

# Data-Driven Methods to Discover Molecular Determinants of Serious Adverse Drug Events

AP Chiang<sup>1,2,3</sup> and AJ Butte<sup>1,2,3</sup>

The dangers of serious adverse drug reactions (SADRs) are well known to clinicians, pharmacologists, and the lay public. Efforts to elucidate the molecular mechanisms behind SADRs have made significant progress through genetics and gene expression measurements. However, as the field of pharmacology adopts the same novel higher-density measurement modalities that have proven successful in other areas of biology, one wonders whether there can be more ways to benefit from the explosion of data created by these tools. The development of analytic tools and algorithms to interpret these biological data to create tools for medicine is central to the field of translational bioinformatics. In this review we introduce some of the types of SADR predictors that are required, and we discuss several databases that are publicly available for the study of SADRs, ranging from clinical to molecular measurements. We also describe recent examples of how bioinformatics methods coupled with data repositories can advance the science of SADRs.

## INTRODUCTION

A particular patient's response to a specific drug-based therapy cannot be easily predicted, because the response can vary from individual to individual. Some patients may not respond to a therapy, whereas others may require only a small dose to achieve therapeutic effects. Side effects are commonly observed with drug-based therapies, however, and adverse drug reactions (ADRs) can have dire consequences. ADRs are defined as unintended and undesired responses to drugs used at normal dosages for therapeutic purposes. They are considered to be serious ADRs (SADRs) if they result in death, hospitalization, or significant or permanent disability, or if they require intervention to prevent permanent and life-threatening conditions. SADRs are a major clinical problem and are estimated to account for more than 2 million incidents requiring hospitalization annually and more than 100,000 deaths in the United States alone.<sup>1</sup> The rate of occurrence of SADRs in hospitalized patients in the United States has been estimated to be 6–7%.<sup>1</sup> Between 0.12 and 0.3% of hospitalized patients in the United States had fatal SADRs.<sup>1</sup> In effect, SADRs rank between fourth and sixth on the list of leading causes of deaths in the United States annually. Besides being a tremendous strain on the health-care industry, SADRs have an enormous impact on the pharmaceutical industry, accounting for most of the drugs withdrawn from market in the past decade.<sup>2</sup> ADRs are also the top reason for drug discontinuation in patients. Moreover, a significant proportion

of drugs under investigation fail because of toxicity resulting in ADRs during clinical trials.

Although susceptibility to ADRs may arise from both genetic and nongenetic factors, our current understanding shows that genetics plays a pivotal role in drug responses, and therefore SADRs are the primary focus of pharmacogenomics. However, despite significant progress in the past 50 years, our knowledge of the genetic factors contributing to ADRs and SADRs is still very limited. This is partly due to low-throughput technologies that prevented global genome-wide analyses. In the current era of sequenced genomes and high-throughput genomic technologies such as DNA microarrays and proteomics, the bottleneck has shifted away from molecular measurement technology and toward our ability to process, analyze, and use these large sets of data. This burgeoning field of translational bioinformatics has been concerned with developing informatics tools to facilitate the capture, storage, management, integration, dissemination, and utility of these large sets of biological data.<sup>3</sup> The development of analytic tools and algorithms to interpret these biological data is central to the field of bioinformatics.

Given the rapid technological advancements of the past few years, this review focuses on existing knowledge bases, resources, and recent methodological developments pertaining to the discovery of molecular factors associated with SADRs. Our goal is to illustrate how methods in bioinformatics have been and will continue to be critical in translating the relevant genomic

<sup>1</sup>Department of Medicine, Stanford Center for Biomedical Informatics, Stanford University School of Medicine, Stanford, California, USA; <sup>2</sup>Department of Pediatrics, Stanford University School of Medicine, Stanford, California, USA; <sup>3</sup>Lucile Packard Children's Hospital, Palo Alto, California, USA. Correspondence: AJ Butte ([abutte@stanford.edu](mailto:abutte@stanford.edu))

Received 20 November 2008; accepted 9 December 2008; advance online publication 28 January 2009. doi:10.1038/clpt.2008.274

discoveries into clinical practice. We start by addressing the kinds of molecular predictors we require that are relevant to ADRs, with a few case examples of gene expression profiles and gene variants that are predictive of adverse drug events and the degree of drug efficacy. After this introduction to the kinds of predictors we need to discover, we cover several of the most useful publicly available repositories of data for the study of ADRs, ranging from clinical data to molecular measurements. We then describe several recent specific cases as examples of how bioinformatics methods, coupled with data repositories, can advance the science relating to SADR. Finally, we end with a discussion of the future challenges in this field.

### MOLECULAR PREDICTORS FOR DRUG TOXICITY

Molecular predictors that are specifically relevant to drug toxicity have traditionally revolved around gene variants. Thiopurine methyltransferase, a phase II metabolizing enzyme, was first associated with the metabolism of 6-mercaptopurine and azathioprine back in 1980 (ref. 4). Lack of knowledge about the genotype of this gene can lead to a 10-fold overdosing of these drugs, potentially leading to fatal hematopoietic toxicity.

Guided largely by an increasing understanding of drug pharmacology and the high degree of variation in drug metabolizing enzymes, genes encoding for drug metabolizing enzymes have been the most studied class in the context of ADRs.<sup>5,6</sup> The highly polymorphic cytochrome P450 (CYP) enzyme system, a class of phase I metabolizing enzymes, is heavily studied because these enzymes are responsible for metabolizing ~67% of all drugs.<sup>7</sup> The Human Cytochrome P450 Allele Nomenclature Committee maintains a complete web-based list of all peer-reviewed alleles of CYP;<sup>8</sup> to date, it lists more than 200 alleles from 29 CYP families.

Candidate gene sequencing approaches have uncovered many of the associations between genetic variants and SADR. As an example, specific alleles in *CYP2C9* have been linked to hemorrhages in patients taking warfarin.<sup>9</sup> Polymorphisms in *CYP2D6* have also been linked to many SADR, including tardive dyskinesia and bradycardia in patients taking antipsychotics and  $\beta$ -blockers, respectively.<sup>10</sup> Similarly, variations in dihydropyrimidine dehydrogenase and UDP-glucuronosyltransferase (encoded by *UGT1A1*) have also been shown to be linked to neurotoxicity<sup>11</sup> and neutropenia<sup>12</sup> in patients taking 5-fluorouracil and irinotecan, respectively.

Although traditional candidate gene approaches have been powerful in yielding specific polymorphisms that are predictive of drug toxicity, they have been limited by progress in biological knowledge. Strategies that target only genes that are known to participate in pharmacokinetics and pharmacodynamics tend to ignore the many other genes that have yet to be biologically linked to SADR. Until recently, the ability to interrogate polymorphisms across the entire genome was hampered by both high costs and limitations in technology. Recently, technologies for genotyping have evolved while the costs have been driven down, so genome-wide association studies (GWASs) to link variants to SADR are now possible.<sup>13</sup> To date, these tools can genotype nearly 2 million variants at a cost of only a few hundred dollars per individual.

The immediate application of these high-density genotyping tools in determining the underlying genetic factors in various disease states is apparent and has demonstrable utility.<sup>13,14</sup> Recently, several GWASs have been successfully applied toward the study of SADR. These include studies of hemorrhage associated with warfarin, hepatic toxicity associated with ximelagatran, and simvastatin-induced myopathy. A GWAS of the most commonly used anticoagulant, warfarin, attempted to identify additional single-nucleotide polymorphisms (SNPs), beyond those in *VKORC1* and *CYP2C9*, that could explain the remaining 50% (estimated) of the variation in stable warfarin dosing.<sup>15</sup> However, in a subsequent validation set of patients, no other SNPs showed a significant enough association.<sup>15</sup> Ximelagatran, another anticoagulant, was slated to replace warfarin but ultimately was not marketed because hepatic toxicity was observed in clinical trials. A GWAS explored why some patients taking ximelagatran exhibited elevated serum alanine aminotransferase levels, a proxy for hepatic injury.<sup>15</sup> Strong genetic associations between elevations in alanine aminotransferase and major histocompatibility locus alleles *DRB\*07* and *DQA1\*02* were identified,<sup>15</sup> implicating antigen-presenting proteins in drug-induced liver toxicity. This finding was consistent with other genetic association studies of drug-induced liver toxicities.<sup>16,17</sup> Another recent GWAS, which included 85 patients with 80-mg simvastatin-induced myopathy and 90 control patients in a case-control design across 300,000 SNPs, revealed a single strong association with the *SLCO1B1* gene.<sup>18</sup> This gene encodes an organic anion transporter that is involved in statin uptake.<sup>19</sup> Findings such as these clearly demonstrate a full reversal in strategy, moving away from the older approach of developing hypotheses of gene variation on the basis of biological findings and knowledge and toward the development of hypotheses of biological processes on the basis of primary gene-variation findings.

### MOLECULAR PREDICTORS OF DRUG EFFICACY

In SADR, the study of drug efficacy is just as important as the study of drug toxicity; indeed, the failure of a drug to have efficacy in a patient with a lethal illness may carry severe consequences. Drug efficacy in cancer therapy is of particular importance, as SADR often develop as a result of cytotoxic chemotherapeutic regimens. For instance, cancer patients treated with cisplatin can suffer from SADR affecting hearing, the nervous system, and the kidneys, but such toxicities may be judged to be an acceptable risk if the therapeutic efficacy against an otherwise fatal disease is high. If SADR are an element of the risk-benefit equation in a chemotherapeutic setting, accurate predictors of efficacy are crucial.

Two prime examples of successful single-target-based predictors of drug efficacy are imatinib mesylate (Gleevec) and trastuzumab (Herceptin). The approval of imatinib mesylate by the US Food and Drug Administration (FDA) for use in the treatment of chronic myelogenous leukemia was given after only 3 months of review, the shortest such interval for any anticancer drug.<sup>20</sup> Almost all patients with chronic myelogenous leukemia were found to have a chromosomal translocation event between chromosomes 9 and 22 (known as the Philadelphia chromosome),

which activates the tyrosine kinase fusion oncoprotein Bcr/Abl.<sup>21</sup> Imatinib mesylate acts by inhibiting Bcr/Abl activation and has revolutionized chronic myelogenous leukemia therapy by dramatically increasing chronic myelogenous leukemia survival rates. On the other hand, only 11% of the breast cancer patients receiving trastuzumab, a monoclonal antibody targeted to human epidermal growth receptor 2, achieved tumor regression in the initial clinical trials.<sup>22</sup> However, with the treatment targeted only to patients with human epidermal growth receptor 2 overexpression, based on immunohistochemistry or cytogenetic assays, a much higher rate (34–50%) of patients demonstrated tumor regression.<sup>23</sup> Trastuzumab is also the first example of an FDA-approved drug with a companion FDA-approved immunohistochemistry diagnostic test, the HercepTest (Dako, Carpinteria, CA), designed to identify the patients most likely to benefit from the drug. This is because human epidermal growth receptor 2 overexpression is observed in only ~25% of patients with breast cancer. Both trastuzumab and imatinib mesylate have acquired additional indications for use—trastuzumab as part of various combination therapies for breast cancer and imatinib for kit (CD117)-positive gastrointestinal stromal tumors—showing the broad utility of single-target drugs.

The single-target drugs imatinib mesylate and trastuzumab are only two examples of several that demonstrate efficacy in a subset of a larger patient population. Over the past 7 years, molecular predictors based on genomic data, particularly gene expression signatures, have paved the way for improved selection of these patients. Successful chemosensitivity predictions—based on gene expression signatures to determine which subsets of patients would best respond to specific chemotherapeutic drugs—have demonstrated their utility in cancer therapies such as breast cancer and childhood leukemia. A review of the 41 studies that utilized gene expression signatures to predict drug chemosensitivity was recently published.<sup>24</sup> One of the earliest examples of these predictions was a 92-gene signature to predict docetaxel response in breast cancer patients.<sup>25</sup> Gene expression signatures have been successfully combined with drug sensitivity data on cancer cell lines to predict how patients will respond.<sup>26</sup> In one study, bioinformatics methods were used to link drug sensitivity data from the National Cancer Institute's (NCI's) 60 cell lines (NCI-60) in order to identify genes that best discriminate responses to etoposide, adriamycin, cyclophosphamide, and 5-fluorouracil.<sup>27</sup> In addition, the patients' gene expression signatures were compared with other gene expression signatures generated from known pathway activations, thereby allowing drugs targeting matching pathways to be the prime therapeutic candidates.<sup>27</sup>

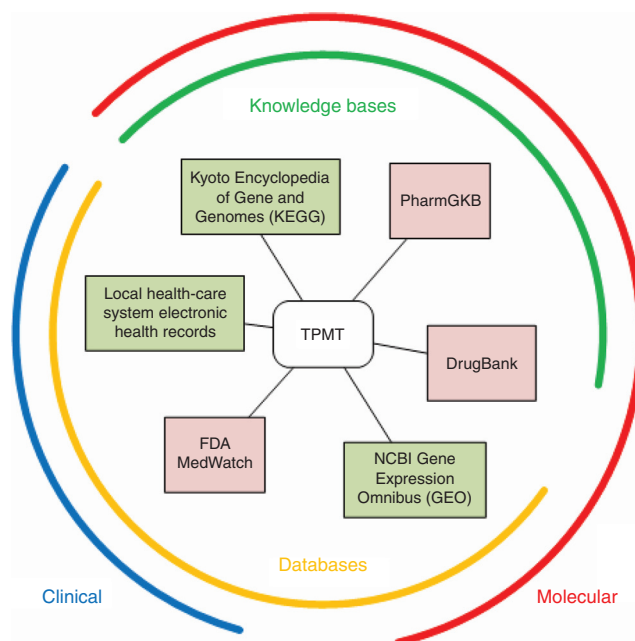
GWAS have also been performed to investigate drug efficacy. Studies examining responses to antitumor necrosis factor therapy in patients with rheumatoid arthritis,<sup>28</sup> interferon- $\beta$  therapy in multiple sclerosis patients,<sup>29</sup> ilperidone,<sup>30</sup> and the diuretic thiazide<sup>31</sup> have enhanced our understanding of why certain individuals fail to respond to certain drugs. A recent GWAS yielded gene variants associated with successful smoking cessation and may aid in the selection of cessation medications.<sup>32</sup> It is becoming clearer that patient stratification based on drug efficacy and toxicity patterns is needed to prevent SADRs.

The National Institutes of Health maintains the most up-to-date list of published high-density GWASs, indicating the phenotype studied and *P* values of significant SNP allele associations, along with the sizes of the studies, for both the discovery and replication phases (Table 1).

### CLINICAL REPOSITORIES AVAILABLE FOR THE STUDY OF SADRs

Although molecular data have been shared across communities of investigators for more than two decades, the number of repositories holding various kinds of biological and molecular measurements has continued to grow exponentially. In 2008, the annual Molecular Biology Database Collection issue of *Nucleic Acids Research* grew by 10% to include more than 1,000 databases.<sup>33</sup> Studies investigating the nature of SADRs have been carried out on both clinical and molecular fronts, with much of the data and results being deposited into various types of databases, some of which are publicly accessible. These many publicly available databases can be used in the study of SADRs, even though they are not labeled as primarily SADR repositories. Here, we give examples of clinical and molecular repositories and their applicability to the study of SADRs (Figure 1).

The focus of the clinical side of SADR studies is primarily detection and prevention. In many countries, pharmacovigilance is a key strategy in detecting and preventing SADRs. Of course, effective pharmacovigilance requires the cooperation of the



**Figure 1** A well-known pharmacogene, thiopurine methyltransferase (TPMT), appears in a variety of publicly accessible information sources. Knowledge bases such as the Kyoto Encyclopedia of Genes and Genomes (KEGG), the Pharmacogenetics and Pharmacogenomics Knowledge Base (PharmGKB), and DrugBank hold existing knowledge on TPMT, whereas databases such as clinical electronic health records, MedWatch, and GEO hold observed measurements or characteristics of TPMT. Both knowledge and data on TPMT can range from the clinical (or whole organism) realm to the molecular realm. Information sources (in pink) are pharmacology specific, whereas the sources shown in green are for general use.

pharmaceutical and biotechnology industries, regulatory agencies, and academic institutions. Clinical trials are typically where ADRs arise for the first time. One example is the identification of a *UGT1A1* variant responsible for susceptibility to tranilast-induced hyperbilirubinemia, found during a phase III clinical trial testing the efficacy of tranilast in reducing the re-stenosis rate after percutaneous *trans*-luminal coronary angioplasty.<sup>34</sup> Clinical trials are excellent for identifying SADRs; however, because of the limited number of participants, short durations of the trials, and the fact that the characteristics of study participants may not reflect those of the patients (e.g., children, individuals of various ethnicities) who may later have the drug prescribed to them, clinical trials often fail to detect SADRs.

Many SADRs are detected through postmarketing spontaneous reporting systems established by many countries following the discovery of the teratogenic effects of thalidomide. Spontaneous reporting systems are valuable tools because they allow the early detection of rare, new SADRs in a cost-effective manner. In fact, such systems are responsible for the majority of drug withdrawals from the market, although some argue that they are insufficient to identify all current SADRs.<sup>35</sup> Global drug safety monitoring efforts spearheaded by the World Health Organization include Vigibase, a database containing drug safety information from 82 nations with more than three million reports.<sup>36</sup> Many countries have also set up their own spontaneous reporting systems, such as the United Kingdom's Yellow Card System, Japan's Pharmaceuticals and Medical Devices Agency, the FDA's MedWatch (also known as the Adverse Event Reporting System), and the Canadian Vigilance Online Database, which contains ADR events dating back to 1965. For SADRs linked to Vioxx, for example, MedWatch lists 608 reported adverse effects from a recent quarter, and the Canadian Vigilance Online Database breaks down its reports for an equivalent period as 511 submitted by pharmacists and 168 submitted by consumers. Although the nomenclature for drugs is not universally standardized within or between these systems, newer vocabularies, such as the National Library of Medicine's RxNorm, could be used across these systems.<sup>37</sup>

Health-care institutions both large and small are moving toward electronic computerized systems as a way to streamline the entire health-care process. This has opened the door for the use of electronic health records to identify ADRs. Indeed, several studies have demonstrated that, for identifying ADRs, the mining of electronic medical records has the advantage of saving time as compared to traditional manual chart reviews. These studies employed a combination of methods to detect ADRs, including rule-based triggers, drug–drug interactions, drug allergy information, natural-language processing, and the Naranjo algorithm, a questionnaire that estimates the probability of an ADR.<sup>38–40</sup> Moreover, all of them revealed that a significant percentage (~25%) of ADRs are preventable. Together, these clinical studies of ADRs have shown that clinical improvements can be made to avoid SADRs. Although clinical records of patients are typically not publicly available and require institutional review board approval for study, the widespread adoption of electronic health records essentially means that this type of data

has at least become locally available in many locales. Toward the goal of improving sensitivity by pooling resources across locales, the FDA recently launched a more active drug safety surveillance system, called the Sentinel Initiative.<sup>41</sup> In cooperation with the Centers for Medicare and Medicaid Services, the Veterans Administration, the Department of Defense, and both public and private organizations, the FDA will have access to 25+ million electronic medical records so that active monitoring of SADRs can be carried out. It is clear that a global, collaborative effort will be required to prevent SADRs. One recent example of this type of collaboration is the establishment of The Predictive Safety Testing Consortium comprising both the public and private sectors. Early findings led to the qualification of seven new biomarkers to monitor renal toxicity.

### MOLECULAR KNOWLEDGE BASES AND REPOSITORIES FOR STUDYING SADRS

The focus of the molecular arm of SADR studies is primarily on determining the underlying molecular mechanisms in SADRs. There are many molecular repositories and databases available to the bioinformatics-enabled SADR researcher (Table 1). Indeed, of the more than 1,000 molecular databases listed earlier,<sup>33</sup> 29 have the word “drug” in them, up from 3–4 years ago. In this review, we first cover the popular knowledge bases containing genes that are known to be related to drug use and effects and then go on to cover the repositories of molecular data that can be exploited to discover new genes associated with drugs.

One of the most comprehensive data repositories of known genes affecting pharmacokinetics and pharmacodynamics is the Pharmacogenetics and Pharmacogenomics Knowledge Base.<sup>42</sup> The curators of this database continually scan the pharmacogenomics literature for the most up-to-date information related to drugs, disease, and genes. As of this writing, a search for “CYP2D6” returns nearly 500 hits with links to various publications, drugs, pathways, phenotypes, and diseases.

DrugBank is a unique knowledge base of drugs (including FDA-approved ones as well as experimental ones), drug actions, and their molecular targets.<sup>43</sup> A DrugCard entry stores all associated information aggregated from various sources, such as GenBank and Chemical Entities of Biological Interest, pertaining to a specific drug. With more than 4,700 drugs, and more than 100 data fields from more than 20 databases, as of this writing, DrugBank effectively bridges the various nomenclatures and identifiers found in diverse drug databases into cohesive DrugCard entries.<sup>43</sup> The recently added GenoBrowse feature in DrugBank summarizes the specific genes and SNP alleles, along with literature references relating to various ADRs and functional effects. To date, GenoBrowse lists adverse effects and drug function information for approximately 60 drugs. For instance, five distinct alleles in four different genes have been linked to ADRs in patients taking 5-fluorouracil.

A smaller knowledge base, the Table of Valid Genomic Biomarkers in the Context of Approved Drug Labels, is maintained by the FDA itself, with a summary webpage on all approved drug label changes related to genomic biomarkers. These include recommendations for physicians prescribing

**Table 1 Translational bioinformatics resources for SADR studies**

|  |   |
|--|---|
| Pharmacogenomic knowledge bases  |   |
| DrugBank–GenoBrowse  | <a href="http://www.drugbank.ca/genobrowse">http://www.drugbank.ca/genobrowse</a>   |
| FDA Table of Valid Genomic Biomarkers in the Context of Approved Drug Labels | <a href="http://www.fda.gov/cder/genomics/genomic_biomarkers_table.htm">http://www.fda.gov/cder/genomics/genomic_biomarkers_table.htm</a> |
| PharmGKB   | <a href="http://www.pharmgkb.org/">http://www.pharmgkb.org/</a>   |
| Chemical Effects in Biological Systems                                       | <a href="http://cebs.niehs.nih.gov">http://cebs.niehs.nih.gov</a>   |
| Genetic databases  |   |
| Catalog of Published Genome-Wide Association Studies                         | <a href="http://genome.gov/gwastudies/">http://genome.gov/gwastudies/</a>   |
| NIH Genetic Association Database   | <a href="http://geneticassociationdb.nih.gov/">http://geneticassociationdb.nih.gov/</a>   |
| NCBI Database of Genotype and Phenotype                                      | <a href="http://www.ncbi.nlm.nih.gov/gap/">http://www.ncbi.nlm.nih.gov/gap/</a>   |
| Molecular databases  |   |
| NCBI Gene Expression Omnibus   | <a href="http://www.ncbi.nlm.nih.gov/geo/">http://www.ncbi.nlm.nih.gov/geo/</a>   |
| EBI ArrayExpress   | <a href="http://www.ebi.ac.uk/microarray-as/ae/">http://www.ebi.ac.uk/microarray-as/ae/</a>   |
| Connectivity Map   | <a href="http://www.broad.mit.edu/cmap/">http://www.broad.mit.edu/cmap/</a>   |
| Pathway and interaction databases  |   |
| Kyoto Encyclopedia of Genes and Genomes                                      | <a href="http://www.genome.ad.jp/kegg/">http://www.genome.ad.jp/kegg/</a>   |
| Comparative Toxicogenomics Database  | <a href="http://ctd.mdibl.org/">http://ctd.mdibl.org/</a>   |
| Chemical databases   |   |
| Chemical Entities of Biological Interest                                     | <a href="http://www.ebi.ac.uk/chebi/">http://www.ebi.ac.uk/chebi/</a>   |
| NCBI PubChem   | <a href="http://pubchem.ncbi.nlm.nih.gov/">http://pubchem.ncbi.nlm.nih.gov/</a>   |
| Developmental Therapeutics Program   | <a href="http://dtp.nci.nih.gov/">http://dtp.nci.nih.gov/</a>   |

EBI, European Bioinformatics Institute; FDA, Food and Drug Administration; NCBI, National Center for Biotechnology Information; NIH, National Institutes of Health.

specific drugs to pay particular attention to subsets of patient populations. These drug labeling changes are slowly altering the nature of clinical care.

There are other knowledge bases that link chemicals with genes and proteins; because these knowledge bases include facts on more than just drugs, they may hold utility for bioinformatics-based SADR studies. The Chemical Effects in Biological Systems knowledge base, a systems toxicogenomic repository founded on systems biology principles, stores toxicological information from gene expression, proteomic, clinical chemistry and histopathology, and metabolic experiments.<sup>44</sup> For instance, clinical and pathology data, along with proteomic and gene expression experiments studying the toxic effects of acetaminophen on liver, are stored within the Chemical Effects in Biological Systems knowledge base. The Comparative Toxicogenomics Database is another curated knowledge base that captures the relationships between chemicals, genes, and diseases, with more than 4,000 chemicals, 14,000 genes, and 3,000 diseases.<sup>45</sup> For example, sirolimus has 75 known interacting genes, associations with 236 diseases, and 162 pathway associations. Along the same lines, the Kyoto Encyclopedia of Genes and Genomes is a foundational knowledge base that stores curated pathway information (molecular interactions and cellular processes) from published literature,<sup>46</sup> giving network context of how genes, diseases, and metabolites are interrelated.

For more than a decade, gene expression microarrays have enabled the measurement of RNA expression levels across the genomes of several organisms.<sup>47,48</sup> With this track record, it is no surprise that whole-genome molecular profiling studies relating to

drug toxicity and responsiveness that use gene expression microarrays outnumber those that use high-density SNP measurements. This is primarily because data on toxicity effects in cell lines and in model organisms such as rat and mouse are much more readily available than samples from human patients. By far the largest repository of gene expression experiments, the Gene Expression Omnibus (GEO), maintained by the National Library of Medicine, contains data from many such drug-related experiments.<sup>49</sup> The largest single toxicogenomics data set in GEO is one that comprises 5,288 microarrays.<sup>50</sup> The study was performed by Iconix Biosciences to explore drug responses in the liver, using 1,695 rats treated with various doses of 344 compounds. The European Bioinformatics Institute also stores similar gene expression experiments in the ArrayExpress database. As of this writing, there are 29 toxicogenomic experiments in ArrayExpress.

One large-scale gene expression study, known as the Connectivity Map, consists of 164 chemicals tested mostly on the MCF7 breast cancer cell line.<sup>51</sup> Pattern-matching algorithms can be used on these gene expression signatures to find hidden similarities in effects across these chemicals. Data from the Connectivity Map have been deposited into GEO. Apart from the data from this single experiment, data from many other gene expression studies are also available in GEO. We have estimated that gene expression data from at least 213 drugs are available in GEO, tested across many doses and tissues, and contained in many separate experiments.<sup>52</sup> As we show later in this review, these public repositories can also serve as resources for larger meta-analyses. Public databases of the bioactivities of small molecules have primarily been the result of government-based

initiatives. The NCI Developmental Therapeutics Program (DTP), established by Congress in 1955 at the Cancer Chemotherapy National Service Center, provides a rich repository of cancer cell lines and measurements made on those cell lines, including the NCI-60 panel.<sup>53</sup> The DTP has an emphasis on cancer, transplantable animal and human tumors, small molecules, and compound screening services. Toward that end, the DTP has made public the data from screening thousands of compounds in those cancer cell lines and animal models, providing invaluable preclinical and research tools. This has resulted in the discovery of 40 chemotherapeutic drugs, including cetuximab (Erbix), a monoclonal antibody that inhibits epidermal growth factors and has been approved for the treatment of colorectal<sup>54</sup> and head and neck cancers.<sup>55</sup> The DTP compound susceptibility measurements have been combined with gene expression measurements, leading to predictors of chemotherapeutic efficacy.<sup>26,56,57</sup>

The Molecular Libraries Initiatives, a generalization of the methods of the DTP, was established by the National Institutes of Health in 2003 as one of the five Roadmap Initiatives designed to expedite the translation of research discoveries to the bedside. The major goal of the Molecular Libraries Initiatives is to acquire and screen 500,000 small molecules in high-throughput bioassays and release the screening results in a database called PubChem.<sup>58</sup> Today, PubChem contains information on 18 million chemical compounds as studied across nearly a thousand bioassays, including those from the DTP. Although PubChem is not specific to the study of SADR, there are bioassay measurements relevant to SADR, including growth inhibition assays that give clues about drug chemosensitivities in various cells or organisms.

The public availability of these knowledge bases and databases directly and indirectly related to SADR is slowly breaking down the barriers to the study of SADR and invites creative integration of data between disparate experimental modalities. Bioinformatics already plays a central role in aggregating and storing these increasingly large sets of data as well as knowledge from both the clinical and molecular domains. More important, the development of the analytic tools and algorithms that are required in order to analyze and interpret these results and knowledge will be pivotal in translating research findings into clinical practice.

### BIOINFORMATICS SUCCESS STORIES FOR THE STUDY OF ADRs

Recent successful studies by several groups highlight the elegant use of disparate data sets in the study of SADR, enabled by bioinformatics methodologies. In each of these cases, two or more knowledge bases and databases were linked because of a commonality across those sets. Drawing on the knowledge stored in DrugBank, which contains information on both approved and experimental drugs and their targets, Yildirim and colleagues constructed a drug-target network and used network properties to describe drugs and how they relate to other data sets.<sup>59</sup> By clustering drugs, along with their targets, on the basis of Anatomical Therapeutic Chemical classification, the drug-target network confirmed that most drugs are “follow-on” drugs,

that is, drugs targeting proteins already targeted by another drug. In comparing the drug-target proteins with essentiality proteins (proteins encoded by genes whose orthologs in model organisms are found to be essential), drug-target proteins show different topographical signatures and have less gene expression/coexpression and higher tissue specificity. Moreover, an additional comparison of drug-target proteins to disease proteins previously implicated in mendelian disorders indicated that most drugs are palliative (targeting proteins that have no causal role in the disease) rather than etiology-based (proteins causing the disease). This last finding is not surprising, because most drugs address the clinical symptoms rather than the underlying disease pathogenesis. Coupled with the lack of a complete understanding of the mechanistic underpinnings of most drugs, often leading to off-target effects, it is not surprising that ADRs are common and SADR are often unanticipated.

An example of a study of drug toxicity is one by Huang and colleagues, who sought to determine the genetic variants associated with cytotoxicity arising from various drug treatments.<sup>60,61</sup> They took advantage of the lymphoblastoid cell lines that have been used to maintain individual DNA samples needed for the Human Genome Project and subsequent HapMap projects. Huang applied various drug treatments to these cell lines and measured the resulting cytotoxicity. Genome-wide gene expression profiling was also performed on the same cell lines. Given that the SNP genotypes had already been determined for these cell lines as part of the HapMap project, Huang was able to borrow from those findings to identify the SNPs responsible for cytotoxicity and sensitivity to cisplatin,<sup>61</sup> carboplatin,<sup>60</sup> and daunorubicin.<sup>62</sup> Although we acknowledge that lymphoblastoid cell lines are not necessarily the best cells in which to probe for drug toxicity, this research approach is an example of how new studies using existing, publicly available data resources can lead to research findings that are greater than the sum of the parts.

The data relating to the activities of chemical compounds on the NCI-60 cancer cell lines are stored in a repository so that they can be used as a data resource for research toward drug development leads. Two studies utilizing data from these screenings exemplify how these data can be extended to make novel discoveries.

(i) While the NCI-60 cell lines comprise cells from various cancers, including leukemias, melanomas, and breast, ovarian, renal, prostate, colon, lung, and central nervous system cancers, other cancer cell lines, such as those from bladder cancers, are not represented. Lee and colleagues sought to take advantage of existing NCI-60 data and apply them to unrepresented cancers in a methodology called “coexpression extrapolation,” or COXEN.<sup>63</sup> This was done by first identifying a common molecular data set or experimental modality between the NCI-60 cells and the bladder cancer cells, which, in this case, were gene expression measurements. Then, chemosensitive or resistant gene expression signatures were linked to the drug activity levels measured in the NCI-60 cells. Based on coexpression patterns common to both the bladder cancer cells and the NCI-60 cells, a multivariate algorithm was used to extrapolate the data and predict the drug’s activity on the bladder cells.<sup>63</sup> In this fashion,

the authors performed *in silico* screening and identified a novel drug candidate for treating bladder cancer.

(ii) Fliri and colleagues had a different goal in mind, namely, to find preclinical markers that could be predictive of postmarketing side effects.<sup>64</sup> Fliri first constructed biological activity spectra (biospectra) from *in vitro* protein binding assays of prescription drugs.<sup>65</sup> Using hierarchical clustering methods, Fliri found regions of similarity between biospectra as well as similarities in structure and side effect profiles from drug information labels across the same drugs. This research methodology shows how coded drug label information and molecular measurements can be linked if these pieces of data apply to the same set of pharmaceuticals.

Lastly, instead of using NCI-60 data, Campillos and colleagues exploited the side effect information from prescription drug labels to identify novel molecular activities of existing drugs.<sup>66</sup> Similarity in side effects was classified according to the Unified Medical Language System (UMLS). The UMLS was created by the National Library of Medicine more than 20 years ago to compile a large, comprehensive, standardized vocabulary for the “language of biomedicine and health,”<sup>67</sup> covering more than 100 biomedical vocabularies. The UMLS contains more than 1 million concepts in biomedicine (e.g., concept *C0206131*, “adipocytes”), unified across lexical variation and terminology (e.g., “adipocyte,” “mature fat cell”), language (e.g., “lipozyten” in German), and original coding system (e.g., *M0026722* is used by the librarians in MeSH, *24826007* by the pathologists in SNOMED-CT).<sup>68</sup> Campillos used the UMLS to represent side effects and a weighting scheme to account for the rareness and interdependence of side effects.<sup>66</sup> Because similarity in side effects correlated with shared target(s) between drugs, Campillos reasoned that side effect similarity could be used to predict novel targets between any two “unexpected” drug pairs. One example was fluoxetine, which was predicted to target dopamine receptor D3 because of its side effects shared with rabeprazole. By combining side effect similarity with chemical similarity, 13 of 20 novel target predictions were experimentally validated,<sup>66</sup> thereby identifying novel off-target effects that could be used to derive novel indications for these drugs.

Several unifying themes related to bioinformatics are common to the success of these studies. First, publicly accessible repositories or databases are crucial in advancing knowledge. All of the studies mentioned used resources freely available to the public, such as the data from the NCI-60 and the HapMap genotypes. Second, novel approaches to data abstraction or representation of existing information, such as biospectra or side effect similarity, can reveal novel relationships. Third, unified vocabularies, such as those found in the Anatomical Therapeutic Chemical classification or the UMLS, are essential not just for data interpretation and analyses but also to enable novel lines of questioning. Finally, profound insights can be gained from creative integration of data from various experimental modalities.

#### NEWER TYPES OF MOLECULAR MEASUREMENTS

Much of our existing knowledge of genetic factors involved in SADR centers on SNPs; however, polymorphisms of other

kinds have also been identified, and they point to the existence of new knowledge that will need to be uncovered for a better understanding of the mechanisms underlying SADR. For instance, copy number variations and deletions have also been shown to contribute to SADR. Some individuals with extra copies of specific variants of CYP2D6—responsible for metabolizing numerous drugs, including opioids such as codeine—can suffer from toxic effects. This has resulted in label changes mandated by the FDA to inform consumers of genetic risk factors. The glutathione transferases GSTM1 and GSTT1, responsible for metabolizing anticancer drugs such as cisplatin and 5-fluorouracil, have been found to be deleted in individuals across many ethnicities.<sup>69</sup> Copy number changes can be surveyed across the genome using the same technologies currently used to survey SNPs.

Besides nucleotide base changes, epigenetic modifications can also result in clinical and molecular phenotypes. For instance, the high CYP1B1 expression in prostate cancer cells as compared with normal cells has been linked to hypomethylation in the promoter and enhancer regions of CYP1B1.<sup>70</sup> Hypomethylation allows transcription factors and enhancers access to promoter/enhancer regions and increases gene expression. Moreover, methylation differences may contribute to variable drug response. For example, the CYP24 methylation pattern is different in tumor-derived endothelial cells as compared with normal endothelial cells.<sup>71</sup> The methylation, or silencing, of CYP24 in tumor cells resulted in reduced responsiveness of CYP24 to a pharmaceutical.

A recently discovered gene regulatory mechanism involving microRNAs (miRNAs) has also been found to have a connection with SADR. Discovered in 1993, miRNAs are small non-coding RNAs that typically bind to the 3'-untranslated region of mRNAs and target them for degradation or translational repression.<sup>72</sup> Tsuchiya and colleagues recently showed how one miRNA, miR-27b, regulates the expression of CYP1B1, which metabolizes polycyclic aromatic hydrocarbons and 17 $\beta$ -estradiol.<sup>73</sup> Because CYP1B1 is found to be highly expressed in cancer tissues, this study essentially links miRNAs to drug chemosensitivity. A separate study found that another miRNA, miR-24, binds to the 3'-untranslated region of dihydrofolate reductase and regulates its level. A SNP in the 3'-untranslated region of dihydrofolate reductase in the presumed binding site of miR-24 results in overexpression of dihydrofolate reductase and increased resistance to methotrexate, which is metabolized by dihydrofolate reductase.<sup>74</sup>

The nascent field of pharmacoproteomics refers to the study of how proteins change in response to drug treatments and has already contributed toward our understanding of SADR. In particular, two-dimensional gel electrophoresis has proved useful in understanding the mechanisms underlying SADR relating to several drugs. One example is the study of the toxic liver effects of methapyrilene, an antihistamine and sleep aid that was eventually withdrawn from the market. The liver toxicity was found to be due to the protein adducts formed by methapyrilene, specifically in mitochondrion of rat liver, which was used to model this effect.<sup>75</sup> In addition, the toxic effects of cyclosporine A, an immunosuppressant, in rat kidney<sup>76</sup> and brain<sup>77</sup> were also

**Table 2 Pharmacovigilance and genetic SADR consortiums**

|  |   |
|--|---|
| Spontaneous reporting systems  |   |
| VigiBase<br>~4 million ADR reports from 82 countries   | <a href="http://www.umc-products.com/DynPage.aspx?id=4910&amp;mn=1107">http://www.umc-products.com/DynPage.aspx?id=4910&amp;mn=1107</a>                     |
| Adverse Event Reporting System<br>ADRs reports maintained by United States   | <a href="http://www.fda.gov/cder/aers/default.htm">http://www.fda.gov/cder/aers/default.htm</a>   |
| Canada Vigilance Online Database<br>Contains ADR reports in Canada dating back to 1965   | <a href="http://www.hc-sc.gc.ca/dhp-mps/medeff/databasdon/index-eng.php">http://www.hc-sc.gc.ca/dhp-mps/medeff/databasdon/index-eng.php</a>                 |
| SADR consortiums   |   |
| Serious Adverse Event Consortium<br>SADR focus: drug-induced liver toxicity, Steven Johnson syndrome   | <a href="http://www.saeconsortium.org">http://www.saeconsortium.org</a>   |
| International Warfarin Consortium<br>Drug focus: warfarin  | <a href="http://www.pharmgkb.org/views/project.jsp?pld=56">http://www.pharmgkb.org/views/project.jsp?pld=56</a>   |
| European collaboration for studying genetic basis of adverse drug reactions (EUDRAGENE)<br>Drug focus: cholesterol-lowering drugs, thyroid drugs                               | <a href="http://www.eudragene.org">http://www.eudragene.org</a>   |
| Canadian Genotype-specific Approaches to Therapy in Childhood program (GATC)<br>Drug focus: amoxicillin, carbamazepine, valproic acid, cefprozil, infliximab, and isotretinoin | <a href="http://www.genomebc.ca/research_tech/research_projects/health/gatc.htm">http://www.genomebc.ca/research_tech/research_projects/health/gatc.htm</a> |
| United States Drug-Induced Liver Injury Network (DILIN)<br>Drug focus: isoniazid, phenytoin, clavulanic acid/amoxicillin, and valproic acid                                    | <a href="http://diln.dcri.duke.edu">http://diln.dcri.duke.edu</a>   |
| Pharmacogenetics of antimicrobial drug-induced liver injury (Diligen)<br>Drug focus: co-amoxiclav, flucloxacillin, antituberculosis drugs                                      | <a href="http://diligen.org">http://diligen.org</a>   |

ADR, adverse drug reaction; SADR, serious adverse drug reaction.

discovered with the aid of two-dimensional gel electrophoresis. Protein arrays, which are conceptually similar to high-density SNP and gene expression measurements, are just emerging as potent tools to aid new studies in drug toxicity.

Together, these studies across the scale of molecular measurements—from genetic copy number variation to miRNA, epigenetics, and proteomics—show that SADRs are the result of variations and defects across many genomic levels and point to the increasing complexities associated with elucidating the underlying mechanisms responsible for SADRs. As more of these data become available, we will undoubtedly see new calls from funding agencies and journals for public disclosure of these data, similar to the calls that led to the current availability of data on gene expression, genetic polymorphisms. The first creative applications of bioinformatics methods to these data sets, especially when “mixed” with previously collected data, are likely to be of high impact.

### PERSPECTIVES AND CHALLENGES

The understanding of the molecular determinants of drug toxicity and efficacy of a few drugs has already transformed clinical care, particularly in cancer therapies. The development of bioinformatics tools is foundational and necessary in order to study and analyze the ever increasing volumes of data and knowledge stored in databases and knowledge bases for studying SADRs. More important, recent studies employing bioinformatics methodologies are paving the way toward elucidating the molecular underpinnings of SADRs. In the near future, bioinformatics will be integral to the continuing movement to bridge the gap

between the molecular and clinical domains. The limitations of existing capabilities hinder a focus on postmarket surveillance; however, a more proactive premarket review of drug toxicity and efficacy is becoming a possibility.

There has never been a more opportune time to realize the potential of pharmacogenomics. We now have at our disposal many genomic technologies that we can leverage toward the identification of genetic risk factors involved in SADRs and toward translation of findings to the clinic. However, there are still many challenges ahead. Determining the molecular risk factors in SADRs, whether or not such factors are genetic, depends fundamentally on the ability to find the patients who suffer from SADRs. Because SADRs are rare events, the identification of patients suffering from SADRs and the collection of their biospecimens is not a trivial problem. The challenge of finding these patients is compounded by the lack of objective, standardized diagnosis of SADRs for all drugs. For example, drug-induced liver toxicity has no established diagnostic criteria and relies instead on diagnosis by exclusion. In addition, patients who suffer from SADRs comprise only a small percentage of the overall patients who use specific drugs, and SADRs sometimes affect only specific populations (e.g., those of certain ethnicities or the elderly).

Finding these patients requires a global effort that spans the pharmaceutical and biotechnology industries, regulatory agencies, and educational institutions. All of these must collaborate in order to identify the appropriate patients. Toward this end, regulatory agencies as well as academic institutions have taken initiatives to achieve better identification of SADR

patients, collect and store biospecimens from these patients, and compile family histories and pertinent information (e.g., diet, geographic location) to enable more fruitful SADR studies.

Specifically, several consortiums have been established to identify genetic factors in SADRs. These include the recently formed Serious Adverse Events Consortium, the International Warfarin Consortium,<sup>78</sup> and the European Consortium, EUDRAGENE,<sup>79</sup> established to study six SADRs (Table 2). Often overlooked, these consortiums are an essential first step toward identifying the drugs causing SADRs and the affected patients. Some of these consortia have committed to releasing data to the public within a defined time frame, and this kind of data will fuel bioinformatics-enabled discoveries about the nature of SADRs.

Patient stratification is another key problem. Stratifying patients on the basis of drug efficacy and toxicity may present an economic challenge to the pharmaceutical and biotechnology industries because the costs associated with drug development may not be offset by sales that are confined to only a subset of the ideal patient population.

Many current efforts are focused on finding SNPs associated with SADRs, with a strong push toward GWASs using high-density SNP measurements to study drug–response variability. There have been only a few successful studies thus far. It is too early to know whether these GWASs are sensitive enough to detect all the genes involved, or whether an SADR represents a single phenotype or a variety of molecular phenotypes that are impossible to distinguish clinically. One major challenge that will need to be addressed is the lack of reproducibility between these high-density genetic association studies. Moreover, variations beyond SNPs, such as copy number variations, miRNAs, and epigenetics, have already been shown to be important in SADRs and will also need to be examined on a genome-wide scale. Additional measurement technologies and tools may be required in order to identify the genes and underlying mechanisms that are important in ADRs, such as protein–protein interactions.

Technological advances must be coupled with development of novel algorithms and analytical tools to evaluate these high-throughput screens in order to yield clinically relevant results. Although investigators are often cautious about sharing data, unfettered access to data is the most efficient way for bioinformaticians to develop better quantitative and analytical methods, from which all will benefit. In addition, bioinformatics-enabled findings are not the end goal of SADR studies but only a means to an end. Accurate, cost-effective, and rapid-turnaround clinical tests must also be developed to enable broad use.

Our existing knowledge of pharmacogenomics has come a long way in the five decades since genetics was first suggested to be involved in drug–response variability, but our available data have greatly surpassed our existing knowledge. Measurements drive analytical tools, tools drive data, and data enable the raising of novel questions. These questions may be asked not by traditional chemists or drug-discovery engineers but by computationally enabled scientists. Continued open-mindedness on the part of academics and industries toward this new group of investigators, toward open sharing of data, and toward pooling of rare

resources will move the field of pharmacology toward the prevention of SADRs and make more individualized, personalized medicine a reality.

#### ACKNOWLEDGMENTS

This work was supported by grants from the Lucile Packard Foundation for Children's Health, the National Institute of General Medical Sciences (R01 GM079719), the National Library of Medicine (T15 LM007033), the Howard Hughes Medical Institute, and the Pharmaceutical Research and Manufacturers of America Foundation.

#### CONFLICT OF INTEREST

The authors declared no conflict of interest.

© 2009 American Society for Clinical Pharmacology and Therapeutics

- Lazarou, J., Pomeranz, B.H. & Corey, P.N. Incidence of adverse drug reactions in hospitalized patients: a meta-analysis of prospective studies. *JAMA* **279**, 1200–1205 (1998).
- Giacomini, K.M., Krauss, R.M., Roden, D.M., Eichelbaum, M., Hayden, M.R. & Nakamura, Y. When good drugs go bad. *Nature* **446**, 975–977 (2007).
- Butte, A.J. Translational bioinformatics: coming of age. *J. Am. Med. Inform. Assoc.* **15**, 709–714 (2008).
- Weinshilboum, R.M. & Sladek, S.L. Mercaptopurine pharmacogenetics: monogenic inheritance of erythrocyte thiopurine methyltransferase activity. *Am. J. Hum. Genet.* **32**, 651–662 (1980).
- Evans, W.E. & Johnson, J.A. Pharmacogenomics: the inherited basis for interindividual differences in drug response. *Annu. Rev. Genomics Hum. Genet.* **2**, 9–39 (2001).
- Pirmohamed, M. & Park, B.K. Genetic susceptibility to adverse drug reactions. *Trends Pharmacol. Sci.* **22**, 298–305 (2001).
- Williams, J.A. et al. Drug–drug interactions for UDP-glucuronosyltransferase substrates: a pharmacokinetic explanation for typically observed low exposure (AUCi/AUC) ratios. *Drug Metab. Dispos.* **32**, 1201–1208 (2004).
- Sim, S.C. & Ingelman-Sundberg, M. The human cytochrome P450 Allele Nomenclature Committee web site: submission criteria, procedures, and objectives. *Methods Mol. Biol.* **320**, 183–191 (2006).
- Higashi, M.K. et al. Association between CYP2C9 genetic variants and anticoagulation-related outcomes during warfarin therapy. *JAMA* **287**, 1690–1698 (2002).
- Meyer, U.A. Pharmacogenetics and adverse drug reactions. *Lancet* **356**, 1667–1671 (2000).
- Wei, X., McLeod, H.L., McMurrugh, J., Gonzalez, F.J. & Fernandez-Salguero, P. Molecular basis of the human dihydropyrimidine dehydrogenase deficiency and 5-fluorouracil toxicity. *J. Clin. Invest.* **98**, 610–615 (1996).
- Iyer, L. et al. Genetic predisposition to the metabolism of irinotecan (CPT-11). Role of uridine diphosphate glucuronosyltransferase isoform 1A1 in the glucuronidation of its active metabolite (SN-38) in human liver microsomes. *J. Clin. Invest.* **101**, 847–854 (1998).
- McCarthy, M.I. et al. Genome-wide association studies for complex traits: consensus, uncertainty and challenges. *Nat. Rev. Genet.* **9**, 356–369 (2008).
- Wellcome Trust Case Control Consortium. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* **447**, 661–678 (2007).
- Cooper, G.M. et al. A genome-wide scan for common genetic variants with a large influence on warfarin maintenance dose. *Blood* **112**, 1022–1027 (2008).
- Sharma, S.K., Balamurugan, A., Saha, P.K., Pandey, R.M. & Mehra, N.K. Evaluation of clinical and immunogenetic risk factors for the development of hepatotoxicity during antituberculosis treatment. *Am. J. Respir. Crit. Care Med.* **166**, 916–919 (2002).
- O'Donohue, J. et al. Co-amoxiclav jaundice: clinical and histological features and HLA class II association. *Gut* **47**, 717–720 (2000).
- Link, E. et al. SLCO1B1 variants and statin-induced myopathy—a genome-wide study. *N. Engl. J. Med.* **359**, 789–799 (2008).
- Konig, J., Seithel, A., Gradhand, U. & Fromm, M.F. Pharmacogenomics of human OATP transporters. *Naunyn Schmiedeberg's Arch. Pharmacol.* **372**, 432–443 (2006).
- FDA approves Gleevec for leukemia treatment. *FDA Consum.* **35**, 6 (2001).
- Rowley, J.D. Letter: a new consistent chromosomal abnormality in chronic myelogenous leukaemia identified by quinacrine fluorescence and Giemsa staining. *Nature* **243**, 290–293 (1973).

22. Baselga, J. *et al.* Phase II study of weekly intravenous recombinant humanized anti-p185HER2 monoclonal antibody in patients with HER2/neu-overexpressing metastatic breast cancer. *J. Clin. Oncol.* **14**, 737–744 (1996).
23. Slamon, D.J. *et al.* Use of chemotherapy plus a monoclonal antibody against HER2 for metastatic breast cancer that overexpresses HER2. *N. Engl. J. Med.* **344**, 783–792 (2001).
24. Minna, J.D., Girard, L. & Xie, Y. Tumor mRNA expression profiles predict responses to chemotherapy. *J. Clin. Oncol.* **25**, 4329–4336 (2007).
25. Chang, J.C. *et al.* Gene expression profiling for the prediction of therapeutic response to docetaxel in patients with breast cancer. *Lancet* **362**, 362–369 (2003).
26. Butte, A.J., Tamayo, P., Slonim, D., Golub, T.R. & Kohane, I.S. Discovering functional relationships between RNA expression and chemotherapeutic susceptibility using relevance networks. *Proc. Natl. Acad. Sci. USA* **97**, 12182–12186 (2000).
27. Potti, A. *et al.* Genomic signatures to guide the use of chemotherapeutics. *Nat. Med.* **12**, 1294–1300 (2006).
28. Liu, C. *et al.* Genome-wide association scan identifies candidate polymorphisms associated with differential response to anti-TNF treatment in rheumatoid arthritis. *Mol. Med.* **14**, 575–581 (2008).
29. Byun, E. *et al.* Genome-wide pharmacogenomic analysis of the response to interferon beta therapy in multiple sclerosis. *Arch. Neurol.* **65**, 337–344 (2008).
30. Lavedan, C. *et al.* Association of the NPAS3 gene and five other loci with response to the antipsychotic iloperidone identified in a whole genome association study. *Mol. Psychiatry* (2008); e-pub ahead of print 3 June 2008.
31. Turner, S.T. *et al.* Genomic association analysis suggests chromosome 12 locus influencing antihypertensive response to thiazide diuretic. *Hypertension* **52**, 359–365 (2008).
32. Uhl, G.R. *et al.* Molecular genetics of successful smoking cessation: convergent genome-wide association study results. *Arch. Gen. Psychiatry* **65**, 683–693 (2008).
33. Galperin, M.Y. The Molecular Biology Database Collection: 2008 update. *Nucleic Acids Res.* **36**, D2–D4 (2008).
34. Danoff, T.M. *et al.* A Gilbert's syndrome UGT1A1 variant confers susceptibility to tranilast-induced hyperbilirubinemia. *Pharmacogenomics J.* **4**, 49–53 (2004).
35. Lenzer, J. FDA is incapable of protecting US "against another Vioxx". *BMJ* **329**, 1253 (2004).
36. Hammond, I.W., Gibbs, T.G., Seifert, H.A. & Rich, D.S. Database size and power to detect safety signals in pharmacovigilance. *Expert Opin. Drug Saf.* **6**, 713–721 (2007).
37. Parrish, F., Do, N., Bouhaddou, O. & Warnekar, P. Implementation of RxNorm as a terminology mediation standard for exchanging pharmacy medication between federal agencies. *AMIA Annu. Symp. Proc.*, 1057 (2006).
38. Seger, A.C., Jha, A.K. & Bates, D.W. Adverse drug event detection in a community hospital utilising computerised medication and laboratory data. *Drug Saf.* **30**, 817–824 (2007).
39. Gurwitz, J.H. *et al.* The incidence of adverse drug events in two large academic long-term care facilities. *Am. J. Med.* **118**, 251–258 (2005).
40. Gandhi, T.K. *et al.* Adverse drug events in ambulatory care. *N. Engl. J. Med.* **348**, 1556–1564 (2003).
41. Kuehn, B.M. FDA turns to electronic "sentinel" to flag prescription drug safety problems. *JAMA* **300**, 156–157 (2008).
42. Hernandez-Boussard, T. *et al.* The pharmacogenetics and pharmacogenomics knowledge base: accentuating the knowledge. *Nucleic Acids Res.* **36**, D913–D918 (2008).
43. Wishart, D.S. *et al.* DrugBank: a knowledgebase for drugs, drug actions and drug targets. *Nucleic Acids Res.* **36**, D901–D906 (2008).
44. Waters, M. *et al.* CEBS—Chemical Effects in Biological Systems: a public data repository integrating study design and toxicity data with microarray and proteomics data. *Nucleic Acids Res.* **36**, D892–D900 (2008).
45. Mattingly, C.J. The comparative toxicogenomics database: a cross-species resource for building chemical-gene interaction networks. *Toxicol. Sci.* **92**, 587–595 (2006).
46. Kanehisa, M. *et al.* KEGG for linking genomes to life and the environment. *Nucleic Acids Res.* **36**, D480–D484 (2008).
47. Chee, M. *et al.* Accessing genetic information with high-density DNA arrays. *Science* **274**, 610–614 (1996).
48. DeRisi, J. *et al.* Use of a cDNA microarray to analyse gene expression patterns in human cancer. *Nat. Genet.* **14**, 457–460 (1996).
49. Barrett, T. *et al.* NCBI GEO: mining tens of millions of expression profiles—database and tools update. *Nucleic Acids Res.* **35**, D760–D765 (2007).
50. Natsoulis, G. *et al.* The liver pharmacological and xenobiotic gene response repertoire. *Mol. Syst. Biol.* **4**, 175 (2008).
51. Lamb, J. *et al.* The Connectivity Map: using gene-expression signatures to connect small molecules, genes, and disease. *Science* **313**, 1929–1935 (2006).
52. Lin, Y.A., Chiang, A., Lin, R., Yao, P., Chen, R. & Butte, A.J. Methodologies for extracting functional pharmacogenomic experiments from international repository. *AMIA Ann. Symp. Proc.*, 463–467 (2007).
53. Weinstein, J.N. *et al.* An information-intensive approach to the molecular pharmacology of cancer. *Science* **275**, 343–349 (1997).
54. New treatments for colorectal cancer. *FDA Consum.* **38**, 17 (2004).
55. Cetuximab approved by FDA for treatment of head and neck squamous cell cancer. *Cancer Biol. Ther.* **5**, 340–342 (2006).
56. Ross, D.T. *et al.* Systematic variation in gene expression patterns in human cancer cell lines. *Nat. Genet.* **24**, 227–235 (2000).
57. Staunton, J.E. *et al.* Chemosensitivity prediction by transcriptional profiling. *Proc. Natl. Acad. Sci. USA* **98**, 10787–10792 (2001).
58. Sayers, E.W. *et al.* Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.* **37**, D5–D15 (2008).
59. Yildirim, M.A., Goh, K.I., Cusick, M.E., Barabasi, A.L. & Vidal, M. Drug-target network. *Nat. Biotechnol.* **25**, 1119–1126 (2007).
60. Huang, R.S., Duan, S., Kistner, E.O., Hartford, C.M. & Dolan, M.E. Genetic variants associated with carboplatin-induced cytotoxicity in cell lines derived from Africans. *Mol. Cancer Ther.* **7**, 3038–3046 (2008).
61. Huang, R.S. *et al.* Identification of genetic variants contributing to cisplatin-induced cytotoxicity by use of a genomewide approach. *Am. J. Hum. Genet.* **81**, 427–437 (2007).
62. Duan, S. *et al.* Mapping genes that contribute to daunorubicin-induced cytotoxicity. *Cancer Res.* **67**, 5425–5433 (2007).
63. Lee, J.K. *et al.* A strategy for predicting the chemosensitivity of human cancers and its application to drug discovery. *Proc. Natl. Acad. Sci. USA* **104**, 13086–13091 (2007).
64. Fliri, A.F., Loging, W.T., Thadeio, P.F. & Volkmann, R.A. Analysis of drug-induced effect patterns to link structure and side effects of medicines. *Nat. Chem. Biol.* **1**, 389–397 (2005).
65. Fliri, A.F., Loging, W.T., Thadeio, P.F. & Volkmann, R.A. Biological spectra analysis: linking biological activity profiles to molecular structure. *Proc. Natl. Acad. Sci. USA* **102**, 261–266 (2005).
66. Campillos, M., Kuhn, M., Gavin, A.C., Jensen, L.J. & Bork, P. Drug target identification using side-effect similarity. *Science* **321**, 263–266 (2008).
67. National Library of Medicine. Unified Medical Language System. About the UMLS Resources. <[http://www.nlm.nih.gov/research/umls/about\\_umls.html](http://www.nlm.nih.gov/research/umls/about_umls.html)>.
68. Bodenreider, O. The Unified Medical Language System (UMLS): integrating biomedical terminology. *Nucleic Acids Res.* **32**, D267–D270 (2004).
69. Geisler, S.A. & Olshan, A.F. GSTM1, GSTT1, and the risk of squamous cell carcinoma of the head and neck: a mini-HuGE review. *Am. J. Epidemiol.* **154**, 95–105 (2001).
70. Tokizane, T. *et al.* Cytochrome P450 1B1 is overexpressed and regulated by hypomethylation in prostate cancer. *Clin. Cancer Res.* **11**, 5793–5801 (2005).
71. Chung, I. *et al.* Epigenetic silencing of CYP24 in tumor-derived endothelial cells contributes to selective growth inhibition by calcitriol. *J. Biol. Chem.* **282**, 8704–8714 (2007).
72. Ruvkun, G. The perfect storm of tiny RNAs. *Nat. Med.* **14**, 1041–1045 (2008).
73. Tsuchiya, Y., Nakajima, M., Takagi, S., Taniya, T. & Yokoi, T. MicroRNA regulates the expression of human cytochrome P450 1B1. *Cancer Res.* **66**, 9090–9098 (2006).
74. Mishra, P.J., Humeniuk, R., Mishra, P.J., Longo-Sorbello, G.S., Banerjee, D. & Bertino, J.R. A miR-24 microRNA binding-site polymorphism in dihydrofolate reductase gene leads to methotrexate resistance. *Proc. Natl. Acad. Sci. USA* **104**, 13513–13518 (2007).
75. Lijinsky, W., Reuber, M.D. & Blackwell, B.N. Liver tumors induced in rats by oral administration of the antihistaminic methapyrilene hydrochloride. *Science* **209**, 817–819 (1980).
76. Steiner, S. *et al.* Cyclosporine A decreases the protein level of the calcium-binding protein calbindin-D 28kDa in rat kidney. *Biochem. Pharmacol.* **51**, 253–258 (1996).
77. Varela, M.C. *et al.* Cyclosporine A-induced decrease in calbindin-D 28 kDa in rat kidney but not in cerebral cortex and cerebellum. *Biochem. Pharmacol.* **55**, 2043–2046 (1998).
78. Owen, R.P., Altman, R.B. & Klein, T.E. PharmGKB and the International Warfarin Pharmacogenetics Consortium: the changing role for pharmacogenomic databases and single-drug pharmacogenetics. *Hum. Mutat.* **29**, 456–460 (2008).
79. Molokhia, M. & McKeigue, P. EUDRAGENE: European collaboration to establish a case-control DNA collection for studying the genetic basis of adverse drug reactions. *Pharmacogenomics* **7**, 633–638 (2006).