CHAPTER 7

Pharmacogenomics

Key points

- Pharmacogenomics relates to the study of how genetic variation influences drug action.
- Genetic variants can have effects on dosing (how much of a drug is required), efficacy (whether the drug will work), and adverse events (what side effects an individual experiences).
- Common pharmacogenomic variants can be directly mapped to personal genomes.
- The functional effects of rare variants in pharmacogenes can be predicted and visualized using a number of methods.
- Information from multiple variants in a single drug pathway can be aggregated to estimate a pathway mutational load, which may indicate a potential pharmacogenomic effect.

7.1 What is pharmacogenomics?

Pharmacogenomics is broadly defined as the study of the relationship between variation in drug effects (pharmacology) and genetic variation (genomics). Pharmacogenomics is often put at the forefront of any discussion about personalized medicine, because it more directly connects genetics with clinical action: knowing that a person with a given genetic disposition is more or less sensitive to a drug enables personalized prescriptions. Since the dawn of medicine, it has been well known that two individuals can respond differently to the same drug. In some cases, the drug might simply fail to work for one patient while exhibiting the expected effect on another. In other cases, a drug might work well in a majority of patients, yet cause a subset of patients to experience unexpected, and sometimes deadly, side effects. These abnormal, harmful side effects of drugs are known in the medical literature as adverse drug reactions (ADR), and they account for millions of medical complications and more than 100,000 deaths per year in the United States alone. Not surprisingly, many ADRs have been associated with particular genetic variants, which are typically only found in a minor proportion of the population. In this chapter, we will focus on the issues of pharmacogenomics for personal genome analysis. For more general information on pharmacogenomics, we refer readers to *Principles of Pharmacogenetics and Pharmacogenomics* (Altman, Flockhart, Goldstein 2012).

Of course, as with any application of personal genomics, the associations and predictions made in this chapter are not determinant of drug response and phenotype. Care must be taken in the interpretation of these variants, which should be discussed with a physician before taking any action.

There are many ways in which the composition of a personal genome sequence can influence drug response. When an individual takes a drug, it goes through a number of steps to produce its effect: for instance, it must be absorbed into the body, find its target, and then act on its target. The main facilitators of these actions are proteins, which are the products of genes, and a drug may interact with dozens of proteins even before it interacts with its intended target. Some of these proteins may aid the drug in reaching its target, while others may be innocent bystanders to which the drug inadvertently binds. Not surprisingly, if two different people have two versions (variants) of one of these proteins, the drug may affect the efficiency or mechanism of that step

Exploring Personal Genomics. First Edition. Joel T. Dudley and Konrad J. Karczewski.
© Joel T. Dudley and Konrad J. Karczewski 2013. Published 2013 by Oxford University Press.

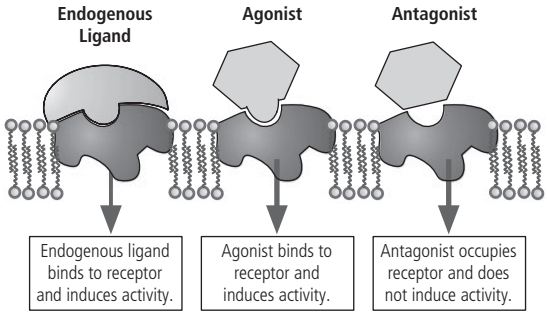


Figure 7.1 The actions of agonist and antagonist drugs: Drug targets, such as the cell surface receptors shown here, typically bind to one or many endogenous ligands to facilitate cellular signaling or other aspects of typical cellular physiology. Agonist drugs attempt to mimic the action of an endogenous ligand by binding to a functional region of the target protein to initiate some kind of downstream activity. Antagonist drugs also bind to target proteins, but instead of activating the target, they occupy a region of the protein to competitively block binding of an endogenous ligand to inhibit downstream activity.

differently, which can change the clinical effects of the drug (intended or otherwise). For instance, one individual may have a protein variant that reduces absorption of the drug, rendering it ineffective. In a much more severe case, another individual may have a variant in a protein that now converts this drug into a toxic compound. Genes whose protein products interact with drugs fall into two general categories, "what the drug does to the body", or pharmacodynamics (PD) and "what the body does to the drug", or pharmacokinetics (PK).

7.1.1 Pharmacodynamics genes

Pharmacodynamics (PD) genes encode proteins that are the targets of the drug: they are the biological molecules with which the drug directly interacts to facilitate the drug's biological effect. Through the course of their treatment, drug compounds will physically bind to one or several specific drug targets, which will either activate (agonist) or inhibit (antagonist) the normal biological function of the drug target (Figure 7.1). Many PD genes are cell surface receptors, such as G-protein coupled receptor (GPCR) or ion channels, which are found on the exterior of the cell membrane. Interactions between a drug with these receptors can have significant downstream effects on the cellular

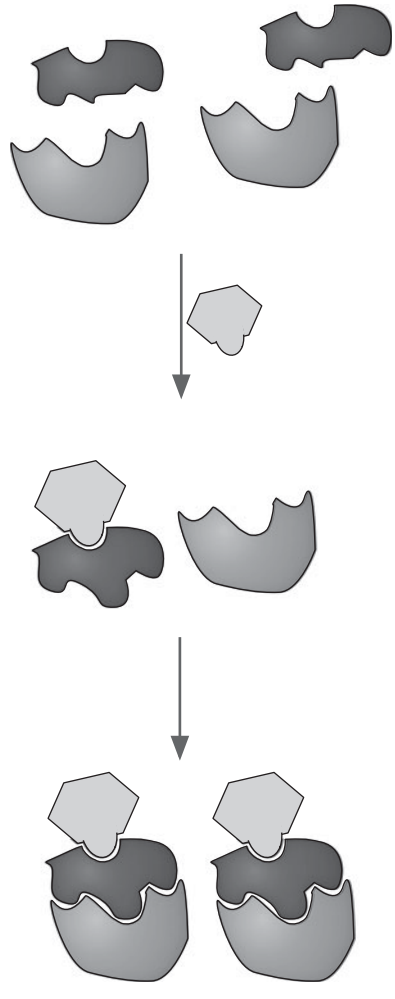


Figure 7.2 Mechanism of action of drugs: After a drug binds to its target, any process downstream of the drug can produce a phenotype or effect. For instance, the drug may cause a conformational change in one protein that allows it to bind to another protein, which in turn causes a physiological effect.

physiology. Other types of PD genes include enzymes and nuclear hormone receptors. Variation in PD genes can alter the structure or chemistry of the target proteins, which might change the chemistry or dynamics of the physical interaction between the drug and the target (Figure 7.2).

7.1.2 Pharmacokinetics genes

Pharmacokinetics (PK) genes encode proteins that are involved in the absorption, distribution, metab-

olism, or elimination (ADME) of a drug (Figure 7.3). PK genes affect the availability of the drug within the body through regulation of concentration of drug in the blood or distribution to various organs and tissues. In some cases, the "active" form of a drug is created within the body only after the drug has gone through some metabolic steps; for example, the chemical form of the drug found in the pill may be inactive within the body until it is chemically modified by PK proteins to produce the active form. Variation in PK genes can alter the structure or chemistry of the encoded PK proteins, which might affect its interaction with the drug, and ultimately, the availability of the drug in the body. For example, a variant of a PK gene might metabolize a drug more quickly than is typical, meaning the drug might be eliminated from the body faster than expected, preventing it from having its intended effect.

Thus far, we have discussed instances of mutations in coding regions, which alter the structure or function of the PD or PK gene. However, it is also possible for genomic variation outside of gene coding regions to affect drug response (See Chapter 9). Genetic variants in non-protein-coding regulatory regions could affect the expression of PD or PK genes, causing less or more of the drug target or metabolizing enzyme to be produced than is typical. Atypical levels of these proteins could unbalance the complex biochemical reactions taking place to facilitate the drug effect. Additionally, the number of copies of a PD or PK gene found in a genome can also influence the action or metabolism of a drug. Note that these issues are even more complicated for an innately personal disease like cancer (see Box 7.1)

As we've discussed previously in Chapter 6, various methods, such as GWAS, can be used to connect traits with genetic factors. Pharmacogenomic response can be considered using the same discovery processes as a "trait", which may include binary (efficacy or adverse event) or quantitative (dose) variables. For instance, genotypes could be measured in a set of cases (individuals that experience myopathy after statin treatment) and compared to those of a set of controls (individuals that respond positively to statins) to discover variants with significant differences that may suggest their role in this adverse event. Such variants can then

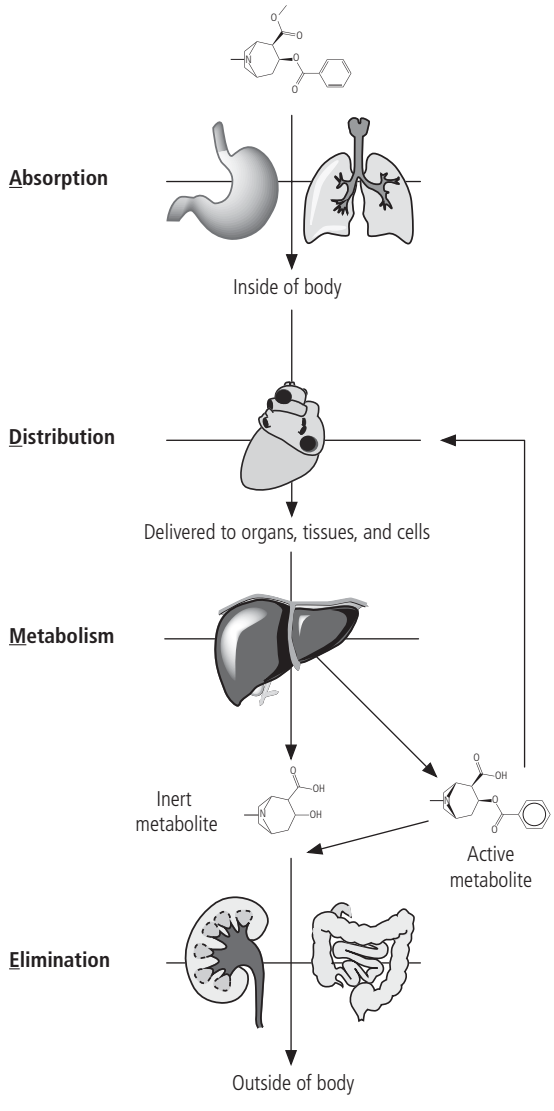


Figure 7.3 Lifecycle of a drug in the human body: The major events of the lifecycle of a drug in the human body are its absorption into the body, distribution throughout the body, metabolism within the body, and elimination from the body. These events are commonly referred to by the acronym ADME. Depending on the route of administration, absorption typically happens in the upper gastrointestinal tract for ingested drugs, or through the lungs for inhaled drugs. Distribution of the drug throughout the body typically occurs through the action of the cardiovascular system, although anatomical structures such as the blood–brain barrier can prevent the drug from being distributed to certain tissues. The metabolism of most drugs is carried out by a large family of redox enzymes collectively referred to as the cytochrome P450 enzymes, which are most prominently produced by the liver. Metabolic transformations by endogenous enzymes can produce both pharmacologically active and inert metabolites. Elimination of drug metabolites will occur primarily through excretion in the urine or as a component of solid waste (see also Plate 10).

Box 7.1 "Cancer Pharmacogenomics"

Cancer can loosely be defined as the genetic reprogramming of an individual's own cells that then causes them grow out of control and become malignant. Because this reprogramming is seemingly random with regard to the regions of DNA affected, it is not surprising that the unique genome of a cancer may be particularly important to its treatment. Thus, pharmacogenomics is emerging as an important consideration in the treatment of cancers, because even tumors of the same clinical type of cancer (e.g. breast cancer) can respond very differently to chemotherapy treatment. Due to somatic mutations and other transformations that occur in carcinogenesis, the genome of tumor cells can be markedly different than that of the individual affected by the cancer. As a result, the tumor could have a very different pharmacogenomic profile than the healthy tissues, meaning the tumor cells may respond differently to a drug than an individual's healthy cells or another cancer of the same type.

One type of mutation that occurs in carcinogenesis is a gene amplification, where errors in DNA replication cause an increase in the number of copies of a particular gene. This can lead to an increase in the expression of the amplified gene, and if the amplified gene is involved in processes such as cell growth or proliferation, the amplification can potentiate the oncogenic transformation of the cell. For example, the gene *ERBB2* (also known at human epidermal growth factor 2, or *HER-2*), is found to be amplified in up to 20% of early stage breast cancers (Cronin et al. 2010). *ERBB2* serves as an important pharmacogenomic marker in the clinic, because the antibody drug trastuzumab (brand name Herceptin), which targets *ERBB2*, was found to be effective against breast cancers over-expressing *ERBB2*, but not other forms of breast cancer.

Another well established cancer pharmacogenomics relationship exists between the antibody drug cetuximab (brand name Erbitux) and mutations in the *KRAS* gene (Cutsem et al. 2009). Cetuximab is an *EGFR* inhibitor that is used to treat head and neck as well as colorectal cancers. Mutations in the *KRAS* gene, which participates in the *EGFR* pathway, are associated with reduced efficacy of cetuximab, and therefore cetuximab is only likely to work against tumors harboring the wild-type *KRAS* gene. Because of these and other many known examples, and the complex genomic transformations that can occur during carcinogenesis, it's likely that pharmacogenomics will play an increasingly important role both in the treatment of cancer and the design of new anti-cancer therapeutics.

be used to infer personal pharmacogenomic response: individuals that have the same genotype at that locus as the "cases" above may be cautious of beginning statin use or may consider alternative treatments. Alternatively, as described later in this chapter, certain variants may be used to predict an optimal dose of certain drugs. These variants can be discovered using similar methods with quantitative variables: individuals with a range of warfarin doses are genotyped and their starting doses predicted using clinical variables combined with significantly predictive variants. These methods embody the heart of pharmacogenomic application: how genetic variation can inform differences in drug response.

Despite the fact that several pharmacogenomic variants are well characterized and validated in multiple clinical studies, pharmacogenomic testing is not yet a routine component of clinical practice. Fortunately, much of the pharmacogenomic knowledge and data is accessible in the public domain, and it is possible to leverage public data resources to conduct our own pharmacogenomic survey of a personal genome. As with any of the applications in this book, these interpretations should not be used to alter any clinical decisions drastically. While these interpretations can provide evidence for the molecular interactions of drugs and genes, they should not be used to make any definitive decisions regarding drug regimens, especially not discontinuing any currently prescribed medications. Such decisions should always be made under the advice of a physician.

7.2 Mapping common pharmacogenomic variants

A fairly large number of "common" pharmacogenomic variants have already been described in the medical literature. Fortunately, many of these variants are being collected from the medical literature and curated into a database of pharmacogenomic variants at Stanford University called the Pharmacogenomics Knowledge Base (PharmGKB; Figure 7.4). PharmGKB has a staff of professional scientists and curators who not only collect the pharmacogenomic variants from the literature, but also interpret the evidence behind each variant

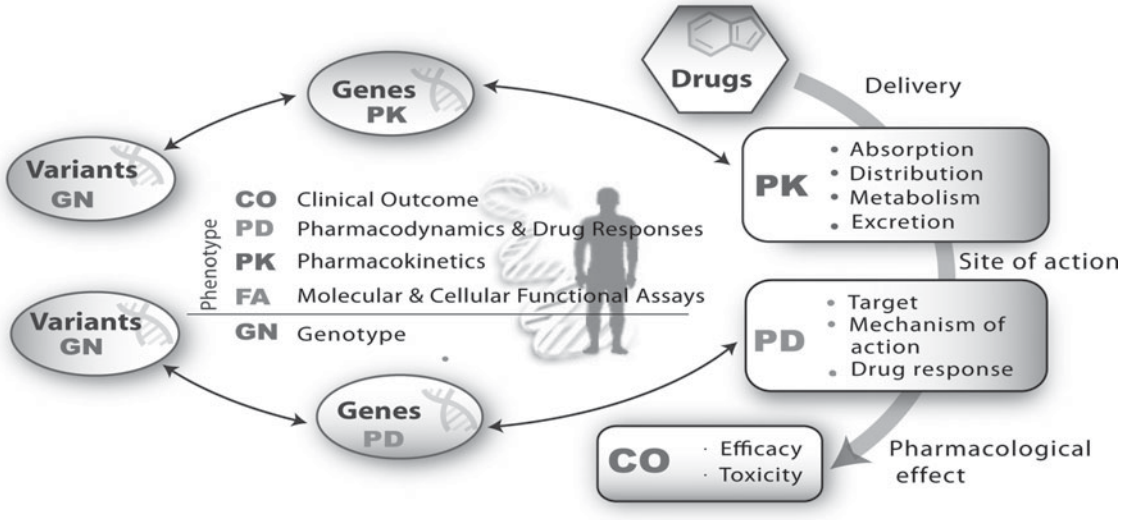


Figure 7.4 Overview and information flow of pharmacogenomics: Genotype variants affect the function of genes involved in pharmacokinetics (PK) and pharmacodynamics (PD). PK genes are typically involved in "what the drug does to the body", or absorption, distribution, metabolism, and excretion. PD genes are often the direct targets of the drug or otherwise related to the mechanism of action of the drug. Each of these factors can alter the pharmacological effect or clinical outcome (efficacy or toxicity of the drug). Reprinted with permission from PharmGKB.

(Klein et al. 2001). In addition, they collect additional information about the genes in which the variants are found and the affected drugs. The contents of the PharmGKB database can be used to investigate a personal genome for genetic variants associated with abnormal drug response.

7.2.1 Exploring PharmGKB data

PharmGKB data can be accessed from the PharmGKB website at <http://www.pharmgkb.org>. Here, we are presented with the PharmGKB web query interface where we can search for pharmacogenomic information from various perspectives. For instance, we can search from the perspective of a drug (Figure 7.4), perhaps one we are considering taking, to find all the associated variant information. Alternatively, we may be interested in the effect of a specific variant which we know we have and want to know the drugs whose action it affects (Figure 7.5).

7.2.2 Mapping pharmacogenomic information

Most of the records in PharmGKB describe single nucleotide polymorphisms (SNPs) that are associated with differential drug response. Like most information concerning SNPs, the variants in PharmGKB are indexed by NCBI dbSNP identifiers (rsIDs). We can use these rsIDs to cross-reference the records in PharmGKB with polymorphic positions in a personal genome. Typically, an association study may report a "risk" allele, where a certain genotype is determined to be the "high sensitivity", "ineffective", or "adverse event associated" allele. For instance, a GG genotype at some locus may be associated with an adverse event, while individuals with AA or AG may be the "low risk" genotypes. In this case, individuals with GG genotypes may closely monitor their reaction to the drug, or prefer to consider alternative medication entirely. As with disease or trait associations, the "risk" allele may not be the minor allele in a given population; occasionally, the majority of the population may have a higher risk for an adverse event or reduced drug efficacy. Accordingly, presence of the "risk" allele does not guarantee a different reaction to a drug, but may simply increase the probability of this event.

Another designation that is occasionally referenced in the literature refers to "haplotype" desig-

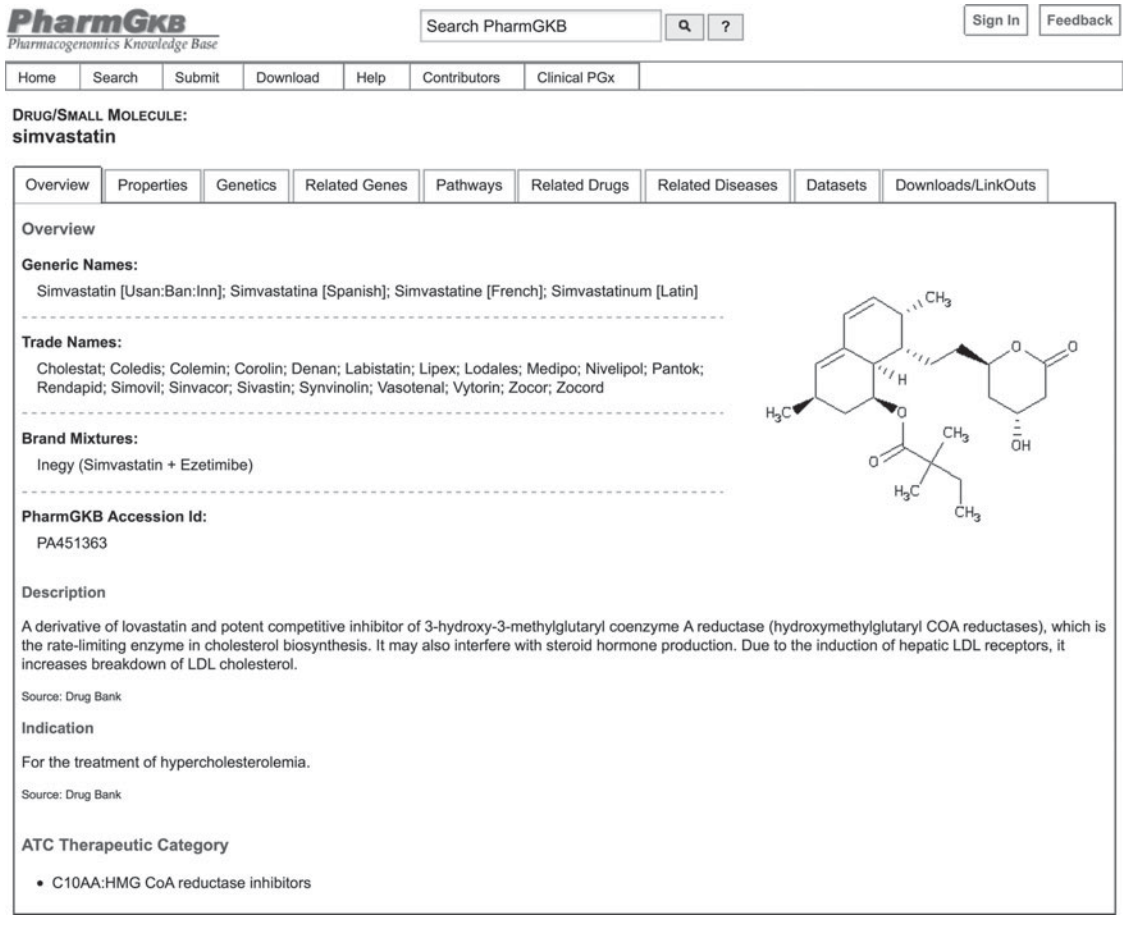


Figure 7.5 PharmGKB drug page: The description of simvastatin (Zocor), with related pharmacogenomic information. Reprinted with permission from PharmGKB.

nations, which are denoted as *1, *2, etc. Associations discovered in older literature based on older genotyping methods use this nomenclature to denote different versions of a given gene. These versions may include a number of polymorphisms that define the whole gene. For instance, *1 may refer to the major allele version of a gene, *2 may indicate the presence of a mutation at one specific position, and *3 may indicate this mutation in addition to a mutation at a second position. These designations are slightly more difficult to apply to personal genomes, as they are typically not directly measured, but involve combinations of SNPs; however, we will apply one such example later in the chapter in the prediction of an optimal warfarin dose.

Given this information, we can now compare our personal genotype to pharmacogenomic information, such as the variant records found in PharmGKB (Figure 7.6). For instance, we may be interested in rs1045642, a variant in ABCB1 (a drug transporter) and compare this to our genotype at rs1045642, using our favorite method, such as a simple lookup on Interpretome. This variant has been very well studied and like many genetic variants, there is conflicting information for its role in some drugs. In the second annotation, one particular study showed that individuals with the TT genotype have increased rhodamine levels in leukocytes.

We can repeat this process for any drug or variant of interest and some other canonical examples are

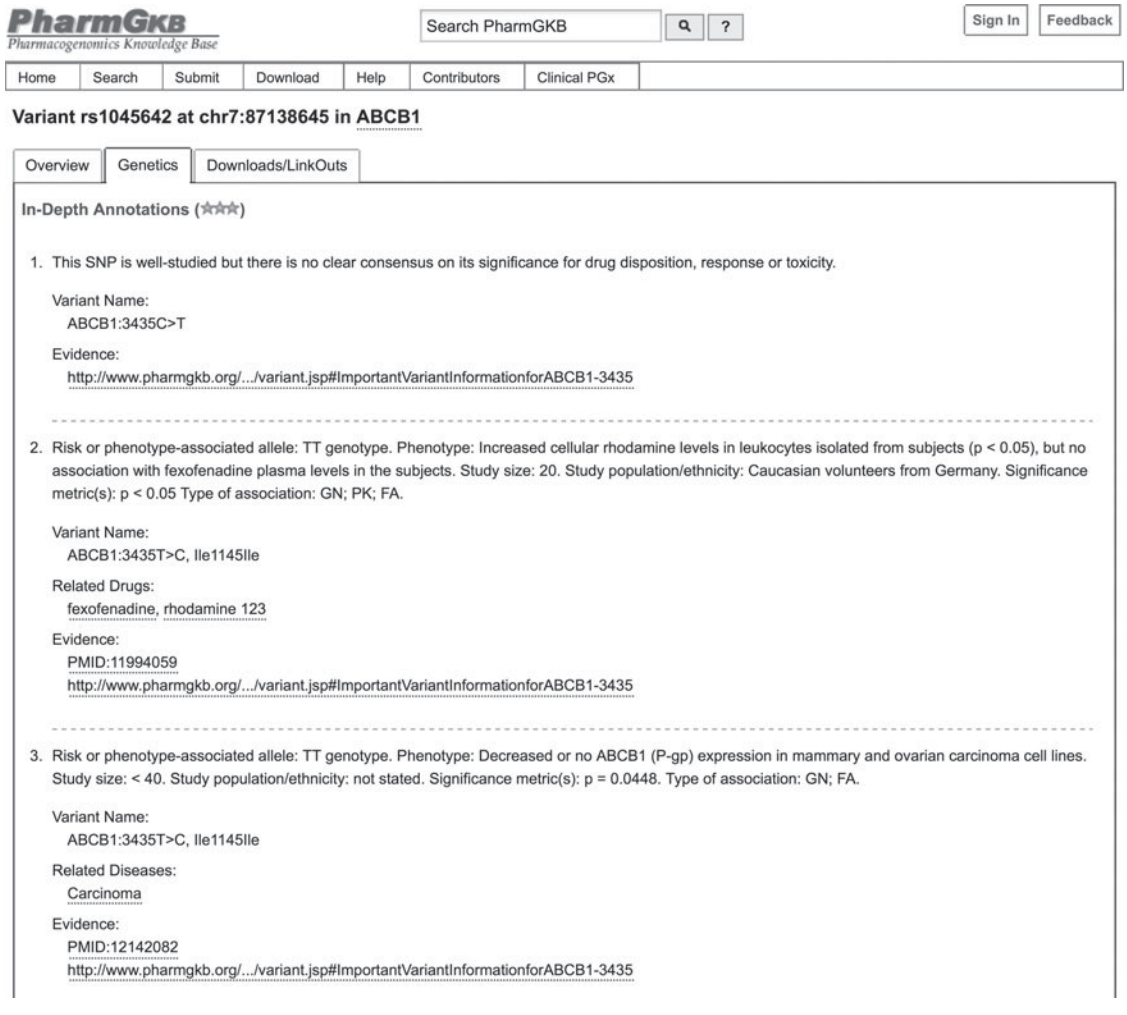


Figure 7.6 PharmGKB variant page: Curated data on a single variant in ABCB1, a transporter gene. In-depth curations and interpretations are provided, as well as all available information and citations. Reprinted with permission from PharmGKB.

provided in the Pharmacogenomics section of Interpretome.

7.2.3 Summarizing and interpreting pharmacogenomic information

In the previous section, we saw how we might connect information from a personal genome with pharmacogenomic information found in the PharmGKB database. Such a mapping can be very useful if we were considering beginning a drug regimen: we would want to be aware of any factors that may cause us to have a different reaction from others to the drug. If we were to do this for a number of drugs, we would be left with a complex set of associations between genetic variants, genes, drugs, and diseases. As is typical for anything pertaining to personal genomics, the challenge is in the representation and interpretation of genetic associations. Methods to summarize these data are under constant development and the exercises later in this chapter will combine data from numerous data sources to highlight important aspects of a personalized pharmacogenomic profile.

7.3 Major applications of pharmacogenomics

As mentioned previously, there are three ways in which pharmacogenomics can be applied to personal genomes, including determining efficacy of drugs, the optimal dose to administer, and potential adverse reactions. Each of these can be mediated by PD or PK genes.

7.3.1 Efficacy and response

The first question to ask when considering a drug treatment, of course, is whether the drug will work in the individual; if an individual's genetic makeup is such that the drug will not reach its target or its target will not be affected, there is no use in even considering the drug. Of course, at present, no set of variants can definitively make such a claim: as with disease or trait variants, pharmacogenomic variants typically adjust the likelihood that the drug will be effective. In other words, in the case of the influence of rs5443 on response to sildenafil, 90% of patients with the TT genotype are likely to respond, while 50% of patients with the CT and CC genotypes are likely to respond. Thus, this SNP is not fully penetrant, and other factors may influence the response to the drug, including environmental factors, other undiscovered genetic factors, or a combination of the two (Chapter 8).

We can intersect a personal genotype with some of these common pharmacogenomic variants by simple exploration and lookup (Table 7.1). For instance, rs10042486 is associated with increased efficacy of rosiglitazone, a drug prescribed to aid in the treatment of Type 2 Diabetes. Individuals with the CC genotype have an average response, while CG is associated with increased response, and GG individuals have an even higher likelihood of response.

7.3.2 Dose prediction

So far, we have only considered pharmacogenomic variants as binary variables for drug effect. Pharmacogenomics SNPs can also be considered as quantitative trait loci (QTLs), where variation at the locus can "explain" some fraction of an observed quantitative drug trait (e.g. blood plasma concentration half-life). For example, genetic variation at a locus might be associated with an increase in the blood serum concentration of a drug one hour after it is administered. If the effect size of the pharmacogenomic QTL is substantial (i.e. genetic variants can explain a significant proportion of the variation of a quantitative drug trait), then it becomes possible to incorporate genetic information into useful predictive models of drug response. If such a model can be translated into a clinical variable such as weekly dose, genotype information for one or several of these loci might then be incorporated into calculations for an appropriate drug dose for a patient given their personal genome. One such example includes the well-studied blood thinner, warfarin.

Warfarin (brand name: Coumadin) is the most commonly prescribed anticoagulant drug in the United States: an estimated 2 million patients per year begin a warfarin regimen to help prevent blood clots, heart attack, and stroke. However, it is somewhat difficult to determine an optimal dose for warfarin, as clinical, genetic, and environmental factors all contribute to the effectiveness of the drug. In addition, the consequences of an incorrect dose are potentially disastrous: an overdose can cause hemorrhaging, while an insufficient dose will not have the desired therapeutic effect. It often takes weeks or months of cycles of dose adjustments followed by follow-up doctor visits and clinical tests before a stable dose is established.

Clinical factors, including age, height, weight, race, and whether the individual is taking amiodarone or any enzyme inducers can explain about 27% of the variation in warfarin dose. In addition, a number of genetic variants have been implicated in the variation in dose of warfarin which, combined with the clinical factors, explain a greater proportion (47%) of the variance. The two major genetic components involved, VKORC1 and CYP2C9, are involved in PD and PK of warfarin, respectively. These factors have been combined into an equation, established by the International Warfarin Pharmacogenomic Consortium (IWPC), to predict an optimal starting warfarin dose (International Warfarin Pharmacogenetics Consortium 2009).

We will use our personal genotype, along with personal clinical factors, to calculate an optimal

Table 7.1 Common pharmacogenomic variants: A set of common variants known to have pharmacogenomic impact, as curated by PharmGKB. Reprinted with permission from PharmGKB and Interpretome

dbSNP	Genotype	Evidence Level	Gene	Interpretation
1057910	AA	1	CYP2C9	Patients with the AA genotype: 1) may require an increased dose of warfarin as compared to patients with the AC or CC genotype 2) may have a decreased risk for adverse events as compared to patients with the AC or CC genotype. Patients with the AA genotype may still be at risk for adverse events when taking warfarin based on their genotype. Other genetic and clinical factors may also influence a patient's risk for adverse events.
9923231	CT	1	VKORC1	Patients with the CT genotype may require a lower dose of warfarin as compared to patients with the CC genotype.
12248560	CC	1	CYP2C19	Patients with the CC genotype: 1) may have decreased activation of clopidogrel 2) may have a decreased risk for bleeding with clopidogrel as compared to patients with the CT or TT genotype 3) may have an increased risk for adverse cardiovascular events as compared to patients with a CT or TT genotype. Patients with the CC genotype may still be at risk for bleeding when taking clopidogrel based on their genotype. Other genetic and clinical factors may also influence a patient's risk for bleeding.
4244285	AG	1	CYP2C19	Patients with the AG genotype: 1) may have poor metabolism of clopidogrel and decreased formation of active drug metabolite 2) may have an increased risk for secondary cardiovascular events when treated with clopidogrel as compared to patients with the GG genotype.
1800460	CT	1	TPMT	Patients with the CT genotype may have an increased risk for toxicity with thiopurine drugs and purine analogues as compared to patients with the CC genotype.
1800462	CC	1	TPMT	Patients with the CC genotype: 1) may have decreased metabolism of thiopurines 2) may have a decreased risk for toxicity with thiopurine drugs as compared to patients with the CG or GG genotype. Patients with the CC genotype may still be at risk for toxicity when taking thiopurine drugs based on their genotype. Other genetic and clinical factors may also influence a patient's risk for toxicity.
4149056	TT	3	SLCO1B1	A person with this genotype may have no increased risk of simvastatin-related myopathy.
5443	CC	3	GNB3	Patients with the CC genotype had reduced response to sildenafil than those with TT genotype.

warfarin dose, acting as though we were to require the drug at this point in time. Of course, these models are constantly being refined, as new genetic factors are associated with drug response. Additionally, these measures should be used as a guide for predicting a starting dose, not to adjust an already established warfarin regimen.

The calculator weights each variable by a known factor (established by the original study) to adjust the predicted dose, as for a personal genome shown in Table 7.2. For instance, since vitamin K absorption decreases with age, older individuals will require less warfarin. Thus, age (in decades) is multiplied by 0.2546 for the clinical algorithm or 0.2614 for the pharmacogenetic algorithm and subtracted from the starting values (4.0376 and 5.6044, respectively). This process is repeated for all the variables (Table 7.2) and the final value is squared to obtain the predicted weekly dose. For this individual, if we compare the clinical factor-only prediction to the pharmacogenomic algorithm (Figure 7.7), we see that the dose may have been over-predicted by the clinical algorithm, which can lead to hemorrhaging.

7.3.3 Adverse events

Finally, pharmacogenomic variants can affect inter-individual variability in adverse reactions to drugs.

Table 7.2 Warfarin dosing: In the clinical algorithm, age, height, weight, race, and other medications are used to calculate a starting warfarin dose. In the pharmacogenomic algorithm, variants in VKORC1 and CYP2C9 are used to adjust the dose further. Reprinted with permission from Interpretome

Feature	Multiplier	Personal factor	Running total	Multiplier	Personal factor	Running total
Age (in decades)	–0.2546	3	3.2738	–0.2614	3	4.8202
Height (in cm)	0.0118	180.34	5.4018	0.0087	180.34	6.3892
Weight (in kg)	0.0134	84.09	6.5286	0.0128	84.09	7.4655
Asian	–0.6752	0	6.5286	–0.1092	0	7.4655
Black	0.406	0	6.5286	–0.276	0	7.4655
Other/Mixed	0.0443	0	6.5286	–0.1032	0	7.4655
Enzyme inducer	1.2799	0	6.5286	1.1816	0	7.4655
Amiodarone	–0.5695	0	6.5286	–0.5503	0	7.4655
VKORC (rs9923231 TT)			6.5286	–1.6974	1	5.7681
VKORC (rs9923231 CT)			6.5286	–0.8677	0	5.7681
VKORC (rs9923231 Unknown)			6.5286	–0.4854	0	5.7681
CYP2C9 (*1/*2)			6.5286	–0.5211	0	5.7681
CYP2C9 (*1/*3)			6.5286	–0.9357	0	5.7681
CYP2C9 (*2/*2)			6.5286	–1.0616	0	5.7681
CYP2C9 (*2/*3)			6.5286	–1.9206	0	5.7681
CYP2C9 (*3/*3)			6.5286	–2.3312	0	5.7681
CYP2C9 Unknown			6.5286	–0.2188	0	5.7681
Clinical dose (mg/week):			42.623	PG × dose(mg/week):		33.2712

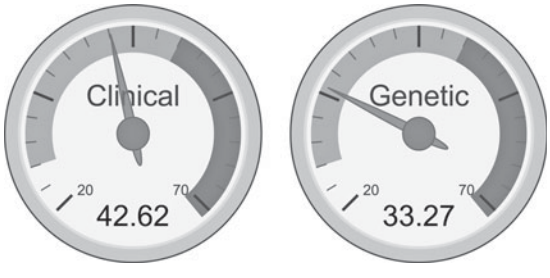


Figure 7.7 Comparison of clinical and pharmacogenomic warfarin doses: We can calculate an optimal starting warfarin dose using clinical factors alone, and genetic factors coupled with clinical factors (Table 7.2). In this individual, the predicted dose from the clinical algorithm would overestimate their dose, compared to the pharmacogenomic algorithm; an excess dose of warfarin can cause hemorrhaging. Reprinted with permission from Interpretome.

Specifically, a drug may work for one group of individuals as intended, but cause adverse events in another, which can range in severity. Perhaps one of the most famous examples of this scenario is statin-related myopathy. Millions of patients in the US are prescribed statins every year to lower their cholesterol; however, in rare cases (a few patients per 100,000 treated), certain statins, including simvastatin (Zocor) and atorvastatin (Lipitor), can cause a severe muscle disease (myopathy) known as rhabdomyolysis. Genetic factors, such as rs4149056, a SNP in a transporter gene SLCO1B1, have been associated with this condition and individuals with the C allele are at higher risk (SEARCH 2008). Additionally, a single variant in the alcohol dehydrogenase gene (ALDH2), which is common in Asian individuals, is associated with the "alcohol flush" reaction, which causes a person to turn red when enjoying a Friday night out at the bar. While this is not as pertinent an example to the healthcare system, it illustrates an "unintended" consequence of a gene-drug interaction. Other drug-variant-adverse event combinations have been documented, with more being constantly discovered, and can be found in the Further Reading section and on PharmGKB.

We can intersect our genetic profiles with PharmGKB in a similar fashion to the "efficacy" section above (Table 7.1). For instance, we can look up our genotype at rs4149056 and determine whether we would be at risk for myopathy if we were to take statins. Additionally, our genotype at rs671 may indicate whether we are likely to experience the

"alcohol flush" interaction, though this is a phenotype that we most likely would not need genetics to predict. Of course, these variants may not exhibit complete penetrance and individuals with the adverse event "risk" genotype may or may not exhibit the adverse event, suggesting there may be other genetic or environmental factors (or combination thereof) involved in variation in drug response.

Thus far, we have explored the effect of common variants in a personal genome on drug response. Common variants have been able to explain a significant portion of this variation in preliminary studies. However, as this field has yet to assign a full model of the effect of genetics on drug response, it is likely (as with disease-trait associations) that some of the variation can be explained by rare or novel variants.

7.4 Assessment of rare variants in PD/PK genes

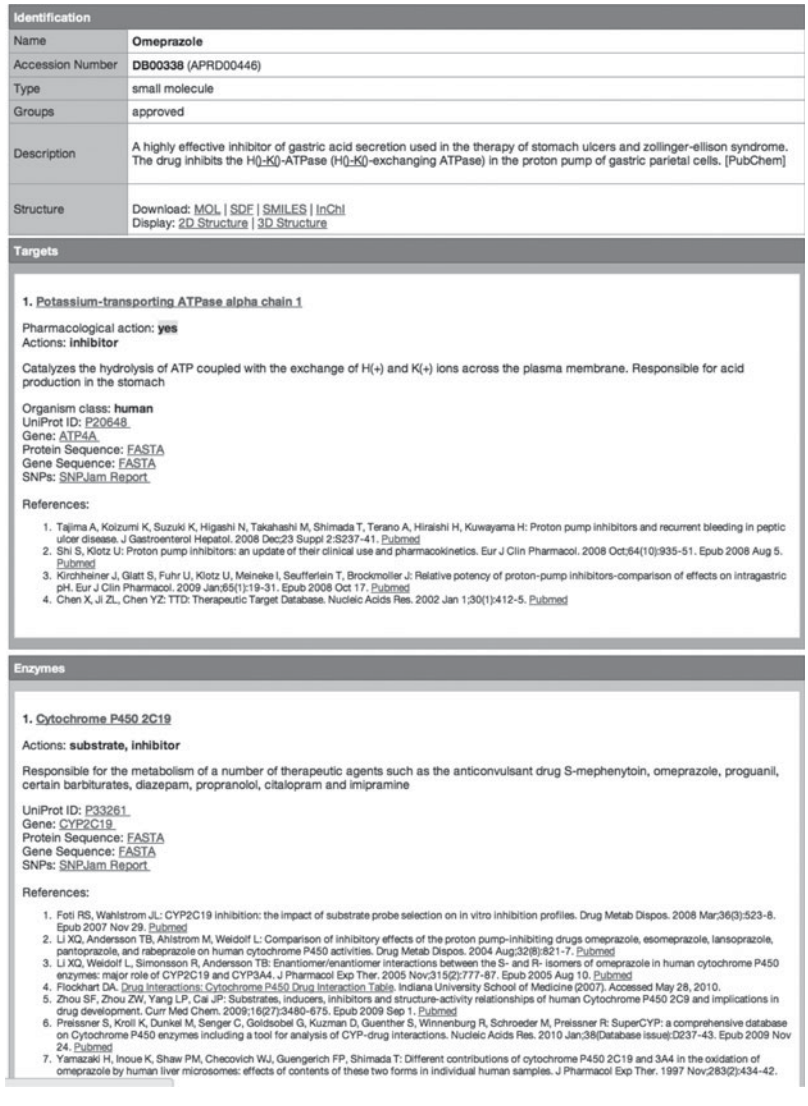
Personal genomes also include tens to hundreds of thousands of rare and novel variants (see Chapter 10), and many of these are likely to be located within or near genes that are known to participate in the therapeutic action (e.g. drug targets) or metabolism of drugs. Rare variants refer to variants that are present below some cutoff frequency in the population (often 1%), while variants that have never been observed in human populations are sometimes called novel or "private" variants. These variants have the potential to alter the biological effect of genes to the degree that an individual exhibits an abnormal response to a drug, however the rarity of such variants means that we cannot use population-based pharmacogenomics studies to characterize their function. Also, just because a variant is rare or novel does not mean it is necessarily harmful to the function of a gene. Unfortunately, there are no simple rules that can be used to assess the biological impact of a variant, however a number of sophisticated methods have been developed towards this aim. The majority of these methods aim to assess the functional impact of non-synonymous (amino-acid changing) SNPs, and typically employ probabilistic machine learning approaches that consider a number of features, such as the evolutionary conservation of the amino acid position, or the physiochemical properties between the normal and mutant amino acid change, to determine the probability that a mutation is damaging. There are fewer tools capable of assessing the functional impact of variants in non-coding regions, however they are becoming more available as our knowledge of the function of non-coding regions increases.

7.4.1 Mapping rare variants to drug-associated genes

In order to assess the pharmacogenomic potential of rare variants in a personal genome, we must first identify those rare variants that are most likely to affect drug response due to their location within, or proximity to genes involved in drug response or metabolism. One of the most comprehensive databases of genetic and chemical information for approved clinical drugs is DrugBank (<http://www.drugbank.ca>). DrugBank contains detailed chemical and genetic information on more than 1,350 FDA-approved small molecule drugs, all of which is made available in the public domain. Information for each drug in DrugBank is represented as a DrugCard, which provides information on genes known to be targeted by a drug, as well as information for any genes whose products are known to participate in the metabolism of a drug (Figure 7.8).

Therefore, we can use the information from each of the DrugCards in DrugBank to construct a list of genes associated with the action or metabolism of the majority of FDA approved drugs. Given a personal genome, we can then look for rare variants within this set of drug-associated genes to identify rare variants with the potential to give rise to abnormal drug response.

Of course, the mere presence of a rare variant within a drug-associated gene does not necessarily mean that the variant will cause the gene to behave abnormally in the presence of a drug. The fact that so many rare variants persist in the genomes of seemingly healthy individuals indicates that most are likely to be benign from a clinical standpoint. To gain further insight into the pharmacogenomic of each rare variant, we must interrogate each SNP individually using functional prediction tools.



Identification

Name	**Omeprazole**
Accession Number	**DB00338** (APRD00446)
Type	small molecule
Groups	approved
Description	A highly effective inhibitor of gastric acid secretion used in the therapy of stomach ulcers and zollinger-ellison syndrome. The drug inhibits the H()-K()-ATPase (H()-K()-exchanging ATPase) in the proton pump of gastric parietal cells. [PubChem]
Structure	Download: MOL | SDF | SMILES | InChI Display: 2D Structure | 3D Structure

Targets

1. Potassium-transporting ATPase alpha chain 1

Pharmacological action: **yes**
Actions: **inhibitor**

Catalyzes the hydrolysis of ATP coupled with the exchange of H(+) and K(+) ions across the plasma membrane. Responsible for acid production in the stomach

Organism class: **human**
UniProt ID: P20648
Gene: ATP4A
Protein Sequence: FASTA
Gene Sequence: FASTA
SNPs: SNPJam Report

References:

1. Tajima A, Koizumi K, Suzuki K, Higashi N, Takahashi M, Shimada T, Terano A, Hiraishi H, Kuwayama H: Proton pump inhibitors and recurrent bleeding in peptic ulcer disease. J Gastroenterol Hepatol. 2008 Dec;23 Suppl 2:S237-41. Pubmed
2. Shi S, Klotz U: Proton pump inhibitors: an update of their clinical use and pharmacokinetics. Eur J Clin Pharmacol. 2008 Oct;64(10):935-51. Epub 2008 Aug 5. Pubmed
3. Kirchheiner J, Glatt S, Fuhr U, Klotz U, Meineke I, Seufferlein T, Brockmoller J: Relative potency of proton-pump inhibitors-comparison of effects on intragastric pH. Eur J Clin Pharmacol. 2009 Jan;65(1):19-31. Epub 2008 Oct 17. Pubmed
4. Chen X, Ji ZL, Chen YZ: TTD: Therapeutic Target Database. Nucleic Acids Res. 2002 Jan 1;30(1):412-5. Pubmed

Enzymes

1. Cytochrome P450 2C19

Actions: **substrate, inhibitor**

Responsible for the metabolism of a number of therapeutic agents such as the anticonvulsant drug S-mephenytoin, omeprazole, proguanil, certain barbiturates, diazepam, propranolol, citalopram and imipramine

UniProt ID: P33261
Gene: CYP2C19
Protein Sequence: FASTA
Gene Sequence: FASTA
SNPs: SNPJam Report

References:

1. Foti RS, Wahlstrom JL: CYP2C19 inhibition: the impact of substrate probe selection on in vitro inhibition profiles. Drug Metab Dispos. 2008 Mar;36(3):523-8. Epub 2007 Nov 29. Pubmed
2. Li XQ, Andersson TB, Ahlstrom M, Weidolf L: Comparison of inhibitory effects of the proton pump-inhibiting drugs omeprazole, esomeprazole, lansoprazole, pantoprazole, and rabeprazole on human cytochrome P450 activities. Drug Metab Dispos. 2004 Aug;32(8):821-7. Pubmed
3. Li XQ, Weidolf L, Simonsson R, Andersson TB: Enantiomer/enantiomer interactions between the S- and R- isomers of omeprazole in human cytochrome P450 enzymes: major role of CYP2C19 and CYP3A4. J Pharmacol Exp Ther. 2005 Nov;315(2):777-87. Epub 2005 Aug 10. Pubmed
4. Flockhart DA. Drug Interactions: Cytochrome P450 Drug Interaction Table. Indiana University School of Medicine (2007). Accessed May 28, 2010.
5. Zhou SF, Zhou ZW, Yang LP, Cai JP: Substrates, inducers, inhibitors and structure-activity relationships of human Cytochrome P450 2C9 and implications in drug development. Curr Med Chem. 2009;16(27):3480-675. Epub 2009 Sep 1. Pubmed
6. Preissner S, Kroll K, Dunkel M, Senger C, Goldsobel G, Kuzman D, Guenther S, Winnenburg R, Schroeder M, Preissner R: SuperCYP: a comprehensive database on Cytochrome P450 enzymes including a tool for analysis of CYP-drug interactions. Nucleic Acids Res. 2010 Jan;38(Database issue):D237-43. Epub 2009 Nov 24. Pubmed
7. Yamazaki H, Inoue K, Shaw PM, Checovich WJ, Guengerich FP, Shimada T: Different contributions of cytochrome P450 2C19 and 3A4 in the oxidation of omeprazole by human liver microsomes: effects of contents of these two forms in individual human samples. J Pharmacol Exp Ther. 1997 Nov;283(2):434-42.

Figure 7.8 DrugCard information from the DrugBank database: Information from the DrugCard for the proton-pump inhibitor drug omeprezole is shown. Each DrugCard contains a wealth of information about a drug, such as its chemical properties, pharmacology, manufacturing, and commercial information, and relevant research publications. Here we see an example of the detailed drug target and metabolizing enzyme information that is provided for each drug. We can use this information to identify genes that are relevant to the drug's pharmacology, and therefore useful for pharmacogenomic analysis using personal genomic information.

7.4.2 Predicting the functional impact of rare variants in drug-associated genes

It is not yet feasible to characterize individual rare variants in a personal genome experimentally (although advances in iPS stem cell technologies are bringing this closer to reality); however, we can use *in silico* (i.e. computer-based) tools to evaluate the functional impact of individual rare variants. We can use these tools to assess the likelihood that an individual variant will be "damaging" to the function of the gene product. Therefore, if one of these tools determines that a rare variant in a drug target is likely to be damaging to the function of the drug tar-

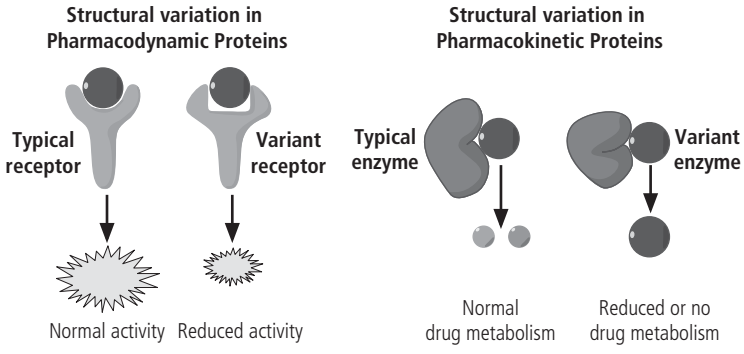


Figure 7.9 Changes in protein structure from genetic variation can alter drug response: Structural changes in pharmacodynamic genes, such as cellular receptors, can change the affinity between the target and the drug ligand, which might result in a reduced physiological response to a drug respective to a given dose. Structural changes in pharmacokinetic genes, such as metabolizing enzymes, can reduce the metabolic activity of the enzyme such that the drug remains in the body longer than intended, or a toxic metabolite of the drug is not eliminated and causes harm. A well known clinical concern of the latter scenario is the increased toxicity from 6-mercaptopurine (6-MP) administration due to variants in the *TPMT* enzyme, which reduces its enzymatic activity and therefore its ability to eliminate 6-MP and its toxic metabolite 6-TG from the body.

get protein, then there is a potential that the individual could respond abnormally to a drug whose targets include the damaged protein target (Figure 7.9). This assumption is admittedly naïve, as proteins typically perform diverse biological functions, and the presence of a novel variant determined to be "damaging" will not necessarily cause detriment to the function facilitating the action or metabolism of a drug (and indeed, certain variants can actually have the opposite effect, opposing other detrimental variants in the same gene). However, the current state-of-the-art functional prediction tools can only offer coarse-grained predictions that differentiate between variants that are likely to be damaging, and those that are not. Furthermore, these tools are generally non-specific in their prediction of functional impact, in that they cannot predict specifically which function of the protein (e.g. drug binding) is likely to be damaged by the variant. The development of tools to predict potential damaging effects of rare variants specifically for drug-related functions is a much needed area of research.

To demonstrate how to predict the functional impact of rare variants found in the coding regions of drug-associated genes, we use the PolyPhen tool (<http://genetics.bwh.harvard.edu/pph2/>), which is one of the most popular and venerable tools for this purpose (Figure 7.10). PolyPhen is based on a statistical approach that integrates evolutionary and physiochemical information for amino acid predictions to predict which non-synonymous variants are likely to cause deleterious changes to the protein (Adzhubei et al. 2007). PolyPhen takes as input a non-synonymous variant in a single gene, and estimates the likelihood that the variant will be deleterious to the function of the gene. Other similar approaches, such as SIFT and Condel, are discussed in Chapter 10.

We can use the Interpretome platform to intersect a personal genome with drug-associated genes from DrugBank and functional assessment of these variants by PolyPhen2 (Table 7.3).

Using the approach detailed in this section, we are able to identify novel variants in a personal genome that have the potential to elicit a pharmacogenomic event, and have used *in silico* approaches to identify those variants that are most likely to be harmful to the function of the drug-associated gene products. However, this analysis is still very preliminary and simply predictive: ultimately, the effects would have to be validated by blood test or enzymatic assay.

7.4.3 3-D structure-based assessment of pharmacogenomic variants

As we have described, tools such as PolyPhen can be used to predict if nsSNPs might be generically damaging to the function of genes. However, what is "damaging" to the protein may not necessarily be

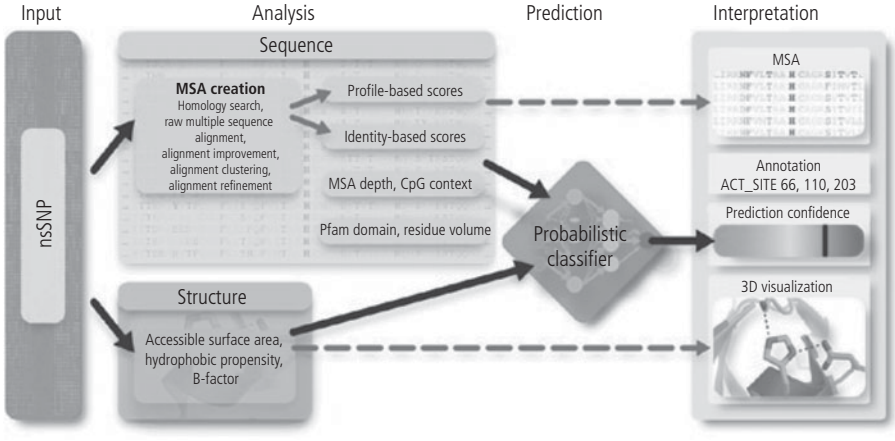


Figure 7.10 PolyPhen: PolyPhen is a common tool to predict the effects of non-synonymous SNPs. The PolyPhen algorithm considers sequence-based features such as evolutionary information, as well as 3-D protein structure features. It then uses a probabilistic classifier to build a statistical model of the protein from the input features to predict the functional effects of amino acid changes. Reprinted by permission from Macmillan Publishers Ltd: Adzhubei, I. A. et al. A method and server for predicting damaging missense mutations. *Nature Methods* 7, 248–9 (2010).

Table 7.3 Rare pharmacogenomic variants: A set of rare variants in genes known to interact with drugs from DrugBank. PolyPhen is then run on these variants to predict whether the SNP is benign, possibly damaging, or probably damaging. Reprinted with permission from Interpretome

dbSNP	Genotype	Minor allele frequency	Gene name	Drug name	PolyPhen class	PolyPhen score
1402954	TT	0.049	Fba	Fructose-1,6-Diphosphate	benign	0.086
6063	CT	0.005	FGG	Sucralfate	probably damaging	1
3213096	CT	0.009	IL12B	Alpha-D-Mannose	possibly damaging	0.407
4997557	CC	0	CYP2A6	2H-1-BENZOPYRAN-2-ONE	benign	0.004

"damaging" to its role in drug response; in other words, from this information alone, it is difficult to draw a conclusion about the impact of the polymorphism on the actual systemic effect of the drug. For example, proteins can often perform many functions, and an nsSNP might not necessarily impact the protein domains, folds, or otherwise functional capacity related to the drug's biochemical effect. Another approach to exploring the pharmacogenomic impact of personal genomic variants is to analyze and visualize these variants using the 3-D structures of drugs and proteins (Figures 7.11, 7.12).

The 3-D protein structures for many drug targets and metabolizing enzymes have been resolved by techniques such as X-ray crystallography or nuclear magnetic resonance (NMR) spectroscopy. Many of these resolved structures have been deposited into the public domain Protein Data Bank (PDB) database (<http://www.pdb.org>). The PDB provides search functionality to find structures for particular proteins and facilitates exporting of the 3-D structure data in a special PDB file format. Since the PDB file format is an open standard file format, there are many different software tools available that can read files in this format and render 3-D visualizations of the protein structure, including the open-source Jmol software (<http://jmol.sourceforge.net>) that is used to render structures on the PDB website. Another popular open-source software designed for this purpose is PyMol (<http://www.pymol.org>), which provides an extensible system and framework for analysis and visualization of 3-D biomolecular structures. The covers of many premier scientific journals such as *Nature* and *Science* have been graced by beautiful renderings of 3-D protein structures created using PyMol. PyMol will be used

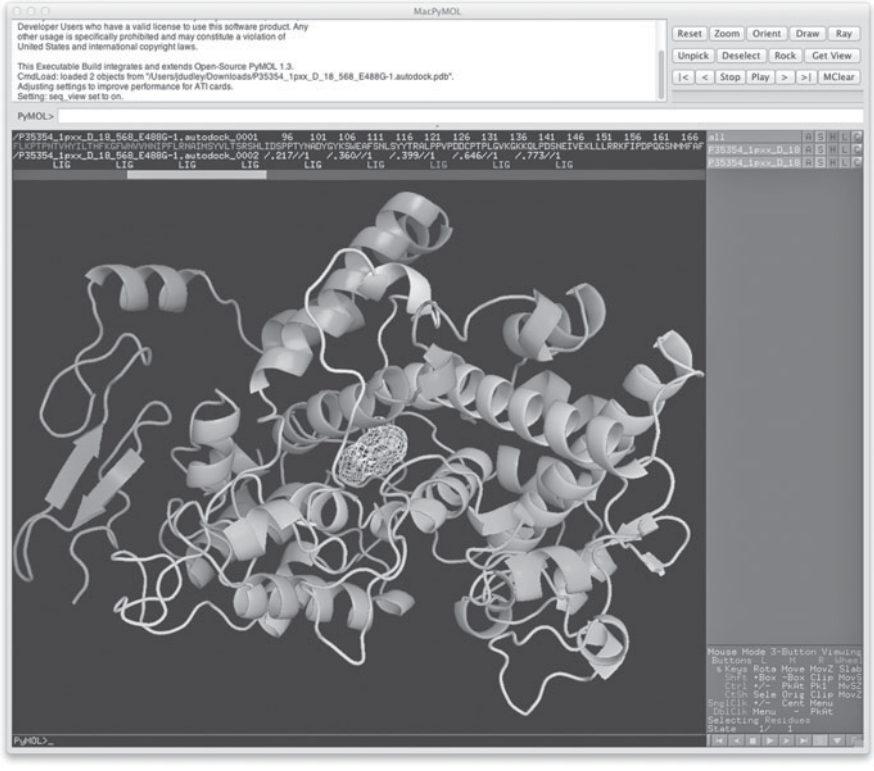


Figure 7.11 The open-source PyMol software for 3-D analysis and visualization: PyMol offers a large array of features for visualizing and analyzing protein 3-D structures as well as the 2-D amino acid sequence underlying the structure. There are a number of extensions available from PyMol that allow for evaluation of sequence mutations, drug-protein interactions, and many other facets of chemistry and structural biology (see also Plate 11).

here to illustrate pharmacogenomic assessment using 3-D structure information, however any software capable of reading the PDB file format could be alternatively used for this purpose.

It should be noted that the tools and techniques required for analysis of 3-D biomolecular structures are still quite specialized, and significant experience and skill in bioinformatics, biochemistry, or structural biology is typically required to carry out such efforts. Therefore, the analysis and results demonstrated in this section serve as advanced examples of personal pharmacogenomic analysis that incorporate such tools. As of the time of writing, there are no generally available, user-friendly tools for facilitating 3-D structure based assessment of personal pharmacogenomic variants; however, we are hopeful that such tools will emerge in the near future.

7.4.4 Using molecular dynamics simulations to assess pharmacogenomic variants

One approach to assessing personal variants in drug-associated genes using 3-D structure information is to take a molecular dynamics approach, in which computer software is used to apply complex algorithms that attempt to model many of the physiochemical properties of drugs and protein molecules existing and interacting in a simulated 3-D space. Although the algorithms and biophysical science behind molecular dynamics simulations are beyond the scope of this book, one key concept in understanding the results of these simulations is *free energy*. All natural systems seek to minimize the amount of free energy within the system, and therefore the output of molecular dynamics simulations typically includes an estimate of the amount of free energy in the system at the

conclusion of the simulation. For example, we can imagine a simulation where a drug ligand is computationally simulated to dock with two or more suspected drug targets. In this simple example, we would expect the target that produces the lowest free energy estimate as the most preferential target of the two.

In order to use a molecular dynamics approach for personal pharmacogenomics assessment, we must first obtain the 3-D structure for the protein product of a gene that exhibits rare or otherwise uncharacterized protein-coding variation in a personal genome. Unfortunately, only approximately 15% of human proteins have resolved 3-D protein structures deposited into the PDB database, which significantly limits the utility of a structure-based approach. However, because of their importance to drug discovery, a number of drug targets have structures available in the PDB. Three-dimensional structures can be located in PBD in a number of ways, ranging from simple keyword search functionality to sequence similarity approaches that compare the amino acid sequence of a protein of interest against the respective sequences of all structures in PDB using a protein sequence alignment algorithm such as BLASTP. In this section, we will use the 3-D structure of the angiotensin I converting enzyme (*ACE*), which is a key regulator of vascular tone, and the target of many drugs used for the regulation of blood pressure. Specifically, we will use 3-D structure of *ACE* defined in the PDB record having the identifier 2C6N.

As a demonstration, we will explore the possible pharmacogenomic effect of a single non-synonymous variant in *ACE* in the context of the the antihypertensive drug lisinopril using a 3-D structure based approach. Due to the high (~30%) prevalence of hypertension in the United States, nearly 90 million prescriptions for lisinopril were written in 2010 alone (source: IMS Health). This indicates that nearly one in three adults in the United States will be prescribed lisinopril in their lifetime, and thus, will be a critical drug from the standpoint of clinical pharmacogenomics. Genetic variants in *ACE* have been associated with different clinical outcomes related to the use of *ACE* inhibitors, such as lisinopril. To evaluate the possible pharmacogenomic effect of a nsSNP in *ACE* the presence of lisinopril, we will use docking simulation software to "dock" the lisinopril ligand with the wild-type 3-D structure of *ACE*, and then again with the variant structure of *ACE* to see if there is a significant change in the free energy of the drug-target interaction. The molecular docking simulation can be carried out by software such as Autodock (<http://autodock.scripps.edu>), which is a popular open-source tool for this purpose.

Let us suppose that we observe a nsSNP in *ACE* that results in a transition from the wild-type glutamic acid (E), a polar and negatively charged amino acid, to alanine (A), a non-polar amino acid with neutral charge, at protein sequence position 388. Figure 7.12 shows the 3-D structure of *ACE* docked with lisinopril using the default parameters of Autodock version 4.0, with the position harboring the nsSNP highlighted in yellow. From visual inspection alone, we can see that the nsSNP is located near the lisinopril binding pocket, and therefore there is good reason to suspect that it might have an influence on the binding of lisinopril. When the docking simulation is run between lisinopril and the wild-type structure of *ACE*, the resulting binding free energy is estimated to be –9.31 kcal/mol, which, as expected, suggests a good binding affinity between lisinopril and *ACE*. However, when the docking simulation is run

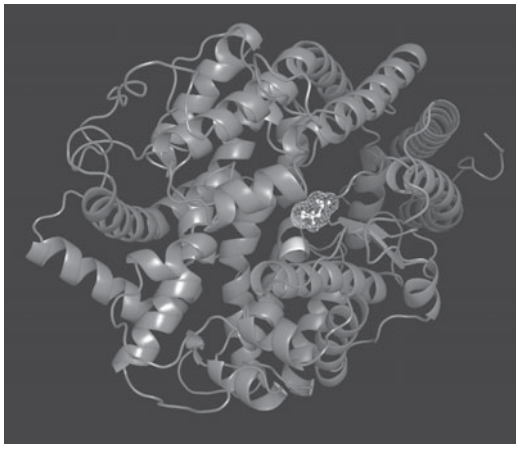


Figure 7.12 3-D structure of the ACE protein with bound lisinopril: The figure shows lisinopril docked into an active binding domain of the ACE 3-D protein structure. The yellow highlight indicates the location of a non-synonymous variant that has replaced a polar amino acid with a non-polar amino acid. Given the change in charge and the proximity of the amino acid change to the drug binding region, we might hypothesize that the variant could affect drug binding and render a pharmacogenomic effect. This hypothesis can be evaluated *in silico* using molecular dynamics simulation tools (see also Plate 12).

between lisinopril and the mutant form of *ACE* with the E388A transition, the resulting free binding energy is estimated to be –1.2 kcal/mol, which represents a shift towards a reduced estimated binding affinity between the two molecules. Although the results of this *in silico* analysis offer evidence that the nsSNP could serve as a pharmacogenomic factor that potentiates an abnormal response to lisinopril, this hypothesis would ultimately require experimental validation by an enzymatic assay.

The molecular-docking approach is just one type of approach that can be taken towards 3-D structure based assessment of pharmacogenomic variants. Extensions of this approach might include additional features, such as the variant's proximity to protein binding pockets, or additional physiochemical properties of the amino acid change. Of course, computational approaches, such as the one demonstrated in this section, would need to be continually validated through experimental studies and clinical trials to validate and ensure their clinical relevance. As more 3-D structures of drug-associated genes are resolved and deposited into PDB, and personal genomes of clinical populations become available, we expect there will be expanded efforts to incorporate 3-D structure data and methods in personal genomics and pharmacogenomics research.

7.5 Variation in pharmacogenomic pathways

Although research efforts in pharmacogenomics have been successful in identifying a large number of pharmacogenomic variants, these studies have revealed that, in many cases, there is rarely a single "smoking gun" genetic variant that explains the majority of the genetic variation of a drug trait. Instead, like many complex disease phenotypes, it appears that multiple genetic variants can contribute towards the same pharmacogenomic trait. Rather than considering only known pharmacogenomic variants reported from large population studies, which are often markers rather than causal variants, we can take a knowledge-guided approach that allows for the incorporation of all genotypes for all genes in the pathways associated with a particular drug. Because the genes and molecular pathways associated with a drug's metabolism and mechanism of action (MOA) have been identified for many drugs, we can use this information *a priori* to guide the assess pharmacogenomic risk. For example, three variants (along with clinical factors) have been associated with differential response to warfarin but this only explains 47% proportion of the variance of dosing among individuals (International Warfarin Pharmacogenetics Consortium 2009). However, we can also consider the aggregation of personal genetic variation in the drug targets and metabolizing enzymes associated with the drug to explore and evaluate any possible deviations from a "normal" genetic landscape across these specific genes.

7.5.1 Estimating mutational load

One way to approach the aggregate use of all genetic information for a particular gene or set of genes is through the concept of *mutational load*. The term *mutational load* has specific meanings in certain domains of biology, such as evolutionary biology, but for our purposes we define mutational load as a quantification of the total genetic deviation for all loci contained within or defined by an organizing genetic unit, such as a gene. In this way, we can incorporate all of the rare and common genetic variation into a single aggregate measure of mutational load that can serve as an indicator of how much a particular gene deviates from the most common allele. This differs significantly from the candidate variant approaches discussed earlier, which seek to evaluate individual variants for their possible effect on gene function, and it allows us to use the full extent of the genetic measurements for a gene represented in a personal genome rather than focusing on a handful of well-characterized variants.

There are a number of approaches that could be taken to define a method for estimating a mutational load score, and several such methods have recently emerged in the scholarly literature (see Further Reading). Here, we will focus on a rather straightforward approach to estimating mutational load, called the pHap method (Tatonetti et al. 2010). In their original work, Tatonetti et al. demonstrated that a mutational load approach based on the pHap method could improve the prediction of a patient's final stable warfarin dose from genotypes when combined with the set of established warfarin pharmacogenomic variants reported from large

population studies. One critical component of any approach for estimating mutational load is to incorporate some measure of deviation from the most common (i.e. wild-type) allele. For this purpose, the pHap method uses population allele frequency data from the HapMap project to quantify the relative frequency of a particular allele within the population. The pHap method can be used to estimate a pHap score for a gene using the following equation.

$$p\mathit{H}ap_i = \sum_{j}^{k} -\log(f_j)$$

For each gene (*i*) in a set of genes associated with a drug, the pHap score is computed for the gene by iterating over all loci (*j*) in the gene for which we have genotype information and compute the negative log of the frequency of the observed genotype in the general population. The pHap score is then reported as the sum of these values. To put the pHap score into a biological context, the method assumes that the presence of low frequency alleles in a gene increases the likelihood of abnormal function, and should therefore increase the mutational load. By taking the negative log of the frequency, low frequency alleles will cause a larger increase in the pHap score.

7.5.2 Evaluating personal mutational load across drug-specific pathways

Now that we are familiar with the pHap method for estimating mutational load scores for genes, we can design an approach to leverage the pHap method to perform personal pharmacogenomic assessment using drug-specific pathway information (Daneshjou et al., in preparation). Here, we will take the approach of estimating the pHap score for all genes in a drug-specific pathway given a personal genome, and then summarize this score across all genes in the pathway to compute a *pathway mutational load (PML)* score for a personal genome. The PML score is simply the sum of individual pHap scores for all genes belonging to a pathway. We will then assess whether the PML score from the individual genome differs significantly from a reference distribution of PML scores, which is computed for the same pathway from the individual genomes from a large population sample.

In order to carry out the PML approach, we need to obtain data on drug-specific pathways and their member genes as well as a large set of individual genomes measured from a representative sample population. There are a number of databases that provide information on genes and pathways associated with either the PD or PK of drugs. Popular public resources for drug pathway information include the KEGG DRUG database (<http://www.genome.jp/kegg/drug/>) and PharmGKB (<http://www.pharmgkb.org>). Here, we will use drug-specific pathway information from PharmGKB because these pathways have been manually curated by trained experts. As a demonstration, we will consider the PharmGKB pathways associated with the drug warfarin (Figure 7.13), a blood-thinning agent known to have a strong component of clinical variability due to genetic variation. For the set of population reference genomes, we will use data from the 1000 Genomes Project (<http://www.1000genomes.org>): note that the reference population genomes should be ethnically matched with the individual genome (Recall from Chapter 5 that ethnic populations exhibit distinct patterns of genetic variation due to differences in migration, admixture, and evolutionary histories).

We begin the PML analysis by obtaining the identifiers of all genes that are members of the warfarin drug pathway as annotated in PharmGKB. If you search for the warfarin pathway on the PharmGKB website, you will find that PharmGKB actually defines two distinct pathways for warfarin; one pathway for the pharmacodynamic (PD) interactions related to warfarin, and another pathway for warfarin pharmacokinetic (PK) interactions. The PD pathway for warfarin is shown in Figure 7.13. The genes associated with this and other pathways can be downloaded directly from PharmGKB. For the purpose of this example we will consider all genes in the union of these two pathways, however an alternative strategy might investigate these pathways separately. From this point forward we'll refer to the union of the warfarin PD and PK pathways as simply the warfarin pathway.

Next, we will compute the warfarin PML score for a personal genome. The first step in this process is to compute the pHap scores for each individual gene in the warfarin pathway. For each gene, we query the personal genome to obtain the genotypes for each nucleotide position measured for that

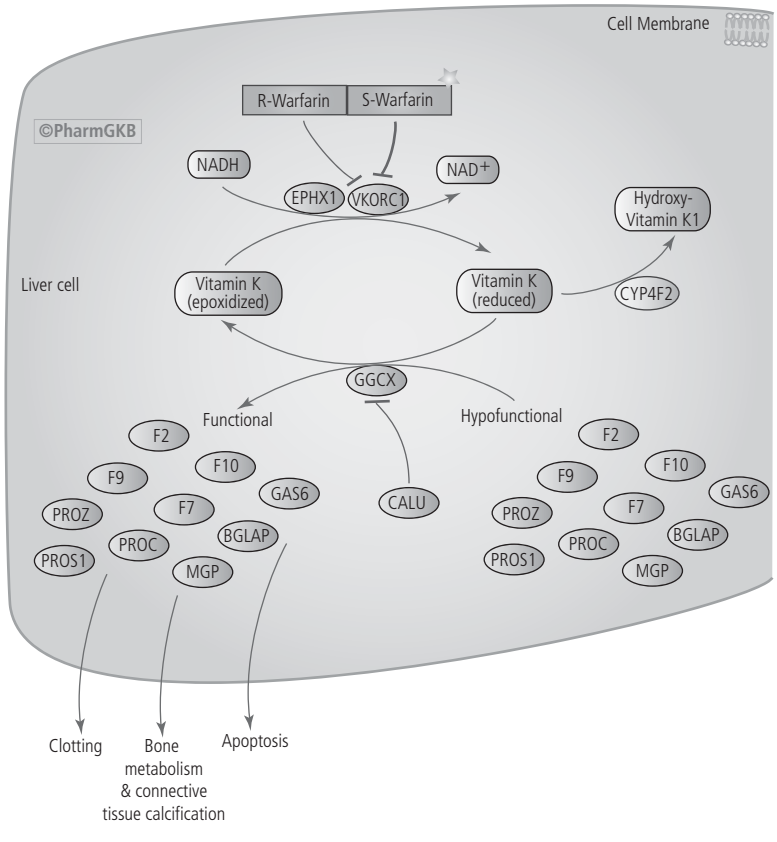


Figure 7.13 Warfarin pharmacodynamics (PD) pathway: The direct target of warfarin is VKORC1, but variation in any of the genes in the pathway can affect the physiological effect of warfarin and its clinical outcome. Reprinted with permission from PharmGKB.

gene. For the purpose of this example, we will limit our analysis to only the exonic regions of the gene (alternative strategies might also include intronic regions, or flagging sequences upstream or downstream of the 5′ or 3′ boundaries of the gene, respectively). We would then compute the pHap score for the gene using the equation shown in section 7.6.1. Finally, we would sum the pHap scores of the individual genes in the warfarin pathway to derive the warfarin PML score for the personal genome. This score will be compared against the background distribution of the warfarin PML scores to test for a significant deviation from the population norm.

To estimate the background distribution of the PML score for the warfarin pathway using the reference set of genomes from the 1000 Genomes Project, we simply repeat the same steps we used to compute the PML score for the personal genome for every individual genome found in our reference population set. If our reference set of genomes contained the genomic sequence data for 400 individuals (the approximate number of northern European individuals sequenced by the 1000 Genomes Project at the time of this writing), then this would produce a set of 400 warfarin PML scores representing the background distribution. We can plot this background PML score distribution to see if the PML score for the personal genome deviates significantly from the background distribution. For better visual assessment of the statistical significance of any possible deviation, we will transform the scores into a standard *z*-score by subtracting the mean of the background PML score distribution from all PML scores and dividing them by the standard deviation of the same background

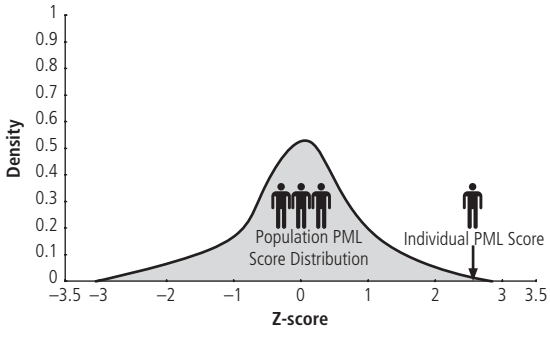


Figure 7.14 Results from a PML analysis for warfarin: The distribution represents the distribution of standardized warfarin PML scores for a large population sample. This distribution is used to determine the range of the typical warfarin pathway mutational load found among the broader population. The arrow indicates the PML score for an individual whose personal genome was evaluated for warfarin PML, indicating that the individual has a rather extreme PML score relative to the rest of the population. This suggests that the individual has a relatively abnormal genetic architecture in their warfarin pathway genes, and could potentially respond abnormally to warfarin.

distribution. A plot of the background PML distribution is shown in Figure 7.14.

As we can see from the plot, the PML score for the personal genome is found near the extreme of the background PML score distribution. Specifically, the PML score for the personal genome has a z-score of 2.5, making it more extreme than ~99.4% of the PML scores in the population distribution. In effect, this result indicates that the genetic composition of the warfarin pathway in the personal genome is significantly different than that of the normal population; therefore, we could hypothesize that the individual might respond abnormally to warfarin. Of course, this hypothesis would ultimately need to be tested by clinical monitoring of the individual should they ever receive warfarin. The PML approach taken in this section is just one possible way to use pathway information for pharmacogenomic assessment of a personal genome. For example, we might modify the PML method used here to add weights to the individual gene pHap scores based on their function (e.g. direct targets versus metabolizing enzymes), or even their degree of separation from the primary drug target in the pathway diagram. Other approaches might include other genes that are known to physically interact with or regulate the genes in the warfarin pathway, or add information about the expression levels of the warfarin pathway genes in relevant tissues or cell types.

7.6 Conclusion

From a clinical perspective, pharmacogenomics is arguably one of the most promising and relevant facets of personal genomics. This is because the environmental component of a drug response trait is easily controlled by a physician (e.g. by choice of drug or dose), and, unlike the majority of common disease variants, a number of pharmacogenomic variants have already demonstrated clinical utility. Just as with other individual traits with roots in a personal genome sequence, genetic variation plays a major role in an individual's response to therapeutic drugs. Variation in pharmacokinetic genes can alter the drug's progression through the body, while variation in pharmacodynamic genes can explain why the drug has or does not have an intended or unintended effect. A number of common variants that affect drug efficacy, dose, and adverse events have been discovered, which we can use directly for pharmacogenomic assessment of a personal genome. However, rare variants presenting in drug-associated genes might also affect drug response, and several tools exist which enable further exploration the pharmacogenomic effect of rare, non-synonymous variants in the context of the protein sequence or 3-D structure of a drug-associated gene. Finally, while many pharmacogenomic analyses tend to focus on the proteins that directly interact with the drug, the combined genetic variation of these genes and other genes in the same drug-related biological pathway (e.g. pharmacodynamic and pharmacokinetic pathways) may also affect an individual's response to the drug.

At the time of this writing, the FDA has updated the drug label information for approximately 100 approved drugs to include information on pharmacogenomic risk markers associated with adverse events or differential response to a drug. While we expect that large, population-based studies of drug-associated traits will continue to yield informative and clinically actionable pharmacogenomic markers, the large clinical repertoire of approved drugs

and likely high prevalence of rare genetic variation in drug-associated genes among the population creates an intractable pharmacogenomic search space for traditional study designs. Therefore, there is a pressing need to develop tools and techniques for providing actionable clinical assessment of rare or otherwise uncharacterized variants presenting in a personal genome that may affect individual drug response. Furthermore, we expect that these methods will expand beyond analysis of drug-associated genes to include relevant functional elements of non-coding regions that may bear on drug response traits, such as variation in regulatory regions encoding miRNA or transcription factor binding domains. The methods described in this section offer several avenues to pursue such efforts.

Further reading

Research publications

Adzhubei, I. A. et al. (2010) A method and server for predicting damaging missense mutations. *Nature Methods* 7, 248–9.

Altman, R. B., Flockhart, D. & Goldstein, D. B. (2012) *Principles of Pharmacogenetics and Pharmacogenomics.* New York, NY: Cambridge University Press.

Bange, J., Zwick, E. & Ullrich, A. (2001) Molecular targets for breast cancer therapy and prevention. *Nature Medicine* 7, 548–52.

Cronin, K. A., Harlan, L. C., Dodd, K. W., Abrams, J. S. & Ballard-Barbash, R. (2010) Population-based estimate of the prevalence of HER-2 positive breast cancer tumors for early stage patients in the US. *Cancer Investigations* 28, 963–8.

Hansen, N. T., Brunak, S. & Altman, R. B. (2009) Generating genome-scale candidate gene lists for pharmacogenomics. Clinical Pharmacology Ther 86, 183–9.

International Warfarin Pharmacogenetics Consortium et al. (2009) Estimation of the warfarin dose with clinical and pharmacogenetic data. *New England Journal of Medicine* 360, 753–64.

Klein, T. E. et al. (2001) Integrating genotype and phenotype information: an overview of the PharmGKB project. Pharmacogenetics Research Network and Knowledge Base. *Pharmacogenomics* J 1, 167–70.

Romond, E. H. et al. (2005) Trastuzumab plus adjuvant chemotherapy for operable HER2-positive breast cancer. *New England Journal of Medicine* 353, 1673–84.

SEARCH Collaborative Group et al. (2008) SLCO1B1 variants and statin-induced myopathy—a genomewide study. *New England Journal of Medicine* 359, 789–99.

Tatonetti, N. P., Dudley, J. T., Sagreiya, H., Butte, A. J. & Altman, R. B. (2010) An integrative method for scoring candidate genes from association studies: application to warfarin dosing. *BMC Bioinformatics* 11 Supplement 9, S9.

Van Cutsem, E. et al. (2009) Cetuximab and chemotherapy as initial treatment for metastatic colorectal cancer. *New England Journal of Medicine* 360, 1408–17.

Yang, J. O. et al. (2010) VnD: a structure-centric database of disease-related SNPs and drugs. *Nucleic Acids Research* 39, D939–44.

Internet resources

<http://www.pharmgkb.org>
<http://www.pdb.org>
<http://www.snpedia.com>
<http://www.drugbank.ca>
<http://genetics.bwh.harvard.edu/pph2/>
<http://autodock.scripps.edu>
<http://www.1000genomes.org>
<http://www.genome.jp/kegg/drug/>