

# A Resource to Acquire and Summarize Pharmacogenetics Knowledge in the Literature

Daniel L. Rubin, Michelle Carrillo, Mark Woon, John Conroy, Teri E. Klein, and Russ B. Altman

Department of Genetics, Stanford University, Stanford, CA 94305-5210, USA

## Abstract

*To determine how genetic variations contribute the variations in drug response, we need to know the genes that are related to drugs of interest. But there are no publicly available databases of known gene-drug relationships, and it is time-consuming to search the literature for this information. We have developed a resource to support the storage, summarization, and dissemination of key gene-drug interactions of relevance to pharmacogenetics. Extracting all gene-drug relationships from the literature is a daunting task, so we distributed a tool to acquire this knowledge from the scientific community. We also developed a categorization scheme to classify gene-drug relationships according to the type of pharmacogenetic evidence that supports them. Our resource (<http://www.pharmgkb.org/home/project-community.jsp>) can be queried by gene or drug, and it summarizes gene-drug relationships, categories of evidence, and supporting literature. This resource is growing, containing entries for 138 genes and 215 drugs of pharmacogenetics significance, and is a core component of PharmGKB, a pharmacogenetics knowledge base (<http://www.pharmgkb.org>).*

## Keywords:

Pharmacogenetics, drugs, genes, knowledge management, databases.

## Introduction

One of the major challenges recently identified for biomedicine in the post-genomic era is to develop genome-based approaches so that we can ultimately predict the drug response in individual patients [1]. Pharmacogenetics, the study of how variation in human genes leads to variation in the response to drugs, is beginning to address this challenge by applying high-throughput DNA sequencing methods to characterize sequence variations in pharmacologically important genes. The availability of a high-quality sequence of the human genome makes it possible to pursue a “genotype to phenotype” approach in which genetic variations are observed, and the phenotypic consequences of these variations are studied. High-throughput sequencing is cost-effective, and the finite size of the human genome with its roughly 30,000 genes makes this approach attractive.

There are difficulties in pursuing a genotype-to-phenotype strategy. For a given drug of interest, one must decide which genes to study first. It is generally desirable to choose genes with known or high likelihood pharmacological associations with the drug, such as genes affecting the drug metabolism, genes for transporters of the drug, and genes that are targets for the drug. Another difficulty is that once genotypic variation is found, it is important to know all the drugs whose pharmacologic action may be affected by that genetic variation. These difficulties could be reduced if there were a curated resource that provides facile access to pharmacogenetically important gene-drug relationships, classified by the type of those relationships.

Community database resources have been tremendously beneficial to biomedical researchers in other domains. The *Saccharomyces Genome Database* ([www.yeastgenome.org](http://www.yeastgenome.org)) curates the yeast literature and summarizes current data related to yeast genes. *FlyBase* ([flybase.bio.indiana.edu](http://flybase.bio.indiana.edu)) has created a similar database for the fly research community. The *Online Mendelian Inheritance in Man* (OMIM, [www.ncbi.nlm.nih.gov/omim](http://www.ncbi.nlm.nih.gov/omim)) is a resource that summarizes known gene-disease relationships, compiling the genes implicated in inherited diseases. These resources catalyze research in their respective fields because they provide a synthesis and compilation of critical information needed by researchers, reducing the need to search and read the literature to find this information.

A similar resource would be useful for pharmacogenetics. Associating genetic variations to drug response is limited by the extent of knowledge of all the genes and drugs that have been shown to be relevant to the particular response of interest. A database that compiled the known drug-gene relationships would be valuable to researchers for determining related genes to drugs of interest. It could also suggest other drug responses to study by showing all drugs related to particular genes.

Our objective was to develop a resource to help us identify, catalog, and disseminate pharmacologically-important drug-gene interactions for the pharmacogenetics community. This is a large task, and requires input from many experts in the field. While other databases have been created using in-house expertise, we chose a distributed community-based approach to build our resource (PharmGKB Community Submission Project, “CBS”), leveraging the Internet to link experts in

pharmacogenetics with CBS to accelerate the acquisition of important gene-drug interactions.

## Materials and Methods

### Relating genes to drugs: categories of pharmacogenetics evidence

In building the CBS resource, we believe it is important to categorize gene-drug associations according to the type of pharmacogenetics evidence that supports them [2]. Pharmacogenetics studies that connect a gene and drug are generally focused on particular aspects of the drug response phenotype: molecular/cellular events related to the drug, physiological events related to the drug getting to the target and being eliminated from the body, manifestations at the whole-organism level (clinical effects), and the ultimate effects of the drug on clinical outcome. Because of this, each gene-drug association has particular types of supporting evidence.

After consulting with pharmacogenetics experts and reviewing articles in the pharmacogenetics literature, we developed a scheme for classifying gene-drug relationships relevant to pharmacogenetics. These “categories of pharmacogenetics evidence” (COE) represent the types of phenotype data that are studied to support gene-drug associations (GN: Genotype information; FA: Molecular and Cellular Functional Assays; PK: Pharmacokinetics; PD: Pharmacodynamics and Drug Response; CO: Clinical Outcome). One can think of the COE as the type of phenotype variation that results for a particular drug given specific variations in a particular gene.

blood warfarin levels. In this case, there is a CYP2C9-warfarin interaction with pharmacokinetics evidence, so this gene-drug association would be classified in the PK category. In general, one or more of these categories is applicable to all pharmacogenetic gene-drug associations.

### Submitting gene-drug interactions

For each gene-drug association submitted to CBS, we collect the gene, drug, related disease (if applicable), and the categories of pharmacogenetics evidence supporting that association. The list of genes was drawn from the HGNC nomenclature [3]. We compiled a drug list from the Department of Veterans Affairs National Drug File (www.va.gov/vdl/Clinical.asp?appID=89), and obtained diseases from the Medical Subject Headings (MeSH, www.nlm.nih.gov/mesh/meshhome.html). In addition to this information, we record the PMID of the article that supports the gene-drug relationship, a description of where in the reference the relationship is described (title, abstract, or in the full-text of the article), pertinent keywords, and the confidence the submitter has about the submitted gene-drug interaction. In addition, a field for comments is provided to capture unstructured information.

A web-based submission page was created to capture this information (Figure 1). Because we use gene, drug, and disease names from controlled vocabularies, the name that a user enters must be matched to a term in these vocabularies. We implemented a form of the agrep algorithm [4] that includes a scoring function based on the number of exact letter matches and the total length of the name. When the user enters a name, a list of top-scoring matches in the vocabularies is re-

#### Submit Relationship Between a Gene, Drug, and/or Disease (Literature Annotation)

Please note that only registered users may submit data. If you are not a registered user and would like to submit data, you will need to [register](#). If you are a registered user, please [sign in](#). For directions on how to fill out this form, please refer to our documentation on [submitting relationship data](#).

<b>Gene:</b>	<input type="text"/>	<a href="#">[Look Up Gene]</a>	<b>Type of Evidence:</b>	<input type="text" value="Select Type of Evidence"/>
<b>Drug:</b>	<input type="text"/>	<a href="#">[Look Up Drug]</a>	<b>Evidence:</b>	<input type="text"/>
<b>Disease:</b>	<input type="text"/>	<a href="#">[Look Up Disease]</a>	<b>Evidence Location:</b>	<input type="checkbox"/> Title <input type="checkbox"/> Abstract <input type="checkbox"/> Body
<b>Categories of Pharmacological Knowledge:</b>	<input type="checkbox"/> CO <input type="checkbox"/> PD <input type="checkbox"/> PK <input type="checkbox"/> FA <input type="checkbox"/> GN			
<b>Degree of Confidence:</b>	<input type="text" value="Select Degree of Confidence"/>			
<b>Type of Evidence:</b>	<input type="text" value="Select Type of Evidence"/>			
	<b>Keywords:</b> (separated by commas)	<input type="text"/>		
	<b>Notes:</b>	<input type="text"/>		
<input type="button" value="Submit Relationship"/>				

Figure 1- Web page for submitting pharmacogenetic gene-drug interactions

For example, a study of warfarin may assess variation in the CYP2C9 gene by studying the effects of this variation on

turned in drop-down box so the proper term can be easily selected. The user can also add new terms to these vocabularies.

We defined an XML schema to specify the information content for all submissions to CBS (<http://www.pharmgkb.org/schema/docs/pharmacogeneticRelationship.html>). This enables us to accept submissions in XML and to validate all submissions to CBS before storing them in the database. The web-based page allows users to submit relationships between a single gene and drug. A particular gene and drug may participate in many gene-drug relationships, or one may wish to submit many gene-drug interactions in bulk. To accommodate these situations, we developed an Excel-based submission mechanism to allow bulk data entry into CBS. Users complete an Excel template (or generate a tab-delimited file programatically), and a software module converts the data to XML and submits it to the database. CBS is integrated into PharmGKB, a pharmacogenetics knowledge base ([www.pharmgkb.org](http://www.pharmgkb.org)). PharmGKB provides search functions to help users locate names for genes, drug, and diseases from the controlled vocabularies that can then be used in making CBS submissions.

Each submission to CBS is given an accession number and stored in a database (Oracle 9.2i running on Solaris 2.8).

### Acquiring submissions and evaluating CBS

We began submitting gene-drug associations to CBS using in-house domain experts initially. We acquired an initial set of submissions by compiling articles from the bibliographies of several major pharmacogenetics reviews and from two years of recent articles from major pharmacogenetics journals. After collecting this initial set of submissions for CBS, we contacted members of the pharmacogenetics community and invited them to participate in the CBS project. Having a partially populated database was important for us to demonstrate the potential utility of this resource and for eliciting community interest in this project.

To evaluate the success of CBS, we obtained feedback from users and submitters to CBS and we tracked submissions. In addition, we compiled lists of genes that pharmacogenetics researchers are interested in (regardless of whether they were also making CBS submissions). These gene lists contain the genes that are relevant to various investigators' research, and covered a variety of lines of investigation in pharmacogenetics. We used these gene lists to measure CBS coverage—the percentage of genes on investigator gene lists for which there are CBS entries.

## Results

### Data query and display

Users can search or browse CBS. There is full-text search, so that if the query string matches anywhere in any field of a CBS submission (including gene symbols, keywords, or anywhere in comment fields), that submission will be returned.

The browse feature is helpful for summarizing the current state of knowledge about gene-drug interactions. A user can browse the list of genes, drugs, or diseases. After selecting a

### ABCB1

ATP-binding cassette, sub-family B (MDR/TAP), member 1

[Add to Watchlist](#)

<b>Alternate Names:</b>	ATP-BINDING CASSETTE, SUBFAMILY B, MEMBER 1; ABCB1; DOXORUBICIN RESISTANCE; GP170; Homo sapiens ATP-binding cassette, sub-family B (MDR/TAP), member 1 (ABCB1), mRNA; P-glycoprotein 1/multiple drug resistance 1; P-GLYCOPROTEIN 1; PGY1; P-glycoprotein-1/multiple drug resistance-1; multidrug resistance 1
<b>Alternate Symbols:</b>	ABC20; GP170; MDR1; NM_000927.1; P-GP; P-gp; PGY1

Related Drugs	Relationship	Details
<a href="#">arsenite</a>	PD GN	
<a href="#">colchicine</a>	FA	
<a href="#">digoxin</a>	CO PD PK FA GN	
<a href="#">efavirenz</a>	PK	
<a href="#">nortriptyline</a>	PD	
<a href="#">pesticides</a>	PD GN	
<a href="#">prednisone</a>	CO PD	
<a href="#">tacrolimus</a>	FA	
<a href="#">vincristine</a>	FA	
<a href="#">xenobiotics</a>	FA GN	

Related Diseases	Relationship	Details
<a href="#">HIV</a>	PK	
<a href="#">Parkinson Disease</a>	PD GN	
<a href="#">liver transplantation</a>	FA	
<a href="#">major depression</a>	PD	

Figure 2 – CBS query result for the ABCB1 gene showing all the ABCB1-drug and ABCB1-disease associations with the applicable categories of pharmacogenetic evidence.

gene or drug, all known relationships relevant to that selection are displayed. For example, if the ABCB1 gene is selected, then all the drugs and diseases associated with ABCB1 are shown (Figure 2). Clicking on the “Details” button brings up the complete CBS submission, including the article that supports the gene-drug association.

For each drug and disease, the categories of pharmacogenetic evidence supporting the association are shown, allowing the user to quickly determine how the various drugs and diseases are related to ABCB1. This display shows that digoxin and prednisone drug actions are related to ABCB1 in studies that assessed patient outcomes. Digoxin is seen to have been extensively studied, having data for all categories of pharmacogenetic evidence. On the other hand, colchicine, tacrolimus, and vincristine have been linked to ABCB1 solely in studies using functional assay methods, but not in pharmacokinetic or pharmacodynamic studies. This may suggest opportunities for research to investigators interested in the clinical impact of ABCB1 variation and these drugs.

### CBS submissions and evaluation

To date, a total of 477 submissions relating to 138 genes and 215 drugs have been made to CBS. Table 1 shows a summary of CBS submissions by study center. Our initial submissions still constitute most of the total submissions, but our resource is still young and awareness of CBS is growing in the pharmacogenetics community. Two investigators believed there is educational value in preparing CBS submissions, and assigned

Submitting Institution	Count
Stanford University	315
Indiana University	87
University of Tennessee-Memphis	34
University of California, San Francisco	21
St. Jude Children's Research Hospital	9
Washington University	6
University of Pittsburgh	2
Mayo Clinic	2
National Institutes of Health	1

Table 1 – Count of CBS submissions by submitting institution

their pharmacology students the task of making these submissions as part of class projects.

Overall feedback from the pharmacogenetics community to CBS has been favorable. Users have found the web entry form easy to use, and have reported that submissions can be done quickly. The vocabulary lookup feature was particularly helpful. A few users have reported being unable to make submissions because of the need to submit multiple genes and drugs together (which is not supported by the web entry form), but such submissions can be made using the Excel template.

A total of 11 study centers submitted lists of genes that are relevant to their pharmacogenetics research. We also created our own list of genes of pharmacogenetic interest. We counted the number of genes for which there are submissions in CBS, and we calculated the percentage of the genes on each study center's list that have at least one entry in CBS (called "percentage of coverage," Table 2). Between 11.6-100% of genes of interest to all study centers had entries in CBS (11.6-68% for study centers not submitting to CBS).

## Discussion

A vast amount of data has been published in the biomedical literature relating genes and drugs. These relationships are key knowledge for pharmacogenetics research, as new studies frequently arise by recognizing gaps or inconsistencies in the prior knowledge. Pharmacogenetics research studies may find a gene in which sequence variations alter drug response, but that does not explain all variation in a patient population because other genetic factors are involved [5]. A resource that shows all genes related to a particular drug could help account for the polygenic nature of drug response and catalyze pharmacogenetics research. Thus, we embarked on this project to create CBS, designed to help us identify and catalog important gene-drug relationships and data sets in pharmacogenetics.

Assembling a catalogue of gene-drug associations is not sufficient; these relationships need to be categorized. Genetic variation and drug response can be studied in different ways: molecular/cellular effects related to the drug, physiological events related to the drug getting to the target and being eliminated from the body, manifestations at the whole-organism level (clinical effects), and ultimately, altered clinical outcome.

We defined categories of pharmacogenetic evidence ("COE") [2] to associate genes and drugs with the types of variations in measurable phenotypes. Defining a classification system for phenotype remains an unsolved challenge [6], and our COE categorization is an initial effort defining a phenotype classification for pharmacogenetics. In our experience, pharmacogenetic studies generally assess phenotypes in one or more of these categories. We believe that our COE classification of phenotypes is general purpose, yet extensible to include as-yet-undefined new characterizations of phenotype.

The COE are also helpful for indexing pharmacogenetic knowledge for storage and retrieval in a database, and these categories appear useful as a summary of the coverage of types of pharmacogenetics research for a particular gene-drug pair. A researcher can see at a glance cases where a gene-drug association has been extensively studied in all categories, or only in a few (Figure 2).

We use controlled vocabularies for gene, drug, and disease names because there are many synonymous terms. When performing queries by gene or drug, it is important to have one name for each entity so that all results for an entity are retrieved. There are no standard vocabularies for pharmacogenetics, but we chose HGNC because it is gaining momentum as a standard for naming genes. It was more difficult selecting standard vocabularies for drugs and diseases. Our choice was made principally based on public availability of terminologies to the academic community.

We have found that sometimes submitters are not certain about a gene-drug association discussed in an article. We use a confidence rating to capture the level of certainty that the submitter has in the association being submitted. In addition, we have occasionally received comments from the pharmacogenetics community questioning gene-drug associations others

Institution	Coverage of Watched Genes (Count (%))
National Institutes of Health	7 (100)
Indiana University	38 (73.1)
University of Chicago*	17 (68)
University of California, Los Angeles*	3 (60)
Mayo Foundation Rochester	8 (34.8)
Lawrence Berkeley National Laboratories*	17 (27.4)
University of California, San Francisco	12 (26.1)
University of California, San Diego*	14 (23.3)
Stanford University	87 (20.1)
Washington University	17 (18.9)
Vanderbilt University Medical Center*	20 (18.2)
Harvard University*	5 (11.6)

Table 2– Percentage coverage of genes of interest to research institutions performing pharmacogenetics research (\*institutions that did not submit to CBS)

have submitted. Because there may be disagreement in the community about particular articles and gene-drug associa-

tions, it is important to highlight such cases and stimulate discussion and prompt new experimentation. We are adding threaded discussions to CBS to identify such controversies and to stimulate discussion and debate in the scientific community.

Our preliminary results suggest that CBS will be a useful resource for pharmacogenetics research. We have been successful in recruiting participation from the community in submitting entries to CBS. While the total number of submissions from the community is much less than our in-house effort, the project has only been recently launched, and interest in CBS is growing. We have also shown that the current content in CBS covers genes that are relevant to pharmacogenetics researchers. Between 11.6-68% of genes of interest to study centers not making submissions to CBS have entries in our database (Table 2). Not surprisingly, centers submitting to CBS had greater coverage of their genes in CBS (18.9-100%). With the continued growth of CBS, we expect this coverage to increase.

Some potential drawbacks of CBS include lack of participation, careless data entry, limited user interface and inadequate curation of submissions. In addition to our manual efforts to alleviate these problems, we work with our community to identify problems and correct them.

Another potential problem is that because different people contribute to CBS, duplications may occur in the database. If the same gene-drug association is submitted, but with a different article that supports it, then this represents additional support for the association. In such cases, we display the gene-drug association as a single entry, but list all supporting evidence when a user expands the entry (clicking on the "Details" button, Figure 2). If two CBS submissions contain the same gene-drug association and the same supporting article, then this is a duplicate submission. We can detect such cases by queries to the database, which we are beginning to implement as part of our quality assurance procedures.

Because the submission process is distributed, CBS submissions could concentrate on certain genes or classes of genes, with paucity of data in other areas. While this could be a limitation of community-based submission to CBS, it may actually be a strength, as submissions from the community reflect the scientific interests of those making the submissions. Consequently, the CBS content may be more relevant to the community it serves.

We believe that there are numerous benefits to be gained by CBS. First, it takes advantage of the Internet to accelerate the acquisition of important pharmacogenetic data sets and bring them to the attention of the scientific community. Second, it presents a way to quickly review the state of knowledge linking genes and drugs, categorized by the type of pharmacogenetic study. Third, it creates a forum for scientists to be introduced to the CBS and to other scientists, through their submissions and discussions. Finally, it creates a tool for identifying new research opportunities in pharmacogenetics. For example, a gene-drug interaction may have good genetic polymorphism information, outcomes information and molecu-

lar/cellular phenotype data, but be relatively unexplored in pharmacokinetics or molecular pharmacodynamics.

Future work will focus on automating the process of identifying articles containing pharmacogenetics data and submitting them to CBS. Progress is being made on using natural language processing methods to identify gene and drug names in the biomedical literature. We are beginning to use these methods to search Medline for articles that describe gene-drug associations. Such methods will assist us curate the pharmacogenetics literature and expand our CBS resource.

If successful, the CBS project will become part of the research infrastructure that accelerates our ability to improve the development and delivery of drugs to patients. It may also serve as a model for other genotype-to-phenotype efforts that need to integrate and summarize knowledge from multiple scientific disciplines.

### Acknowledgments

This work is supported by grants from the National Institute of General Medical Sciences (NIGMS), Human Genome Research Institute (NHGRI), National Library of Medicine (NLM), and the NIH/NIGMS Pharmacogenetics Research Network and Database (U01GM61374). We wish to thank Caroline Thorn for contributing submissions to CBS.

### References

- [1] Collins FS, Green ED, Guttmacher AE, and Guyer MS. A vision for the future of genomics research. *Nature* 2003; 422(6934): 835-47.
- [2] Altman RB, Flockhart DA, Sherry ST, Oliver DE, Rubin DL, and Klein TE. Indexing pharmacogenetic knowledge on the World Wide Web. *Pharmacogenetics* 2003; 13(1): 3-5.
- [3] Povey S, Lovering R, Bruford E, Wright M, Lush M, and Wain H. The HUGO Gene Nomenclature Committee (HGNC). *Hum Genet* 2001; 109(6): 678-80.
- [4] Wu S, and Manber U. Fast text searching allowing errors. *Communications of the ACM* 1992; 35(10): 83-91.
- [5] Katz DA. Associating genes to drug response. *Drug Information Journal* 2002; 36: 751-761.
- [6] Freimer N, and Sabatti C. The human phenome project. *Nat Genet* 2003; 34(1): 15-21.

### Address for correspondence

Russ B. Altman (altman@helix.stanford.edu), Department of Genetics, 300 Pasteur Drive, Mail code: 5120, Stanford, CA 94305, USA.