Integrative genomics strategies to elucidate the complexity of drug response

Pharmacogenomic investigation from both genome-wide association studies and experiments focused on candidate loci involved in drug mechanism and metabolism has yielded a substantial and increasing list of robust genetic effects on drug therapy in humans. At the same time, reasonably comprehensive molecular data such as gene expression, proteomic and metabolomic data are now available for collections of hundreds to thousands of individuals. If these data are structured in a statistically robust and computationally tractable way, such as a network model, they can aid in the analysis of new pharmacogenomics studies by suggesting novel hypotheses for the regulation of genes involved in drug metabolism and response. Similarly, hypotheses taken from these same models can direct genome-wide association studies by focusing the genome-wide association studies analysis on a number of specific hypotheses informed by the relationships customarily seen between a gene's expression or protein activity and genetic variation at a particular locus. Network models based on other sorts of systematic biological data such as cell-based surveys of drug effect on gene expression and mining of literature and electronic medical records for associations between clinical and molecular phenotypes also promise similar utility. Although surely primitive in comparison with what will be developed, these model-based approaches to leveraging the increasing volume of data generated in the course of patient care and medical research nevertheless suggest a huge opportunity to improve our understanding of biological systems involved in pharmacogenomics and apply them to questions of medical relevance.

KEYWORDS: abacavir ▪ Bayesian network ▪ coexpression network ▪ connectivity map ▪ electronic medical records ▪ flucloxacillin ▪ gene networks ▪ genetics of gene expression ▪ genome sequencing ▪ genome-wide association study ▪ integrative genomics ▪ interferons ▪ lumiracoxib ▪ pharmacogenomics ▪ ribavirin ▪ statins ▪ warfarin

Andrew Kasarskis*[1,2‡], Xia Yang[3,4‡] & Eric Schadt[1,2]

[1]*Pacific Biosciences, 1380 Willow Road, Menlo Park, CA 94025, USA*
[2]*Department of Genetics & Genomics Sciences, Mount Sinai School of Medicine, One Gustave L. Levy Place, Box 1498, New York, NY 10029, USA*
[3]*Sage Bionetworks, 1100 Fairview Ave N, Seattle, WA 98109, USA*
[4]*Department of Integrative Biology & Physiology, Univeristy of California, Los Angeles, Terasaki Life Sciences Building, 610 Charles E. Young Dr. East, Los Angeles, CA 90095-7239, USA*
**Author for correspondence: andrew.kasarskis@mssm.edu*
‡These authors contributed equally

Interindividual variability in drug response, adverse events and toxicity represents a common observation in clinical therapeutics, and strong genetic components underlying drug metabolism and clinical drug response have been implicated through family and twin studies [1–7]. Understanding the genetic factors that predispose patients to poor treatment outcome, adverse events and/or toxicity will help guide therapeutic strategies for individual patients to obtain maximal efficacy and safety. With genome-wide genotyping of common variation becoming routine, and whole exome and even whole-genome sequencing beginning to emerge as an affordable practice in research settings, the study of the relationship between genetic variations (primarily SNPs) and drug response or toxicity has shifted from limited numbers of candidate variations (pharmacogenetics) to whole-genome variation that covers hundreds of thousands to millions of SNPs simultaneously (pharmacogenomics). Whole-genome scanning provides a far more objective and comprehensive view of the genetic susceptibility of drug response than previous knowledge-based candidate-gene approaches, just as it has in the genetic study of common human diseases.

Pharmacogenomic studies have identified genetic markers that are strong predictors of drug effects [8]. For instance, common and rare genetic variations in *CYP2C9* and *VKORC1* are found to predict warfarin dosage or response [9–13], the *HLA-B*5701* variant is associated with abacavir hypersensitivity [14,15], *HLA-B*5701*, *HLA-DRB1*0107-DRB1*0103*, and *TNF*-α-238G/A are linked to flucloxacillin-induced liver injury [16], and SNPs at the *IL28A/IL28B* locus are associated with antihepatitis C treatment response to interferons and ribavirin [17–20]. These findings highlight the power of pharmacogenomics to identify genetic risk factors. However, the identification of susceptibility SNPs or loci does not directly lead to mechanisms, as shown by a survey of genetic loci associated with common human disease [21]. In fact, drug response is a complex trait just like common human disease traits, and so interpretations of genetic associations with drug

future medicine part of fsg

response and disease traits are necessarily similar. In both cases, genome-wide association studies (GWAS) have revealed a definitive relationship between a segment of DNA and a phenotype of interest, a key fact, which is very hard to establish for common genetic variation before GWAS were possible, and that points the way toward fine mapping, functional, and other experiments needed to determine exactly how the DNA variation influences phenotype. Typically, given an association between SNP genotypes from a GWAS and drug response, the gene that is driving variations in response is often unknown. In addition, whether activation or inactivation of that gene leads to increased drug response is often unknown and neither the context in which the associated gene operates nor the mechanism by which the SNP (or causal variants linked to the associated SNP) confers risk of poor response, adverse events or toxicity is clear; therefore there remains ample further experimental work to be done.

According to the GWAS catalog maintained by the National Human Genome Research Institute [22,201], a total of 48 pharmacogenomic GWAS have reported the pharmacogenomics of various therapeutic agents in terms of dosage, therapeutic efficacy, adverse events or hepatotoxicity. These studies cover a variety of therapeutic agents including antipsychotics [23–29], antidepressants [30–33], anticoagulants [11,34–37], anti-hepatitis C drugs [18–20], statins [38,39] and others (Table 1). In contrast to these successful cases, the results from many other pharmacogenomics studies, especially those for the antidepressant and antipsychotic agents, appear to involve multiple loci, with many loci supported by only a single study [201], suggesting more complex genetic susceptibility and/or a more complex phenotype affected by many distinct biological processes. In addition, many of the genetic loci are located in intergenic regions with no obvious candidate genes directly implicated, which is consistent with the previous observation that almost 40% of disease/trait-associated SNPs are located in intergenic regions and another 40% are intronic, whereas only 12% are located in or are close to protein-coding regions of genes [21]. Therefore, interpreting pharmacogenomic findings is rarely a straightforward task

Recently, integrative genomics approaches that leverage functional genomics and network biology have been developed to identify genes, pathways and gene networks that underlie GWAS findings for various diseases and traits [40–47]. These methods also exploit a large volume of valuable information gathered in a GWAS, which is underutilized when the GWAS is analyzed in isolation. Typically, without a prior hypothesis with which to focus the analysis on a subset of gene regions, the multiple testing correction utilized in genome-wide statistical analysis allows for the detection of only the strongest effects, and penalizes weaker associations that are biologically meaningful. In this review, we discuss several integrative genomics methodologies that are just as applicable to the pharmacogenomics field as they are in the study of common human disease and systems biology more generally, where they were largely developed.

Functional genomics

As discussed above, genetic association studies such as pharmacogenomics or GWAS investigate the association between genetic variations and clinical diseases or traits to uncover genetic risk factors of phenotypic traits. However, although the identification of significant genetic loci from a genetic association study does imply that they are functionally relevant to the associated phenotype, it does not directly imply functional relevance of the associated SNPs or provide the underlying genes and molecular mechanisms, except in those cases where the genetic polymorphism has an obvious path to mechanism such as structural effect on a protein. In order to tackle the functional role of the risk loci, it is necessary to not only understand the corresponding genes involved, but also to understand the molecular consequences of the sequence variations. To this end, the relationship between the genome and intermediate molecular traits such as gene expression, alternative splicing and protein products can be explored via functional genomic studies to identify molecular quantitative trait loci (QTL).

The study of genetic variations that are associated with gene expression, the most commonly characterized molecular trait, is coined genetics of gene expression (GGE). Through GGE, one can simultaneously scan for associations between millions of SNPs and tens of thousands of gene-expression traits (including alternative splicing) corresponding to tens of thousands of genes in the human genome, and thereby determine which genetic loci are linked to the expression traits. These genetic loci are termed expression QTL (eQTL), and individual SNPs under eQTL are named expression SNPs (eSNPs). These eQTL or eSNPs linking to intermediate molecular traits are useful in elucidating the functions of the genetic variations. The correlation between a SNP and a molecular trait implies that either the SNP itself or a DNA variant in

Table 1. Confirmed or replicated genetic loci in pharmacogenomic genome-wide association studies based on National Human Genome Research Institute genome-wide association studies catalog†.

PubMed ID	Disease or medical condition	Therapy	End point	Loci at association $p < 1 \times 10^{-5}$	Reported candidate genes	Ref.
18650507	Dyslipidemia	Statin	AE – myopathy	12p12.1	*SLCO1B1*	[84]
21149285	Epilepsy, mood disorders	Carbamazepine	AE – cutaneous	6p21.33	*HLA-A*	[173]
21428769	Epilepsy, mood disorders	Carbamazepine	AE – carbamazepine-induced SJS and TEN	6p21.33	*HLA-A*	[174]
18535201	Blood clots	Warfarin	Dosage	10q23.33, 12p13.33, 16p11.2	*CYP2C9*, *CACNA1C*, *VKORC1*	[35]
19300499	Blood clots	Warfarin	Dosage	10q23.33, 16p11.2, 19p13.12	*CYP2C9*, *VKORC1*, *CYP4F2*	[11]
20833655	Blood clots	Warfarin	Dosage	10q23.33, 13q21.1, 16p11.2, 19p13.12, 7q22.3	*CYP2C9*, *NR*, *VKORC1*, *CYP4F2*	[34]
19578179	Blood clots	Acenocoumarol	Dosage	10q23.33, 19p13.12, 3q22.3	*CYP2C18*, *CYP2C19*, *CYP4F2*, *CNTN4*	[36]
19706858	Blood clots	Clopidogrel	Efficacy – antiplatelet, cardiovascular outcomes	10q23.33	*CYP2C18, CYP2C19, CYP2C9, CYP2C8*	[37]
19684573	Hepatitis C	Three treatment regimens involving PegIFN-2β or PegIFN-2α combined with ribavirin	Efficacy – sustained virological response	19q13.2, 4q34.3, 6q21	*IL28B, AKD2*	[18]
19749757	Chronic hepatitis C	Pegylated IFN-α and ribavirin therapy	Efficacy – sustained virological response	19q13.2	*IL28B*	[20]
19749758	Chronic hepatitis C	IFN-α and ribavirin therapy	Efficacy – sustained virological response	19q13.2	*IL28A, IL28B*	[19]
19483685	Infection	Flucloxacillin	Hepatotoxicity	12q12, 15q26.2, 3q11.2, 3q27.3, 6p21.33, 9p21.2	*ALG10B*, *MCTP2*, *OR5H2*, *ST6GAL1*, *HCP5, HLA-B*, *C9org82*	[16]
20639878	Osteoarthritis	Lumiracoxib	Hepatotoxicity	6p21.32	*HLA-DRB1*	[83]
21570397	Infection	Amoxicillin-clavulanate	Hepatotoxicity	6p21.32, 6p21.33	*HLA-DRB1*, *HLA-A*	[175]

†Data downloaded from the National Human Genome Research Institute genome-wide association studies catalog on 17 July 2011 [201].
AE: Adverse event; SJS: Stevens–Johnson syndrome; TEN: Toxic epidermal necrolysis.

linkage disequilibrium (LD) with the observed SNP is functional. When the same genetic loci or SNPs are found to be associated with a phenotypic trait, the genes whose expression levels are correlated with these QTL or SNPs represent more plausible candidate causal genes for the phenotypic traits than uncorrelated genes [48–56]. Put simply, it is difficult to imagine how a DNA variation could cause a phenotype without modifying the expression of at least one gene in one tissue, so genes for which eQTL are detected should have a better than random chance of causing the phenotype.

GGE studies have been conducted in lymphocytes or lymphoblastoid cell lines [57–65], monocytes [66], fibroblasts [67], T-cells [67], brain [68], liver [42,69] and adipose tissues [41,58,69]. These studies support the presence of both shared eQTL across tissues or cell types and tissue- or cell-type-specific eQTL. The overlap of eQTL across tissues has been found to depend on both sample size and similarity between tissues. As reported by Greenawalt *et al.* [69] with a sample size of approximately 100, approximately 30% of eQTL overlap between liver and adipose, and 49% overlap between subcutaneous adipose and

omental adipose. When the sample size increases to approximately 800, the overlap increases to 60% between liver and adipose and to 80% between the two adipose depots. These results suggest that a relatively large proportion of eQTL is shared between tissues.

It is important to note that eQTL identification is subject to confounders that produce nonreplicable false discoveries and that replication between studies is therefore critical in order to select eQTL of biological significance. In a recent study, Innocenti *et al.* investigated eQTL in two independent liver cohorts and found that confounders that affect liver gene expression, such as drug exposure, clinical descriptors and unknown factors associated with tissue ascertainment and analysis, and all contribute to the ability to replicate eQTL between studies [70]. Therefore, thorough quality control and advanced statistical methods that take hidden cofounding variables into consideration are necessary to identify reliable signals.

In addition to eQTL identified in GGE studies, QTL for other intermediate molecular traits such as alternative splicing [60,71–73] allelic expression [74,75], DNA methylation [76–81], and liver enzyme activity [82] have also been pursued and similar tissue specificities have been observed for these molecular QTL types. Additional types of molecular QTL associated with traits such as miRNA, protein and metabolite levels, histone modification and chromatin structure can also be detected to further annotate the function of genetic loci.

The power of utilizing QTL of molecular traits identified from functional genomic studies to identify candidate genes as well as the mechanisms underlying GWAS findings of complex human diseases have been demonstrated in a number of recent studies [40–42,72]. For example, the *HLA-DRB1* locus identified for Type 1 diabetes is strongly associated with the expression levels of the *HLA-DRB1* gene [42,69]. In the case of warfarin treatment, a SNP approximately 50 kb upstream of the gene *VKORC1*, rs10871454, was found to be associated with warfarin dosage with an association p-value of 4.7×10^{-34} in a meta-analysis of index and replication cohorts in a recent pharmacogenomics GWAS [35]. As this SNP is in perfect LD ($r^2 = 1.0$) with the promoter SNP rs9923231, it could act as a surrogate for the promoter SNP, which functions in regulating the expression of *VKORC1*. In order to elucidate the potential mechanism, we searched our human liver eSNP database and found that this SNP is strongly associated with the expression level of *VKORC1* at $p = 1.07 \times 10^{-69}$ [42,69]. Specifically, the minor allele of rs10871454 is associated with decreased *VKORC1* expression (Figure 1). This finding suggests that rs10871454 itself or one or more functional DNA variants in LD with this SNP regulate the expression levels of the warfarin target gene, thus providing a mechanistic explanation for the observed association between the SNP and warfarin response. The liver eSNPs from two cohorts, an unselected liver cohort [42] and a morbidly obese cohort [69], have been published and are publicly accessible. The expression data for both cohorts are available at Gene Expression Omnibus [202] with accession numbers GSE9588 and GSE24335. The genotype data can be obtained upon request and will be available at dbGaP [203]. Between the two cohorts, 66% of eSNPs overlap, suggesting consistent signals. We chose to report the SNP–gene association p-values from the larger liver cohort due to its higher statistical power [70].

When intersecting the genetic loci identified from other recent pharmacogenomics GWAS listed in Table 1 with published liver GGE or eQTL data [42,69], we were able to identify or confirm additional candidate genes for a number of loci via eQTL mapping. For example, the *HLA-DRB1* locus for lumiracoxib-related liver injury [83] is found to be a liver eQTL for the *HLA-DRB1* gene ($p = 3.15 \times 10^{-33}$ for association of SNP rs3129934 and *HLA-DRB1* gene expression). In addition, several loci for statin response or adverse effects [38,39,84] are liver eQTL for *SORT1* ($p = 5.20 \times 10^{-88}$ with rs646776)/*PSRC1* ($p = 3.05 \times 10^{-86}$ with rs646776)/*CELSR2* ($p = 6.27 \times 10^{-68}$ with rs646776), *FADS1* ($p = 7.00 \times 10^{-19}$ with rs1535)/*FADS2* ($p = 5.63 \times 10^{-8}$ with rs1535)/*FADS3* ($p = 2.96 \times 10^{-6}$ with rs1535) and *SLCO1B1* ($p = 4.44 \times 10^{-6}$ with rs12371604, a SNP in LD with the reported SNP rs4149056); and a locus linked to flucloxacillin-induced liver injury (represented by the SNP rs2395029) [16] is a liver eQTL for the MHC-I polypeptide-related sequence B protein ($p = 2.68 \times 10^{-9}$), HLA-B ($p = 5.66 \times 10^{-7}$) and HLA-C ($p = 1.90 \times 10^{-6}$). Although these eSNP–gene associations were identified from published eQTL studies, these associations have not been highlighted previously in the context of pharmacogenomics, with the exception of the *VKORC1* association. Tables 1 & 2 list all of the genes significantly associated with pharmacogenomics GWAS SNPs as this association provides empirical evidence to support them as candidate genes at these loci. Whether or not these eSNP-nominated candidate genes do in fact mediate the association with phenotype

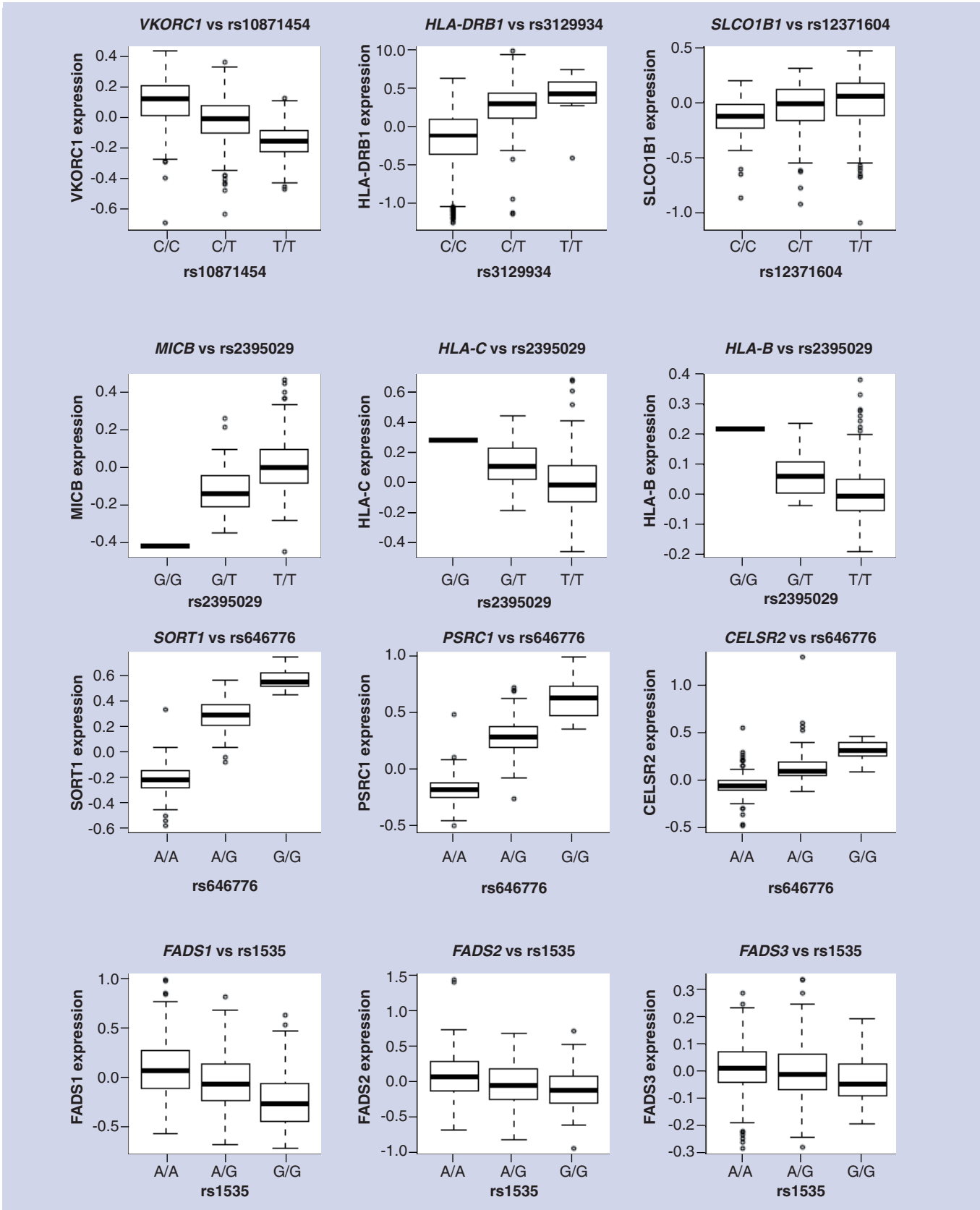

Figure 1. Pharmacogenomic risk SNPs that are associated with gene-expression levels in human liver. Each panel shows box plots of the normalized expression of the gene in individuals of the three given genotypes at that SNP.
Data taken from [69].

will, of course, await further confirmatory experiments of the type performed to nominate *SORT1* as the causal gene underlying the 1p13 human dyslipidemia locus [85].

It is worth noting that many loci could not be successfully mapped via GGE in these liver data sets. One possible explanation for this is that the effect of these loci on the expression of genes in the liver is small, and that the loci negative for regulation of gene expression in the liver might demonstrate that regulation in other tissues. Another possibility is that many loci may affect the function of a gene via post-transcriptional mechanisms that are not detected at the level of gene expression or eQTL. Still another possibility is that many of the pharmacogenomic associations that have been reported may represent false discoveries, as is suggested for those that have not been reproduced in independent studies. All of these effects demonstrate an opportunity to improve the use of functional genomics in interpreting pharmacogenomics studies.

Network biology

While helpful when inferring the candidate genes underlying individual genetic loci via QTL mapping as demonstrated in the previous section, functional genomics cannot, in general, directly address the molecular mechanisms of each locus as well as the potential interactions among genetic loci. In recent years, network biology has increasingly come to provide a systems view of how individual molecular traits such as genes, metabolites or proteins interact with one another and their relationship with various clinical phenotypes (including drug response) within a cell, a tissue or an organ via integration of large-scale genetic, genomic, transcriptomic, metabolomic and proteomic data [86]. As shown in Figure 2, networks are represented graphically as nodes and edges, where nodes are individual molecular traits or higher-order phenotypes (e.g., clinical traits associated with disease) and edges represent the interactions such as physical binding or statistical correlation among molecular traits or between molecular traits and higher-order phenotypes. Networks can be directed or undirected. In a directed network, the direction of an edge between any two nodes represents a causal relationship or a sequential event, whereas in an undirected network the edge between two nodes represents reciprocal association such as correlation.

A variety of network methodologies have been developed to construct networks of various types. The most common networks are protein–protein interaction networks and gene regulatory networks. In this review, we will focus initially on two types of gene regulatory networks, namely, weighted gene coexpression network analysis (WGCNA) [87,88] and Bayesian network (BN) (Figure 3) [89–96]. We have constructed such networks in various species including human, mouse and yeast to help identify gene sub-networks associated with a variety of common human diseases such as obesity and atherosclerosis [16,97–100,204].

WGCNA is a correlation-based method and focuses on the coregulation pattern among genes. The advantage of coexpression networks is that they allow one to look at the overall gene–gene correlation structure at a high level via the construction of gene modules comprised of highly interconnected sets of genes (a schematic analysis flow is shown in Figure 3A). A number of studies have demonstrated that coexpression network modules are generally enriched for genes involved in known biological pathways, for genes that are linked to common genetic loci, and for genes associated with diseases [42,51,87,97–105]. Using coexpression networks, one can identify key groups of genes that are modulated by genetic loci or regulated by key transcription factors and that in turn lead to disease, and therefore define disease states at the molecular level [89].

BNs on the other hand are probabilistic graphical models comprised of molecular and higher-order traits and constructed by assessing the conditional dependencies between all of the variables under consideration (a schematic analysis flow is shown in Figure 3B). A BN provides a natural and mathematically elegant framework for integrating diverse large-scale, high-dimensional genetic, transcriptomic, proteomic and metabolomic data sets to decipher the biological function of individual genes and pathways. Prior information such as genetic regulation, transcription factor binding and protein–protein interaction can be incorporated to help infer directionality between genes. Compared with WGCNA, BNs are more sparse since they penalize complexity and therefore only keep primary interactions, however, they allow a more granular view of the relationships and directional predictions between genes than can be obtained with WGCNA [95,102]. A number of studies performed by us and others in a variety of species have demonstrated that predictive networks such as BN can capture fundamental properties of complex systems in states that give rise to complex phenotypes [42,82,89–91,95,96,102,106–110]. Both types of networks provide objective views

Table 2. Suggestive genetic loci in pharmacogenomic genome-wide association studies based on National Human Genome Research Institute genome-wide association studies catalog†.

PubMed ID	Disease or medical condition	Therapy	End point	Loci at association $p < 1 \times 10^{-5}$	Reported candidate genes	Ref.
21130132	Attention-deficit/ hyperactivity disorder	Methylphenidate	AE – BP	‡	‡	[176]
17505501	Blood clots	Oral-direct thrombin inhibitor ximelagatran	AE – elevated levels of serum alanine aminotransferase	6p21.3	*HLA-DRB1*	[177]
19680635	Schizophrenia	Antipsychotics	AE – antipsychotic-induced parkinsonism severity	2q24.3	*FIGN*	[25]
18521091	Schizophrenia	Iloperidone	AE – QT prolongation	10q23.1, 14q12, 15q26.1, 18q12.2, 2q31.3, 4q32.3	*NRG3*, *NUBPL*, *SLCO3A1*, *BRUNOL4*, *CERKL*, *PALLD*	[29]
20195266	Schizophrenia	Antipsychotics	AE – metabolic	10p11.22, 10p12.33, 11q23.1, 11q24.2, 12p12.1, 13q12.11, 14q32.13, 15q14, 16p13.12, 16p13.13, 16q23.3, 16q23.3, 18q12.2, 18q22.2, 1p21.2, 1p31.1, 20q13.2, 2p11.2, 2p12, 2p16.1, 2p24.1, 2p25.1, 2p25.3, 2q33.1, 4q23, 4q24, 5q14.3, 5q31.3, 6p21.31, 6q14.3, 7p21.1, 7q22.3, 8q22.3, 9q31.1, 9q33.1	*KIRREL3*, *SOX5*, *CLMN*, *MEIS2*, *LOC729993*, *ATF7IP2*, *CDH13*, *FHOD3*, *RNF144A*, *GPR98*, *PPARD*, *PRKAR2B*, *ASTN2*	[24]
19875103	Schizophrenia	Antipsychotics	AE – movement related	11p13, 11q24.1, 14q11.2, 14q32.2, 16p13.2, 1p32.1, 1q41, 1q41, 20q13.32, 2p12, 2q37.3, 4q22.1, 4q24, 8p23.1, 9p21.3, 9q33.1, 9q33.2	*TRIM44*, *ZNF202*, *A2BP1*, *FGGY*, *MOSC2*, *ZNF831*, *KIAA0914*, *intergenic*	[23]
20876420	Early breast cancer	Aromatase inhibitors	AE – musculoskeletal	14q32.13	*TCL1A*	[178]
18650507	Dyslipidemia	Statin	AE – myopathy	12p12.1	*SLCO1B1*	[84]
19724244	Depression	Citalopram	AE – suicidal ideation	‡	‡	[32]
21386754	Cardiovascular disease	Cerivastatin	AE – rhabdomyolysis	1q43	*RYR2*	[179]
21659334	Chronic hepatitis C	Pegylated interferon and ribavirin (PEG-IFN/RBV)	AE – thrombocytopenia	Chr20	*DDRGK1*, *ITPA*	[180]
21396408	Influenza	Influenza vaccine	AE – wheezing, influenza	1q23.2, 7p11.2		
18195134	Multiple sclerosis	IFN-β therapy	Efficacy – relapse, expanded disability status score	‡	‡	[181]

†Data downloaded from the National Human Genome Research Institute genome-wide association studies catalog on 17 July 2011 [201].
‡National Human Genome Research Institute genome-wide association studies catalog has no information, proabably because no statistically significant loci were reported.
AE: Adverse event; BP: Blood pressure; DAS28: Disease activity score; HDL: High-density lipoprotein; LDL: Low-density lipoprotein; NAPQI: N-acetyl-p-benzoquinone imine.

Table 2. Suggestive genetic loci in pharmacogenomic genome-wide association studies based on National Human Genome Research Institute genome-wide association studies catalog† (cont.).

PubMed ID	Disease or medical condition	Therapy	End point	Loci at association $p < 1 \times 10^{-5}$	Reported candidate genes	Ref.
19667218	Multiple sclerosis	IFN-β therapy	Efficacy – relapse, expanded disability status score	‡	‡	[182]
21502966	Multiple sclerosis	IFN-β therapy	Efficacy – development of neutralizing antibodies	8q24.3, 6p21.32	*DENND3, PTK2, HLA-DQA1, HLA-DRB1*	[183]
18591461	Hypertension	Thiazide diuretic	Efficacy – diastolic BP response	12q15	*LYZ, YEATS4, FRS2*	[184]
18615156	Rheumatoid arthritis	Anti-TNF treatment	Efficacy – change in DAS28	1p22.3, 20p11.21, 20q12, 2q24.3, 4p15.1, 6q26, 7q21.3, 9p21.2	*LMO4*, *CST5*, *MAFB*, *LASS6*, *CENTD1*, *QKI*, *PON1*, *IFNK*	[185]
21061259	Rheumatoid arthritis	Anti-TNF treatment	Efficacy – change in 28 joint count DAS28			[186]
19176441	Childhood acute lymphoblastic leukemia	Multiple	Efficacy – minimal residual disease	10p12.33, 10p14, 10q26.12, 11p15.1, 11q21, 20q13.12, 2q33.1, 4q31.21, 5p13.2, 6q25.3, 7p14.2, 7p21.2	*ST8SIA6*, *intergenic*, *intergenic*, *intergenic*, *MAML2*, *NCOA3*, *C2orf47*, *IL15*, *LMBRD2*, *intergenic*, *ELMO1*, *DGKB*	[187]
19448189	Recurrence in bipolar disorder	Lithium	Efficacy – hazard for mood episode recurrence	‡	‡	[188]
19721433	Schizophrenia	Antipsychotics	Efficacy – Positive and Negative Syndrome Scale	12q23.1, 15q13.3, 1q21.3, 2q14.3, 3q28, 4p15.1, 6p21.33, 6p24.1, 9q33.3	*ANKS1B*, *TRPM1*, *CNTNAP5*	[28]
18521090	Schizophrenia	Iloperidone	Efficacy – antipsychotic response	‡	‡	[26]
21107309	Schizophrenia	Antipsychotics	Efficacy – neurocognitive measures	11p13, 1q32.1	*EHF*, *SLC26A9*	[27]
20360315	Depression	Antidepressants	Efficacy – improvement of depression severity	10p12.31, 18q12.1, 19p13.11, 19q13.42, 1p22.2, 1q25.3, 6q25.1	*SLC27A1*, *IL11*, *RGL1*, *UST*	[33]
19736353	Depression	Antidepressants	Efficacy – beneficial treatment outcome	‡	‡	[31]
19846067	Major depression	Citalopram	Efficacy – remission	15q22.2, 18q12.1, 20q13.31, 21q21.3, 7q36.3	*RORA*, *NOL4*, *BMP7*, *EIF4A1P*, *UBE3C*	[30]
20031582	Dyslipidemia	Statin	Efficacy – HDL, LDL, triglyceride	‡	‡	[38]

†Data downloaded from the National Human Genome Research Institute genome-wide association studies catalog on 17 July 2011 [201].
‡National Human Genome Research Institute genome-wide association studies catalog has no information, proabably because no statistically significant loci were reported.
AE: Adverse event; BP: Blood pressure; DAS28: Disease activity score; HDL: High-density lipoprotein; LDL: Low-density lipoprotein; NAPQI: N-*acetyl*-p-*benzoquinone imine.*

Table 2. Suggestive genetic loci in pharmacogenomic genome-wide association studies based on National Human Genome Research Institute genome-wide association studies catalog (cont.).

PubMed ID	Disease or medical condition	Therapy	End point	Loci at association $p < 1 \times 10^{-5}$	Reported candidate genes	Ref.
20339536	Dyslipidemia	Statin	Efficacy – changes in LDL, total cholesterol, HDL and triglyceride	10p15.3, 10q21.3, 11q12.2, 11q12.2, 12p12.1, 13q31.1, 14q32.13, 16q24.1, 1p13.3, 1p32.3, 1p36.32, 1q31.2, 1q42.2, 20q13.31, 21q21.3, 2p21, 2p24.1, 2q14.3, 2q22.3, 2q24.1, 2q24.3, 2q37.2, 3p26.3, 3p26.3, 4q35.2, 4q35.2, 6p22.3, 6q13, 6q15, 7p21.1, 7p21.2-p21.1, 7q11.21, 8p12, 8q24.13, 9q21.33, 9q33.2	*LOC727878, LOC728209, LOC441546, ANXA2P3, LOC645084, RPL7AP1, FADS2, FADS1, FADS3, FADS2, FADS1, FEN; SOX5, FLJ32894, BCAT1, LOC390415, LOC647298, C13orf7, CLMN, FLJ45244, DICER1, LOC401864, LOC283904, LOC729464, CELSR2, PSRC1, SORT1, GLIS1, DMRTB1, FLJ40434, PRDM16, ARHGEF16, MEFGF6, FAM5C, LOC647132, DISC1, SIA1L2, DISC2, HMG1L1, CTCFL, RBM38, GRIK1, CLDN17, CLDN8, TTC7A,MCFD2, FLJ40172, APOB, FLJ21820, GDF7, CNTNAP5, ACVR2A, ORC4L, LOC647065, NR4A2, GPD2, LOC728038, COBLL1, LOC728184, GRB14, ASB18, LOC728087, IQCA, CNTN6, LOC402123, CHL1, LOC402123, CNTN6, CHL1, LOC544042, F11, KLKB1, LOC644282, LOC644325, MRPS36P2, LOC729105, ID4, MBOAT1, BAI3, LMBRD1, COL19A1, RNGTT, ACTBP8, LOC644119, LOC402642, ABCB5, SP8, SOSTDC1, LOC442511, LOC729920, ZNF679, LOC728927, LOC442320, LOC646909, DUSP4, KIF13B, TRIB1, NSMCE2, KIAA0196, SLC28A3, RMI1, LOC729388, MIRN147, CDK5RAP2, MEGF9*	[39]
20463552	Small-cell lung cancer	Platinum-based chemotherapy	Efficacy	‡	‡	[189]
21483023	Advanced stage non-small-cell lung cancer	Platinum-based chemotherapy	Efficacy – overall survival	12q23.3	*CMKLR1*	[190]
20923822	Human lymphoblastoid cell lines	Radiation	Efficacy – radiation cytotoxicity	10p15.3, 14q32.2, 16q23.1, 1q43, 2p16.3, 6p12.1, 6p22.3, 6q25.1, 7q21.3, 8q22.1, 9p24.1	*ADARB2, LOC100132612, WWOX, WDR64, FSHR, TINAG, KIAA0319, ULBP1, DLX6AS, PLEKHF2, KDM4C*	[191]
21186350	Type 2 diabetes	Metformin	Efficacy – glycemic response	11q22.3	*ATM, C111orf65*	[192]
21659360	Chronic lymphocytic leukemia	Fludarabine, chlorambucil and fludarabine with cyclophosphamide	Efficacy – progression-free survival	Chr1, 3, 4, 8, 9, 14	*CENPF, DISC1, intergenic, C4orf23, ACOX3, intergenic, intergenic, intergenic*	[193]
21177773	Pain	Acetaminophen-NAPQI	Hepatotoxicity	1q23.3, 1q32.1, 3p11.1, 4q34.1, Xq28	*LMX1A, ETNK2, C3orf38, KIAA1712, LOC100129661*	[194]

†*Data downloaded from the National Human Genome Research Institute genome-wide association studies catalog on 17 July 2011* [201].
‡*National Human Genome Research Institute genome-wide association studies catalog has no information, proabably because no statistically significant loci were reported.*
AE: Adverse event; BP: Blood pressure; DAS28: Disease activity score; HDL: High-density lipoprotein; LDL: Low-density lipoprotein; NAPQI: N-*acetyl*-p-*benzoquinone imine.*

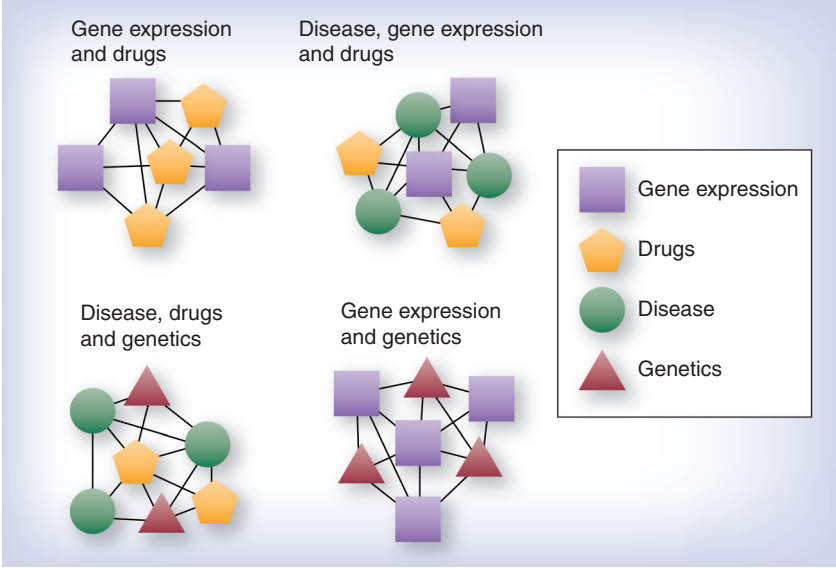

Figure 2. Networks from different data types useful in pharmacogenomics. Drug, disease, gene expression and genetic data generated in different combinations in coherent data sets produce reference networks that can be accessed by unique queries and inform on distinct biological relationships. Specific examples of each network type, methods used to generate them, and references are provided in the text. Gene expression and genetics networks include the weighted gene coexpression network analysis and Bayesian network approaches described under 'Network biology', and the remainder are described in the 'Other integrative genomics approaches to data' section.

of the biological systems based on the data, and priors, in the case of BN, that they are given. Hence, they are an ideal compliment to the specific hypotheses based on prior experience that all scientists carry into an analysis.

Integrative genomics of the human liver

As liver is the most relevant tissue for the pharmacology of many drugs, functional variations and molecular networks of this tissue are of particular interest for pharmacogenomics. In two recent studies, Schadt *et al.* [42] and Yang *et al.* [82] investigated the genetic architecture of liver gene expression, the enzyme activities of CYPs, and the network properties using approximately 500 human liver samples. Through GGE analysis, more than 3000 eQTL for over 6000 distinct liver genes were reported, among which hundreds of genes encode drug-metabolizing enzymes and transporters. In addition to GGE analysis, the genetics of the activity measures of nine key drug-metabolizing P450 enzymes (CYP1A2, CYP2A6, CYP2B6, CYP2C8, CYP2C9, CYP2C19, CYP2D6, CYP2E1 and CYP3A4) were also studied. With the exception of CYP2E1, each of these P450 enzymes showed variation in activity as a function of genotype in this sample, and these relationships defined 54 activity SNPs in this cohort. Many eSNPs and activity SNPs identified represent novel discoveries. For instance, three novel long-range SNPs were found to be associated with both the expression and the enzyme activity measurements of CYP2D6 in this liver cohort and were confirmed to be associated with the metabolism of dextromethorphan *in vivo* in an independent human cohort. These eSNPs and activity SNPs discovered from the human liver studies are of importance for pharmacogenomics as they help to understand the impact of individual genetic variants on drug-metabolizing enzymes, transporters and liver drug targets and in addition, help identify plausible candidate genes underlying the genetic loci associated with pharmacogenomic outcomes.

Using the WGCNA and BN network methodologies and the genetic and gene-expression data, coexpression and BN networks have been constructed from the same human liver cohort to illustrate the gene regulatory network structure in the human liver [82]. A total of eight coexpression network modules comprised of genes with similar biological functionalities were identified and genes involved in particular functional categories such as immune response, cell cycle and metabolic pathways were enriched in several modules. Four modules enriched for P450 genes, genes involved in oxidative stress and apoptosis, acute-phase response genes and translation-related genes were found to be highly correlated with P450 expression and activity. Through integration of GGE, two of the four modules were implicated as upstream regulatory modules for P450s. These modules and the genes within them provide additional insights into the regulation of P450s. In addition, BN subnetworks which incorporated both genetic and gene-expression information helped identify a P450 regulatory BN subnetwork. Novel candidate P450 regulatory genes including *EHHADH*, *SLC10A1* and *AKR1D1* were highlighted from the analyses of both networks. Although *in silico* validation of the P450 regulatory subnetwork has been presented, the novel regulators nominated are still under experimental validation.

In addition to gene regulatory networks, metabolic networks of the liver and hepatocytes have also been recently constructed based on literature, transcriptomic, proteomic, metabolomic and phenotypic data to help understand liver metabolism and physiology [111–113]. These liver networks provide a framework for understanding how a given gene interacts with other genes and how together these genes may impact

biological functions such as P450 activity and hepatic metabolism of compounds, thereby providing mechanistic insights into individual genes and pathways.

Other integrative genomics approaches to data

As these liver metabolism network studies demonstrate, networks useful in an integrative genomics strategy such as described for human liver are by no means limited to GGE and genetics of drug response traits. In fact, almost any combination of sufficiently rich, coherent and orthogonal data in comparable subjects may be used to derive a biological network model that could be useful when referenced to interpret a pharmacogenomic experiment (FIGURE 2). Most obvious are networks that incorporate the effect of drugs on the expression of genes. Early surveys of the expression of compendia of compounds on gene expression have utility and describe basic networks, even if their results were at first not expressly modeled as networks. Examples of this category include a compendium of chemical signatures as well as individual toxin signatures in yeast [114,115] and the effect of xenobiotic compounds in rat liver [116,117]. These early efforts clearly demonstrated that drugs have an impact on gene expression and that they could be classified by that expression. Subsequent publications referencing these early networks demonstrated the value of referencing them to draw conclusions regarding basic biology and drug action [55,118–120].

More recently, systematic efforts to screen drugs in mammalian cells and model them in network-based ways have become a reality. The NCI made substantial efforts in this area early on by developing a panel of 60 human tumor cell lines representing tumors from nine tissues and funding diverse experiments to characterize them with many molecular techniques under diverse experimental conditions [121,122]. These cell lines were also SNP genotyped, allowing genome-wide association studies with all the molecular traits measured in the NCI60 cell lines, and similar association studies have been performed with cellular phenotypes measured in the HapMap human cell lines used to benchmark human genetic variation [123]. The end result of all of this investment has been networks encompassing phenotypic response

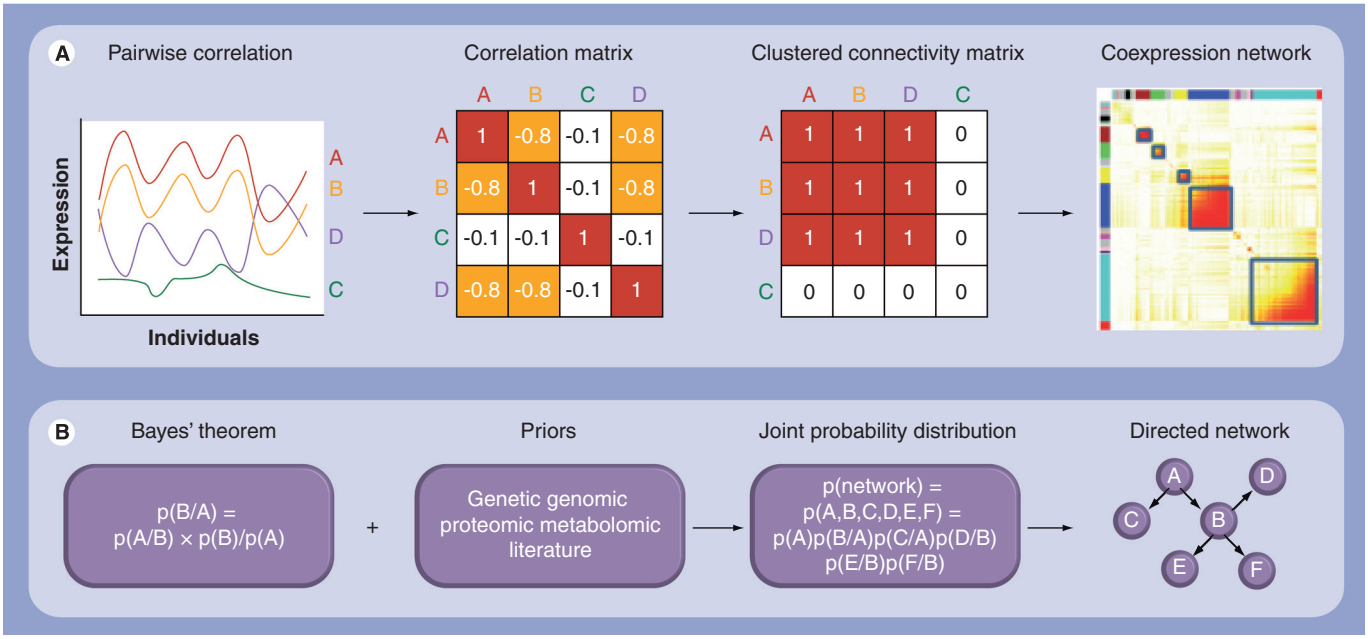

Figure 3. Weighted gene coexpression networks analysis and Bayesian network analysis. (A) Weighted gene coexpression network analysis (WGCNA) workflow. Pairwise correlations between genes (shown as genes A, B, C and D in this example) are first calculated to construct a correlation matrix across all gene pairs. A connection matrix is then derived from the correlation matrix by defining connectivity using a correlation threshold. For example, at a correlation coefficient of absolute value of over 0.6, all gene pairs that reach this threshold are defined as being connected. The connectivity matrix is then clustered and a dynamic cut-tree algorithm applied to construct the coexpression network and define network modules. Each network module consists of genes that are highly interconnected with one another. **(B)** Bayesian network (BN) workflow. Joint probability distribution of genes based on the conditional dependencies is calculated to identify a maximum likelihood network given the observed data. Prior information such as genetic regulation, transcription factor binding and protein–protein interaction can then be incorporated to infer directionality between genes. Examples of WGCNA and BN analysis applied to biological data sets and references to the analysis methods themselves are given in the text.

to many natural products, experimental compounds, US FDA-approved anticancer agents, gene expression and genetics that has in turn shown relationships between genetic loci, proliferation and viability phenotypes, and drugs such as statins and paclitaxel in these cells [124,125].

Among the most popular of these efforts is the connectivity map (cmap), a systematic effort to profile the effect of compounds in human cell lines and then make the results accessible to a broad community of bioinformaticians and bench laboratory researchers [118,126]. A current download of the cmap database contains information on 1309 compounds in five cancer cell lines, for a total of 6100 combinations, each assayed by gene-expression microarray. Although the genetic and phenotypic diversity of the cell lines is limited, the number of compounds profiled and the accessibility of the data are unprecedented. Accordingly, the cmap has featured in publications seeking to understand the mechanisms by which drugs act and their linkage to disease phenotypes and biological processes. For example, the map has been used to support mechanistic studies of trastuzumab desensitization in breast cancer and extended to identify combinatorial effects between retinoids and histone deacetylase inhibitors in promoting differentiation and increasing survival in xenograft models of neuroblastoma [127,128]. However, the cmap networks are valuable outside of cancer as well. For instance, Loboda *et al.* intersected signatures of natural diurnal variation in metabolism, the effects of the weight loss drug sibutramine and genes repressed by two mTOR inhibitors in the cmap, supporting the relationship between sibutramine and diurnal variation in metabolism and identifying a previously unappreciated relationship between mTOR signaling and circadian variation in metabolism [129]. This intersection of drug signatures from cmap with other gene-expression signatures and networks is one of the most promising applications of the cmap data since it can test the ability of the highly controlled, cell line cmap data for a broad collection of drugs to generalize to whole organisms under less controlled conditions. One can imagine that the combination of modern, comprehensive cmap drug signatures with eQTL linking to a particular locus should yield results even more informative than those discriminating the effects of rosiglitazone and a *Alox5*-targeted mutation through intersection with metabolic QTL and associated eQTL in a mouse cross [55]. Similar benefit would be expected by intersecting cmap, NCI60 and HapMap cell line data with many of the other recently developed network models discussed in this review. Since many cmap-derived results have yet to be widely replicated, it will be interesting over time to catalog which cellular phenotypes and networks underlying them are robust across laboratories and which consistently reproduce biology of whole tissues and organisms; key questions that define the utility of any laboratory or computational model of biology.

Networks potentially useful for pharmacogenomics have also been built around diseases and symptoms. Initially, these were text-mining applications on semi-structured databases such as Online Mendelian Inheritance in Man [130,131], and Medical Subject Headings [132]. Because of their less quantitative nature they were useful in generating hypotheses regarding links between diseases, disease mechanisms and potential therapeutic approaches. However, as more quantitative and systematic data has become accessible for each disease, more quantitative approaches to describe relationships between diseases that leverage annotated gene-expression profiles and combinations of annotated gene-expression profiles with protein–protein interaction data have become available [133,134]. These most recent efforts show robust statistical relationships between networks associated with particular diseases and the drugs used to treat them, and provide testable, quantitative hypotheses about which drugs will be useful for which diseases.

These disease networks, however, at present lack a description of the individual-level variability we customarily think about when discussing pharmacogenomics since they, by definition, treat each disease as a coherent whole. This is beginning to change, however, in two ways. First, as more molecular and clinical data are available for patients on each disease, it is becoming possible to subtype diseases more precisely. For instance, cardiovascular health is now routinely assessed through the measurement of high-density lipoprotein cholesterol, low-density lipoprotein cholesterol, C-reactive protein, blood glucose and HBA1c, triglycerides and blood pressure [135,136]. It may even include an imaging component if disease is suspected [137]. In total, this data, measured over time, represents far more information than was routinely available on patients in the past, and can in theory be used to identify patient populations with distinct subtypes of cardiovascular disease. However, it is clear that most studies intended to test enhancements to standard risk prediction tools such as the Framingham Risk Score have to date not

been performed with designs rigorous enough to exploit this information on its own [138], underscoring the potential value of analyzing clinical measurements in the context of genetic information and network models derived from molecular measurements across populations. In cancer, where important molecular subtypes of cancers have long been recognized, there is a similar growth in the molecular and other information available on tumors [139–141]. Second, obtaining a good picture of a patient's genetic information is becoming increasingly routine. This has long been instrumental in the characterization of Mendelian disorders, but is now routine for common diseases such as Alzheimer's disease, where the *ApoE* genotype has long been appreciated to be strongly associated with age of onset and course of disease [142], and recent GWAS have now implicated a number of loci that appear relevant to immune function, cell membrane function and lipid metabolism [143,144]. The recent advent of inexpensive genome sequencing will also make such determinations possible for many diseases and be a significant aid to diagnosis, as has already been shown in the genetic differential diagnosis of Bartter's syndrome and congenital chloride diarrhea [145]. A combination of multifactorial molecular, clinical and genetic parameters measured in an individual patient generates a comprehensive profile of that patient. As data and network models grows more complete and better organized, this suggests an opportunity to compare that individual's profile to a reference collection of network models built from population-level data to derive increasingly tailored lifestyle and therapeutic interventions for that individual (FIGURE 4).

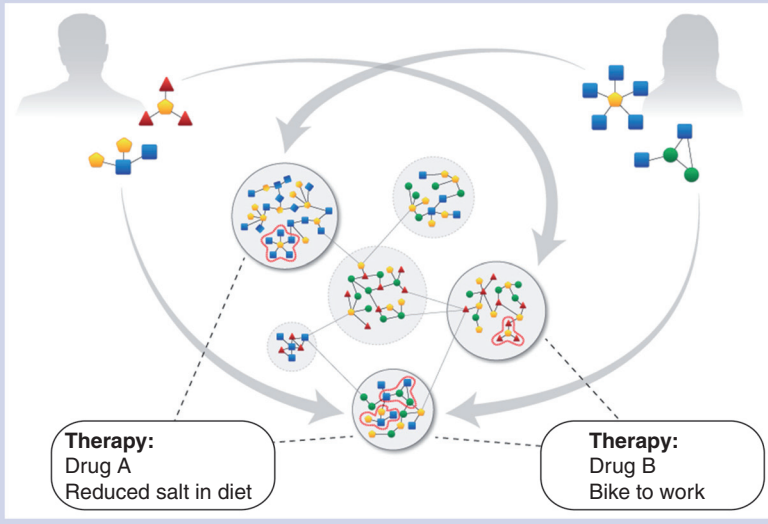

Figure 4. Using genetic, clinical and environmental data for individualized risk assessment and therapy. Measurements in individuals of biological network states using acute laboratory tests or longitudinal mining of clinical data will reveal motifs that map to known disease- and treatment-associated network modules determined from population-based studies. Here, the motifs are shown as the small networks next to the individuals in the upper corners of the figure, and they map to subnetworks outlined in red within the encircled reference modules at the center of the figure. These reference modules were derived from studies of populations that did not include the individuals shown in this figure. This mapping of motif to module subnetwork can then in principal be used to guide individualized recommendations for pharmacological and lifestyle interventions.

Nowhere is the potential impact of integrative genomics approaches to pharmacogenomics greater than in cancer. Tumors are frequently accessible for molecular investigation, and their great diversity in symptoms and response to therapy provide strong motivation to apply pharmacogenomics approaches. Hence, many projects are underway across the globe to characterize, at the molecular level, many of these individual tumor types [146,147]. A hint of where this rich information may lead is provided by a case study in glioblastoma. Early work based entirely on gene-expression profiling of high-grade gliomas led to the identification of three subtypes, of which the mesenchymal subtype had the worst prognosis, with post-treatment survival as the end point [148,149]. Application of the Algorithm for the Reconstruction of Accurate Cellular Networks to define modules of genes related by coexpression, followed by the Master Regulator Inference algorithm to search for transcriptional regulators of these modules identified a hierarchical grouping of transcription factors that controlled more than 74% of the transcription in the mesenchymal signature [150]. This regulatory structure was then supported by chromatin immunoprecipitation in a human glioma cell line, gain and loss of function experiments in cell lines and tumor production in mice. Importantly, high expression of master regulators of this mesenchymal signature, Stat3 and C/EBPb, increased markers of tumor aggressiveness in all these assays and decreased survival of human glioma patients [150]. With several other glioblastoma data sets containing coherent molecular and clinical information now available, there will be many opportunities to refine the structure of these predictive networks and test their reproducibility across different sets of patients [149,151,152]. These genes will undoubtedly also be investigated by copy number variation analysis in glioblastomas, and the Cancer Targeted Discovery and Development Network (CTD2) program is seeking small molecule validation of these targets as well [146]. Impressive as these results are, the data on which they are based is scant and imprecise in comparison to

what is coming. Already, it is cost effective to sequence adjacent normal and tumor tissue for some research studies [205], something that was a noteworthy achievement as recently as 2010 [153], and collections of hundreds of cancer genomes have been analyzed in combination to identify networks of genes not previously implicated in distinct cancer types [154,155]. This rich information will surely allow us to make much better predictions about which of an increasing array of targeted cancer therapies are appropriate for a given cancer patient. The difficulty of getting clinical benefit from such predictions is not to be underestimated, however. First, these results have not yet been widely confirmed in many independent cohorts. Second, many factors combine to make introduction of genomic information into clinical settings a challenging and time consuming process in general, and the huge diversity of individual tumors make each case a unique challenge to each individual patient and his or her physicians that require detailed and comprehensive investigation to understand, interpret and present [156,157].

Conclusion & future perspective

At present, there are no real examples of moving prospectively from biological network approaches to diversity of drug response in humans. This contrasts with many strong examples of prospective drug-target discovery by network methods and is certainly a reflection of the fact that the clinical studies take time to complete. However, our ability to access large amounts of clinical data in a structured way to rigorously test hypotheses stemming from integrative genomics approaches in the clinic is improving rapidly and should become routine in at least some cases within the next 5–10 years. Even modest improvements in data capture, such as those investigating the pharmacogenomics of cisplatin-induced deafness, hydrocodone toxicity and other drugs through a focused effort in Canada clearly shows the benefit of a focused search for drug reactions in an ambulatory population [158–161]. Although moving more slowly than many would like, there is a global trend towards increased electronic data capture into electronic medical records (EMRs) and higher quality of the coding in those EMRs [162,163]. As more and more EMR data become available, it will be a well-powered resource for mining in combination with the increasingly precise and voluminous molecular data made cheaply available by advances in sequencing technology.

One widely quoted estimate is that over 10,000 human genomes will have been sequenced worldwide by the end of 2011, in comparison with tens of human genomes at the end of 2009 [164], and that genome sequencing will start to make an appearance in clinical diagnosis [145,165,166]. In combination with what is likely to be a substantially even greater number of RNA-Seq experiments as well as existing volumes of other high information content medical data such as images, EEGs and ECGs, this represents a vast set of data to be mined. Add to that the likely advances in all manner of sensor technologies and potentially better descriptions of social and environmental interactions that leverage our increasing documentation of our activities online, and there is clearly a demand to apply both the lessons in data management and mining that have been learned from large internet firms and to develop analytical and data mining methods specific to biological and health data [167–169,206,207]. Although they will surely be viewed as infant steps a decade from now, we are nevertheless seeing progress in both areas. For instance, an analysis optimized to minimize the multiple-testing issues of a GWAS demonstrated well over an order of magnitude improvement in the ratio of true to false-positive marker-trait associations [170]. Similarly, mining even a relatively small hospital's EMR demonstrated a clear retrospective association between rosiglitazone and COX-2 inhibitors and adverse cardiovascular outcomes [171,172]. Clearly, the potential benefits of postmarket surveillance in the coming information-rich heathcare environment are on par with the potential benefits in discovery of new therapeutic modalities and their individualized and cost-effective application to appropriate patient subpopulations. Properly harnessed, better capture and open sharing of clinical measurements, environmental conditions, social circumstances and molecular phenotypes of patients hold out the possibility of better health and safer, more effective therapies.

What stands in the way of this vision? Although there are technological challenges to delight legions of clever people in fields as diverse as computer science, clinical medicine, pharmacology, genetics, economics, epidemiology, nanotechnology and statistics, it is our opinion that those will be overcome in a reasonably efficient way if sufficient data is openly available to address challenges as they appear. Rather, the main barriers to realizing the promise of truly integrating genomic technologies with information-rich medical practice will be social. Data from research studies need to be captured and accessible for reuse

by future investigators, but this is neither widely instilled across the academic community nor funded by most research sponsors. The security and rights of individuals cannot be compromised if their medical, molecular and other data are used in research, and this is not easy to achieve, let alone harmonize internationally to support the sort of global collaborative efforts that are a natural consequence of intersecting data with whatever other information is most constructive in gaining new insight. Perhaps most difficult is the fact that humans have never before been confronted with so much definite information on health and prognosis as the coming data and decision support systems built on it will eventually provide. Figuring out what interventions to pursue as a consequence of this information and determining who pays for it all in the fragmented healthcare delivery systems across much of the world are nontrivial undertakings. These social challenges do not have to be resolved all at once or immediately; however, the benefits of a systems approach to pharmacotherapy will be paced by how successfully they are addressed.

Acknowledgements

The authors would like to thank Kristin Robertshaw for help in the design and construction of the figures.

Financial & competing interests disclosure

X Yang has no relevant affiliations or financial involvement with any organization or entity with a financial interest in or financial conflict with the subject matter or materials discussed in the manuscript. A Kasarskis and E Schadt are employees of Pacific Biosciences and own stock in the company. The authors have no other relevant affiliations or financial involvement with any organization or entity with a financial interest in or financial conflict with the subject matter or materials discussed in the manuscript apart from those disclosed.

No writing assistance was utilized in the production of this manuscript.

Executive summary

Full mechanisms behind pharmacogenomic effects are often opaque

- Although there are well-understood examples of pharmacogenetic effects with drugs such as clopidogrel, warfarin, abacavir, flucloxacilin and others, it is frequently difficult to obtain a solid understanding of the full mechanism by which genetics affects the action of a drug.
- Even when a mechanism is understood, there is other variation in drug response that is not explainable by that mechansim, so there is an opportunity to apply integrative genomics approaches that leverage relationships frequently observed between genes, drugs, diseases and other concepts in population studies.

Fuctional genomics & network biology

- Genetics of gene expression and other techniques for relating genetic variation and gene level or activity often suggest mechanisms underlying associations between genetic variation and drug effects in genome-wide association studies. This has been observed for warfarin, lumiracoxib, flucloxacilin and statins, demonstrating that the genetics of gene expression has general utility to probe pharmacogenetic associations.
- Incorporation of Bayesian and weighted gene coexpression network analysis methods routinely used to structure the analysis and application of functional genomics data enhances the interpretation of disease-related genome-wide association study results.

Integrative genomics of the human liver

- Networks derived from genetics of gene expression in human liver demonstrate many genes robustly associated with the function and activity of CYP450s.
- Human and mouse networks, used in combination, identify previously unappreciated potential regulators of CYP450s and can be used to address other liver-expressed genes of pharmacogenomic interest in an analogous manner.

Other integrative genomics approaches to data

- Networks built from other data combinations are also tools with the potential to improve our understanding of any given drug response.
- Notable examples include systematic cell-based connectivity map screening approaches that marry drug effects on cells with gene expression in those cells, copy number variation analysis of tumors with gene expression and clinical end points and database mining of clinical records and the published literature in combination with public molecular data repositories.

Conclusion & future perspective

- Although functional genomics approaches have shown benefit in understanding pharmacogenetic associations for several drugs, at present we lack examples of going prospectively from biological network models to diversity of drug response in humans. This contrasts with many examples of target discovery, and we anticipate that prospective evidence for the utility of these approaches to drug response will emerge as the lengthy process of clinical testing progresses.
- Data relevant to pharmacogenomics will become ubiquitous with improved clinical data capture in electronic medical records, affordable personal genomes and the continued development of new technologies for assaying and imaging patients in a near-continuous fashion.
- Fortunately, better analytics coupled to increasingly available computing power will enable better discrimination of informative patterns in all this data and their application to clinical practice.
- If coupled with open-data availability and sharing of materials needed to replicate laboratory results, these trends promise not only more effective use of existing therapeutics, but also less expensive and more rapid development of new therapies, including conventional drugs as well as, or in combination with, other therapeutic modalities.

References

1 Evans DA, Manley KA, McKusick VA. Genetic control of isoniazid metabolism in man. *Br. Med. J.* 2(5197), 485–491 (1960).

2 Vesell ES, Page JG. Genetic control of drug levels in man. Antipyrine. *Science* 161(836), 72–73 (1968).

3 Vesell ES, Page JG. Genetic control of drug levels in man. Phenylbutazone. *Science* 159(3822), 1479–1480 (1968).

4 Pare CM, Rees L, Sainsbury MJ. Differentiation of two genetically specific types of depression by the response to antidepressants. *Lancet* 2(7270), 1340–1343 (1962).

5 Pare CM, Mack JW. Differentiation of two genetically specific types of depression by the response to antidepressant drugs. *J. Med. Genet.* 8(3), 306–309 (1971).

6 O'Reilly RL, Bogue L, Singh SM. Pharmacogenetic response to antidepressants in a multicase family with affective disorder. *Biol. Psychiatry* 36(7), 467–471 (1994).

7 Franchini L, Serretti A, Gasperini M, Smeraldi E. Familial concordance of fluvoxamine response as a tool for differentiating mood disorder pedigrees. *J. Psychiatr. Res.* 32(5), 255–259 (1998).

8 Wang L, McLeod HL, Weinshilboum RM. Genomics and drug response. *N. Engl. J. Med.* 364, 1144–1153 (2011).

9 Zambon CF, Pengo V, Padrini R *et al.* *VKORC1*, *CYP2C9* and *CYP4F2* genetic-based algorithm for warfarin dosing. an Italian retrospective study. *Pharmacogenomics* 12(1), 15–25 (2010).

10 Esmerian MO, Mitri Z, Habbal MZ *et al.* Influence of *CYP2C9* and *VKORC1* polymorphisms on warfarin and acenocoumarol in a sample of Lebanese people. *J. Clin. Pharmacol.* 51(10), 1418–1428 (2011).

11 Takeuchi F, Mcginnis R, Bourgeois S *et al.* A genome-wide association study confirms *VKORC1*, *CYP2C9*, and *CYP4F2* as principal genetic determinants of warfarin dose. *PLoS Genet.* 5(3), e1000433 (2009).

12 Limdi NA, Arnett DK, Goldstein JA *et al.* Influence of CYP2C9 and VKORC1 on warfarin dose, anticoagulation attainment and maintenance among European–Americans and African–Americans. *Pharmacogenomics* 9(5), 511–526 (2008).

13 Scott SA, Edelmann L, Kornreich R, Desnick RJ. Warfarin pharmacogenetics. *CYP2C9* and *VKORC1* genotypes predict different sensitivity and resistance frequencies in the Ashkenazi and Sephardi Jewish populations. *Am. J. Hum. Genet.* 82(2), 495–500 (2008).

14 Hetherington S, Hughes AR, Mosteller M *et al.* Genetic variations in *HLA-B* region and hypersensitivity reactions to abacavir. *Lancet* 359(9312), 1121–1122 (2002).

15 Mallal S, Nolan D, Witt C *et al.* Association between presence of HLA-B*5701, HLA-DR7, and HLA-DQ3 and hypersensitivity to HIV-1 reverse-transcriptase inhibitor abacavir. *Lancet* 359(9308), 727–732 (2002).

16 Daly AK, Donaldson PT, Bhatnagar P *et al.* *HLA-B*5701* genotype is a major determinant of drug-induced liver injury due to flucloxacillin. *Nat. Genet.* 41, 816–819 (2009).

17 Ahlenstiel G, Booth DR, George J. *IL28B* in hepatitis C virus infection. translating pharmacogenomics into clinical practice. *J. Gastroenterol.* 45(9), 903–910 (2010).

18 Ge D, Fellay J, Thompson AJ *et al.* Genetic variation in *IL28B* predicts hepatitis C treatment-induced viral clearance. *Nature* 461(7262), 399–401 (2009).

19 Suppiah V, Moldovan M, Ahlenstiel G *et al.* IL28B is associated with response to chronic hepatitis C interferon-alpha and ribavirin therapy. *Nat. Genet.* 41(10), 1100–1104 (2009).

20 Tanaka Y, Nishida N, Sugiyama M *et al.* Genome-wide association of *IL28B* with response to pegylated interferon-alpha and ribavirin therapy for chronic hepatitis C. *Nat. Genet.* 41(10), 1105–1109 (2009).

21 Manolio TA. Genomewide association studies and assessment of the risk of disease. *N. Engl. J. Med.* 363(2), 166–176 (2010).

22 Hindorff LA, Sethupathy P, Junkins HA *et al.* Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proc. Natl Acad. Sci. USA* 106(23), 9362–9367 (2009).

23 Aberg K, Adkins DE, Bukszar J *et al.* Genomewide association study of movement-related adverse antipsychotic effects. *Biol. Psychiatry* 67(3), 279–282 (2010).

24 Adkins DE, Aberg K, McClay JL *et al.* Genomewide pharmacogenomic study of metabolic side effects to antipsychotic drugs. *Mol. Psychiatry* 16(3), 321–332 (2011).

25 Alkelai A, Greenbaum L, Rigbi A, Kanyas K, Lerer B. Genome-wide association study of antipsychotic-induced parkinsonism severity among schizophrenia patients. *Psychopharmacology (Berl.)* 206(3), 491–499 (2009).

26 Lavedan C, Licamele L, Volpi S *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* 14(8), 804–819 (2009).

27 McClay JL, Adkins DE, Aberg K *et al.* Genome-wide pharmacogenomic study of neurocognition as an indicator of antipsychotic treatment response in schizophrenia. *Neuropsychopharmacology* 36(3), 616–626 (2011).

28 McClay JL, Adkins DE, Aberg K *et al.* Genome-wide pharmacogenomic analysis of response to treatment with antipsychotics. *Mol. Psychiatry* 16(1), 76–85 (2011).

29 Volpi S, Heaton C, Mack K *et al.* Whole genome association study identifies polymorphisms associated with QT prolongation during iloperidone treatment of schizophrenia. *Mol. Psychiatry* 14(11), 1024–1031 (2009).

30 Garriock HA, Kraft JB, Shyn SI *et al.* A genomewide association study of citalopram response in major depressive disorder. *Biol. Psychiatry* 67(2), 133–138 (2010).

31 Ising M, Lucae S, Binder EB *et al.* A genomewide association study points to multiple loci that predict antidepressant drug treatment outcome in depression. *Arch. Gen. Psychiatry* 66(9), 966–975 (2009).

32 Laje G, Allen AS, Akula N, Manji H, John Rush A, McMahon FJ. Genome-wide association study of suicidal ideation emerging during citalopram treatment of depressed outpatients. *Pharmacogenet. Genomics* 19(9), 666–674 (2009).

33 Uher R, Perroud N, Ng MY *et al.* Genome-wide pharmacogenetics of antidepressant response in the GENDEP project. *Am. J. Psychiatry* 167(5), 555–564 (2010).

34 Cha PC, Mushiroda T, Takahashi A *et al.* Genome-wide association study identifies genetic determinants of warfarin responsiveness for Japanese. *Hum. Mol. Genet.* 19(23), 4735–4744 (2010).

35 Cooper GM, Johnson JA, Langaee TY *et al.* A genome-wide scan for common genetic variants with a large influence on warfarin maintenance dose. *Blood* 112(4), 1022–1027 (2008).

36 Teichert M, Eijgelsheim M, Rivadeneira F *et al.* A genome-wide association study of acenocoumarol maintenance dosage. *Hum. Mol. Genet.* 18(19), 3758–3768 (2009).

37 Shuldiner AR, O'Connell JR, Bliden KP *et al.* Association of cytochrome P450 2C19 genotype with the antiplatelet effect and clinical efficacy of clopidogrel therapy. *JAMA* 302(8), 849–857 (2009).

38 Thompson JF, Hyde CL, Wood LS *et al.* Comprehensive whole-genome and candidate gene analysis for response to statin therapy in the Treating to New Targets (TNT) cohort. *Circ. Cardiovasc. Genet.* 2(2), 173–181 (2009).

39 Barber MJ, Mangravite LM, Hyde CL *et al.* Genome-wide association of lipid-lowering response to statins in combined study populations. *PLoS ONE* 5(3), e9763 (2010).

40 Zhong H, Yang X, Kaplan LM, Molony C, Schadt EE. Integrating pathway analysis and genetics of gene expression for genome-wide association studies. *Am. J. Hum. Genet.* 86(4), 581–591 (2010).

41 Zhong H, Beaulaurier J, Lum PY *et al.* Liver and adipose expression associated SNPs are enriched for association to Type 2 diabetes. *PLoS Genet.* 6, e1000932 (2010).

42 Schadt EE, Molony C, Chudin E *et al.* Mapping the genetic architecture of gene expression in human liver. *PLoS Biol.* 6(5), e107 (2008).

43 Torkamani A, Schork NJ. Pathway and network analysis with high-density allelic association data. *Methods Mol. Biol.* 563, 289–301 (2009).

44 Holmans P. Statistical methods for pathway analysis of genome-wide data for association with complex genetic traits. *Adv. Genet.* 72, 141–179 (2010).

45 Patel SD, Le-Niculescu H, Koller DL *et al.* Coming to grips with complex disorders. genetic risk prediction in bipolar disorder using panels of genes identified through convergent functional genomics. *Am. J. Med. Genet. B. Neuropsychiatr. Genet.* 153B(4), 850–877 (2010).

46 Peng G, Luo L, Siu H *et al.* Gene and pathway-based second-wave analysis of genome-wide association studies. *Eur. J. Hum. Genet.* 18(1), 111–117 (2010).

47 Segre AV, Groop L, Mootha VK, Daly MJ, Altshuler D. Common inherited variation in mitochondrial genes is not enriched for associations with Type 2 diabetes or related glycemic traits. *PLoS Genet.* 6(8), e1001058 (2010).

48 Fransen K, Visschedijk MC, van Sommeren S *et al.* Analysis of SNPs with an effect on gene expression identifies *UBE2L3* and *BCL3* as potential new risk genes for Crohn's disease. *Hum. Mol. Genet.* 19(17), 3482–3488 (2010).

49 Farber CR, Aten JE, Farber EA *et al.* Genetic dissection of a major mouse obesity QTL (Carfhg2). Integration of gene expression and causality modeling. *Physiol. Genomics* 37(3), 294–302 (2009).

50 Guhathakurta D, Xie T, Anand M *et al.* *Cis*-regulatory variations. a study of SNPs around genes showing *cis*-linkage in segregating mouse populations. *BMC Genomics* 7, 235 (2006).

51 Lum PY, Chen Y, Zhu J *et al.* Elucidating the murine brain transcriptional network in a segregating mouse population to identify core functional modules for obesity and diabetes. *J. Neurochem.* 97(Suppl. 1), 50–62 (2006).

52 Yaguchi H, Togawa K, Moritani M, Itakura M. Identification of candidate genes in the Type 2 diabetes modifier locus using expression QTL. *Genomics* 85(5), 591–599 (2005).

53 Meng H, Vera I, Che N *et al.* Identification of *Abcc6* as the major causal gene for dystrophic cardiac calcification in mice through integrative genomics. *Proc. Natl Acad. Sci. USA* 104(11), 4530–4535 (2007).

54 Drake TA, Schadt EE, Davis RC, Lusis AJ. Integrating genetic and gene expression data to study the metabolic syndrome and diabetes in mice. *Am. J. Ther.* 12(6), 503–511 (2005).

55 Mehrabian M, Allayee H, Stockton J *et al.* Integrating genotypic and expression data in a segregating mouse population to identify 5-lipoxygenase as a susceptibility gene for obesity and bone traits. *Nat. Genet.* 37, 1224–1233 (2005).

56 Schadt EE, Lamb J, Yang X *et al.* An integrative genomics approach to infer causal associations between gene expression and disease. *Nat. Genet.* 37(7), 710–717 (2005).

57 Dixon AL, Liang L, Moffatt MF *et al.* A genome-wide association study of global gene expression. *Nat. Genet.* 39(10), 1202–1207 (2007).

58 Emilsson V, Thorleifsson G, Zhang B *et al.* Genetics of gene expression and its effect on disease. *Nature* 452(7186), 423–428 (2008).

59 Goring HH, Curran JE, Johnson MP *et al.* Discovery of expression QTLs using large-scale transcriptional profiling in human lymphocytes. *Nat. Genet.* 39(10), 1208–1216 (2007).

60 Montgomery SB, Sammeth M, Gutierrez-Arcelus M *et al.* Transcriptome genetics using second generation sequencing in a Caucasian population. *Nature* 464(7289), 773–777 (2010).

61 Pickrell JK, Marioni JC, Pai AA *et al.* Understanding mechanisms underlying human gene expression variation with RNA sequencing. *Nature* 464(7289), 768–772 (2010).

62 Stranger BE, Nica AC, Forrest MS *et al.* Population genomics of human gene expression. *Nat. Genet.* 39(10), 1217–1224 (2007).

63 Veyrieras JB, Kudaravalli S, Kim SY *et al.* High-resolution mapping of expression-QTLs yields insight into human gene regulation. *PLoS Genet.* 4(10), e1000214 (2008).

64 Schadt EE, Monks SA, Drake TA *et al.* Genetics of gene expression surveyed in maize, mouse and man. *Nature* 422(6929), 297–302 (2003).

65 Monks SA, Leonardson A, Zhu H *et al.* Genetic inheritance of gene expression in human cell lines. *Am. J. Hum. Genet.* 75(6), 1094–1105 (2004).

66 Zeller T, Wild P, Szymczak S *et al.* Genetics and beyond – the transcriptome of human monocytes and disease susceptibility. *PLoS ONE* 5(5), e10693 (2010).

67 Dimas AS, Deutsch S, Stranger BE *et al.* Common regulatory variation impacts gene expression in a cell type-dependent manner. *Science* 325(5945), 1246–1250 (2009).

68 Myers AJ, Gibbs JR, Webster JA *et al.* A survey of genetic human cortical gene expression. *Nat. Genet.* 39(12), 1494–1499 (2007).

69 Greenawalt DM, Dobrin R, Chudin E *et al.* A survey of the genetics of stomach, liver, and adipose gene expression from a morbidly obese cohort. *Genome Res.* 21(7), 1008–1016 (2011).

70 Innocenti F, Cooper GM, Stanaway IB *et al.* Identification, replication, and functional fine-mapping of expression quantitative trait loci in primary human liver tissue. *PLoS Genet.* 7(5), e1002078 (2011).

71 Coulombe-Huntington J, Lam KC, Dias C, Majewski J. Fine-scale variation and genetic determinants of alternative splicing across individuals. *PLoS Genet.* 5(12), e1000766 (2009).

72 Heinzen EL, Ge D, Cronin KD *et al.* Tissue-specific genetic control of splicing. implications for the study of complex traits. *PLoS Biol.* 6(12), e1 (2008).

73 Nembaware V, Lupindo B, Schouest K, Spillane C, Scheffler K, Seoighe C. Genome-wide survey of allele-specific splicing in humans. *BMC Genomics* 9, 265 (2008).

74 Ge B, Pokholok DK, Kwan T *et al.* Global patterns of *cis* variation in human cells revealed by high-density allelic expression analysis. *Nat. Genet.* 41(11), 1216–1222 (2009).

75 Heap GA, Yang JH, Downes K *et al.* Genome-wide analysis of allelic expression imbalance in human primary cells by high-throughput transcriptome resequencing. *Hum. Mol. Genet.* 19(1), 122–134 (2010).

76 Gibbs JR, van der Brug MP, Hernandez DG *et al.* Abundant quantitative trait loci exist for DNA methylation and gene expression in human brain. *PLoS Genet.* 6(5), e1000952 (2010).

77 Kerkel K, Spadola A, Yuan E *et al.* Genomic surveys by methylation-sensitive SNP analysis identify sequence-dependent allele-specific DNA methylation. *Nat. Genet.* 40(7), 904–908 (2008).

78 Ollikainen M, Smith KR, Joo EJ *et al.* DNA methylation analysis of multiple tissues from newborn twins reveals both genetic and intrauterine components to variation in the human neonatal epigenome. *Hum. Mol. Genet.* 19(21), 4176–4188 (2010).

79 Schalkwyk LC, Meaburn EL, Smith R *et al.* Allelic skewing of DNA methylation is widespread across the genome. *Am. J. Hum. Genet.* 86(2), 196–212 (2010).

80 Shoemaker R, Deng J, Wang W, Zhang K. Allele-specific methylation is prevalent and is contributed by CpG-SNPs in the human genome. *Genome Res.* 20(7), 883–889 (2010).

81 Zhang D, Cheng L, Badner JA *et al.* Genetic control of individual differences in gene-specific methylation in human brain. *Am. J. Hum. Genet.* 86(3), 411–419 (2010).

82 Yang X, Zhang B, Molony C *et al.* Systematic genetic and genomic analysis of cytochrome P450 enzyme activities in human liver. *Genome Res.* 20(8), 1020–1036 (2010).

83 Singer JB, Lewitzky S, Leroy E *et al.* A genome-wide study identifies *HLA* alleles associated with lumiracoxib-related liver injury. *Nat. Genet.* 42(8), 711–714 (2010).

84 Link E, Parish S, Armitage J *et al. SLCO1B1* variants and statin-induced myopathy – a genomewide study. *N. Engl. J. Med.* 359(8), 789–799 (2008).

85 Musunuru K, Strong A, Frank-Kamenetsky M *et al.* From noncoding variant to phenotype via SORT1 at the 1p13 cholesterol locus. *Nature* 466, 714–719 (2010).

86 Barabasi AL, Oltvai ZN. Network biology. understanding the cell's functional organization. *Nat. Rev. Genet.* 5(2), 101–113 (2004).

87 Zhang B, Horvath S. A general framework for weighted gene co-expression network analysis. *Stat. Appl. Genet. Mol. Biol.* 4, Article 17 (2005).

88 Langfelder P, Horvath S. WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics* 9, 559 (2008).

89 Margolin AA, Nemenman I, Basso K *et al.* ARACNE: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. *BMC Bioinformatics* 7(Suppl. 1), S7 (2006).

90 Pe'er D. Bayesian network analysis of signaling networks: a primer. *Sci. STKE* 2005(281), pl4 (2005).

91 Pe'er D, Regev A, Elidan G, Friedman N. Inferring subnetworks from perturbed expression profiles. *Bioinformatics* 17(Suppl. 1), S215–S224 (2001).

92 Battle A, Jonikas MC, Walter P, Weissman JS, Koller D. Automated identification of pathways from quantitative genetic interaction data. *Mol. Syst. Biol.* 6, 379 (2010).

93 Manichaikul A, Moon JY, Sen S, Yandell BS, Broman KW. A model selection approach for the identification of quantitative trait loci in experimental crosses, allowing epistasis. *Genetics* 181(3), 1077–1086 (2009).

94 Banerjee S, Yandell BS, Yi N. Bayesian quantitative trait loci mapping for multiple traits. *Genetics* 179(4), 2275–2289 (2008).

95 Zhu J, Lum PY, Lamb J *et al.* An integrative genomics approach to the reconstruction of gene networks in segregating populations. *Cytogenet. Genome Res.* 105(2–4), 363–374 (2004).

96 Zhu J, Wiener MC, Zhang C *et al.* Increasing the power to detect causal associations by combining genotypic and expression data in segregating populations. *PLoS Comput. Biol.* 3(4), e69 (2007).

97 Chen Y, Zhu J, Lum PY *et al.* Variations in DNA elucidate molecular networks that cause disease. *Nature* 452(7186), 429–435 (2008).

98 Gargalovic PS, Imura M, Zhang B *et al.* Identification of inflammatory gene modules based on variations of human endothelial cell responses to oxidized lipids. *Proc. Natl Acad. Sci. USA* 103(34), 12741–12746 (2006).

99 Ghazalpour A, Doss S, Zhang B *et al.* Integrating genetic and network analysis to characterize genes related to mouse weight. *PLoS Genet.* 2(8), e130 (2006).

100 Horvath S, Zhang B, Carlson M *et al.* Analysis of oncogenic signaling networks in glioblastoma identifies ASPM as a molecular target. *Proc. Natl Acad. Sci. USA* 103(46), 17402–17407 (2006).

101 Zhu J, Zhang B, Schadt EE. A systems biology approach to drug discovery. *Adv. Genet.* 60, 603–635 (2008).

102 Zhu J, Zhang B, Smith EN *et al.* Integrating large-scale functional genomic data to dissect the complexity of yeast regulatory networks. *Nat. Genet.* 40(7), 854–861 (2008).

103 Miller JA, Horvath S, Geschwind DH. Divergence of human and mouse brain transcriptome highlights Alzheimer disease pathways. *Proc. Natl Acad. Sci. USA* 107(28), 12698–12703 (2010).

104 Keller MP, Choi Y, Wang P *et al.* A gene expression network model of Type 2 diabetes links cell cycle regulation in islets with diabetes susceptibility. *Genome Res.* 18(5), 706–716 (2008).

105 Farber CR, Bennett BJ, Orozco L *et al.* Mouse genome-wide association and systems genetics identify Asxl2 as a regulator of bone mineral density and osteoclastogenesis. *PLoS Genet.* 7(4), e1002038 (2011).

106 Jansen R, Yu H, Greenbaum D *et al.* A Bayesian networks approach for predicting protein-protein interactions from genomic data. *Science* 302(5644), 449–453 (2003).

107 Lee I, Date SV, Adai AT, Marcotte EM. A probabilistic functional network of yeast genes. *Science* 306(5701), 1555–1558 (2004).

108 Yi N, Banerjee S, Pomp D, Yandell BS. Bayesian mapping of genomewide interacting quantitative trait loci for ordinal traits. *Genetics* 176(3), 1855–1864 (2007).

109 Sachs K, Perez O, Pe'er D, Lauffenburger DA, Nolan GP. Causal protein-signaling networks derived from multiparameter single-cell data. *Science* 308(5721), 523–529 (2005).

110 Friedman N, Linial M, Nachman I, Pe'er D. Using Bayesian networks to analyze expression data. *J. Comput. Biol.* 7(3–4), 601–620 (2000).

111 Gille C, Bolling C, Hoppe A *et al.* HepatoNet1: a comprehensive metabolic reconstruction of the human hepatocyte for the analysis of liver physiology. *Mol. Syst. Biol.* 6, 411 (2010).

112 Jerby L, Shlomi T, Ruppin E. Computational reconstruction of tissue-specific metabolic models. application to human liver metabolism. *Mol. Syst. Biol.* 6, 401 (2010).

113 Vodovotz Y, Constantine G, Faeder J *et al.* Translational systems approaches to the biology of inflammation and healing. *Immunopharmacol. Immunotoxicol.* 32, 181–195 (2010).

114 Hughes TR, Marton MJ, Jones AR *et al.* Functional discovery via a compendium of expression profiles. *Cell* 102, 109–126 (2000).

115 Gasch AP, Huang M, Metzner S, Botstein D, Elledge SJ, Brown PO. Genomic expression responses to DNA-damaging agents and the regulatory role of the yeast ATR homolog Mec1p. *Mol. Biol. Cell* 12, 2987–3003 (2001).

116 Natsoulis G, Pearson CI, Gollub J *et al.* The liver pharmacological and xenobiotic gene response repertoire. *Mol. Syst. Biol.* 4, 175 (2008).

117 Slatter JG, Cheng O, Cornwell PD *et al.* Microarray-based compendium of hepatic gene expression profiles for prototypical *ADME* gene-inducing compounds in rats and mice *in vivo*. *Xenobiotica* 36(10–11), 902–937 (2006).

118 Lamb J, Crawford ED, Peck D *et al.* The Connectivity Map: using gene-expression signatures to connect small molecules, genes, and disease. *Science (New York)* 313, 1929–1935 (2006).

119 Anderson JB, Sirjusingh C, Parsons AB *et al.* Mode of selection and experimental evolution of antifungal drug resistance in *Saccharomyces cerevisiae*. *Genetics* 163, 1287–1298 (2003).

120 Green SR, Johnson AD. Promoter-dependent roles for the Srb10 cyclin-dependent kinase and the Hda1 deacetylase in Tup1-mediated repression in *Saccharomyces cerevisiae*. *Mol. Biol. Cell* 15, 4191–4202 (2004).

121 Holbeck SL, Collins JM, Doroshow JH. Analysis of Food and Drug Administration-approved anticancer agents in the NCI60 panel of human tumor cell lines. *Mol. Cancer Ther.* 9, 1451–1460 (2010).

122 Su G, Burant CF, Beecher CW, Athey BD, Meng F. Integrated metabolome and transcriptome analysis of the NCI60 dataset. *BMC Bioinformatics* 12(Suppl. 1), S36 (2011).

123 Welsh M, Mangravite L, Medina MW *et al.* Pharmacogenomic discovery using cell-based models. *Pharmacol. Rev.* 61, 413–429 (2009).

124 Eng L, Ibrahim-Zada I, Jarjanazi H *et al.* Bioinformatic analyses identifies novel protein-coding pharmacogenomic markers associated with paclitaxel sensitivity in NCI60 cancer cell lines. *BMC Med. Genomics* 4, 18 (2011).

125 Savas S, Azorsa DO, Jarjanazi H *et al.* NCI60 cancer cell line panel data and RNAi analysis help identify EAF2 as a modulator of simvastatin and lovastatin response in HCT-116 cells. *PLoS ONE* 6, e18306 (2011).

126 Lamb J. The Connectivity Map: a new tool for biomedical research. *Nat Rev Cancer.* 7, 54–60 (2007).

127 Wang SE, Xiang B, Guix M *et al.* Transforming growth factor beta engages TACE and ErbB3 to activate phosphatidylinositol-3 kinase/Akt in ErbB2-overexpressing breast cancer and desensitizes cells to trastuzumab. *Mol. Cell. Biol.* 28, 5605–5620 (2008).

128 Hahn CK, Ross KN, Warrington IM *et al.* Expression-based screening identifies the combination of histone deacetylase inhibitors and retinoids for neuroblastoma differentiation. *Proc. Natl Acad. Sci. USA* 105, 9751–9756 (2008).

129 Loboda A, Kraft WK, Fine B *et al.* Diurnal variation of the human adipose transcriptome and the link to metabolic disease. *BMC Med. Genomics* 2, 7 (2009).

130 Goh K-I, Cusick ME, Valle D, Childs B, Vidal M, Barabási A-L. The human disease network. *Proc. Natl Acad. Sci. USA* 104, 8685–8690 (2007).

131 Van Driel MA, Bruggeman J, Vriend G, Brunner HG, Leunissen JA. A text-mining analysis of the human phenome. *Eur. J. Hum. Genet.* 14, 535–542 (2006).

132 Liu YI, Wise PH, Butte AJ. The 'etiome': identification and clustering of human disease etiological factors. *BMC Bioinformatics* 10(Suppl. 2), S14 (2009).

133 Hu G, Agarwal P. Human disease-drug network based on genomic expression profiles. *PLoS ONE* 4, e6536 (2009).

134 Suthram S, Dudley JT, Chiang AP, Chen R, Hastie TJ, Butte AJ. Network-based elucidation of human disease similarities reveals common functional modules enriched for pluripotent drug targets. *PLoS Comput. Biol.* 6, e1000662 (2010).

135 Smith SC, Allen J, Blair SN *et al.* AHA/ACC guidelines for secondary prevention for patients with coronary and other atherosclerotic vascular disease. 2006 update: endorsed by the National Heart, Lung, and Blood Institute. *Circulation* 113, 2363–2372 (2006).

136 Windgassen EB, Funtowicz L, Lunsford TN, Harris LA, Mulvagh SL. C-reactive protein and high-sensitivity C-reactive protein. an update for clinicians. *Postgrad. Med.* 123, 114–119 (2011).

137 Sanz J, Fayad ZA. Imaging of atherosclerotic cardiovascular disease. *Nature* 451, 953–957 (2008).

138 Tzoulaki I, Liberopoulos G, Ioannidis JPA. Assessment of claims of improved prediction beyond the Framingham risk score. *JAMA* 302, 2345–2352 (2009).

139 Reeves GK, Travis RC, Green J *et al.* Incidence of breast cancer and its subtypes in relation to individual and multiple low-penetrance genetic susceptibility loci. *JAMA* 304, 426–434 (2010).

140 Carey LA, Perou CM, Livasy CA *et al.* Race, breast cancer subtypes, and survival in the Carolina Breast Cancer Study. *JAMA* 295, 2492–2502 (2006).

141 Wei X, Walia V, Lin JC *et al.* Exome sequencing identifies *GRIN2A* as frequently mutated in melanoma. *Nat. Genet.* 43(5), 442–446 (2011).

142 Corder EH, Saunders AM, Strittmatter WJ *et al.* Gene dose of apolipoprotein E type 4 allele and the risk of Alzheimer's disease in late onset families. *Science (New York)* 261, 921–923 (1993).

143 Hollingworth P, Harold D, Sims R *et al.* Common variants at *ABCA7*, *MS4A6A*/*MS4A4E*, *EPHA1*, *CD33* and *CD2AP* are associated with Alzheimer's disease. *Nat. Genet.* 7, 1–8 (2011).

144 Naj AC, Jun G, Beecham GW *et al.* Common variants at *MS4A4*/*MS4A6E*, *CD2AP*, *CD33* and *EPHA1* are associated with late-onset Alzheimer's disease. *Nat. Genet.* 43(5), 436–441 (2011).

145 Choi M, Scholl UI, Ji W *et al.* Genetic diagnosis by whole exome capture and massively parallel DNA sequencing. *Proc. Natl Acad. Sci. USA* 106, 19096–19101 (2009).

146 Schreiber SL, Shamji AF, Clemons PA *et al.* Towards patient-based cancer therapeutics. *Nat. Biotechnol.* 28, 904–906 (2010).

147 International Cancer Genome Consortium, Hudson TJ, Anderson W *et al.* International network of cancer genome projects. *Nature* 464, 993–998 (2010).

148 Basso K, Margolin AA, Stolovitzky G, Klein U, Dalla-Favera R, Califano A. Reverse engineering of regulatory networks in human B cells. *Nat. Genet.* 37, 382–390 (2005).

149 Phillips HS, Kharbanda S, Chen R *et al.* Molecular subclasses of high-grade glioma predict prognosis, delineate a pattern of disease progression, and resemble stages in neurogenesis. *Cancer Cell* 9, 157–173 (2006).

150 Carro MS, Lim WK, Alvarez MJ *et al.* The transcriptional network for mesenchymal transformation of brain tumours. *Nature* 463(7279), 318–325 (2010).

151 Sun L, Hui A-M, Su Q *et al.* Neuronal and glioma-derived stem cell factor induces angiogenesis within the brain. *Cancer Cell* 9, 287–300 (2006).

152 Cancer Genome Atlas Research Network. Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature* 455(7216), 1061–1068 (2008).

153 Lee W, Jiang Z, Liu J *et al.* The mutation spectrum revealed by paired genome sequences from a lung cancer patient. *Nature* 465, 473–477 (2010).

154 Kan Z, Jaiswal BS, Stinson J *et al.* Diverse somatic mutation patterns and pathway alterations in human cancers. *Nature* 466(7308), 869–873 (2010).

155 Chapman MA, Lawrence MS, Keats JJ *et al.* Initial genome sequencing and analysis of multiple myeloma. *Nature* 471, 467–472 (2011).

156 McDermott U, Downing JR, Stratton MR. Genomics and the continuum of cancer care. *N. Engl. J. Med.* 364, 340–350 (2011).

157 Scheuner MT, Sieverding P, Shekelle PG. Delivery of genomic medicine for common chronic adult diseases: a systematic review. *JAMA* 299, 1320–1334 (2008).

158 Carleton B, Poole R, Smith M *et al.* Adverse drug reaction active surveillance: developing a national network in Canada's children's hospitals. *Pharmacoepidemiol. Drug Saf.* 18, 713–721 (2009).

159 Madadi P, Ross CJD, Hayden MR *et al.* Pharmacogenetics of neonatal opioid toxicity following maternal use of codeine during breastfeeding. a case–control study. *Clin. Pharmacol. Ther.* 85, 31–35 (2009).

160 Ross CJD, Katzov-Eckert H, Dubé M-P *et al.* Genetic variants in *TPMT* and *COMT* are associated with hearing loss in children receiving cisplatin chemotherapy. *Nat. Genet.* 41, 1345–1349 (2009).

161 Ross CJD, Visscher H, Sistonen J *et al.* The Canadian Pharmacogenomics Network for Drug Safety: a model for safety pharmacology. *Thyroid* 20, 681–687 (2010).

162 Pearson JF, Brownstein CA, Brownstein JS. Potential for electronic health records and online social networking to redefine medical research. *Clin. Chem.* 57(2), 196–204 (2010).

163 Wilke RA, Xu H, Denny JC *et al.* The Emerging role of electronic medical records in pharmacogenomics. *Clin. Pharmacol. Ther.* 89, 379–386 (2011).

164 Human genome. Genomes by the thousand. *Nature* 467, 1026–1027 (2010).

165 Roach JC, Glusman G, Smit AFA *et al.* Analysis of genetic inheritance in a family quartet by whole-genome sequencing. *Science* 328(5978), 636–639 (2010).

166 Worthey EA, Mayer AN, Syverson GD *et al.* Making a definitive diagnosis. Successful clinical application of whole exome sequencing in a child with intractable inflammatory bowel disease. *Genet. Med.* 13(3), 255–262 (2011).

167 Vaddiraju S, Burgess DJ, Tomazos I, Jain FC, Papadimitrakopoulos F. Technologies for continuous glucose monitoring. current problems and future promises. *J. Diabetes Sci. Technol.* 4, 1540–1562 (2010).

168 Soleymani S, Borzage M, Seri I. Hemodynamic monitoring in neonates. advances and challenges. *J. Perinatol.* 30(Suppl.), S38–S45 (2010).

169 Hoeks LBEA, Greven WL, De Valk HW. Real-time continuous glucose monitoring system for treatment of diabetes. a systematic review. *Diabet. Med.* 28, 386–394 (2011).

170 Zhang W, Zhu J, Schadt EE, Liu JS. A Bayesian partition method for detecting pleiotropic and epistatic eQTL modules. *PLoS Comput. Biol.* 6, e1000642 (2010).

171 Brownstein JS, Murphy SN, Goldfine AB *et al.* Rapid identification of myocardial infarction risk associated with diabetes medications using electronic medical records. *Diabetes Care* 33, 526–531 (2010).

172 Brownstein JS, Sordo M, Kohane IS, Mandl KD. The tell-tale heart. population-based surveillance reveals an association of rofecoxib and celecoxib with myocardial infarction. *PLoS ONE* 2, e840 (2007).

173 Ozeki T, Mushiroda T, Yowang A *et al.* Genome-wide association study identifies *HLA-A*3101* allele as a genetic risk factor for carbamazepine-induced cutaneous adverse drug reactions in Japanese population. *Hum. Mol. Genet.* 20(5), 1034–1041 (2011).

174 McCormack M, Alfirevic A, Bourgeois S *et al. HLA-A*3101* and carbamazepine-induced hypersensitivity reactions in Europeans. *N. Engl. J. Med.* 364(12), 1134–1143 (2011).

175 Lucena MI, Molokhia M, Shen Y *et al.* Susceptibility to amoxicillin-clavulanate-induced liver injury is influenced by multiple HLA class I and II alleles. *Gastroenterology* 141(1), 338–347 (2011).

176 Mick E, McGough JJ, Middleton FA, Neale B, Faraone SV. Genome-wide association study of blood pressure response to methylphenidate treatment of attention-deficit/hyperactivity disorder. *Prog. Neuropsychopharmacol. Biol. Psychiatry* 35(2), 466–472 (2010).

177 Kindmark A, Jawaid A, Harbron CG *et al.* Genome-wide pharmacogenetic investigation of a hepatic adverse event without clinical signs of immunopathology suggests an underlying immune pathogenesis. *Pharmacogenomics J.* 8(3), 186–195 (2008).

178 Ingle JN, Schaid DJ, Goss PE *et al.* Genome-wide associations and functional genomic studies of musculoskeletal adverse events in women receiving aromatase inhibitors. *J. Clin. Oncol.* 28(31), 4674–4682 (2010).

179 Marciante KD, Durda JP, Heckbert SR *et al.* Cerivastatin, genetic variants, and the risk of rhabdomyolysis. *Pharmacogenet. Genomics* 21(5), 280–288 (2011).

180 Tanaka Y, Kurosaki M, Nishida N *et al.* Genome-wide association study identified *ITPA/DDRGK1* variants reflecting thrombocytopenia in pegylated interferon and ribavirin therapy for chronic hepatitis C. *Hum. Mol. Genet.* 20(17), 3507–3516 (2011).

181 Byun E, Caillier SJ, Montalban X *et al.* Genome-wide pharmacogenomic analysis of the response to interferon beta therapy in multiple sclerosis. *Arch. Neurol.* 65(3), 337–344 (2008).

182 Comabella M, Craig DW, Morcillo-Suarez C *et al.* Genome-wide scan of 500,000 single-nucleotide polymorphisms among responders and nonresponders to interferon beta therapy in multiple sclerosis. *Arch. Neurol.* 66(8), 972–978 (2009).

183 Weber F, Cepok S, Wolf C *et al.* Single-nucleotide polymorphisms in *HLA*- and non-*HLA* genes associated with the development of antibodies to interferon-beta therapy in multiple sclerosis patients. *Pharmacogenomics J.* doi:10.1038/tpj.2011.14 (2011) (Epub ahead of print).

184 Turner ST, Bailey KR, Fridley BL *et al.* Genomic association analysis suggests chromosome 12 locus influencing antihypertensive response to thiazide diuretic. *Hypertension* 52(2), 359–365 (2008).

185 Liu C, Batliwalla F, Li W *et al.* Genome-wide association scan identifies candidate polymorphisms associated with differential response to anti-TNF treatment in rheumatoid arthritis. *Mol. Med.* 14(9–10), 575–581 (2008).

186 Plant D, Bowes J, Potter C *et al.* Genome-wide association study of genetic predictors of anti-tumour necrosis factor (TNF) treatment in rheumatoid arthritis (RA) identifies associations with polymorphisms at seven loci. *Arthritis Rheum.* 63(3), 645–653 (2010).

187 Yang JJ, Cheng C, Yang W *et al.* Genome-wide interrogation of germline genetic variation associated with treatment response in childhood acute lymphoblastic leukemia. *JAMA* 301(4), 393–403 (2009).

188 Perlis RH, Smoller JW, Ferreira MA *et al.* A genomewide association study of response to lithium for prevention of recurrence in bipolar disorder. *Am. J. Psychiatry* 166(6), 718–725 (2009).

189 Wu C, Xu B, Yuan P *et al.* Genome-wide examination of genetic variants associated with response to platinum-based chemotherapy in patients with small-cell lung cancer. *Pharmacogenet. Genomics* 20(6), 389–395 (2010).

190 Wu X, Ye Y, Rosell R *et al.* Genome-wide association study of survival in non-small cell lung cancer patients receiving platinum-based chemotherapy. *J. Natl Cancer Inst.* 103(10), 817–825 (2011).

191 Niu N, Qin Y, Fridley BL *et al.* Radiation pharmacogenomics: a genome-wide association approach to identify radiation response biomarkers using human lymphoblastoid cell lines. *Genome Res.* 20(11), 1482–1492 (2010).

192 Zhou K, Bellenguez C, Spencer CC *et al.* Common variants near ATM are associated with glycemic response to metformin in Type 2 diabetes. *Nat. Genet.* 43(2), 117–120 (2011).

193 Wade R, Di Bernardo MC, Richards S *et al.* Association between SNP-genotype and chronic lymphocytic leukemia outcome in a randomized chemotherapy trial. *Haematologica* 96(10), 1496–503 (2011).

194 Moyer AM, Fridley BL, Jenkins GD *et al.* Acetaminophen-NAPQI hepatotoxicity. a cell line model system genome-wide association study. *Toxicol. Sci.* 120(1), 33–41 (2011).

Websites

201 Hindorff L, Junkins H, Hall P, Mehta J, Manolio T. A catalog of published genome-wide association studies. www.genome.gov/gwastudies (Accessed 17 July 2011)

202 Gene Expression Omnibus. www.ncbi.nlm.nih.gov/geo

203 dbGaP. www.ncbi.nlm.nih.gov/gap

204 Rubin J, Donald J. The New Biology (2010). www.pacificbiosciences.com/newbio

205 Wetterstrand KA. DNA sequencing costs. Data from the NHGRI large-scale genome sequencing program. www.genome.gov/sequencingcosts (Accessed 1 August 2011)

206 Sutter JD. 'Smart dust' aims to monitor everything (2 May 2010). http://edition.cnn.com/2010/TECH/05/03/smart.dust.sensors/index.html (Accessed 31 July 2011)

207 Guo S, Wang M, Leskovec J. The role of social networks in online shopping. Information passing, price of trust and consumer choice (2011). http://arxiv.org/abs/1104.0942