# PharmGKB: A Resource to Link Genotype and Phenotype in Pharmacogenetics

**Daniel L. Rubin, Mark Woon, Michelle Carrillo, Steve Lin, Feng Liu, John Conroy, Winston Gor, Caroline Thorn, Micheal Hewett, Teri E. Klein, and Russ B. Altman**

*Department of Genetics, Stanford University, Stanford, CA, USA*

*The PharmGKB (http://www.pharmgkb.org/) is a pharmacogenetics and pharmacogenomics knowledge base built to support the representation, storage, analysis and dissemination of genotype and phenotype data. We have collected genetic sequence data from research centers in which polymorphic variations in genes of pharmacogenetic interest have been characterized in particular populations. PharmGKB also contains phenotype data sets that are linked to particular genotypes as well as information about key gene-drug interactions of relevance to pharmacogenetics obtained by curating the literature. PharmGKB integrates these data with relevant information from other databases and provides browsing and analytical functions to help scientists discover connections between genetic variations and alterations in drug effects and related phenotypes. Our objective is to catalyze research in the pharmacogenetics scientific community, and to discover the genetic basis for variations in the response to drugs.*

High throughput experimental methods are producing massive amounts of genetic information, but very little of it is linked to phenotype data that can be used to ascertain the functional significance of varying genotypes. Researchers in pharmacogenetics need a resource that compiles known polymorphisms in different subject populations and links these genetic variations to observed phenotypes. This resource could help them define new experiments by observing genetic variants that have not been phenotypically studied, or it could generate new hypotheses by revealing certain polymorphisms that are associated with significant phenotypes. We are designing PharmGKB to become such a resource for pharmacogenetics.

PharmGKB is built around key pharmacogenetics concepts that serve to organize all the content within the database: genes, drugs, diseases, and pathways. These concepts support search and browsing because genetic data, phenotype data sets, and literature references are annotated with relevant genes, drugs, and diseases. We use standard nomenclatures for these classes (HGNC names for genes, VA/NDFRT vocabulary for drugs, and MeSH hierarchy for diseases). PharmGKB uses a graphical web-based display to show the location of introns, exons, and polymorphisms on the gene. It also has many tabular displays modeled after data tables in journal publications that summarize polymorphic variants in populations, SNP positions, and amino acid changes.

We have developed a standard model of genetic information in XML Schema (www.pharmgkb.org/schema/) that is used for exchanging detailed genetic polymorphism data. PharmGKB uses rule-based methods to validate submitted data, and many of the rules are represented within the XML Schema so users can validate large parts of their submissions locally.

Developing standard models for representing phenotype data is challenging because of the great diversity of types of experimental studies (from cell-based experiments to clinical studies), a variety of parameters measured, and many different study endpoints. We have begun addressing this by developing categories of pharmacogenetics information to classify data in PharmGKB: studies focused on genotype (GN), molecular and cellular functional assays (FA), pharmacokinetics (PK), pharmacodynamics and drug response (PD), and clinical outcome (CO). We have also begun developing an ontology of experimental methods for measuring phenotype (the "phenotype ontology", or "PO"), inspired by the Gene Ontology project that has helped organize genetic functional annotations. By annotating phenotype data sets and the pharmacogenetics literature with terms drawn from the PO, we can aggregate data from multiple studies that were performed using similar methods. We also use PO annotations for searching PharmGKB, matching user queries to these terms and retrieving the appropriate data sets. These annotations can also be useful for surveying the state of knowledge in pharmacogenetics. For example, there may be many results for polymorphisms and clinical outcome for a particular drug and gene, but little known about pharmacokinetics. Such information could be useful to generate new hypotheses and suggest new research directions.

An additional organizing feature of PharmGKB is pathways of bioprocesses (e.g., metabolic, regulatory, or signaling pathways). In complex biological domains such as pharmacogenetics, systems biology approaches are becoming increasingly important. We are building representations of key pathways in pharmacogenetics. These pathways relate a set of genes, drugs, and phenotypes with key pharmacological or physiological systems that are relevant to drug effects. The entities in these pathways link together the core pharmacogenetics data in PharmGKB. For example, the pathway for irinotecan illustrates the relationships among the drug, critical genes and metabolites, link to the genetic variants studied, and see the different types of phenotype studies performed as well as the actual phenotype data reported. We believe that these pathways will become a key launching point for browsing data in PharmGKB, as well as for providing an overview of the state of knowledge for key genes related to drug responses of interest.