Bioinformatics Advance Access published May 19, 2011

Bioinformatics Challenges for Personalized Medicine

Guy Haskin Fernald[1], Emidio Capriotti[2,3], Roxana Daneshjou[4], Konrad J. Karczewski[1,5], and Russ B. Altman[2,5,*]

[1]Biomedical Informatics Training Program, Stanford University School of Medicine, Stanford, CA, USA.

[2]Department of Bioengineering, Stanford University, Stanford, CA, USA.

[3]Department of Mathematics and Computer Sciences, University of Balearic Islands, Palma de Mallorca, Spain.

[4]Stanford University School of Medicine, Stanford, CA, USA.

[5]Department of Genetics, Stanford University, Stanford, CA, USA.

Associate Editor: Dr. Jonathan Wren

ABSTRACT

Motivation: Widespread availability of low-cost, full genome sequencing will introduce new challenges for bioinformatics.

Results: This review outlines recent developments in sequencing technologies and genome analysis methods for application in personalized medicine. New methods are needed in four areas to realize the potential of personalized medicine: 1) processing large-scale robust genomic data; 2) interpreting the functional effect and the impact of genomic variation; 3) integrating systems data to relate complex genetic interactions with phenotypes; and 4) translating these discoveries into medical practice.

Contact: russ.altman@stanford.edu

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 INTRODUCTION

We are on the verge of the genomic era: doctors and patients will have access to genetic data to customize medical treatment. Consumers can already get 500,000 to 1,000,000 variant markers analyzed with associated trait information (Hindorff, et al., 2009), and soon full genome sequencing will cost less than $1000 (Drmanac, et al., 2010). One group has performed a complete clinical assessment of a patient using a personal genome (Ashley, et al., 2010), and the 1000 Genomes Project is sequencing 1000 individuals (1000 Genomes Project Consortium, et al., 2010). In the coming years, the bioinformatics world will be inundated with individual genomic data. This flood of data introduces significant challenges that the bioinformatics community needs to address. This review outlines the developments that led to these challenges, the previous work that can address them, and the need for new methods to address them. The challenges fall into four main areas: 1) processing large-scale robust genomic data; 2) interpreting the functional impacts of genomic variation; 3) integrating data to relate complex interactions with phenotypes; and 4) translating these discoveries into medical practices.

2 THE PROMISE OF PERSONALIZED MEDICINE

In the last decade, molecular science has made many advances to benefit medicine, including the Human Genome project, International HapMap project, and Genome Wide Association Studies (GWAS) (International HapMap Consortium, 2005). Single Nucleotide Polymorphisms (SNPs) are now recognized as the main cause of human genetic variability and are already a valuable resource for mapping complex genetic traits (Collins, et al., 1997). Thousands of DNA variants have been identified that are associated with diseases and traits (Hindorff, et al., 2009). By combining these genetic associations with phenotypes and drug response, personalized medicine will tailor treatments to the patients' specific genotype (see Fig 1). Although whole genome sequences are not used in regular practice today (McGuire and Burke, 2008), there are already many examples of personalized medicine in current practice. Chemotherapy medications such as trastuzumab and imatinib target specific cancers (Gambacorti-Passerini, 2008; Hudis, 2007), a targeted pharmacogenetic dosing algorithm is used for warfarin (International Warfarin Pharmacogenetics Consortium, et al., 2009; Sagreiya, et al., 2010), and the incidence of adverse events is reduced by checking for susceptible genotypes for drugs like abacavir, carbamazepine, and clozapine (Dettling, et al., 2007; Ferrell and McLeod, 2008; Hetherington, et al., 2002).

Despite all of these advances, many challenges need to be addressed to make personalized medicine a reality. Today, a patient's genetics are consulted only for a few diagnoses and treatment plans and only in certain medical centers. Even if doctors had access to their patients' genomes today, only a small percentage the genome could even be used (Yngvadottir, et al., 2009). Many of the annotations come from association studies, which tend to identify variants with small effect sizes and have limited applications for healthcare (Moore, et al., 2010). By addressing the challenges outlined in this review bioinformatics will create the tools to tailor medical care to each individual genome, rather than rely on blanket therapies (Ginsburg and Willard, 2009).

*To whom correspondence should be addressed.

3 CHALLENGE 1: PROCESSING LARGE-SCALE ROBUST GENOMIC DATA

Sequencing technologies are becoming affordable and are replacing the microarray based genotyping methods, which were limited to interrogating regions of known variation (Ng, et al., 2010). Now a whole genome or a few dozen exomes can be sequenced in less than two weeks with an error rate of approximately 1 error per 100 kilobases (Drmanac, et al., 2010). Even such low error rates can lead to a significant number of errors; a 3-gigabase human genome would have approximately 30,000 erroneous variant calls.

The error rate from these technologies is a source of significant challenges in applications, including discovering novel variants. Each newly sequenced genome is expected to have between 100,000-300,000 previously undiscovered SNPs and less than 1,000 somatic mutations per generation (1000 Genomes Project Consortium, et al., 2010). The number of expected mutations may decrease as new genomes are sequenced, however, such a high number of errors turns variant discovery into a "needle in a haystack" problem. Whenever a novel variant is identified it will still have to be verified due to this false positive rate. In addition, other classes of variation, such as short insertion-deletion variants (indels), as well as copy number variants (CNVs) and structural variants (SVs), are even more difficult to detect using high-throughput sequencing. New algorithms for calling indels, CNVs, and SVs from read data will be crucial in detecting these types of variations for clinical applications.

Even high-quality sequence reads must be placed into their genomic context to identify variants, which is an active area of research since, for example, different mapping and alignment algorithms often yield different results. Because *de novo* assembly (Shendure and Ji, 2008) is slow and complicated by repetitive elements, sequences are usually mapped to a genomic reference sequence instead. Algorithms such as BLAST (Altschul, et al., 1990) or Smith-Waterman (Smith TF, 1981) have been traditionally used, but their execution speed depends on genome size. While individual queries may only take seconds per CPU, aligning 100 million of them would require more than 3 CPU years.

As a result, new algorithms are being developed to address this problem. BLAT is similar to standard sequence alignment, but also incorporates an indexed version of the genome instead of linear search (Kent, 2002). Many packages like BLAT have been optimized for the alignment of short reads by using hashing, prefix and suffix trees, or other heuristics (Li and Homer, 2010). BWA, used for the 1000 Genomes Project, is highly accurate with < 0.1% errors for simulated data and can map ~7 Gb of short reads per CPU day (Li and Durbin, 2009; Li and Homer, 2010). To achieve the standard 30X coverage would still require 13 CPU days and so is ideally performed on a cluster or by using a cloud computing environment (Dudley and Butte, 2010), which can be used for efficient computational analysis of secure clinical data.

A remaining challenge for short read assemblers is reference sequence bias: reads that more closely resemble the reference sequence are more likely to successfully map as compared with reads that contain valid mismatches. Proper care must be taken to avoid

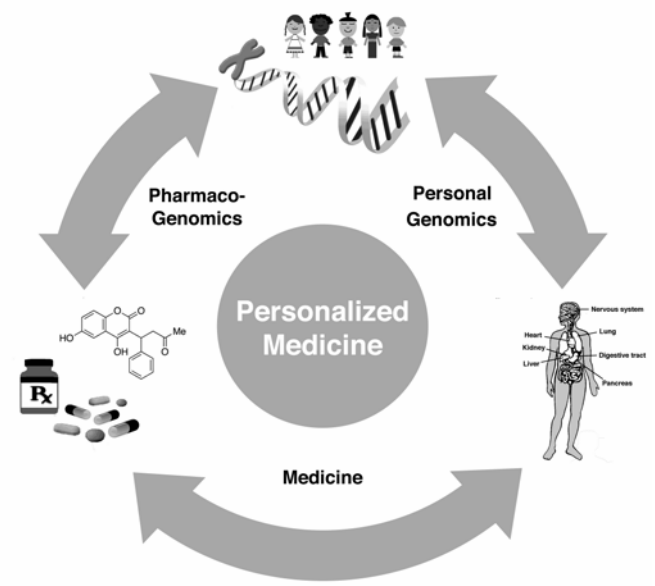

Fig 1. Personalized medicine. Personal genomics connect genotype to phenotype and provide insight into disease. Pharmacogenomics connect connects genotype to patient specific treatment. Traditional medicine defines the pathologic states and clinical observations to evaluate and adjust treatments.

errors in these alignments, and is discussed in a recent review (Pool, et al., 2010). There is an inherent trade-off in allowing mismatches: the program must allow for mismatches without resulting in false alignments. Reference sequence bias is important when making heterozygous SNP calls and when analyzing allele-specific expression using RNA-Seq data (Degner, et al., 2009). The problem is exacerbated with longer reads: allowing for one mismatch per read is acceptable for 35 base pair reads, but insufficient for 100 base pair reads.

When the diploid sequence is known, reference sequence bias can be avoided by mapping the reads to both strands, as can be done when mapping RNA-Seq reads to a sequenced genome. An alternative approach is to use ambiguous base codes to avoid the requirement of storing redundant sequences, such as with MOSAIK, developed by the Marth Lab (Michael Stromberg, Boston University). Using this approach, a C/T SNP can be represented as Y. This representation increases the storage requirements: because the genome is often stored in a hashed data structure, the number of keys and mappings increases to accommodate the new codes.

Another challenge is developing new methods for novel SNP discovery: while the calling of common variants can be aided by their presence in a database such as dbSNP, accurate detection of rare and novel variants will require increased confidence in the SNP call. *De novo* alignment methods require too much computation time to be feasible and reference alignment methods are biased. The challenge is to develop new algorithms that are computationally tractable and still avoid reference sequence bias.

Finally, there is a pressing need to improve quality control metrics. We can judge mapping and SNP call qualities by the ratio of tran-

sition (purine/purine or pyrimidine/pyrimidine) substitutions to transversion (purine/pyrimidine) substitutions. These ratios were established during previous sequencing efforts and we expect to see similar ratios (~2-2.1) for newly human genomes (Zhang and Gerstein, 2003). When working with genomes from families we can estimate errors with the Mendelian inheritance error (MIE) rate: impossible combinations of inheritance most likely represent errors (Ewen, et al., 2000). Transition/transversion ratio and MIE metrics are useful for measuring the quality of a dataset and are used by most large projects, such as the 1000 Genomes project (1000 Genomes Project Consortium, et al., 2010). At the individual SNP level, we must rely on relative quality scores, so in order to confidently identify novel variants we must be verify them with an independent method. Variants can be validated with targeted resequencing or genotyping arrays. Alternatively, whole genome resequencing by an orthogonal sequencing platform can be performed, but is expensive and time consuming.

4 CHALLENGE 2: INTERPRETATION OF THE FUNCTIONAL EFFECT AND THE IMPACT OF GENOMIC VARIATION

After genomic data has been processed, the functional effect and the impact of the genetic variations must be analyzed. Genome-wide association studies (GWAS) have been used to assess the statistical associations of SNPs with many important common diseases (WTCC Consortium, 2007). These methods are providing new insights, but only a limited number of variants have been characterized, and understanding the functional relationship between associated variants and phenotypic traits has been difficult (Frazer, et al., 2009).

In the strictest definition, a SNP is a single nucleotide variant where the allele frequency in the human population is higher then 1%. In this review, we use the term SNP in a broader sense to also include rare variants that occur in a smaller fraction of the population. Important issues for predicting the impact of SNPs are data management, retrieval, and quality control. During the last few years, the number of known SNPs has increased at an exponential rate (Fig 2). The dbSNP database (Sherry, et al., 2001) is the most comprehensive repository of SNPs data from different organisms. At the time of writing this review, the database contains about 20 million validated human SNPs (Build 132, September 2010). The Human Gene Mutation Database (HGMD) is a comprehensive collection of germline mutations in genes that are associated with human inherited diseases. The free version for academic and non-profit users contains more than 76,000 mutations from about 2,900 genes. The SwissVar is a database of manually annotated missense SNPs (mSNPs) and contains 56,000 mSNPs from more than 11,000 genes.

Another important resource for SNP data is the Online Mendelian Inheritance in Man (OMIM) database (Amberger, et al., 2009) of human SNPs and their associations with Mendelian disorders. The PharmGKB database contains manually curated associations between genes and drugs and a catalog of genetic variations with known impact on drug response, including more than 40 very important pharmacogenes (VIPs) and over 3,400 annotated drug-response variants. The Catalogue of Somatic Mutations in Cancer (COSMIC) at the Sanger Institute stores ~25,000 unique mutations somatic mutation data related to human cancer extracted from the literature. A selection of the most significant SNP data sources is reported in supplemental table 1.

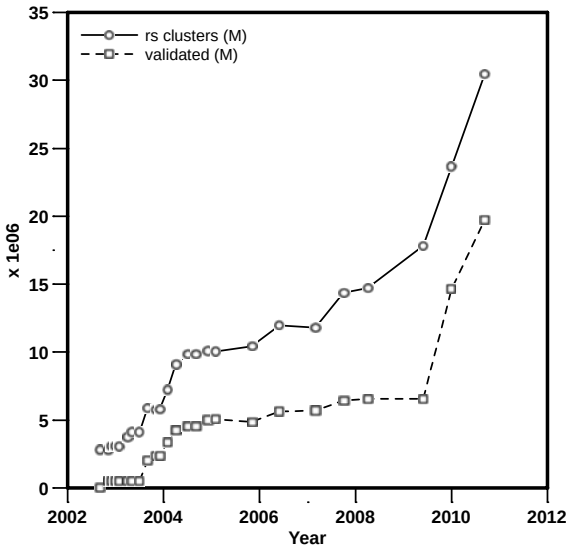

Fig 2. Number of validated human SNPs in dbSNP over time.

In the last few years, several computational methods have been developed to predict deleterious missense SNPs. (Karchin, 2009; Mooney, 2005; Tavtigian, et al., 2008). These methods have used different approaches such as empirical rules (Ng and Henikoff, 2003; Ramensky, et al., 2002), Hidden Markov Models (Thomas and Kejariwal, 2004), Neural Networks (Bromberg, et al., 2008; Ferrer-Costa, et al., 2005), Decision Trees (Dobson, et al., 2006; Krishnan and Westhead, 2003), Random Forests (Bao and Cui, 2005; Carter, et al., 2009; Kaminker, et al., 2007; Li, et al., 2009; Wainreb, et al., 2010) and Support Vector Machines (Calabrese, et al., 2009; Capriotti, et al., 2008; Capriotti, et al., 2006; Karchin, et al., 2005; Yue and Moult, 2006).

The prediction algorithms input features generally include amino acid sequence, protein structure, and evolutionary information. The amino acid sequence features rely on the physico-chemical properties of the mutated residues such as hydrophobicity, charge, polarity, and bulkiness. Protein structural information describes the structural environment of the mutation and has been successfully used to predict the protein stability change upon mutation (Capriotti, et al., 2004; Capriotti, et al., 2005; Schymkowitz, et al., 2005; Zhou and Zhou, 2002). Some of the most important features for the prediction of the impact of missense SNPs are derived from evolutionary analysis: critical amino acids are often conserved in protein families and so changes at conserved positions tend to be deleterious.

New algorithms that include knowledge-based information are being developed (Alexiou, et al., 2009; Calabrese, et al., 2009; Kaminker, et al., 2007). Methods based on evolutionary information for the prediction of mSNPs include SIFT (Ng and Henikoff, 2003) and PolyPhen (Ramensky, et al., 2002). SIFT scores the

normalized probabilities for all possible substitutions using a multiple sequence alignment between homolog proteins, and PolyPhen predicts the impact of mSNPs using different sequence-based features and a Position Specific Independent Counts (PSIC) matrix from multiple sequence alignment. The PANTHER algorithm (Thomas, et al., 2003) uses a library of protein family HMM models to predict deleterious mutations. Recent work shows that three-dimensional structural features improve the prediction of disease-related mSNPs (Bao and Cui, 2005; Karchin, et al., 2005; Yue and Moult, 2006). Knowledge-based information has been used to increase the accuracy of prediction algorithms to over 80%. For example, SNPs&GO (Calabrese, et al., 2009) is an algorithm based on functional information that takes in input log-odd scores calculated using Gene Ontology (GO) annotation terms. MutPred (Li, et al., 2009) evaluates the probabilities of gain or loss of structure and function upon mutations and predicts their impact using a Random Forest based approach. Selected methods for the prediction of deleterious mSNPs are listed in supplemental table 2 and more details about mSNP predictors have been recently reviewed (Cline and Karchin, 2011; Thusberg, et al., 2011)

Prediction methods do not provide any information about the pathophysiology of the diseases and so experimental tests are required to validate genetic predictions. Laboratory validation is expensive and time consuming and so there is a need for fast and accurate methods for gene prioritization. Currently, the most effective strategy uses the concept of similarity to genes that are linked to the biological process of interest (guilt-by-association). The input data for the available gene prioritization methods are derived from functional annotation, protein-protein interaction data, biological pathways, and literature.

The SUSPECT algorithm prioritizes genes by comparing sequence features, gene expression data, Interpro domains, and functional terms (Adie, et al., 2006). ToppGene combines mouse phenotype data with human gene annotations and literature. MedSim uses functional information from human disease genes or proteins and their orthologs in mouse modeles (Schlicker, et al., 2010). Endeavour is trained on genes involved in a known biological process and ranks candidate genes after considering several genomic data sources (Tranchevent, et al., 2008). G2D prioritization strategy is based on a combination of data mining on biomedical databases and sequence features (Perez-Iratxeta, et al., 2005). PolySearch analyzes biomedical databases to build relationships between diseases, genes, mutations, drugs, pathways, tissues, organs and metabolites in humans (Cheng, et al., 2008). MimMiner ranks phenotypes using text mining by comparing the human phenome and disease phenotypes (van Driel, et al., 2006). PhenoPred detects gene–disease associations using the human protein–protein interaction network, known gene–disease associations, protein sequences, and protein functional information at the molecular level (Radivojac, et al., 2008). GeneMANIA (Andersen, et al., 2008) generates hypotheses about gene function, analyzing gene lists and prioritizing genes for functional assays. The method takes in input genes from six organisms and analyzes them using information from different general and organism-specific functional genomics data sets. For more details about gene prioritizing tools, a recently published review (Tranchevent, et al., 2010), and the Gene Prioritization Portal provide comprehensive descriptions of available predictors.

The methods for the analysis of SNPs are mainly limited to the prediction of the impact of missense SNPs. New methods are needed to evaluate the impact of insertion, deletion, and synonymous SNPs. In addition, there is a need to detect functional regions in the genome so that the effect of intronic SNPs can be analyzed, such as those in promoter regions and splicing sites. For non-coding regions, conservation across species is more difficult to detect. Fortunately, with the fast growth of functionally annotated genomes our ability to predict the impact of non-coding variants will increase. For example, SNPs occurring in transcriptional motifs can affect transcription factor binding, which suggests functional consequences for variants in regulatory regions (Kasowski, et al., 2010). Recently a method to identify possible genetic variations in regulatory regions (is-rSNP) has been developed (Andersen, et al., 2008). Is-rSNP combines phylogenetic information and transcription factor binding site prediction to identify variation in candidate cis-regulatory elements. The detection of variants affecting splicing site is also an important task. The Skippy algorithm (Woolfe, et al., 2010) analyzes the genomic region surrounding the variant to predict severe effects on gene function through disruption of splicing. A more exhaustive description of the methods for the prediction of deleterious variants in non-coding has been recently published (Cline and Karchin, 2011).

Last year, the first edition of the Critical Assessment of Genome Interpretation (CAGI) was organized to assess the available methods for predicting phenotypic impact of genomic variation and to stimulate future research. In the first year of CAGI (http://genomeinterpretation.org/) the organizers provided six different sets of data for six different tasks. The majority of the participating groups submitted predictions for just two classes of experiments related to the detection of disease-related and function modifying variants. A few groups submitted predictions for the other categories: evaluation of risky SNPs from GWAS studies, interpretation of the Personal Genome Project data, prediction of mutations to P53 function, and the response of breast cancer cell lines to different drugs. Several available predictors performed well for disease and functional predictions and there were promising results in the other categories. In the future, competitions such as CAGI will improve the quality of the available prediction methods and will renew the challenge for the understanding of genomic variation data.

5 CHALLENGE 3: INTEGRATING SYSTEMS AND DATA TO CAPTURE COMPLEXITY

Given the complex phenotypes involved in personalized medicine, the simple “one-SNP, one-phenotype” approach taken by most studies is insufficient. Most medically relevant phenotypes are thought to be the result of gene-gene and gene-environment interactions (Manolio, et al., 2009). For example, drug response often depends on multiple pharmacokinetic and pharmacodynamic interactions, which form a robust and tolerant system with highly polymorphic enzymes and many interaction partners (Wilke, et al., 2005). As a result of this complexity, a drug response phenotype

of interest is likely to depend on many genes and environmental factors.

Basic GWAS approaches for pharmacogenomics have had some success, including studies of warfarin that have linked the majority of variation in response to just two genes, CYP2C9 and VKORC1 (Limdi and Veenstra, 2008). These and other studies of warfarin have even led to an improved dosing algorithm with improvements over the traditional clinical algorithm (International Warfarin Pharmacogenetics Consortium, et al., 2009). Clopidogrel response has similarly been associated with variants of CYP2C19 (Shuldiner, et al., 2009).

Despite this success, there is debate over whether or not traditional techniques will be successful for pharmacogenomics. There is concern that pharmacogenomics GWAS themselves are susceptible to many limitations: insufficient sample size, selection biases for genetic variants, environmental interactions that may affect the outcome measures, and multiple gene-gene interactions which may underlie unexplained effects (Motsinger-Reif, et al., 2010). These limitations become particularly difficult when researching rare events such as the pharmacogentics of adverse events.

The methods for GWAS are designed for single marker associations and are known to have limitations in explaining the heritability of disease (Manolio, et al., 2009). It is unlikely that these same methods will do any better with pharmacogenetics. In fact, if these methods are parameterized for the multiple-marker associations necessary for pharmacogenetics then they will suffer from the "curse of dimensionality" and lose a significant amount of statistical power (Bellman and Kalaba, 1959). For example, to evaluate all combinations of 2 SNPs for 1 million SNPs in a genome requires examining nearly 500 billion possibilities. The challenge for bioinformatics is to address this complexity by developing methods that combine multiple data sources without losing statistical power.

Several groups have already tried to deal with this kind of complexity in GWAS for disease (Motsinger, et al., 2007). Exhaustive search (Storey, et al., 2005) and forward search (Consortium, et al., 2007) have both been applied, however, the former can still lose statistical power and the later may miss some associations. Model selection methods have been successful with disease and trait GWAS studies by using selection techniques to choose multifactorial models that balance the false positive rate, statistical power, and computational requirements of the search (Lee, et al., 2008; Wray, et al., 2007; Wu and Zhao, 2009).

Given the size of the genomic data sets, dimensionality reduction methods such as principal components analysis, information gain, and Multifactor Dimensionality Reduction will be essential to make complexity algorithms tractable (Hahn, et al., 2003; Statnikov, et al., 2005; Yeung and Ruzzo, 2001). Some of these methods have proven successful for finding multi-locus associations with diseases such as hypertension and familial amyloid polyneuropathy type I (Soares, et al., 2005; Williams, et al., 2004). Many more feature selection techniques for bioinformatics are classified and discussed in a recent review (Saeys, et al., 2007). These methods can be very effective when dealing with large datasets, however they do not integrate with any external knowledge sources or inform the biology behind the interactions.

Systems biology and network approaches address to the problem of complexity by integrating molecular data at multiple levels of biology including genomes, transcriptomes, metabolomes, proteomes, and functional and regulatory networks (Kohl, et al., 2010). We can view a disease or a drug response phenotype as a global perturbation of networks from their stable state (Auffray, et al., 2009). This approach integrates biological knowledge from networks to make inferences about what genes or combinations of genes and other biological markers are more likely to be associated.

Combining disparate data sources can result in novel associations and provide insight into gene-gene and gene-environment interactions. One group created a disease-gene network by combining the diseases and associated genes available in OMIM (Goh, et al., 2007). Analyzing this network showed that disease genes are often non-essential and not necessarily hub-genes. The same group created a Drug-Target network and integrated that network with a protein-protein interaction (PPI) network. The network shows that similar drugs cluster together, palliative and etiological drugs show different topologies, and newer and experimental drugs tend towards polypharmacology (Yildirim, et al., 2007). A global mapping of pharmacalogical space can be made using chemical structure, disease indication, and protein sequence and can be used to make predictions of polypharmacology (Paolini, et al., 2006). Another suggestion is to integrate epigenetic information to further our understanding of drug phenotypes (Zhang and Dolan, 2009).

Pathway and gene set methods can also be applied to GWAS, where a set of genes is identified that is suspected to be associated. These methods are similar to Gene Set Enrichment Analysis (GSEA) for microarray expression data (Subramanian, et al., 2005). Usually a standard statistical test is used to determine if a set of genes is associated (Chasman, 2008; Wang and Li, 2007; Yu, et al., 2009), but other more specialized metrics have been created. The SNP Ratio Test compares the number of SNPs in a pathway to permuted sets and the Prioritizing Risk Pathways method, combines pathway and genetic data into a single metric (Chen, et al., 2009; O'Dushlaine, et al., 2009).

Many groups hypothesize that the integrative approach of systems biology will successfully link genomic measurements with clinical applications (Atkinson and Lyster, 2010; Berg, et al., 2010; Hopkins, 2007). Indeed, one group has integrated chemical similarity metrics, pharmacogenomic interactions, and protein-protein interaction to predictive method for pharmacogenes (Hansen, et al., 2009). Another group has used similarity of drug ligand sets to predict and validate novel "off-target" interactions (Keiser, et al., 2007).

These systems approaches are encouraging, but bioinformaticians need to be careful of a few pitfalls as they proceed. Methods need to be based on high quality data to avoid the "garbage-in, garbage-out" phenomenon, especially when one incorrect assumption can propagate through multiple data source and magnify the error. For example, transferring annotations based on similarity works some-

times, but could easily associate a paralog with an incorrect function. Chemical similarity poses the same risk; two similar molecules may behave very differently biochemically. Finally, assumptions must also be examined carefully, for example, a method that relates gene expression with drug targets must bear in mind than most drugs bind proteins, not DNA or RNA.

6 CHALLENGE 4: MAKING IT ALL CLINICALLY RELEVANT

The ultimate challenge for this research is to apply the results for improved patient care. Much of this research has yet to be translated to the clinic. In fact, many physicians are unprepared to incorporate personal genetic testing into their practice and it is unclear how to best apply research results to improve patient care (McGuire and Burke, 2008). One of the areas where bioinformatics can have the greatest clinical impact is in pharmacogenomics.

Most pharmaceutical development addresses medical problems with a "one drug fits all" approach (Ehrlich, 1906). Genetic variation has been shown to influence drug selection, dosing, and adverse events (Giacomini, et al., 2007), and the therapeutic benefits of taking a genetically tailored approach to drug development is now recognized (Foot, et al., 2010; Roses, 2004). One study found that a hypothetical pharmacogenetically-driven clinical trial of the anti-coagulant warfarin could save up to 60% of the cost and reduce possible adverse events (Ohashi and Tanaka, 2010). There are already many examples of drugs which have retrospectively been found to have strong pharmacogenomic interactions, including thiopurines for cancer (Weinshilboum, 2001) and the anti-coagulant clopiogrel (Shuldiner, et al., 2009).

A trial for using rosiglitazone, an approved Type II diabetes drug, for Alzheimer's disease is an early example of prospective application of pharmacogenomics. The hypothesis was that ApoE4 non-carriers would have a better response than ApoE4 carriers. The initial Phase II pharmacogenetic-based results appeared to show that non-ApoE4 carriers showed improvement over placebo (Roses, 2009). A later study of ApoE4-stratified patients showed that no significant benefits, however, the idea of prospective gene-based stratification for drug trials still holds future promise (Gold, et al., 2010). Prospective gene-stratification hypotheses need to be generated for future trials and will require new bioinformatics methods (Roses, 2009). Since new drugs will not have any known gene interactions, tools for predicting drug-target or drug-gene interactions will be essential (Hansen, et al., 2009; Keiser, et al., 2009).

Pharmacogenomics has already been successful in improving drug prescription and dosing. Most prescriptions are written with a "one dose fits all" approach with adjustments based on gender, weight, liver and kidney functions, or allergies. Some drugs have more laborious dosing calculations such as the anti-coagulant warfarin (Gage and Lesko, 2008; Wysowski, et al., 2007). Warfarin dosing is traditionally determined by a time-intensive "guess and test" method, until the coagulation tests stabilize. Pharmacogenomics identified several SNPs affecting dosing, including-CYP2C9 and VKORC1 (Higashi, et al., 2002; Rieder, et al., 2005; Rost, et al., 2004). Similar studies have been applied to clopidogrel, tramadol, anti-psychotics, and many other drugs (Wilffert, et al., 2010). Ultimately, pharmacogenomic prescription and dosing algorithms need to be accessible to physicians, like the new warfarin dosing algorithms from the International Warfarin Pharmacogenomic Consortium (IWPC) (International Warfarin Pharmacogenetics Consortium, et al., 2009). Moreover, the current state of medical practice needs to be updated to include routine pharmacogenetic testing, educating and training physicians in personalized medicine, and further clinical trials to prove the efficacy of pharmacogenetic based prescriptions.

Bioinformatics also translates discoveries to the clinic by disseminating discoveries through curated, searchable databases like PharmGKB, dbGaP, PacDB, and the FDA AERS (Gamazon, et al., 2010; Mailman, et al., 2007; Thorn, et al., 2010). A major bottleneck for these databases is manual curation of the data. Biologically and medically focused text mining algorithms can speed the collection of this structured data, such as methods that use sentence syntax and natural language processing to derive drug-gene and gene-gene interactions from scientific literature (Coulet, et al., 2010; Garten, et al., 2010). These databases and methods need to be developed and used carefully. All of these data sources are susceptible to errors and so validation of data is essential, especially before the information is applied in the clinic.

Finally, there are challenges and opportunities for bioinformatics to integrate with the electronic medical record (EMR) (Busis, 2010). For example, the BioBank system at Vanderbilt links patient DNA with a de-identified EMRs to provide a rich research database for additional translational research in disease-gene and drug-gene associations (Denny, et al., 2010; Roden, et al., 2008). Some health care companies and HMOs have also begun to collect genetic information from their patients. In order to even implement such genome-based systems, the medical infrastructure will have to shift from paper to electronic medical records, in order to be compatible with bioinformatics portals for data delivery and interpretation. Ultimately, bioinformatics needs to develop methods that interrogate the genome in the clinic and allow physicians to use personalized medicine in their daily practice.

ACKNOWLEDGEMENTS

EC would like to acknowledge Dr Laura Kerov-Ghiglianovich who helped to draw the figure 1.

Funding: GHF and KJK are supported by training grant NIH LM007033. RD is supported by Stanford Medical Scholars. EC is supported by the Marie Curie International Outgoing Fellowship program (PIOF-GA-2009-237225). RBA is supported by LM05652 and the NIH/NIGMS Pharmacogenetics Research Network and Database and the PharmGKB resource (NIH U01GM61374).

REFERENCES

1000 Genomes Project Consortium, *et al.* (2010) A map of human genome variation from population-scale sequencing, *Nature*, **467**, 1061-1073.

Adie, E.A., *et al.* (2006) SUSPECTS: enabling fast and effective prioritization of positional candidates, *Bioinformatics*, **22**, 773-774.

Alexiou, P., *et al.* (2009) Lost in translation: an assessment and perspective for computational microRNA target identification. *Bioinformatics*. pp. 3049-3055.

Altschul, S.F., *et al.* (1990) Basic local alignment search tool, *J Mol Biol*, **215**, 403-410.

Amberger, J., *et al.* (2009) McKusick's Online Mendelian Inheritance in Man (OMIM), *Nucleic Acids Res*, **37**, D793-796.

Andersen, M.C., *et al.* (2008) In silico detection of sequence variations modifying transcriptional regulation, *PLoS Comput Biol*, **4**, e5.

Ashley, E.A., *et al.* (2010) Clinical assessment incorporating a personal genome. *Lancet*. pp. 1525-1535.

Atkinson, A. and Lyster, P. (2010) Systems Clinical Pharmacology. *Clinical Pharmacology & Therapeutics*. pp. 3-6.

Auffray, C., Chen, Z. and Hood, L. (2009) Systems medicine: the future of medical genomics and healthcare. *Genome*.

Bao, L. and Cui, Y. (2005) Prediction of the phenotypic effects of non-synonymous single nucleotide polymorphisms using structural and evolutionary information, *Bioinformatics*, **21**, 2185-2190.

Bellman, R. and Kalaba, R. (1959) A MATHEMATICAL THEORY OF ADAPTIVE CONTROL PROCESSES. *Proc Natl Acad Sci USA*. pp. 1288-1290.

Berg, J., Rogers, M. and Lyster, P. (2010) Systems Biology and Pharmacology. *Clinical Pharmacology & Therapeutics*. pp. 17-19.

Bromberg, Y., Yachdav, G. and Rost, B. (2008) SNAP predicts effect of mutations on protein function, *Bioinformatics*, **24**, 2397-2398.

Busis, N.A. (2010) How can I choose the best electronic health record system for my practice?, *Neurology*, **75**, S60-64.

Calabrese, R., *et al.* (2009) Functional annotations improve the predictive score of human disease-related mutations in proteins, *Hum Mutat*, **30**, 1237-1244.

Capriotti, E., *et al.* (2008) Use of estimated evolutionary strength at the codon level improves the prediction of disease-related protein mutations in humans, *Hum Mutat*, **29**, 198-204.

Capriotti, E., Calabrese, R. and Casadio, R. (2006) Predicting the insurgence of human genetic diseases associated to single point protein mutations with support vector machines and evolutionary information, *Bioinformatics*, **22**, 2729-2734.

Capriotti, E., Fariselli, P. and Casadio, R. (2004) A neural-network-based method for predicting protein stability changes upon single point mutations, *Bioinformatics*, **20 Suppl 1**, I63-I68.

Capriotti, E., Fariselli, P. and Casadio, R. (2005) I-Mutant2.0: predicting stability changes upon mutation from the protein sequence or structure, *Nucleic Acids Res*, **33**, W306-310.

Carter, H., *et al.* (2009) Cancer-specific high-throughput annotation of somatic mutations: computational prediction of driver missense mutations, *Cancer Res*, **69**, 6660-6667.

Chasman, D. (2008) On the utility of gene set methods in genomewide association studies of quantitative traits. *Genet Epidemiol*.

Chen, L., *et al.* (2009) Prioritizing risk pathways: a novel association approach to searching for disease pathways fusing SNPs and pathways. ….

Cheng, D., *et al.* (2008) PolySearch: a web-based text mining system for extracting relationships between human diseases, genes, mutations, drugs and metabolites, *Nucleic Acids Res*, **36**, W399-405.

Cline, M.S. and Karchin, R. (2011) Using bioinformatics to predict the functional impact of SNVs, *Bioinformatics*, **27**, 441-448.

Collins, F.S., Guyer, M.S. and Charkravarti, A. (1997) Variations on a theme: cataloging human DNA sequence variation, *Science*, **278**, 1580-1581.

Consortium, I.H., *et al.* (2007) A second generation human haplotype map of over 3.1 million SNPs, *Nature*, **449**, 851-861.

Coulet, A., *et al.* (2010) Using text to build semantic networks for pharmacogenomics. *J Biomed Inform*. pp. 1009-1019.

Degner, J.F., *et al.* (2009) Effect of read-mapping biases on detecting allele-specific expression from RNA-sequencing data, *Bioinformatics*, **25**, 3207-3212.

Denny, J.C., *et al.* (2010) PheWAS: demonstrating the feasibility of a phenome-wide scan to discover gene-disease associations. *Bioinformatics*. pp. 1205-1210.

Dettling, M., *et al.* (2007) Clozapine-induced agranulocytosis in schizophrenic Caucasians: confirming clues for associations with human leukocyte class I and II antigens. *Pharmacogenomics J*. pp. 325-332.

Dobson, R.J., *et al.* (2006) Predicting deleterious nsSNPs: an analysis of sequence and structural attributes, *BMC Bioinformatics*, **7**, 217.

Drmanac, R., *et al.* (2010) Human genome sequencing using unchained base reads on self-assembling DNA nanoarrays, *Science*, **327**, 78-81.

Dudley, J.T. and Butte, A.J. (2010) In silico research in the era of cloud computing, *Nat Biotechnol*, **28**, 1181-1185.

Ehrlich, P. (1906) Die aufgaben der chemotherapie., *Frankfurter Zeitung und Handelsblatt: Zweites Morgenblatt*.

Ewen, K.R., *et al.* (2000) Identification and analysis of error types in high-throughput genotyping, *Am J Hum Genet*, **67**, 727-736.

Ferrell, P.B. and McLeod, H.L. (2008) Carbamazepine, HLA-B*1502 and risk of Stevens-Johnson syndrome and toxic epidermal necrolysis: US FDA recommendations. *Pharmacogenomics*. pp. 1543-1546.

Ferrer-Costa, C., *et al.* (2005) PMUT: a web-based tool for the annotation of pathological mutations on proteins, *Bioinformatics*, **21**, 3176-3178.

Foot, E., Kleyn, D. and Palmer Foster, E. (2010) Pharmacogenetics--pivotal to the future of the biopharmaceutical industry. *Drug Discov Today*. pp. 325-327.

Frazer, K.A., *et al.* (2009) Human genetic variation and its contribution to complex traits, *Nat Rev Genet*, **10**, 241-251.

Gage, B.F. and Lesko, L.J. (2008) Pharmacogenetics of warfarin: regulatory, scientific, and clinical issues. *J Thromb Thrombolysis*. pp. 45-51.

Gamazon, E.R., *et al.* (2010) PACdb: a database for cell-based pharmacogenomics. *Pharmacogenet Genomics*. pp. 269-273.

Gambacorti-Passerini, C. (2008) Part I: Milestones in personalised medicine--imatinib. *Lancet Oncol*. pp. 600.

Garten, Y., Coulet, A. and Altman, R.B. (2010) Recent progress in automatically extracting information from the pharmacogenomic literature. *Pharmacogenomics*. pp. 1467-1489.

Giacomini, K.M., *et al.* (2007) The pharmacogenetics research network: from SNP discovery to clinical drug response. *Clin Pharmacol Ther*. pp. 328-345.

Ginsburg, G.S. and Willard, H.F. (2009) Genomic and personalized medicine: foundations and applications. *Transl Res*. pp. 277-287.

Goh, K.-I., *et al.* (2007) The human disease network. *Proc Natl Acad Sci USA*. pp. 8685-8690.

Gold, M., *et al.* (2010) Rosiglitazone monotherapy in mild-to-moderate alzheimer's disease: results from a randomized, double-blind, placebo-controlled phase III study, *Dement Geriatr Cogn Disord*, **30**, 131-146.

Hahn, L.W., Ritchie, M.D. and Moore, J.H. (2003) Multifactor dimensionality reduction software for detecting gene-gene and gene-environment interactions. *Bioinformatics*. pp. 376-382.

Hansen, N.T., Brunak, S. and Altman, R.B. (2009) Generating genome-scale candidate gene lists for pharmacogenomics. *Clin Pharmacol Ther*. pp. 183-189.

Hetherington, S., *et al.* (2002) Genetic variations in HLA-B region and hypersensitivity reactions to abacavir, *Lancet*, **359**, 1121-1122.

Higashi, M.K., *et al.* (2002) Association between CYP2C9 genetic variants and anticoagulation-related outcomes during warfarin therapy. *JAMA*. pp. 1690-1698.

Hindorff, L.A., *et al.* (2009) Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proc Natl Acad Sci USA*. pp. 9362-9367.

Hopkins, A. (2007) Network pharmacology. *Nat Biotechnol*. pp. 1110-1111.

Hudis, C.A. (2007) Trastuzumab--mechanism of action and use in clinical practice. *N Engl J Med*. pp. 39-51.

International Warfarin Pharmacogenetics Consortium, *et al.* (2009) Estimation of the warfarin dose with clinical and pharmacogenetic data. *N Engl J Med*. pp. 753-764.

Kaminker, J.S., *et al.* (2007) CanPredict: a computational tool for predicting cancer-associated missense mutations, *Nucleic Acids Res*, **35**, W595-598.

Karchin, R. (2009) Next generation tools for the annotation of human SNPs, *Brief Bioinform*, **10**, 35-52.

Karchin, R., *et al.* (2005) LS-SNP: large-scale annotation of coding non-synonymous SNPs based on multiple information sources, *Bioinformatics*, **21**, 2814-2820.

Kasowski, M., *et al.* (2010) Variation in transcription factor binding among humans, *Science*, **328**, 232-235.

Keiser, M., *et al.* (2007) Relating protein pharmacology by ligand chemistry. *Nat Biotechnol*. pp. 197-206.

Keiser, M.J., *et al.* (2009) Predicting new molecular targets for known drugs. *Nature*. pp. 175-181.

Kent, W.J. (2002) BLAT--the BLAST-like alignment tool, *Genome Res*, **12**, 656-664.

Kohl, P., *et al.* (2010) Systems biology: an approach. *Clinical Pharmacology & Therapeutics*. pp. 25-33.

Krishnan, V.G. and Westhead, D.R. (2003) A comparative study of machine-learning methods to predict the effects of single nucleotide polymorphisms on protein function, *Bioinformatics*, **19**, 2199-2209.

Lee, S.H., *et al.* (2008) Predicting unobserved phenotypes for complex traits from whole-genome SNP data. *PLoS Genet*. pp. e1000231.

Li, B., *et al.* (2009) Automated inference of molecular mechanisms of disease from amino acid substitutions, *Bioinformatics*, **25**, 2744-2750.

Li, H. and Durbin, R. (2009) Fast and accurate short read alignment with Burrows-Wheeler transform, *Bioinformatics*, **25**, 1754-1760.

Li, H. and Homer, N. (2010) A survey of sequence alignment algorithms for next-generation sequencing, *Brief Bioinformatics*, **11**, 473-483.

Limdi, N.A. and Veenstra, D.L. (2008) Warfarin pharmacogenetics. *Pharmacotherapy*. pp. 1084-1097.

Mailman, M.D., *et al.* (2007) The NCBI dbGaP database of genotypes and phenotypes. *Nat Genet*. pp. 1181-1186.

Manolio, T.A., *et al.* (2009) Finding the missing heritability of complex diseases. *Nature*. pp. 747-753.

McGuire, A.L. and Burke, W. (2008) An unwelcome side effect of direct-to-consumer personal genome testing: raiding the medical commons. *JAMA*. pp. 2669-2671.

Mooney, S. (2005) Bioinformatics approaches and resources for single nucleotide polymorphism functional analysis, *Brief Bioinform*, **6**, 44-56.

Moore, J.H., Asselbergs, F.W. and Williams, S.M. (2010) Bioinformatics challenges for genome-wide association studies. *Bioinformatics*. pp. 445-455.

Motsinger, A.A., Ritchie, M.D. and Reif, D.M. (2007) Novel methods for detecting epistasis in pharmacogenomics studies. *Pharmacogenomics*. pp. 1229-1241.

Motsinger-Reif, A.A., *et al.* (2010) Genome-wide association studies in pharmacogenomics: successes and lessons. *Pharmacogenetics and Genomics*. pp. 1.

Ng, P.C. and Henikoff, S. (2003) SIFT: Predicting amino acid changes that affect protein function, *Nucleic Acids Res*, **31**, 3812-3814.

Ng, S.B., *et al.* (2010) Exome sequencing identifies the cause of a mendelian disorder, *Nat Genet*, **42**, 30-35.

O'Dushlaine, C., Kenny, E. and Heron…, E. (2009) The SNP ratio test: pathway analysis of genome-wide association datasets. ….

Ohashi, W. and Tanaka, H. (2010) Benefits of pharmacogenomics in drug development-earlier launch of drugs and less adverse events. *J Med Syst*. pp. 701-707.

Paolini, G.V., *et al.* (2006) Global mapping of pharmacological space. *Nat Biotechnol*. pp. 805-815.

Perez-Iratxeta, C., *et al.* (2005) G2D: a tool for mining genes associated with disease, *BMC Genet*, **6**, 45.

Pool, J.E., *et al.* (2010) Population genetic inference from genomic sequence variation, *Genome Res*, **20**, 291-300.

Radivojac, P., *et al.* (2008) An integrated approach to inferring gene-disease associations in humans, *Proteins*, **72**, 1030-1037.

Ramensky, V., Bork, P. and Sunyaev, S. (2002) Human non-synonymous SNPs: server and survey, *Nucleic Acids Res*, **30**, 3894-3900.

Rieder, M.J., *et al.* (2005) Effect of VKORC1 haplotypes on transcriptional regulation and warfarin dose. *N Engl J Med*. pp. 2285-2293.

Roden, D., *et al.* (2008) Development of a Large-Scale De-Identified DNA Biobank to Enable Personalized Medicine. *Clin Pharmacol Ther*. pp. 362-369.

Roses, A.D. (2004) Pharmacogenetics and drug development: the path to safer and more effective drugs. *Nat Rev Genet*. pp. 645-656.

Roses, A.D. (2009) The medical and economic roles of pipeline pharmacogenetics: Alzheimer's disease as a model of efficacy and HLA-B(*)5701 as a model of safety. *Neuropsychopharmacology*. pp. 6-17.

Rost, S., *et al.* (2004) Mutations in VKORC1 cause warfarin resistance and multiple coagulation factor deficiency type 2. *Nature*. pp. 537-541.

Saeys, Y., Inza, I. and Larranaga, P. (2007) A review of feature selection techniques in bioinformatics, *Bioinformatics*, **23**, 2507-2517.

Sagreiya, H., *et al.* (2010) Extending and evaluating a warfarin dosing algorithm that includes CYP4F2 and pooled rare variants of CYP2C9. *Pharmacogenetics and Genomics*. pp. 407-413.

Schlicker, A., Lengauer, T. and Albrecht, M. (2010) Improving disease gene prioritization using the semantic similarity of Gene Ontology terms, *Bioinformatics*, **26**, i561-567.

Schymkowitz, J., *et al.* (2005) The FoldX web server: an online force field, *Nucleic Acids Res*, **33**, W382-388.

Shendure, J. and Ji, H. (2008) Next-generation DNA sequencing. *Nat Biotechnol*. pp. 1135-1145.

Sherry, S.T., *et al.* (2001) dbSNP: the NCBI database of genetic variation, *Nucleic Acids Res*, **29**, 308-311.

Shuldiner, A.R., *et al.* (2009) Association of cytochrome P450 2C19 genotype with the antiplatelet effect and clinical efficacy of clopidogrel therapy, *JAMA*, **302**, 849-857.

Shuldiner, A.R., *et al.* (2009) Association of cytochrome P450 2C19 genotype with the antiplatelet effect and clinical efficacy of clopidogrel therapy. *JAMA*. pp. 849-857.

Smith TF, W.M. (1981) Identification of Common Molecular Subsequences. *Journal of Molecular Biology*. pp. 195-197.

Soares, M.L., *et al.* (2005) Susceptibility and modifier genes in Portuguese transthyretin V30M amyloid polyneuropathy: complexity in a single-gene disease. *Hum Mol Genet*. pp. 543-553.

Statnikov, A., *et al.* (2005) A comprehensive evaluation of multicategory classification methods for microarray gene expression cancer diagnosis, *Bioinformatics*, **21**, 631-643.

Storey, J.D., Akey, J.M. and Kruglyak, L. (2005) Multiple locus linkage analysis of genomewide expression in yeast. *PLoS Biol*. pp. e267.

Subramanian, A., *et al.* (2005) Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci USA*. pp. 15545.

Tavtigian, S.V., *et al.* (2008) In silico analysis of missense substitutions using sequence-alignment based methods, *Hum Mutat*, **29**, 1327-1336.

Thomas, P.D. and Kejariwal, A. (2004) Coding single-nucleotide polymorphisms associated with complex vs. Mendelian disease: evolutionary evidence for differences in molecular effects, *Proc Natl Acad Sci U S A*, **101**, 15398-15403.

Thomas, P.D., *et al.* (2003) PANTHER: a browsable database of gene products organized by biological function, using curated protein family and subfamily classification, *Nucleic Acids Res*, **31**, 334-341.

Thorn, C.F., Klein, T.E. and Altman, R.B. (2010) Pharmacogenomics and bioinformatics: PharmGKB, *Pharmacogenomics*, **11**, 501-505.

Thusberg, J., Olatubosun, A. and Vihinen, M. (2011) Performance of mutation pathogenicity prediction methods on missense variants, *Hum Mutat*, **32**, 358-368.

Tranchevent, L., *et al.* (2010) A guide to web tools to prioritize candidate genes, *Briefings in Bioinformatics*.

Tranchevent, L.C., *et al.* (2008) ENDEAVOUR update: a web resource for gene prioritization in multiple species, *Nucleic Acids Res*, **36**, W377-384.

van Driel, M.A., *et al.* (2006) A text-mining analysis of the human phenome, *Eur J Hum Genet*, **14**, 535-542.

Wainreb, G., *et al.* (2010) MuD: an interactive web server for the prediction of non-neutral substitutions using protein structural data, *Nucleic Acids Res*, **38 Suppl**, W523-528.

Wang, K. and Li, M. (2007) Pathway-based approaches for analysis of genomewide association studies. *The American Journal of Human Genetics*.

Weinshilboum, R. (2001) Thiopurine pharmacogenetics: clinical and molecular studies of thiopurine methyltransferase, *Drug Metab Dispos*, **29**, 601-605.

Wilffert, B., *et al.* (2010) From evidence based medicine to mechanism based medicine. Reviewing the role of pharmacogenetics, *Pharm World Sci*.

Wilke, R.A., Reif, D.M. and Moore, J.H. (2005) Combinatorial pharmacogenetics. *Nat Rev Drug Discov*. pp. 911-918.

Williams, S.M., *et al.* (2004) Multilocus analysis of hypertension: a hierarchical approach. *Hum Hered*. pp. 28-38.

Woolfe, A., Mullikin, J.C. and Elnitski, L. (2010) Genomic features defining exonic variants that modulate splicing, *Genome Biol*, **11**, R20.

Wray, N.R., Goddard, M.E. and Visscher, P.M. (2007) Prediction of individual genetic risk to disease from genome-wide association studies. *Genome Res*. pp. 1520-1528.

WTCC Consortium (2007) Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls, *Nature*, **447**, 661-678.

Wu, Z. and Zhao, H. (2009) Statistical power of model selection strategies for genome-wide association studies. *PLoS Genet*. pp. e1000582.

Wysowski, D.K., Nourjah, P. and Swartz, L. (2007) Bleeding complications with warfarin use: a prevalent adverse effect resulting in regulatory action. *Arch Intern Med*. pp. 1414-1419.

Yeung, K.Y. and Ruzzo, W.L. (2001) Principal component analysis for clustering gene expression data, *Bioinformatics*, **17**, 763-774.

Yildirim, M.A., *et al.* (2007) Drug-target network. *Nat Biotechnol*. pp. 1119-1126.

Yngvadottir, B., *et al.* (2009) The promise and reality of personal genomics. *Genome Biology*. pp. 237.

Yu, K., *et al.* (2009) Pathway analysis by adaptive combination of P-values. *Genetic* ….

Yue, P. and Moult, J. (2006) Identification and analysis of deleterious human SNPs, *J Mol Biol*, **356**, 1263-1274.

Zhang, W. and Dolan, M.E. (2009) Use of cell lines in the investigation of pharmacogenetic loci. *Curr Pharm Des*. pp. 3782-3795.

Zhang, Z. and Gerstein, M. (2003) Patterns of nucleotide substitution, insertion and deletion in the human genome inferred from pseudogenes, *Nucleic Acids Res*, **31**, 5338-5348.

Zhou, H. and Zhou, Y. (2002) Distance-scaled, finite ideal-gas reference state improves structure-derived potentials of mean force for structure selection and stability prediction, *Protein Sci*, **11**, 2714-2726.