PharmGKB: a logical home for knowledge relating genotype to drug response phenotype

To the Editor:
The Pharmacogenomics and Pharmacogenetics Knowledge Base (PharmGKB, found at http://www.pharmgkb.org/) is devoted to cataloging information about pharmacogenes—those genes involved in modulating the response to drugs. Genes may be pharmacogenes because they are involved in the pharmacokinetics of a drug (how the drug is absorbed, distributed, metabolized and eliminated) or the pharmacodynamics of a drug (how the drug acts on its target and its mechanisms of action). PharmGKB's goal is to be a comprehensive resource on pharmacogenes, their variations, their pharmacokinetics and pharmacodynamic pathways and their effects on drug-related phenotypes. Whereas the older field of pharmacogenetics often focused on the effect of single dominant genes on drug response, pharmacogenomics connotes the study of the multigenic influences on drug response, often using modern high-throughput experimental techniques.

The sequencing of the human genome and the study of human genetic variation have opened up great opportunities for understanding the association between genotypes and phenotypes. There is an active debate about the merits of single, centralized databases to hold all genetic variation information and associated phenotypes. Recently, there has been a movement toward this model with the introduction of the National Center for Biotechnology Information (NCBI) dbGAP databases to hold genotype and phenotype data for large genome association trials (http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=gap).The success of GenBank and the suite of NCBI data resources has demonstrated clearly that the NCBI is capable of meeting the demand for high volumes of raw data and serving as a reliable repository.

In addition to the need for providing raw data in standard formats, there is a major, continuing need for integration, aggregation and curation of the information contained within these data stores for the purpose of supporting specific areas of scientific enquiry. For example, locus-specific databases (LSDBs) discussed in the accompanying correspondence[1] are thriving, because their curators carefully scrutinize the raw genotype and phenotype measurements and present the data along with summaries of the literature and additional information about rare mutants—particularly those with important phenotypes. In addition, model organism databases (MODs) provide extensive curation of the genomic sequence and functional information associated with those organisms. Curators comb the literature and create summaries of gene function in both textual and controlled terminologies, and they integrate high-throughput data sets relevant to their mission.

Although not an LSDB or a MOD, PharmGKB has elements of both. Like the LSDB curators, the PharmGKB curators constantly survey the literature and other databases for reports of important genetic variations in known pharmacogenes. PharmGKB accepts and/or integrates information from the central data warehouses about sequence polymorphisms in order to provide a single location for high-quality information about the location and population frequencies of variations in pharmacogenes. In particular, curators look for variations that have functional phenotypic consequences related to drug response. They create summaries of the pharmacogenomics literature to create a definitive list of gene-drug interactions and characterize those interactions. Like MOD curators, the PharmGKB curators attempt to provide annotations of the functions and phenotypes of pharmacogenes. Thus, PharmGKB also accepts and/or integrates information from the central warehouses about drug-related phenotypes. Curators integrate high-throughput data sets relevant to drug response. They work to define these phenotypes with controlled terminologies to facilitate indexing, searching and aggregation of these data. They also work with members of the US National Institutes of Health (NIH) Pharmacogenetics Research Network (PGRN) to generate summaries of important genes and their phenotypes and create pathway diagrams relevant to drug response.

Today, the PharmGKB has curated evidence for 1,994 genes involved in drug response. PharmGKB has high-quality genotype variation data (in many cases with population frequencies) for 240 genes, and 1,671 literature entries have been curated to create gene-drug associations that are labeled with respect to the type of information contained in the papers. There are 38 manually created drug-related pathways created in collaboration with PGRN investigators and others. Finally, we have introduced a new Very Important Pharmacogenome (VIP) initiative to create structured summaries of key pharmacogenes, their important polymorphisms, phenotypes, haplotypes and alternative splices (if relevant). Sixteen such summaries are available today. Usage statistics (as summarized on my 'PharmGKBlog') show more than 2,000 registered users (who can gain access to individual-level data) with more than 50,000 unique internet address visitors per month. All data and knowledge contents are available for download. PharmGKB provides a cross-reference file that associates all SNPs in PharmGKB with identifiers from the Golden Path human genome browser, dbSNP, HapMap, jSNP, Illumina, Affymetrix, SeattleSNP and ALFRED.

Thus, PharmGKB is a hybrid resource that provides both primary data as well as curated knowledge about pharmacogenomics. It benefits from the existence of data warehouses, because although it is capable of accepting and presenting primary data, it will increasingly depend on reliable archival resources for basic primary data storage. More importantly, PharmGKB will provide the aggregation, integration of literature and summaries of knowledge (through pathways and VIP genes) that still require PhD-level human curation. Our software developers are charged with building tools to help curators work more effectively in managing the increasing volume of pharmacogenomics science and to help users in searching, visualizing and analyzing the data and knowledge contained in the knowledge base.

The five-year goal of PharmGKB is to be a comprehensive store of information about pharmacogenes and their associated phenotypes and to catalyze research in pharmacogenomics. In the longer term, the knowledge contained within PharmGKB will be used as a starting point for implementing genome-informed drug prescribing decisions. With sufficient information, we may be able to move toward predictive pharmacogenomics, where patterns of variation in pharmacokinetics and pharmacodynamics for existing drugs are used to predict the variation in response to new drugs.

Russ B Altman

Department of Bioengineering and Department of Genetics, Stanford University, Stanford, California 94305-5120, USA.
e-mail: russ.altman@stanford.edu

1. Horaitis, O. *et al. Nat. Genet.* **39**, 425 (2007).