PharmGKB, a Centralized Resource for Pharmacogenomic Knowledge and Discovery

Li Gong and Teri E. Klein

Differential response to the standard dose of drug therapy in patients is commonly seen in clinical practice. Environmental factors, diet, age, gender, disease severity, interacting drugs, and genetic variation all contribute to the variability in drug response. The importance of genetic variation in drug response has become more prominent with the emergence of pharmacogenomics and personalized medicine. Pharmacogenomics is the study of how genetic differences affect responses to drugs. By knowing more about a person’s genetic makeup, clinicians will be better able to assess the risks and benefits associated with medications to maximize treatment success. Pharmacogenomics has the potential to have an impact on many steps of medical care, from diagnosis to tailored drug prescription, and from basic drug discovery to clinical trial design.

In the past two decades, substantial knowledge has accumulated about genetic events contributing to differences in drug responses, some resulting in drug-labeling changes and clinical practices (1, 2). U.S. drug labels for 6-mercaptopurine, warfarin, and irinotecan contain dosage adjustments based on TPMT, CYP2C9, and VKORC1, and UGT1A1 genotypes (3–5). As the U.S. Food and Drug Administration (FDA) requires more pharmacogenomic information to be included on drug labels, pharmacogenomics will be increasingly accepted and integrated into mainstream clinical practice. We are already observing a steady shift from the “trial and error” approach to a more knowledge-guided personalized approach toward drug therapy. To realize the full potential of pharmacogenomics, many formidable challenges still need to be overcome. These challenges include the ethical, economical, and legal issues associated with the ever expanding field of genetic testing, as well as the strong need to increase genetic literacy among patients and health care providers. Easier access to well-characterized clinical outcome and biometric data on patients under treatment are also crucial to identify and track genotype-phenotype relationships. With the combined efforts of researchers and health care providers to use the knowledge of pharmacogenomics in drug treatment and diagnosis, the barriers between scientific discovery and the clinical application of pharmacogenomics will diminish over time to realize the full benefits of the field.

The completion of the human genome sequence and the availability of affordable high-throughput genotyping technology have enabled pharmacogenomic researchers to study drug response in the context of the entire genome. The amount of data accumulated from various genome-wide studies grows exponentially each year. With data scattered throughout the literature and more than a thousand biological databases, as well (6), it is a significant challenge for researchers to find relevant data and to turn those data into distilled knowledge. Establishing centralized domain-specific knowledge bases is a more efficient way to manage and query information within specific biomedical fields. The Pharmacogenomics Knowledge Base (PharmGKB, http://www.pharmgkb.org) is a centralized Web-based resource for pharmacogenomic data and knowledge. Originated in 2000 as a National Institutes of Health (NIH)-sponsored pharmacogenetics knowledge base for the scientific community (7), PharmGKB has become a preeminent resource that disseminates comprehensive and up-to-date knowledge and data in pharmacogenomics (Figure 5.1). The knowledge domain includes extensively reviewed information on genes and important variants that are implicated in drug response, and it presents condensed information in the form of drug-centered pathways, very important pharmacogenomic summaries (VIPs), and annotation of functionally important variants and relationships among gene, drug, and disease. The primary data domain comprises both genotype and phenotype data submitted by the scientific community (7). As of June 2009, PharmGKB contained information on more than 800 genes with genotype information (20,970 subjects that have been genotyped with 3,401,867
PharmGKB, a Centralized Resource for Pharmacogenomic Knowledge and Discovery

Knowledge domain:

- Functional annotation for variants that impact drug-response phenotypes (2,374 concise summaries)
- Major repository for pathways of drug action and metabolism (60 manually curated pathways)
- Publication-quality in-depth summaries for “very important pharmacogenes” (VIPS) (39 in-depth summaries)
- Gene-drug-disease interactions from the literature, tagged with relevance to pharmacokinetics (PK) and pharmacodynamics (PD) (3,626 manually curated articles)

Data domain:

- Data repository for Pharmacogenetics Research Network (PGRN) (810 genes with genotyping information, 20,970 subjects that have been genotyped with 3,401,867 variants reported)
- Data repository for international data-sharing consortia (International Warfarin Pharmacogenetics Consortium [IWPC] and International Tamoxifen Pharmacogenomics Consortium [ITPC])
- Searchable summary and links to high-impact datasets of relevance to pharmacogenomics

Figure 5.1. The PharmGKB content overview. PharmGKB is composed of two complementary domains: a knowledge domain and a primary data domain. The knowledge domain includes very important pharmacogene (VIP) summary, annotation of variants of pharmacogenomic interest, drug response pathways, and relationships among genes, drugs, and diseases. The primary data domain contains both genotype and phenotype data. Establishing correlations between genotypes and phenotypes will generate novel pharmacogenomics relationships, and the knowledge gained can in turn be used to catalyze scientific research and discovery.

PharmGKB HOME PAGE AND BASIC QUERY

PharmGKB’s home page (www.pharmgkb.org) is the entry point for most users (Figure 5.2). It was designed to highlight the interests of its users and to ensure that the primary contents of the knowledge base are easily accessible to them. The menu tabs at the top of the page provide access to the top-level section of the PharmGKB site. Home is the front page where we highlight the knowledge and data content, mission, contact information, and registration; Search is the main search page where the user can search by either free text or canned queries, or browse information by domain; Submit describes how a user can submit genotype, phenotype, pathway, or literature data; Help includes an extensive list of background information, downloads, and educational and technical references; PGRN lists all members involved in the NIH Pharmacogenetics Research Network, their research interests, submissions to PharmGKB, and information in support of the PGRN; Contributors is the section where people are listed who have contributed data to PharmGKB; Clinical PGx catalogs drugs with pharmacogenomic information in the context of FDA-approved drug labels; My PharmGKB is the section for our registered users to view and edit their profile, submission, and Web site statistics and only appears after a user has logged in. A user can ask questions of the PharmGKB Team regarding any aspect of PharmGKB by clicking the Feedback button, which is located at the top right corner of every PharmGKB Web page. To orient our users to the most important

variants reported), 2,374 variants with functional annotation, 594 drugs with supporting data, 60 drug pathways, 39 in-depth pharmacogene summaries, and more than 3,626 curated articles. Whether searching for candidate genes for a specific drug treatment or disease susceptibility, or a specific genetic variation and its functional consequence regarding drug response, PharmGKB’s highly structured knowledge content and flexible query interface provide the infrastructure to easily retrieve information of interest and advanced tools to support hypothesis generation.

This chapter describes PharmGKB’s knowledge and data contents, basic navigation and search of the database, and Web site contents. Acute lymphoblastic leukemia (ALL) is used as an example to demonstrate how clinicians can use PharmGKB to find pharmacogenomic information related to a specific disease. A study design workflow is also presented on how PharmGKB can be used to facilitate study design and downstream analysis to stimulate scientific discovery.
knowledge content pertinent to pharmacogenomics, we present four featured types of data in the center of the home page: Genes with pharmacogenetic significance; genetic variants associated with drug response; pathways for drug transport, metabolism, and action; and drugs for which there is pharmacogenomic knowledge. The current “favorite” papers selected biweekly by our curators and a dynamically generated news feature are presented on the right side of the home page.

Users may query and browse PharmGKB in many ways. The general search box above the menu bar allows free text search across all domains in the PharmGKB database, whereas the focused search embedded in the four main data types narrows the search results by data categories. For example, a search in the genes search box (e.g., drug names, variants, and gene names) will return a list of genes related to that search text. Similarly, a search under the variant search box will only return a list of variants whose documentation contains the input query. If a user is looking for information on “which genetic variants play a role in response to gefitinib?” she can enter “gefitinib” in the variant search box, which will return the list of variants in genes EGFR, ABCG2, and CYP2D6 that are associated with gefitinib response (Figure 5.3). The highly integrated nature of our database and focused search strategy provide users with flexibility in how they find information and allow them to ask mechanistic questions about the relationships between drugs and genes and genetic variation.

PharmGKB gene, drug, and disease pages are structured similarly with a tab system. Figure 5.4 demonstrates the general layout of a PharmGKB gene page for VKORC1, which encodes a key enzyme in the vitamin K cycle with a significant contribution to the variable response of warfarin therapy. The Overview tab for the gene lists the official and alternative names and symbol for the gene, its genomic boundaries, and associated OMIM phenotypes. The VIP tab includes links to...
PharmGKB, a Centralized Resource for Pharmacogenomic Knowledge and Discovery

Figure 5.3. Example of focused search for variants associated with response to gefitinib.

Figure 5.4. Example of PharmGKB Gene page (VKORC1, vitamin K epoxide reductase complex, subunit 1). The PharmGKB gene page is organized by a tab system and provides detailed information on synonyms, phenotype, genomic location, variant information, associated pathways, related drugs and diseases from the literature, as well as download/cross-references to other complementary genomic databases.
Figure 5.5. Example of PharmGKB Variant Gene Page (VKORC1) with variant browser and variant table. The PharmGKB variant page contains a graphic browser on the top and a variant table below. The variant browser displays all the polymorphisms across the span of the gene of interest from various resources such as PharmGKB primary data, SNP arrays (Illumina and Affymetrix), dbSNP, and jSNP. Each tick on the browser represents a variant from the respective resource. Gene features are color coded to differentiate exons, introns, promoters, and UTRs. The variant table lists the details for nonarray genotype data in PharmGKB, such as their genomic positions, functional roles, frequencies, and assay types.

the in-depth pharmacogene summary with introductory information describing the gene, important variants/haplotypes, and their impact in drug response. The Variant tab includes a graphic browser on the top and variant table below (Figure 5.5). The Datasets tab contains the phenotype data stored in PharmGKB for the gene of interest. The Pathway tab lists all related pathways for the gene. Gene-drug-disease relationships can be found under the Related Drugs and Related Diseases tabs. Links for the download of experimental data for the gene and cross-references to other complementary genomic databases are under the Download/LinkOuts tab.

KNOWLEDGE DOMAIN: SIGNIFICANT PHARMACOGENES/VARIANTS, DRUG PATHWAYS, AND GENE-DRUG-DISEASE RELATIONSHIPS

PharmGKB collects and summarizes knowledge on genes and genetic variations related to the safety and efficacy of
Significant Pharmacogenes/Variants

Variability in drug response is often related to functional changes in genes/proteins that are involved in the pharmacokinetic (PK) or pharmacodynamic (PD) processing of the drugs. The PK genes generally include drug metabolism enzymes such as cytochrome P450s (or CYPs) and sulfotransferases, and drug-uptake or efflux transporters such as members of the OATP and p-glycoprotein families. The PD genes usually involve the drug target, its downstream signaling molecules, and other molecules that might modulate the biological context where the drug-target interaction happens. Variation in any element that controls the PK or PD process may lead to variability in drug efficacy or safety. PharmGKB presents knowledge for key genes and important variants in the forms of (1) in-depth summary for VIPs (three stars) and (2) the abbreviated summary for variants of pharmacogenomic interests (Annotations, two stars) such as functional variants, tagging SNPs, and variants in association studies. The VIPs are structured summaries that include detailed information about a given gene, including its important polymorphisms, haplotypes, phenotypes, interacting drugs and complete mapping information, and supporting literature references. An allele frequency table may also be included if the specific variant has been studied extensively in different populations. To significantly increase the breadth of coverage for important variants, PharmGKB has begun to catalog and highlight variants of pharmacogenomic significance with the use of concise summaries. In contrast to the VIPs, which are encyclopedia-like summaries that require large amounts of manual curation, the two-star annotated variants are minimally curated, yet still bring key phenotypic consequences of individual variants to our users. Two variants (VKORC1:rs9923231 and CYP4F2:rs2108622) that are important for warfarin dosing illustrate these two different levels of curation (Figure 5.6). Because nomenclature inconsistencies for genes and variants present a major challenge when searching for relevant pharmacogenomic information, PharmGKB includes lists of current as well as historical terms used for genes and their variants in both our VIPs and variants of interest efforts. For example, rs9923231, a common promoter variant for VKORC1, has been mentioned in the literature as “G3673A” and “-1639G>A” (Figure 5.4). By including the alternative names that have been used in the literature as references, our users can quickly reconcile the multiple naming issues when conducting comprehensive searches in biomedical literature for genetic variations.

Drug Pathways

Many of the early successful examples in pharmacogenetics focused on one gene and its candidate SNPs. Although studies of individual genes are valuable for understanding their function in drug response, it is well accepted that genes and their proteins do not act in isolation, but rather as components of larger pathways or networks. Similarly, most drug affects are determined by the interplay of multiple gene products in biological and pharmacological pathways that modulate the pharmacokinetics and pharmacodynamics of the drug (10, 11). PharmGKB drug-centered pathways aim to provide an overview of how genes are involved in the pharmacokinetics and pharmacodynamics of drugs. The pathway diagrams use standard shapes and colors to represent genes, drugs, metabolites, and interactions. Users can click on each of these objects to go directly to an individual drug and gene page to access more specific information. Each pathway has a textual summary description and a downloadable evidence spreadsheet that provides literature support for each interaction depicted on the pathway diagram. Unlike pathway resources that primarily focus on common biological and physiological processes (e.g., KEGG, Reactome, Biocarta, GenMAPP), PharmGKB is the only resource with a primary focus on drug-centered pathways, in particular, PK pathways (12–15). The genes included in the PK and PD pathways can serve as candidate genes for a pharmacogenetic study of the drug and can be used to help interpret findings from a genome-wide association study. At present, PharmGKB has sixty pathways that cover drugs used in a wide variety of disease classes such as asthma, cancer, cardiovascular disease, diabetes, depression, HIV, and inflammatory diseases. PharmGKB pathways and VIPs are now published monthly in Pharmacogenetics and Genomics (16). Figure 5.7 shows an example of a published drug response pathway for the chemotherapy agent etoposide. This pathway illustrates the complex gene-drug interaction network.
A. 

Figure 5.6. Example of PharmGKB variant annotation (VKORC1: rs9923231 in-depth annotation and CYP4F2: rs2108622 brief annotation). PharmGKB collects and annotates variants of pharmacogenomic interest at two curation levels: in-depth VIP annotation (three stars) and brief annotation (two stars). (A) rs9923231, a VKORC1 variant important for warfarin dosing, is annotated as an in-depth VIP variant. The VIP variant annotation includes a detailed functional summary, associated haplotypes, allele frequency, interacting drugs, complete mapping, and supporting literature references. (B) rs2108622, a CYP4F2 variant that has recently been implicated in warfarin dosing, is annotated with a brief summary, key phenotypic consequences, alternative names, and supporting evidence.
that is involved in the metabolism, cellular disposition, mechanism of action, and delayed toxicity of etoposide (17).

**Gene-Drug-Disease Relationships**

Capturing the complex interactions between genes, drugs, and diseases is a continuing focus in pharmacogenetic and pharmacogenomic research. Most of the accumulated knowledge is only found in journal articles and books. PharmGKB curators routinely scan the primary literature to annotate gene-drug-disease relationships and present the knowledge at PharmGKB in the Related Genes, Related Drugs, and Related Diseases sections. Figure 5.8 shows an example of gene-drug-disease relationships for the antidiabetic drug rosiglitazone (Avandia®). Relationships between the drug-gene and drug-disease are tagged with Categories of Pharmacogenetic Knowledge (PK, PD, GN [Genotype], FA [Molecular & Cellular Functional Assays], CO [Clinical Outcome]) for easy classification and retrieval, and are linked to the original journal article for detailed information. Recently, we have implemented an automated literature pipeline that identifies papers through text-mining approaches (18) and then highlights potentially relevant terms for quick orientation to pertinent information (19). The automatic scan of applicable literature can be found in the Non-curated Information section, along with the metabolizing enzymes and drug target information automatically retrieved from DrugBank (20).
Figure 5.8. Example of PharmGKB curated publications (rosiglitazone). PharmGKB annotates gene-drug-disease relationship and presents the knowledge in the Curated Publications section. Relationships between the drug-gene and drug-disease are tagged with categories of pharmacogenetic knowledge (PK, PD, GN, FA, CO) for easy classification and retrieval, and linked to the original journal article for detailed information.

**PRIMARY DATA DOMAIN**

In addition to our knowledge mission, PharmGKB is the central primary data repository for pharmacogenetic and pharmacogenomic studies. Initially, the data repository was seeded by studies from the Pharmacogenetics Research Network (PGRN) and focused on a handful of genes. As whole-genome scans become more widely used, PharmGKB has expanded its capacity to accommodate large-scale high-throughput data that may involve a large number of samples assayed across the entire genome. For example, SNP array experiments on pharmacogenomics of statin therapy can now be viewed and downloaded from PharmGKB (https://www.pharmgkb.org/search/browseAlpha.action?browseKey=snpArray). As of June 2009, more than 20,972 subjects had been genotyped with 3,461,867 variants reported in PharmGKB. In addition, there were 461,324 distinct phenotype measurements. The large volume of integrated primary data in PharmGKB offers potential for meta-analysis between related pharmacogenomic phenotypes. More recently, PharmGKB has moved into a leadership role in creating data-sharing consortia for high-impact pharmacogenomics collaborations. The goal of the consortia is to create merged international datasets from a large multicenter, multinational cohort of patients to develop the best strategy for predicting drug efficacy and toxicity. Two example consortia are the International Warfarin Pharmacogenetics Consortium (IWPC) (http://www.pharmgkb.org/views/project.jsp?pid=56) (21) and the International Tamoxifen Pharmacogenomics Consortium (ITPC) (https://www.pharmgkb.org/views/project.jsp?pid=63). The IWPC pooled data from more than 5,000 patients and derived a globally relevant warfarin pharmacogenetic dosing equation (22). The IWPC dataset is carefully curated, stored, and managed at PharmGKB. The detailed demographic and phenotype information (e.g., ethnicity, age, height, weight, dose, and target International Normalized Ratio or INR) for all subjects in the study as
well as associated genotypes is available for download to researchers worldwide (Figure 5.9A). Besides in-house primary data, PharmGKB regularly surveys literature and external archival databases to identify important datasets of relevance to pharmacogenomics, and to provide access to these datasets for our users through a searchable summary of the study tagged with pertinent genes, drugs, and diseases by the use of standardized vocabularies. Figure 5.9B illustrates an example dataset summary for the genome-wide association study of addiction housed at the National Center for Biotechnology Information database of Genotypes and Phenotypes (dbGaP). The experimental results of the study from dbGaP can be accessed through the link under External Data Links tab.

CLINICAL PHARMACOGENOMICS (PGx)

In April 2009, PharmGKB introduced the Clinical PGx domain to bring accurate and up-to-date pharmacogenomic information in the context of FDA-approved drug labels to clinicians and the general public (Figure 5.10). This page includes information for all the drugs for which the FDA drug label has been revised to include specific pharmacogenomics recommendations, as well as drugs with mounting pharmacogenomic evidence. Clicking on a drug on this page opens a new window displaying pharmacogenomic drug-labeling information, FDA-approved/cleared diagnostic tests for the drug, and links to relevant PharmGKB resources. A summary table of pharmacogenomic diagnostic tests for many of the most-studied genetic variant-drug interactions is also available from the useful link section on the home page (https://www.pharmgkb.org/resources/forScientificUsers/pharmacogenomic_diagnostic_tests.jsp).

DISEASE CASE STUDY: HOW TO FIND IMPORTANT PHARMACOGENOMIC INFORMATION ASSOCIATED WITH A DISEASE AND ITS THERAPEUTICS (THE ALL EXAMPLE)

Cancer has a strong genetic component, making it an ideal field for pharmacogenomic research. Anticancer drugs, in general, have a narrow therapeutic index and exhibit significant interindividual variability in their efficacy and toxicity. With multiple treatments available for many cancer types, there is a pressing need for tools to aid decision making on drug selection. Methods to identify genetic variants that contribute to variable clinical outcomes associated with chemotherapy drugs have been an area of active research for many decades. Variants in the drug-metabolizing enzymes (e.g., TPMT, UGT1A1, and CYP2D6) have been identified to be associated with response to chemotherapy drugs (23–25). Gene expression profiling has also revealed gene signatures for classifying disease and predicting treatment outcome (26, 27). The great strides made in the field of cancer pharmacogenomics generated deeper insights into the pathogenesis of cancer and the mechanisms of drug sensitivity, as well, to provide clinicians the opportunities to personalize chemotherapy to maximize efficacy and reduce toxicity for each patient.

Research in childhood ALL is a classic example of how pharmacogenomic information can be used to benefit clinical practices. Without the introduction of any new chemotherapy drugs in the past 30 years, the overall survival of patients with ALL has improved from less than 10 percent in the 1960s to more than 80 percent at present (28). One of the key reasons behind the advance in cure rate is the improved ability to identify patients early on who are at high risk of treatment failure or severe toxicity. A well-characterized pharmacogenetic marker for chemotherapy agents used to treat ALL is human thiopurine methyltransferase (TPMT). TPMT is a phase II metabolizing enzyme catalyzing the metabolism and intracellular inactivation of thiopurine drugs such as 6-mercaptopurine and azathiopurine. TPMT activity is highly variable, with approximately 10 percent of the patients having intermediate enzymatic activity and 0.3 percent being deficient for TPMT activity. Numerous studies have shown that patients with inherited TPMT deficiency are exposed to a higher concentration of the active drug, and are at risk for severe or fatal hematopoietic toxicity when given the normal dose of the drug. TPMT deficiency has also been linked to a higher risk of second malignancies among patients with ALL. Several genetic polymorphisms in TPMT (notably TPMT*2, TPMT*3A, and TPMT*3C) have been discovered that lead to the reduced activity of the enzyme. A genetic test on TPMT genotype has been developed to identify patients who are at higher risk of severe bone marrow toxicity and thus require significant dose reductions to avoid life-threatening toxicity. This example demonstrates the clear utility of using pharmacogenomic knowledge for dose selection and optimizing drug therapy. The early groundbreaking research in ALL pharmacogenomics was conducted by PGRN scientists; the relevant knowledge and significant datasets are archived at PharmGKB; and knowledge is represented in the form of drug pathways, significant variant annotation, and literature annotation.

Searching for Specific Information on the PharmGKB

To search for information relevant to a disease in PharmGKB, simply type the disease name in the general search box, for example, “ALL,” and then open the
Figure 5.9. Example of high-impact datasets with relevance to pharmacogenomics, housed at PharmGKB or at external archival databases. PharmGKB identifies and curates high-impact datasets from pharmacogenomic studies. (A) Example of a carefully curated dataset housed at PharmGKB (IWPC dataset). (B) Example dataset summary for important pharmacogenomic study housed at external archival databases (study of addiction: genetics and environment from dbGAP, http://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs000092.v1.p1).
PharmGKB disease page for ALL (precursor cell lymphoblastic leukemia-lymphoma) (Figure 5.11). PharmGKB Disease names are imported from Medical Subject Headings (MeSH, http://www.nlm.nih.gov/mesh/). Precursor cell lymphoblastic leukemia-lymphoma is the official disease name in MeSH for ALL. On the ALL disease page, the primary phenotype data files associated with ALL can be found under the Datasets tab. These include PK studies on drugs used in ALL treatments and gene expression studies done to correlate drug resistance in children with ALL. The pharmacogenomic knowledge accumulated from ALL studies can be found under the Genetics, Pathways, Related Genes, and Related Drugs tabs. Under the Genetics tab, users can find lists of genetic variations that have been associated with treatment response in ALL, including ones associated with end-of-induction minimal residual disease in ALL. In the Pathways section, there are five pathways on pharmacokinetics and/or pharmacodynamics of the drugs used to treat ALL (doxorubicin, etoposide, methotrexate, thiopurine, and vinca alkaloids). Clicking on the pathway name opens the pathway diagram that provides a quick overview of genes that are involved in the metabolism, transport, and action of the drug of interest. The genes and drugs on the pathways are all hyperlinked to the specific gene and drug page for more detailed information. For example, TPMT is shown on the thiopurine pathway diagram to convert mercaptopurine into an inactive metabolite called methylmercaptopurine (meT-GMP), thereby decreasing the formation of the active thioguanine nucleotides. To learn more about polymorphisms for TPMT, clicking on the TPMT gene from the pathway image will display detailed variant annotations from the TPMT variant page and VIP page. Variants that lead to decreased enzymatic activity of TPMT can be easily identified by viewing the star annotations column in the variant table. Under the Related Genes and Related Drugs tabs, related genes and drugs from the literature for ALL are presented and tagged with our categories of knowledge. Clicking on View under the Details column in related genes/drugs from the literature section will lead to links to the original paper documenting the relationship between ALL and respective genes and drugs.
Adverse drug reactions (ADRs) are a major concern in current drug therapy across all diseases. Each year, millions of people in the United States experience an ADR using marketed drugs, and ADR is ranked as one of the leading causes of illness and death. ADRs are also the top reason for drug withdrawal from the market and are responsible for the termination of approximately 20 percent of investigational drugs in the drug development pipeline (29). Given the significant impact, lowering ADR rates is a major goal for both the medical field and the pharmaceutical industry. Although ADRs can result from a variety of factors (e.g., age, organ function, drug interactions), genetic factors play a significant role in the incidence and severity of ADRs (2). Identifying genetic markers to predict which individuals are at greater risk for ADRs has been a high-priority area of pharmacogenomic research (30). Once identified and validated, the genetic variant information can be incorporated into a diagnostic test that will help predict a patient’s response to a specific drug and guide treatment and dosage selections. Currently, the two principle approaches to study ADRs are the candidate gene approach and the genome-wide approach. The candidate gene approach selects genes for study based on knowledge of the target or metabolic pathways of the drugs used. Alternatively, the genome-wide approach uses a broad discovery-based method by examining hundreds of thousands of SNPs simultaneously. These two approaches are complementary, each with its advantages and disadvantages. The rational selection of genes in the candidate gene approach normally leads to biologically meaningful results with a low-cost, moderate sample size and a low false-discovery rate. However, novel genes might be missed. The genome-wide approach allows identification of new candidate genes that were previously unknown to be important in response to a drug. This later type of study, however, normally requires increased expense, has a large sample size, and has significantly higher risk of false discovery because of multiple comparisons. Each of the novel genes or variants identified from the genome scan will need to be evaluated individually, and the positive findings need to be replicated in independent datasets. The public SNP databases now
contain more than 18 million validated SNPs, and the number is continuing to grow (31, 32). These SNPs are of various degrees of data quality, with only a small portion of them characterized in terms of population frequencies, and an even smaller fraction that have been functionally characterized. It becomes an increasing challenge to sort through a vast number of SNPs to select appropriate ones for pharmacogenetic studies. The extensive knowledge content stored at PharmGKB can help scientists quickly identify sets of genes and important variants that might reasonably be expected to modulate pharmacogenetic phenotypes. Scientists conducting studies using the candidate gene approach can use PharmGKB knowledge to front-load a study with relevant PGx candidate genes and variants. Scientists pursuing the genome-wide approach can use knowledge extracted from PharmGKB to filter and prioritize results by looking for genes and variants with a pharmacogenomic rationale. We will use ADRs associated with statin treatments to demonstrate how PharmGKB can be used to guide the knowledge-based study design with the use of the candidate gene approach.

The β-hydroxy-β-methylglutaryl-coenzyme A (HMG-CoA) reductase inhibitors, collectively known as statins, are currently one of the most effective medications for managing elevated concentrations of low-density lipoprotein cholesterol. Statins reduce the frequency of coronary heart diseases by as much as 21 percent to 43 percent and are the most prescribed class of drugs in America and worldwide. Six statins are currently marketed in the United States, including atorvastatin, simvastatin, rosuvastatin, pravastatin, fluvastatin, and lovastatin. These statins all target HMG-CoA reductase, but differ in terms of their potencies and PK properties. Although statins are generally well tolerated, the response to treatment varies greatly among individuals. Up to 7 percent of patients receiving statins develop muscle-related complications (ranging from aches, cramps, mild myalgia, to severe rhabdomyolysis) and require either dosage reduction or discontinuation of therapy. Cervastatin (Baycol) was withdrawn from the U.S. market in 2001 because of nearly 100 cases of rhabdomyolysis-related death. Various mechanistic hypotheses have been proposed to explain statin-induced muscle injury. These hypotheses range from pharmacodynamics (e.g., interference with interactions with target or mitochondrial function) to pharmacokinetics (e.g., variability in drug metabolism and transport) (33–36). However, the exact role of genetic polymorphisms in predicting adverse events associated with statin use is still unknown. A knowledge-guided candidate gene study may help us gain significant insight into the mechanism of statin-induced ADRs. PharmGKB can be used in the following fashion to help select candidate genes and variants in the study design for statin-induced ADRs (Figure 5.12).

Step 1. Candidate Gene Selection
Genes involved in pharmacokinetics or pharmacodynamics of the drugs are prime candidates for selection in a pharmacogenomic study. PharmGKB has multiple statin PK/PD pathways that summarize the complex multigenic influences on statin drug response. The PK pathways describe genes involved in the absorption, distribution, metabolism, and excretion of statins, whereas the PD pathways illustrate the physiological effects of the drug, its mechanism of action, and possible side effects. Each gene in the PK or PD pathway can serve as a candidate gene for a pharmacogenetic study of that drug. The detailed information on a specific gene can be retrieved by clicking on the gene symbol. The pathway for each drug can be found on the drug page under the Pathways tab. If no pathway is currently available for the drug of interest, another place to retrieve the gene-drug relationship is under the Related Genes tab. This section includes genes that have an established relationship to the drug of interest from published articles. Similarly, we can start from the Related Diseases page to look for candidate genes to follow up in the ADR study. For a statin-induced ADR, we can search for diseases such as “Myalgia” or “Muscular diseases” and then go to the specific disease page to find pathways and related genes under the Pathways and Related Genes tabs. This will potentially offer additional candidate genes from a different perspective that may not have been previously known to be related to statin response.

In addition to the method using only curated information available at PharmGKB, we also provide a link to an expanded “candidate gene finder tool” under the “PGx tool box” (http://www.cbs.dtu.dk/services/PGx_pipeline-1.0) on the home page. This tool was developed by using PharmGKB gene-drug relationships as well as external protein-protein interaction information and drug-drug similarity measurements to help prioritize pharmacogene candidates. By comparing an input drug (or a set of input drugs) and its (their) putative indications with a repository of known gene-drug interactions, the PGx pipeline ranks the entire genome for their likelihood of being involved in the pharmacogenetic response to the inputted drug(s) (37).

Step 2. Significant Variants Selection
Once a candidate gene list is assembled, the next step is to generate a list of functionally significant SNPs for the genes of interests. If a PharmGKB VIP page is available for the gene, the VIP will describe the most important variants and haplotypes for the gene of interest. Alternatively, the scientist can go to the PharmGKB gene page, under the Variants tab, to browse through the variant table to look for variants with “star” annotation under the Annotated Variant curation level column. In addition, SNPs that lead to changes in amino acid composition, activity of the protein, or expression of the gene
and SNPs that reside in the regulatory regions of the gene are good candidates to be included for the study.

Step 3. Selection of Variants with Desirable Population Frequency
If the association study will only include a certain ethnic population (e.g., Asian or white), variants with extremely low frequency in the desired population can be further screened out by using frequency information on PharmGKB. The allele frequency information can be found under the frequency column of the variant table via the Expanded Variants View button. Clicking on the frequency value will display the breakdown of frequencies by racial categories. Additional frequency information
can also be found in allele frequency databases such as ALFRED (http://alfred.med.yale.edu/).

Step 4. Find Assay and Primer Information for the Chosen Variants
If PharmGKB has genotype data submitted for the chosen variant, our users can find information such as assay methods and primers by clicking on the variants in the PharmGKB variant table. For instance, in many cases, if the TaqMan assay was used to genotype a specific drug-metabolizing enzyme variant, PharmGKB has a direct link to ordering information at Applied Biosystems to help the user locate the material required for the study.

Study design is an important aspect of any scientific investigation. By iterating through these steps, a scientist can rapidly compile a list of candidate genes and important variants to follow up in the candidate gene study to explain and predict the adverse effect profiles of the drug. This can potentially save scientists tremendous time and effort in their literature-mining process during the study design stage.

CONCLUSION
Pharmacogenomics is a rapidly evolving field with great potential to help clinicians select the most appropriate drug and doses in treatment of their patients. Advances in high-throughput genotyping and sequencing technologies and the availability of international publicly available databases are enabling new advances in pharmacogenomics and bring us closer to the long-awaited era of personalized medicine. PharmGKB is designed to provide centralized access to comprehensive knowledge and significant datasets that are central to pharmacogenomic research and to provide links out to other relevant resources. We will continue to curate knowledge about how human genetics has an impact on drug response phenotypes. We will also create and provide access to more informatics tools that will be useful for understanding the mechanism of drug response. By aggregating, integrating, and annotating the latest findings in pharmacogenomic research, PharmGKB serves as a valuable resource, not only for basic science researchers, but also for the broader community, including clinicians, pharmacists, students, educators, and the general public, as well.

ACKNOWLEDGMENTS
This work is supported by the National Institutes of Health, National Institute of General Medical Sciences (U01GM61374).

REFERENCES


