

# A Personalized and Automated dbSNP Surveillance System

Shuo Liu, Steve Lin, Mark Woon, Teri E. Klein, and Russ B. Altman  
*Department of Genetics, Stanford Medical Informatics*  
251 Campus Drive, MSOB X-215, Stanford, CA, 94305-5479, USA  
Corresponding Email: russ.altman@stanford.edu

## Abstract

*The development of high throughput techniques and large-scale studies in the biological sciences has given rise to an explosive growth in both the volume and types of data available to researchers. A surveillance system that monitors data repositories and reports changes helps manage the data overload. We developed a dbSNP surveillance system (URL: <http://www.pharmgkb.org/do/serve?id=tools.surveillance.dbsnp>) that performs surveillance on the dbSNP database and alerts users to new information. The system is notable because it is personalized and fully automated. Each registered user has a list of genes to follow and receives notification of new entries concerning these genes. The system integrates data from dbSNP, LocusLink, PharmGKB, and Genbank to position SNPs on reference sequences and classify SNPs into categories such as synonymous and non-synonymous SNPs. The system uses data warehousing, object model-based data integration, object-oriented programming, and a platform-neutral data access mechanism.*

## 1. Introduction

Surveillance systems are used throughout the biomedical domain and beyond to provide critical monitoring and reporting functions[1-10]. Examples of such systems include the public health surveillance systems[5], the HIV surveillance systems [2, 3], MedLine alert systems[6], and security monitoring systems[7]. With the explosive growth in both the volume and types of data available to researchers, such systems are becoming essential to help researchers manage data overload.

Single nucleotide polymorphisms (SNPs) are the most common genetic variations. The dbSNP database[11], hosted by the National Center of Biotechnology Information (NCBI), is a central repository for SNPs. Because SNPs are expected to

facilitate large-scale association genetics studies, there has recently been great interest in SNP discovery and detection. The Jan 2003 build of dbSNP contains more than 4.89 million human SNP submissions, in addition to SNPs for many other organisms.

The Pharmacogenetics and Pharmacogenomics Knowledge Base (PharmGKB; <http://www.pharmgkb.org/>) [12] provides data and tools to support pharmacogenetics research. Because SNP information is an important component of pharmacogenetics studies, it is critical that investigators have timely knowledge of identified SNPs in order to avoid duplicated effort and optimize experimental design. We designed the dbSNP surveillance system to help manage this information about SNPs.

The system, in contrast to some similar applications, is personalized and fully automated. Instead of presenting a long list of genes or requiring query of genes for each new session, a custom gene list is maintained for each registered user. Our system requires little human intervention or supervision, and programmatically integrates data from dbSNP, LocusLink, PharmGKB, and Genbank to add value to the basic SNP data, including positioning SNPs on reference sequences and classifying SNPs into categories such as synonymous and non-synonymous SNPs. After each data update, it can send email notifications to highlight new entries for genes of interest. Somewhat unexpectedly, we have found that this system also provides a useful feedback mechanism for investigators to confirm that their submissions have appeared correctly in dbSNP.

## 2. Design

The surveillance system is structured in three layers: the user interface layer, the application layer, and the data repository layer. The user interface consists of web pages available over the Internet. JavaServer Pages (JSP) technology developed by Sun allows the presentation of dynamic content.

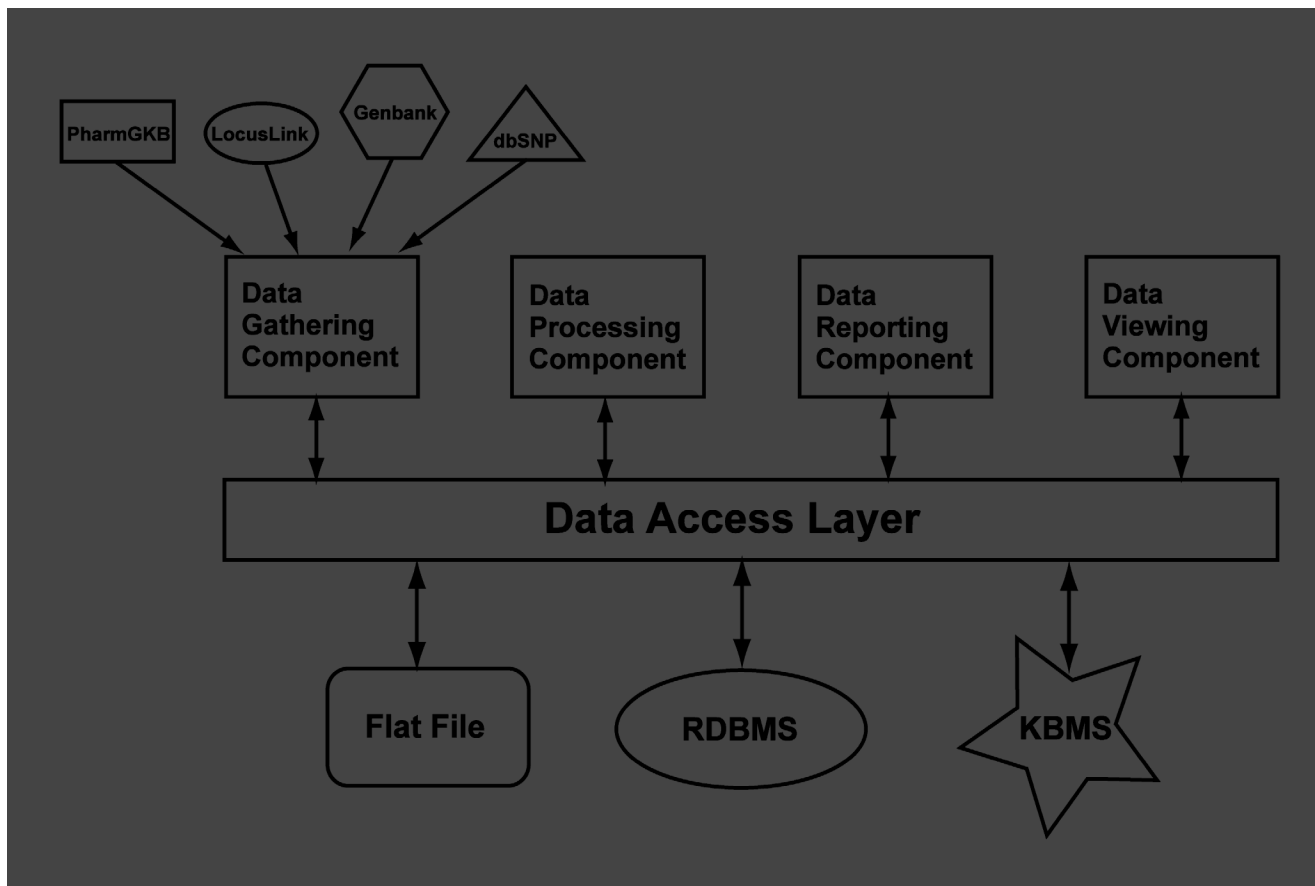


Figure 1. An architecture diagram showing the components of the system.

In the application layer, a common object model that encompasses Gene and SNP classes allows for manipulation of information without awareness of its source and storage format. A data access layer provides the mechanism to support the storage and retrieval of the data from storage systems including relational, object-oriented, frame-based and flat file sources. The data access layer works directly with objects for data storage and retrieval and bridges objects and repositories through mappings between data model and object model. The existence of the data access layer allows flexibility and portability of the system with regard to the storage system.

For the data repository layer, a critical requirement is the ability to integrate data from several data repositories: Locus Link, Genbank, PharmGKB, and dbSNP. Two major categories of data integration are data federation and data warehousing. In data federation, data in heterogeneous data sources remain where they are, whereas in data warehousing, data in heterogeneous data sources are extracted and consolidated in a database called a data warehouse. A data warehouse is optimized

for reporting and analysis. The data warehousing approach was chosen because of the needs to timestamp data, to save derived results, and to optimize performance, scalability, and reliability. With the data federation approach, it is not easy to satisfy the first two needs and the system will not be as scalable and reliable because of the dependence on the external data sources.

Figure 1 shows the component view of the system architecture. The top left corner has the four external data repositories from which the Data Gathering component retrieves data from. Data Processing component processes the raw data and stores the derived results. Data Reporting component handles the report generation and notification to the users. Data Viewing component represents the web-based user interface. Below the Data Access Layer are the three representative types of data repositories that can be used for the system: Flat File, Relational Database Management System (RDBMS), and Knowledge Base Management System (KBMS).

For each gene in PharmGKB, curated information such as the official HUGO Gene Nomenclature Committee (HGNC) gene name, gene symbol, and

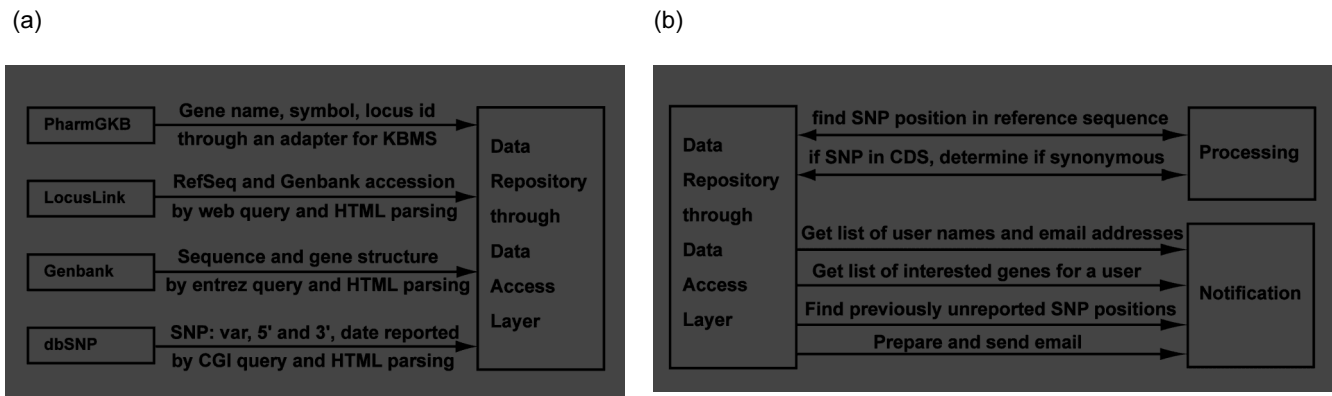


Figure 2. Data flow diagrams for the data gathering, data processing, and data reporting functions.

associated sequence accessions is downloaded from Locus Link. RefSeq sequence(s) and coding sequence (CDS) positions are downloaded via Entrez from Genbank. The RefSeq and Genbank accessions are then used to perform dbSNP queries to download SNP records associated with each sequence. This data flow is represented in Figure 2(a). The flanking sequences of each SNP record are aligned to each RefSeq sequence to position it and classify it into the following categories: non-coding, coding synonymous, and coding non-synonymous for the specific RefSeq and SNP combination. For non-synonymous coding SNPs, amino acid changes are determined by comparing mutant with wild type sequences. The upper part of Figure 2(b) shows this data flow.

PharmGKB maintains a personalized gene list for each registered user. A user can put genes of interest on this list and have convenient access to SNP information on these genes. At the end of each data update cycle, the system determines if any new SNP discoveries have been made on the genes of interest for each user and sends out a personalized email alert. This part of the data flow is shown in the lower part of Figure 2(b).

### 3. Results

A typical result page showing the SNP information for a list of genes is shown in Figure 3. The format allows a researcher to locate a SNP on a reference sequence and

determine whether it affects the protein product. The amino acid change information allows a researcher to assess the potential impact of the change on the protein product. A hyperlink connects each SNP to the corresponding record at dbSNP for further investigation. Newly downloaded SNPs are flagged with red asterisks.

### 4. Discussion

The surveillance system is useful for managing publicly available SNP information. The alignment of SNP flanking sequences to reference sequences is currently somewhat conservative. Perfect matches are required to position a SNP on a sequence. Less stringent criteria would allow the positioning of more SNPs, perhaps with less confidence.

RefSeq sequences assigned by LocusLink are currently used as reference sequences. Some researchers have their "favorite" sequence for a gene that differs from RefSeq sequences and we currently do not support alternative reference sequences.

We designed a personalized and fully automated dbSNP surveillance system to help researchers stay informed about SNP data hosted in dbSNP. Our architecture is general and can support similar surveillance systems for other resources. These capabilities may become increasingly important, in order to assist researchers in coping with the explosive growth of biomedical data.



## Acknowledgement

We thank Mike Hewett and Farhad Shafa for support and suggestions. RBA, TEK, S. Lin, MW and S. Liu are supported by NIH GM61374. S. Liu was a Stanford Graduate Fellowship James Clark Fellow and is also supported by NIH LM07033. We also thank SUN, Inc. for a hardware grant.

## References

[1] B. Foot, M. Stanford, J. Rahi, and J. Thompson, "The British Ophthalmological Surveillance Unit: an evaluation of the first 3 years," *Eye*, vol. 17, pp. 9-15, 2003.

[2] A. K. Nakashima and P. L. Fleming, "HIV/AIDS Surveillance in the United States, 1981-2001," *J Acquir Immune Defic Syndr*, vol. 32 Suppl, pp. S68-85, 2003.

[3] M. Kihara, M. Ono-Kihara, M. D. Feldman, S. Ichikawa, S. Hashimoto, A. Eboshida, T. Yamamoto, and M. Kamakura, "HIV/AIDS Surveillance in Japan, 1984-2000," *J Acquir Immune Defic Syndr*, vol. 32 Suppl, pp. S55-62, 2003.

[4] K. Osaka, H. Takahashi, and T. Ohyama, "Testing a symptom-based surveillance system at high-profile gatherings as a preparatory measure for bioterrorism," *Epidemiol Infect*, vol. 129, pp. 429-34, 2002.

[5] L. R. Jorm, S. V. Thackway, T. R. Churches, and M. W. Hills, "Watching the Games: public health surveillance for the Sydney 2000 Olympic Games," *J Epidemiol Community Health*, vol. 57, pp. 102-8, 2003.

[6] B. B. Cavanaugh and A. S. Horne, "Ovid's evidence-based medicine reviews: a review of a clinical information product," *Med Ref Serv Q*, vol. 18, pp. 1-14, 1999.

[7] S. A. Velastin, M. A. Vicencio-Silva, B. Lo, J. Sun, and L. Khoudour, "A distributed surveillance system for improving security in public transport networks," *Measurement and Control*, vol. 35, pp. 209-13, 2002.

[8] A. C. M. Fong and S. C. Hui, "Web-based intelligent surveillance system for detection of criminal activities," *Computing & Control Engineering Journal*, vol. 12, pp. 263-70, 2001.

[9] K. Y. K. Ng and A. Ghanmi, "An automated surface surveillance system," *Journal of the Operational Research Society*, vol. 53, pp. 697-708, 2002.

[10] G. Roberts and I. Darney, "Novel rapid deployment surveillance system," *Computing & Control Engineering Journal*, vol. 13, pp. 21-5, 2002.

[11] S. T. Sherry, M. H. Ward, M. Kholodov, J. Baker, L. Phan, E. M. Smigielski, and K. Sirotkin, "dbSNP: the NCBI database

of genetic variation," *Nucleic Acids Res*, vol. 29, pp. 308-11, 2001.

[12] M. Hewett, D. E. Oliver, D. L. Rubin, K. L. Easton, J. M. Stuart, R. B. Altman, and T. E. Klein, "PharmGKB: the Pharmacogenetics Knowledge Base," *Nucleic Acids Res*, vol. 30, pp. 163-5, 2002.