Published in final edited form as:

Clin Pharmacol Ther. 2009 August ; 86(2): 183–189. doi:10.1038/clpt.2009.42.

Generating Genome-Scale Candidate Gene Lists for Pharmacogenomics

NT Hansen^{1,2,3}, S Brunak¹, and RB Altman^{2,3}

¹ Center for Biological Sequence Analysis, Technical University of Denmark, Lyngby, Denmark

² Department of Bioengineering, Stanford University, Stanford, California, USA

³ Department of Genetics, Stanford University, Stanford, California, USA

Abstract

A critical task in pharmacogenomics is identifying genes that may be important modulators of drug response. High-throughput experimental methods are often plagued by false positives and do not take advantage of existing knowledge. Candidate gene lists can usefully summarize existing knowledge, but they are expensive to generate manually and may therefore have incomplete coverage. We have developed a method that ranks 12,460 genes in the human genome on the basis of their potential relevance to a specific query drug and its putative indications. Our method uses known gene–drug interactions, networks of gene–gene interactions, and available measures of drug–drug similarity. It ranks genes by building a local network of known interactions and assessing the similarity of the query drug (by both structure and indication) with drugs that interact with gene products in the local network. In a comprehensive benchmark, our method achieves an overall area under the curve of 0.82. To showcase our method, we found novel gene candidates for warfarin, gefitinib, carboplatin, and gemcitabine, and we provide the molecular hypotheses for these predictions.

Pharmacogenomics is the large-scale study of how genes and their variations impact drug response. A critical step in pharmacogenomics is identifying genes that are important for drug response. A given drug may have pharmacogenes—genes important for its pharmacology—that are involved in its absorption, distribution, metabolism, and excretion (pharmacokinetics) or that are involved in its mechanism of action (pharmacodynamics). However, it may take many years of study to identify these pharmacokinetic or pharmacodynamic genes completely. This delay in understanding impedes our ability to identify, evaluate, and use genetics to optimize drug selection and dose.¹

Recently, high-throughput genomic technologies have been successful in identifying new pharmacogenetic (PGx) interactions.² RNA expression arrays can measure differential gene expression related to drug response;³ genes showing significant changes in expression are more likely to be relevant to the drug response. Similarly, genome-wide association studies (GWASs) analyze data related to patients with different drug-response phenotypes and seek polymorphisms (most often single-nucleotide polymorphisms, SNPs) that correlate with drug response;⁴ genes with variants that correlate with drug response are more likely to be pharmacogenetically relevant.

Correspondence: RB Altman (russ.altman@stanford.edu).

CONFLICT OF INTEREST

The authors declared no conflict of interest.

SUPPLEMENTARY MATERIAL is linked to the online version of the paper at <http://www.nature.com/cpt>

One of the challenges with these high-throughput technologies is the statistical analysis required for the assessment of genome-scale data. The large number of hypotheses that undergo testing and the multiple sources of experimental noise may lead to the overlooking of important genes or the embracing of irrelevant ones.^{5,6} In expression studies, some genes may show expression changes that are not pertinent to the drug response. In GWASs, many of the correlated polymorphisms may be chance occurrences, not indicative of any biological connection. There is therefore an acute need for principled ways of applying current biological knowledge to the interpretation and ranking of these data sets.⁵ By combining the weak statistical signals in these data sets with an existing model favoring genes for which we have some molecular evidence, we can find pharmacogenes more reliably.

One way to use current biological knowledge is to focus on candidates—genes for which there is some biological or chemical reason to hypothesize their possible involvement in the response to a particular drug. Candidate gene lists are typically compiled from pathway databases and gene annotations as well as by mining the literature.⁷ These methods can be successful, but their quality varies depending on the care with which they are created and our understanding of the relevant biology.^{8–11} Nevertheless, the notion of using automated and systematic methods to create candidate gene lists is very appealing. Given a candidate gene list, we can (using Bayes' Rule) mathematically encode methods for combining our candidate list with high-throughput data.

In this article, we describe a method that takes as input (i) a query drug and (ii) a query indication for its use and creates a ranking of 12,460 genes in the human genome with respect to the likelihood that each gene is an important modulator for the input drug and indication for its use. Our method can, in principle, be applied to the entire genome, but this requires for all genes that at least some information be available about their primary protein–protein interactions. The method uses a high-quality protein–protein interaction network to describe the context (subnetwork of genes) in which each gene operates. For each gene-specific subnetwork, our method compares the drugs and indications associated with its subnetwork with the query drug and indication. If the gene and its subnetwork interact with drugs that are similar to the query drug/indication, then the gene is assigned a high score as a potential pharmacogene. This assessment is made for every gene in the protein network in order to produce a rank-ordered list. The list is imperfect and does not replace the need to perform experiments, but it provides an independent source of information that can be used to inform the analysis of the data from such experiments.

Our method is logically derived from the following assumptions: (i) proteins tend to work collectively in modules to carry out cellular tasks; (ii) drugs interacting with these modules provide a clue to the chemistry and disease–treatment phenotypes pertinent to that module; (iii) chemical similarity provides information about similarity with respect to pharmacology; and (iv) similarity in indication of use provides information about similarity in the molecular mechanisms (within modules or pathways) targeted by the drugs. These four assumptions are not universally true, but their degree of truth determines the accuracy and utility of our method.

Figure 1 shows the inputs and the subnetwork we use to derive features for each gene–drug combination. We determine the interacting partners for a gene using the InWeb human gene–gene interaction network.¹² We associate each gene with particular drugs on the basis of a corpus of genetic interactions registered in the Pharmacogenetics and Pharmacogenomics Knowledge Base¹³ (PharmGKB) and a corpus of known physical drug–target interactions (DTIs) from DrugBank.¹⁴ We define a similarity metric based on chemical structure that allows us to assess whether the query drug is similar to the drugs

interacting with the gene or its network neighbors. We also define a metric for similarity of indication of use that allows us to assess whether the query indication is similar to the indications of use for drugs that interact with the query gene or its network neighbors. We used the set of known pharmacogenes in PharmGKB as a benchmark to train a logistic regression that takes the similarity of chemistry and indication for both PharmGKB and DrugBank interactions and produces an overall score. We apply this regression to all genes and rank them according to score. We have validated our model on a fivefold cross-validation, a genome-scale prediction, and an independent test set, and we demonstrate sample results for warfarin, gefitinib, carboplatin, and gemcitabine.

RESULTS

We extracted 2,488 primarily genetic interactions from PharmGKB and complemented these with 3,600 physical DTIs extracted from DrugBank. A total of 1,139 drugs and 1,546 genes were involved in the extracted interactions. Each of the genes was mapped to the InWeb interactome that encodes 313,524 physical interactions among 12,460 human gene products. We mapped all drugs and drug classes to chemical structures and indications of use as described in the Methods section.

In a set of preliminary investigations, described in the Supplementary Data online, we showed that (i) gene products that interact in a protein–protein interaction network are more likely to interact with a common small molecule (Supplementary Tables S1–S4 and Figures S1 and S2 online) and (ii) that drugs with a similar indication of use are more likely to interact with each other in a protein–protein interaction network (Supplementary Figure S3 online). These two observations are not perfectly predictive, but they adjust probabilities. This is all that our method requires, because the goal is simply to rank genes on the basis of higher and lower aggregate probabilities.

Construction of four features from the gene–gene–drug network

Integrating the data from InWeb, PharmGKB, and DrugBank, we created a gene–gene–drug network (an example of the data structure can be seen in Figure 1). On average, each gene interacts with 24.9 other genes. The 844 pharmacogenes each interact with an average of 2.95 drugs. The structural similarity between any two drugs in the data set is measured by the Tanimoto coefficient (ranging between 0 and 1) and has an average value of 0.31 in our data set. The similarity in indication of use (also measured by the Tanimoto coefficient) spanned the same range but had an average coefficient of 0.02 (distributions of pairwise similarities can be found in Supplementary Figures S4 and S5 online). We defined a set of four features that can be computed for any combination of a drug and an indication of application, given a gene (see Figure 1 and the Methods section for more details). These features capture structural similarity of drugs with respect to both genetic and physical interactions and similarity in indication of use with respect to both genetic and physical interactions. We calculated these features for each drug–gene combination in the gene–gene–drug network.

Training and validating a logistic classifier

We used PharmGKB as a resource for training gene–drug interactions. As described in the Methods section, we trained a logistic regression model using fivefold cross-validation. The resulting model correctly classified 81.2% of the instances and had an area under the receiver operating characteristic curve of 0.82.

In order to test our model, we used it to make whole-genome predictions for all drugs with PGx interactions registered in PharmGKB. We removed the known interactions for each

drug as it was tested. The known interacting genes for each drug constituted the gold standard. We scored all genes using the logistic regression algorithm and assessed the rank of the gold-standard genes. We also assessed the ability to recall these gold-standard genes with different cutoffs. Figure 2 shows sensitivity (true-positive detection rate) of the model as a function of the specificity (true-negative detection rate). The area under the curve is 0.86. Importantly, our method associates 81% of the drugs in PharmGKB with at least one gene scoring above zero. At that score, we can recall 41% of the known pharmacogenes in PharmGKB. Overall, nearly 7,000 genes have scores above zero, suggesting that there are many potential candidate genes with a relatively high likelihood of interacting with the tested drugs. Therefore, as expected, our method orders the human genome so as to produce a list that is enriched at the top for true pharmacogenes.

While we built our model, PharmGKB staff registered 225 new articles manually, thereby providing an independent set of 199 gene–drug interactions for testing. We ranked the genome for each of the drugs in this new set and made a receiver operating characteristic curve for performance on this external validation set (Figure 2). The area under the curve for the external validation was 0.81. Figure 2 shows that the performance on the external validation set is comparable with the performance on the cross-validation training set, and that we have not overfitted our model to the available examples.

As an illustration, consider warfarin. The literature on the pharmacogenetics of warfarin is quite mature.^{15,16} This anticoagulant drug is difficult to dose because it depends strongly on a patient's genetic background. In particular, two genes affect the appropriate dosage: *vitamin K epoxide reductase complex 1* (*VKORC1*) and *cytochrome P450 2C9*.^{17,18} Leaving out all known warfarin–gene interactions, we ranked all genes in the genome using warfarin and all its indications, as recorded in the PharmGKB. *VKORC1* and *cytochrome P450 2C9* ranked 11 and 16, respectively (out of 12,460). The data for *VKORC1* is shown in Figure 3. The recently discovered warfarin pharmacogene *cytochrome P450 4F219* had weaker support, ranking in the top 500 (4%) of the genome.

Experimental validation of the model

Experimental evidence for novel candidates provides independent validation of our method. To show the broad spectrum of applicability of our method, we obtained: (i) a GWAS of warfarin by Cooper *et al.*,²⁰ (ii) an expression study of gefitinib-treated cancer cell lines by Coldren *et al.*,²¹ (iii) a study of carboplatin response in cancer cell lines from Huang *et al.*,²² (iv) a study of the correlation between gene expression and cytotoxicity of gemcitabine from Li *et al.*,²³ and (v) another study of gene expression and cytotoxicity of gemcitabine by Jarjanazi *et al.*²⁴ For each of these studies, we ranked genes and compared the top five predictions with the experimental data (Supplementary Table S7 online); we also combined the overall rank to the experimental ranks to generate a new, combined ranking (Supplementary Table S8 online). We summarize our findings here.

Warfarin with the indications myocardial infarction, venous thrombosis, thrombolytic disease, and pulmonary embolism—First, one of our top five genes, *NSUN6*, has never before been described in the literature as a pharmacogene for warfarin, but it is associated with response to warfarin in a GWAS ($P < 3.1 \times 10^{-3}$). Second, we combined our rankings with the experimental rank, which put the well-known *VKORC1* and *cytochrome P450 2C9* right at the top; more interestingly perhaps, the third gene is the novel candidate *HSD3B7* that scores highly in both our ranking and the GWAS data. Finally, high-scoring candidates rank significantly higher in the experimental rank than may be expected on the basis of chance ($P < 1.1 \times 10^{-2}$).

Gefitinib with the indication non-small-cell lung cancer—First, of the top five genes, *SH3BGRL* was significantly underexpressed ($P < 3.0 \times 10^{-2}$) in cells showing sensitivity to gefitinib. *AMH* and *PTPRS* were significantly overexpressed ($P < 4.1 \times 10^{-3}$ and $P < 2.9 \times 10^{-2}$). Second, when we combined our ranks and the experimental data, the gene with the best rank sum (i.e., supported by both) was *FAM96B* (evidence shown in Figure 3). Finally, high-scoring genes in our ranking had overall significantly higher expression in the experiment ($P < 6.5 \times 10^{-8}$).

Carboplatin with the indications esophageal, head and neck, lung, and ovarian cancers—Of the top five genes, only *MAK3* was expressed in the cell lines studied by Huang *et al.*²² The expression was correlated with a sensitivity to carboplatin in the Yoruban population ($P < 5.6 \times 10^{-5}$), but not in the Ceph population.

Gemcitabine with the indications breast, kidney, ovarian, lung, and pancreatic cancers—First, of the top five genes, the gene expressions of *AICDA*, *SLC29A1*, and *SH2D5* were significantly correlated with cytotoxicity according to the study by Li *et al.*²³ ($P < 6.8 \times 10^{-4}$, $P < 1.9 \times 10^{-3}$, and $P < 3.2 \times 10^{-2}$, respectively). To our knowledge, of these three validated top-ranking candidates, *AICDA* and *SH2D5* are novel pharmacogenes not previously associated with drug–response phenotypes. The indirect evidence supporting the rank of *AICDA* is represented in Figure 3. Second, when we combined the experimental ranks and the ranks obtained by our method, the novel pharmacogene candidate *NT5C3* had the best combined rank (supporting evidence shown in Figure 3). In addition, we found novel genes, *GPM6A* and *BNIP3*. Finally, an overall comparison showed high correlation between the predicted and experimental ranks ($P < 4.2 \times 10^{-5}$).

We compared our novel predictions with a second data set provided by Jarjanazi *et al.*²⁴ Although there were no data for *NT5C3*, we were able to replicate one of the remaining two, *BNIP3* ($P < 4.9 \times 10^{-3}$). This prediction therefore looks solid and replicable.

Finally, we assessed the impact of improved (i.e., larger) training sets on our performance. Figure 4a shows the anticipated rate of accumulation of drug–gene interactions in the literature, and Figure 4b shows the correlation between the performance of our method and the size of the training set. The performance improves substantially as our training knowledge increases.

DISCUSSION

Our validation experiments demonstrate that our method can rank 12,460 genes in the human genome (i.e., those appearing in the InWeb protein–protein interaction map) and that the top of the list is enriched for true pharmacogenes. Because the PharmGKB is manually curated, the quality of the annotations is fairly high, but the volume is low. In fact, PharmGKB associates only 844 genes with PGx effects. Using protein–protein interactions, we have been able to extend predictions to cover all of the InWeb interactome with 12,460 genes.

Recently, Yamanishi *et al.* successfully developed a method for predicting physical DTIs.²⁵ Although physical interactions could indicate a phenotypic drug–gene relationship, this method did not attempt to predict PGx interactions that extend beyond the physical interactions. A recently published drug–target network shows how drug targets and disease genes cluster in protein–protein interaction networks.²⁶ The drug–target network gives a comprehensive framework for analysis of networks related to drug response (see comparison with the PGx pipeline data structure in Supplementary Table S11 online), but it does not quantify, automate, or focus on the process of prioritizing pharmacogene candidates.

While scientists carrying out pharmacogenomics studies are likely to use any or all available informatics resources to assess “hits,” we have automated this process to allow quantitative genome-scale integration of our rankings (and associated scores) with experimental data sets. This is, to our knowledge, the first automated method for prediction of pharmacogenes.

Our method is most useful as a source of information that can be integrated with other data, as illustrated by combining our ranking results with the ranks from the warfarin GWAS and the gefitinib expression data. By examining Figure 3, one can see that the predictions include both expected and novel candidates. For expected candidates, the evidence comes from related drugs interacting directly with the gene of interest, as with the interaction between warfarin and *VKORC1*. In contrast, for novel predictions, the candidate gene often links a set of known interactants together, as with gemcitabine and *AICDA* in Figure 3. Importantly, our method provides molecular explanations (on the basis of interacting genes and drug structure) that can generate new hypotheses and guide further experimentation.

Our method requires as input both a drug and an indication, and it can yield very different results for the same drug, given different indications. For example, consider the nucleoside analog gemcitabine, a cancer chemotherapeutic. In addition to targeting cancer, gemcitabine has side effects such as neutropenia, hypertension, and proteinuria. Ranking the genome for gemcitabine in combination with the cancer indication yields a very different result from ranking the genome for gemcitabine and its side effects. Whereas the cancer query ranks *CDA* and *ABCB1* in the top 20, the side-effect query includes both *ACE* and *ACE2*, important parts of the blood pressure control pathways (Supplementary Tables S9 and S10 online).

There are important limitations to this approach to pharmacogene prediction. First, the predictions are limited to the parts of the genome that have protein–protein interactions. Second, the indication-for-use associations for some drugs are not complete. Third, we do not have sufficient data about the extensive off-label use of drugs and general-use approved indications.²⁷ Fourth, our structural-similarity methods assume small organic molecular structures and do not work with biological agents such as antibodies or RNA therapeutics. Finally, our method does not explicitly predict whether genes are involved in the pharmacokinetics or pharmacodynamics of a drug. In some cases, this can be inferred by inspection, on the basis of the types of genes that are implicated. We are investigating ways to distinguish pharmacokinetics and pharmacodynamics predictions.

Because our method is modular, the system can be adapted when better protein–protein interactions become available. Likewise, novel DTIs or PGx interactions can easily be added to improve the performance of the model. The method is made available online at <http://pgxpipeline.stanford.edu/>. The validation of the PGx pipeline shows that it is capable of recovering known pharmacogenes as well as predicting novel candidates. Most important, we can apply the method to novel molecules and their putative indications in order to generate an initial genome-scale candidate gene list.

METHODS

Full details of the method are provided in the Supplementary Data online.

Data sources

From the PharmGKB core tables, we extracted 2,488 PGx (“genetic”) interactions between drugs and genes. Although PharmGKB encodes some direct physical interactions, it is dominated by genetic interactions, and so we believe DrugBank to contain fundamentally complementary information about physical interactions. Similarly, we extracted 3,600 drug–

target (“physical”) interactions and 1,173 chemical structures from DrugBank. The InWeb interactome¹² is a human protein–protein interaction network created with data from experiments on both humans and model organisms. In total, the interactome includes data on 12,278 proteins. We added proteins with corresponding gene–drug interactions in PharmGKB as unconnected nodes, bringing the total number of proteins in the interactome to 12,460.

Construction of four features from the gene–gene–drug networks

For each protein in the InWeb, we looked for direct interaction partners (first-degree partners)—for a virtual pull-down experiment.¹² For each of the proteins in the pull-down, we looked for interactions between the corresponding gene and drugs. The gene–drug interactions can be classified in two categories: genetic interactions (PGx) derived from PharmGKB and physical DTIs derived from DrugBank. The resulting networks link the query gene to a group of first-degree interacting genes that, in turn, connect to a set of associated drugs (as shown in Figure 1). We separately considered both the structural similarity and the indication similarity of direct physical interactions (from DrugBank) and indirect genetic interactions (from PharmGKB), leading to the use of four features, as shown in Figure 1. We computed the structural similarity of drugs using MACCS fingerprints and a Tanimoto coefficient over 166 predefined features.^{28–29} We computed similarity in indications of use, using a similar bit string of 2,000 diseases in PharmGKB and the Tanimoto coefficient. We defined four features of the gene–gene–drug network that describe the similarity between a drug (**D**) and indication (**I**) and drugs/indications interacting with the local neighborhood of a gene (**G**) (Figure 1).

Feature 1 (F1): Structural similarity of D to all drugs in the gene–gene–drug network based on their genetic interactions—We take the average score of structural similarities in the subnetwork to the query drug for associated drugs that come from PharmGKB gene–drug interactions.

Feature 2 (F2): Similarity of I to the indications of all drugs in the gene–gene–drug network based on their PGx interactions—We take the average score of similarity of indication for all the drugs in the subnetwork that come from PharmGKB.

Feature 3 (F3): Structural similarity of D to all drugs in the gene–gene–drug network based on their physical DTIs—We take the average score of structural similarities in the subnetwork to the query drug for associated drugs that come from DrugBank gene–drug interactions.

Feature 4 (F4): Similarity of I to the indications of all drugs in the gene–gene–drug network based on their physical DTIs—We take the average score of similarity of indication for all the drugs in the subnetwork that come from DrugBank.

Training and validating a logistic classifier

The PharmGKB has evidence of 2,412 interactions between drugs and genes. In order to provide negative interactions, each positive interaction from PharmGKB was matched with interactions between the same drug and three randomly selected genes that were not present in the positive set. We trained a logistic model on the other examples, using four-way cross-validation in the WEKA software.³⁰ A summary of the performances of other classifiers can be found in Supplementary Tables S5 and S6 online. We validated the method by running the classifier on all drugs with at least one PGx interaction manually registered in PharmGKB and for all 12,460 genes in the InWeb network. During all validations, we

always removed all information related to the drug from the interaction database before running our prediction algorithm.

As an independent validation set, we tested the method on recently discovered drug–gene interactions that were added to the PharmGKB after the training data set was downloaded. This validation set included 84 drugs, 94 genes, and a total of 199 interactions. We used receiver operating characteristic curves to compare the genome-wide performance on the training and external validation set. We found that a cutoff of -1.075 gave the best separation of positives and negatives of the genome-wide validation.

Experimental validation of the model

We used three published prospective data sets and one database example to validate the PGx pipeline. For warfarin, we used the P values of all 500,000 SNPs in the dose-dependence study of warfarin by Cooper *et al.*²⁰ The SNPs were mapped to associated genes and, where there were multiple SNPs, a gene was associated with the SNP with the lowest P value. For gemcitabine, we used two prospective experimental data sets.^{23,24} For gefitinib, we used normalized data from a study of gene expression in 45 non-small-cell lung cancer cell lines with varying resistance to gefitinib, through the Gene Expression Omnibus.^{21,31} For each of the drugs warfarin, carboplatin, gemcitabine, and gefitinib, we carried out the method and evaluated the top five ranking genes. Supplementary Table S7 online includes the five highest scoring predictions for each of the four drugs. For gemcitabine, we summed the ranks of the predictions with the experimental results from Li *et al.*²³ and validated three genes with the lowest summed ranks in the experiment by Jarjanazi *et al.*²⁴ For each of the drugs warfarin, gemcitabine and gefitinib, the 10 genes with the highest summed ranks are summarized in Supplementary Table S8 online.

From the genes in Supplementary Tables S7 and S8 online, we chose four examples to illustrate the molecular data we integrate for each PGx pipeline prediction. In Figure 3, we split each feature into components showing the contributions both from genes and from drugs to a given prediction. The contributions were calculated as percentages of the total logistic score, after correcting for the intercept.

Relationship between performance and database size

The drug–gene interaction data for calculation of the four features stem from both PharmGKB and DrugBank. We simulated more sparse data by removing 0 to 100% of the PGx interactions from PharmGKB in 5% intervals. For each simulation, we evaluated the performance using whole-genome ranks, calculating the enrichment at recall rates of 0 to 100% in 5% intervals. We fitted a curve using linear regression.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

We thank G. Cooper for providing us with genome-wide data for warfarin, M.E. Dolan and A. Stark for data on carboplatin, K. Lage for access to InWeb interactome, R. Whaley for PharmGKB data, and B. Daigle and Y. Garten for their valuable input on the manuscript. We also acknowledge support from the National Institutes of Health (grant GM61374 and LM05652), the National Science Foundation (award CNS-0619926), the Villum Kann Rasmussen Foundation, and the NABIIT (Nanoscience and Technology, Biotechnology and IT) program.

References

1. Roses AD. Genome-based pharmacogenetics and the pharmaceutical industry. *Nat Rev Drug Discov.* 2002; 1:541–549. [PubMed: 12120260]
2. Grant SF, Hakonarson H. Recent development in pharmacogenomics: from candidate genes to genome-wide association studies. *Expert Rev Mol Diagn.* 2007; 7:371–393. [PubMed: 17620046]
3. Downward J. Cancer biology: signatures guide drug choice. *Nature.* 2006; 439:274–275. [PubMed: 16421553]
4. Byun E, et al. Genome-wide pharmacogenomic analysis of the response to interferon beta therapy in multiple sclerosis. *Arch Neurol.* 2008; 65:337–344. [PubMed: 18195134]
5. McCarthy MI, et al. Genome-wide association studies for complex traits: consensus, uncertainty and challenges. *Nat Rev Genet.* 2008; 9:356–369. [PubMed: 18398418]
6. Döllery CT. Beyond genomics. *Clin Pharmacol Ther.* 2007; 82:366–370. [PubMed: 17851575]
7. Goldstein DB, Tate SK, Sisodiya SM. Pharmacogenetics goes genomic. *Nat Rev Genet.* 2003; 4:937–947. [PubMed: 14631354]
8. Zhu M, Zhao S. Candidate gene identification approach: progress and challenges. *Int J Biol Sci.* 2007; 3:420–427. [PubMed: 17998950]
9. Yu W, Gwinn M, Clyne M, Yesupriya A, Khoury MJ. A navigator for human genome epidemiology. *Nat Genet.* 2008; 40:124–125. [PubMed: 18227866]
10. Günther S, et al. SuperTarget and Matador: resources for exploring drug-target relationships. *Nucleic Acids Res.* 2008; 36:D919–D922. [PubMed: 17942422]
11. Kuhn M, von Mering C, Campillos M, Jensen LJ, Bork P. STITCH: interaction networks of chemicals and proteins. *Nucleic Acids Res.* 2008; 36:D684–D688. [PubMed: 18084021]
12. Lage K, et al. A human phenome-interactome network of protein complexes implicated in genetic disorders. *Nat Biotechnol.* 2007; 25:309–316. [PubMed: 17344885]
13. Hernandez-Boussard T, et al. The pharmacogenetics and pharmacogenomics knowledge base: accentuating the knowledge. *Nucleic Acids Res.* 2008; 36:D913–D918. [PubMed: 18032438]
14. Wishart DS, et al. DrugBank: a knowledgebase for drugs, drug actions and drug targets. *Nucleic Acids Res.* 2008; 36:D901–D906. [PubMed: 18048412]
15. Shurin SB, Nabel EG. Pharmacogenomics—ready for prime time? *N Engl J Med.* 2008; 358:1061–1063. [PubMed: 18322288]
16. Elias DJ, Topol EJ. Warfarin pharmacogenomics: a big step forward for individualized medicine: enlightened dosing of warfarin. *Eur J Hum Genet.* 2008; 16:532–534. [PubMed: 18301451]
17. Rieder MJ, et al. Effect of VKORC1 haplotypes on transcriptional regulation and warfarin dose. *N Engl J Med.* 2005; 352:2285–2293. [PubMed: 15930419]
18. Higashi MK, et al. Association between CYP2C9 genetic variants and anticoagulation-related outcomes during warfarin therapy. *JAMA.* 2002; 287:1690–1698. [PubMed: 11926893]
19. Caldwell MD, et al. CYP4F2 genetic variant alters required warfarin dose. *Blood.* 2008; 111:4106–4112. [PubMed: 18250228]
20. Cooper GM, et al. A genome-wide scan for common genetic variants with a large influence on warfarin maintenance dose. *Blood.* 2008; 112:1022–1027. [PubMed: 18535201]
21. Coldren CD, et al. Baseline gene expression predicts sensitivity to gefitinib in non-small cell lung cancer cell lines. *Mol Cancer Res.* 2006; 4:521–528. [PubMed: 16877703]
22. Huang RS, Duan S, Kistner EO, Hartford CM, Dolan ME. Genetic variants associated with carboplatin-induced cytotoxicity in cell lines derived from Africans. *Mol Cancer Ther.* 2008; 7:3038–3046. [PubMed: 18765826]
23. Li L, et al. Gemcitabine and cytosine arabinoside cytotoxicity: association with lymphoblastoid cell expression. *Cancer Res.* 2008; 68:7050–7058. [PubMed: 18757419]
24. Jarjanazi H, et al. Discovery of genetic profiles impacting response to chemotherapy: application to gemcitabine. *Hum Mutat.* 2008; 29:461–467. [PubMed: 18330920]
25. Yamanishi Y, Araki M, Gutteridge A, Honda W, Kanehisa M. Prediction of drug-target interaction networks from the integration of chemical and genomic spaces. *Bioinformatics.* 2008; 24:i232–i240. [PubMed: 18586719]

26. Yildirim MA, Goh KI, Cusick ME, Barabasi AL, Vidal M. Drug-target network. *Nat Biotechnol.* 2007; 25:1119–1126. [PubMed: 17921997]
27. Stafford RS. Regulating off-label drug use—rethinking the role of the FDA. *N Engl J Med.* 2008; 358:1427–1429. [PubMed: 18385495]
28. Molecular Operating Environment. Chemical Computing Group; Quebec, Canada: 2005.
29. Tanimoto TT. IBM Internal Report. 1957
30. Witten, IH.; Frank, E. *Data Mining: Practical Machine Learning Tools and Techniques*. 2. Morgan Kaufmann; San Francisco, CA: 2005.
31. Barrett T, et al. NCBI GEO: mining millions of expression profiles—database and tools. *Nucleic Acids Res.* 2005; 33:D562–D566. [PubMed: 15608262]

Figure 1.

Overview of the strategy for deriving features describing current knowledge about the interaction between a drug (D) and a gene (G) when treating an indication (I). Current knowledge is captured through three types of interactions: protein–protein interactions (PPI), pharmacogenetic (PGx) interactions derived from PharmGKB, and drug–target interactions (DTIs) from DrugBank. For each gene, a subnetwork is constructed with the input gene product (G), its physical interaction partners, and all drugs genetically or physically interacting with any of the proteins. Four features are derived from the subnetwork. F1: structural similarity of D to all drugs in the network based on their PGx interactions. F2: similarity of putative indications of D to the indications of all drugs in the network based on their PGx interactions. F3: structural similarity of D to all drugs in the network based on their DTIs. F4: similarity of putative indications of D to the indications of all drugs in the network based on their DTIs.

Figure 2.

Receiver operating characteristic (ROC) curves for genome-wide predictions with the pharmacogenetic (PGx) pipeline on the training genome-wide and external validation sets. The specificity is the proportion of negatives classified as negatives, and the sensitivity is the proportion of positives classified as positives at a given cutoff. By varying the cutoff on the logistic model, we can estimate specificity and sensitivity and draw the ROC curve for the genome-wide prediction on the training and external validation sets. The area under the curve (AUC) is a performance measure independent of the logistic cutoff. The AUC is given in parentheses in the figure.

Figure 3.

Evidence underlying predicted pharmacogenetic interactions. The query gene is shown at center right, in white. The size of the surrounding circles corresponds to the contribution to the overall score of the gene stated by the percentage under the gene symbol. The colors of the circles reflect the proportions of the gene contribution attributable to a drug interaction. The total contribution of each drug is summarized on the left as a percentage after the drug name in the figure. The interaction between warfarin and *VKORC1* is well known and serves as an example of how the pharmacogenetic (PGx) pipeline predicts within the pharmacogene space without knowledge of warfarin and its interactions. The other three examples show how the PGx pipeline associates novel pharmacogenes with known drugs. These images provide a quick summary of the interacting genes and their associated interacting drugs that cause the method to rank the query gene highly in the context of the input drug.

Figure 4.

The relationship between model performance and database size: **(a)** the number of PubMed hits on the keywords “pharmacogenomics” and “pharmacogenetics” over the past 12 years and **(b)** the correlation between the size of the knowledge base and the multiple of enrichment that can be achieved at 75, 50, and 25% recall rates. The curve at 25% recall is fitted using a logarithmic function and has a Pearson’s R^2 of 0.93. The lines at 50 and 75% recall are fitted using a linear regression, yielding Pearson’s R^2 of 0.96 and 0.94, respectively.