, and over time Genapha expanded from a static results repository into an interactive site and we have incorporated many of these features into BRISK.

3 DISCUSSION

BRISK was built taking into account the growing collaborative nature of science. Data management in collaborative research involves working with restrictions and protocols that are not routinely found in research carried out in a single laboratory. Researchers must consider how the data will be stored and distributed so anyone that needs to access the data can do so without difficulty. On the other hand, they must also restrict access to sensitive and undisclosed data. Data integrity is ensured through a multi-tier login system; each user account can only access data in their account tier. By providing efficient data storage and features that cater to the needs of researchers, BRISK sets itself apart from other database management tools. There is a similar open-sourced data management package called SIMBioMS (Krestyaninova *et al*., 2009). The feature set offered by SIMBioMS is oriented toward storage and retrieval-related functions used in collaborative studies; many of these features are also present in BRISK. However, in addition to these basic features, BRISK contains tools that help researchers organize, mine and analyze study data. Furthermore, there are tools in BRISK that allow researchers to test and explore additional data for further analysis. These features enable BRISK to act as a centralized hub where users can go to access information.

BRISK is an advance from BioWarehouse (Lee *et al*., 2006) in that BRISK unifies information from clinical (phenotypes), laboratory (SLIMS biological samples) and GWAS association results with electronic online database resources. BioWarehouse compiles information from a variety of electronic resources such as ENZYME, KEGG and BioCyc and integrates the information from these sources using a mulitdatabase approach. BioWarehouse does not incorporate analysis tools, clinical or genotype information. BioWharehouse retrieves and stores information locally. The advantages of this approach are that users can select which version of a database they wish to interrogate and they are not limited in the types of databases that can be queried. The disadvantage is that the data is stored locally and may not be up to date.

Similar to BioMediator and the BioConductor platforms that have been developed for the analysis of gene expression arrays (Mei *et al*., 2003) BRISK uses a semantic architecture such that the data stays at the source (NCBI resources). This architecture has several advantages; the data is stored, managed and maintained at the source rather than locally. This avoids data warehousing, which helps to manage and maximize local resources by cutting down on the storage space necessary for BRISK. It also ensures that at query time, when the data is accessed, the data is up to date and accurate. We are able to do this because the NCBI resources can be queried via the internet, although the number of queries per user per minute is limited. Genetics is a rapidly evolving and changing discipline and information is being changed, compiled and updated daily, given these conditions dynamic access is preferred over data warehousing. Also similar to BioMediator and BioConductor, BRISK is an open source project, we anticipate that ongoing development will be collaborative. BRISK has been proven in heavy use and has been developed and maintained by a professional development team. To facilitate the ongoing development of BRISK we have published the project on Google Code see (http://code.google.com/hosting/). Each service has its own project site Investigator Services (http://code.google.com/p/brisk-investigator-services/), Web Information Services (http://code.google.com/p/brisk-web-information-services/) and BRISK Sample-based Laboratory Information Management System (http://code.google.com/p/brisk-slims/). This will allow BRISK to further evolve to meet the needs of its users and developers world-wide.

Limitations: BRISK currently manages and maintains genotype, clinical and laboratory information. It works well with GWAS level data but we acknowledge that BRISK does not address all the needs of 'omics' studies. We have not addressed gene-expression, methylation or sequence level data concerns in this release as these utilities are currently in development. Data integration platforms are an emerging endeavor and there is no 'one size fits all' approach at this time. This is likely because the development of these platforms has been driven by the research needs of the groups they serve. BRISK and Genapha (www.genapha.ca) were developed to meet the needs of an emerging genetics consortium whose needs were the management of laboratory samples, clinical phenotypes, exposure and questionnaire data and genotypes from candidate gene and GWAS. BRISK has numerous utilities to address these needs.

However, sequence level data will soon be the 'norm' for genetic consortium studies, and new data models will need to be developed that can handle the complexity and size of genome-wide sequence data. Data import and exporting of genome-wide sequence data within a MySQL database framework is unlikely to be feasible, as it will be prohibitively time consuming. Data management needs are growing rapidly and the need for new tools that can organize and integrate data from diverse data structures and sources are urgently needed.

ACKNOWLEDGEMENTS

We would like to thank everyone who's used and taken the time to comment on the functionalities offered by BRISK.

Funding: AllerGen NCE Inc. (the Allergy, Genes and Environment Network), a national multidisciplinary research network supporting research, networking, commercialization, knowledge mobilization and capacity building activities that contribute to reducing the morbidity, mortality and socio-economic impact of allergic disease.

Additional funding provided by the Canadian Institutes of Health Research (CIHR), D.D. is the recipient of a Michael Smith Foundation for Health Research (MSFHR) Career Scholar Award and holds a Tier II Canadian Research Chair appointment.

Conflict of Interest: none declared.

REFERENCES

Becker,K.G. *et al.* (2004) The genetic association database. *Nat. Genet.*, **36**, 431–432.

Daley,D. *et al.* (2009) Analyses of associations with asthma in four asthma population samples from Canada and Australia. *Hum. Genet.*, **125**, 445–459.

Kanehisa,M. (2002) The KEGG database. *Novartis Found. Symp.*, **247**, 91–101; discussion 101–103, 119–128, 152–244.

Kanehisa,M. *et al.* (2008) KEGG for linking genomes to life and the environment. *Nucleic Acids Res.*, **36**, D480–D484.

Kanehisa,M. and Goto,S. (2000) KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.*, **28**, 27–30.

Kanehisa,M. *et al.* (2006) From genomics to chemical genomics: new developments in KEGG. *Nucleic Acids Res.*, **34**, D354–D357.

Krestyaninova,M, *et al*. (2009) A system for information management in BioMedical studies–SIMBioMS. *Bioinformatics*, **25**, 2768–2769.

Lee,T.J. *et al.* (2006) BioWarehouse: a bioinformatics database warehouse toolkit. *BMC Bioinformatics*, **7**, 170.

Mei,H. *et al.* (2003) Expression array annotation using the BioMediator biological data integration system and the BioConductor analytic platform. *AMIA Annu. Symp. Proc.*, 445–449.

Schadt,E.E. *et al.* (2010) Computational solutions to large-scale data management and analysis. *Nat. Rev. Genet.*, **11**, 647–657.

Sonnenwald,D.H. (2007) Scientific collaboration. *Ann. Rev. Inform. Sci.*, **41**, 643–681.

Van Rossum,T. *et al*. (2010) SLIMS–a user-friendly sample operations and inventory management system for genotyping labs. *Bioinformatics*, **26**, 1808–1810.

Zamar,D. *et al.* (2009) Path: a tool to facilitate pathway-based genetic association analysis. *Bioinformatics*, **25**, 2444–2446.