Published in final edited form as:

Clin Pharmacol Ther. 2012 June ; 91(6): 994–1000. doi:10.1038/clpt.2012.49.

Translational Bioinformatics: Linking the Molecular World to the Clinical World

RB Altman^{1,2,3}

¹Department of Bioengineering, Stanford University, Stanford, California, USA

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

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

Abstract

Translational bioinformatics represents the union of translational medicine and bioinformatics. Translational medicine moves basic biological discoveries from the research bench into the patient-care setting and uses clinical observations to inform basic biology. It focuses on patient care, including the creation of new diagnostics, prognostics, prevention strategies, and therapies based on biological discoveries. Bioinformatics involves algorithms to represent, store, and analyze basic biological data, including DNA sequence, RNA expression, and protein and small-molecule abundance within cells. Translational bioinformatics spans these two fields; it involves the development of algorithms to analyze basic molecular and cellular data with an explicit goal of affecting clinical care.

Simply put, translational bioinformatics research integrates information about molecular entities (DNA, RNA, proteins, small molecules, and lipids) with information about clinical entities (patients, diseases, symptoms, laboratory tests, pathology reports, clinical images, and drugs) to improve patient care and our understanding of biology.

Three important technical capabilities enable translational bioinformatics. First, the revolution in computing has brought advanced capabilities in data storage, analysis, and visualization. Second, the revolution in high-throughput biological measurements allows us to sequence entire genomes, and characterize the transcriptome, the proteome, and the metabolome. Finally, the revolution in population-based health data and electronic medical records (EMRs) offers access to clinical data on a large scale. Researchers in translational bioinformatics must therefore have expertise in the algorithms for managing basic biological and cellular data as well as a good understanding of how to connect the molecular scale to the clinical scale. In the past five years, a cadre of specialists has begun to create a scholarly community with the required infrastructure, emerging journals, and conferences. One such

conference is the annual American Medical Informatics Association Summit on Translational Bioinformatics.

As for all informatics, data form the basic raw material for translational bioinformatics. The past few years have seen the introduction of a powerful network of data resources. Although each has its own characteristic noise and bias, together they make it possible to find the links between molecular/cellular entities and clinical entities. Table 1 presents some of the key data resources containing clinical information. Table 2 presents the key data resources containing molecular information. Importantly, some resources include both clinical data and biological data, thereby providing important links between these worlds, as shown in Table 3. Without these linkages, it would be impossible to connect clinical to basic data. This review covers seven active areas in bioinformatics, as well as recent representative work within each of these areas.

INFORMATICS FOR HEALTH-RELATED GENOMICS

Our ability to measure DNA sequences has undergone remarkable changes in the past decade. Not only can we measure genetic variations accurately and cheaply (through expression arrays, for example), but we can also sequence entire genomes (6 billion bases) and entire exomes (the 2% of the genome that is transcribed), with continually dropping costs (currently ~\$4,000) and increasing levels of accuracy. This revolution in our ability to measure human genetic information creates opportunities for translational bioinformatics. The most basic tasks include creating the information infrastructure to store and process the raw data from these technologies. The ability to sequence DNA has actually outpaced the growth in computer power for the past several years, and therefore DNA sequencing has begun to stress the infrastructure for storage and analysis. There are also opportunities to create algorithms for associating genetic variations with clinically important outcomes such as disease risk, disease prognosis, and therapeutic response. This can be a straight-forward task if the variations are sufficiently common to allow cohorts to be enrolled in studies with good statistical power. However, there is increasing concern that common genetic variations may not explain all of the observed genetic variability in health-relevant phenotypes; therefore, the frontier is at the interpretation of rare variants, in particular the interpretation of variants within noncoding “control” regions of the genome. The translational promise for clinical medicine, of course, is the vision of genome-informed medicine. This vision envisages that the genomes of all patients will be sequenced and that health-care workers would use the sequence information to prioritize disease surveillance and prevention, and choose therapies with maximal efficacy and minimal side effects. Recent illustrative work includes efforts in genome annotation, phenotype capture, and epidemiologic and social impact.

Ashley *et al.* reported a preliminary full clinical annotation of a whole-genome sequence.¹ Using the sequenced genome of a 40-year-old man with a family history of heart disease, they assessed his risk for common diseases, based on available information about alleles associated with genetic risk, the presence of rare alleles associated with Mendelian disease or likely Mendelian disease, and his anticipated response to hundreds of drugs. The analysis recommended early initiation of treatment with a statin for heart disease, based on the

personalized risk/benefit profile derived from this analysis. Although this paper provides an early look at genomic medicine, it also raises issues of how to include genomics into routine clinical workflow and how physicians will integrate genetic risk with other risk information.

Ritchie *et al.* showed that EMRs have information sufficient to identify cases and controls for 21 previously studied phenotypes and to replicate the genome-wide association study (GWAS) results created in much more controlled trials.² These results indicate that EMR-based phenotyping and its use in genetic discovery are feasible and much less expensive than specially created study cohorts.

Patient self-reports can also be a valuable phenotyping tool. Eriksson *et al.* showed that GWAS studies can be successful in identifying single-nucleotide polymorphisms (SNPs) associated with phenotypes even when the phenotypes are not collected by investigators but are reported by the study subjects.³ They found GWAS hits for hair morphology, photic sneeze reflex, freckling, and asparagus urinary metabolites.

Molecular networks show great promise in helping to interpret genetic data. Iossifov *et al.* proposed a novel computational framework for linking molecular interaction networks to classical genetic linkage models to improve performance.⁴ They applied these to autism, bipolar disorder, and schizophrenia and found several shared gene targets, supporting the hypothesis that these disorders exist as a spectrum rather than as distinctly separate entities.

Several ongoing large-scale epidemiologic studies have enrolled very large cohorts and are starting to acquire genetic information. Chang *et al.* analyzed the allele frequencies of 90 important genetic variants in 50 genes from six pathways in 7,159 participants in the National Health and Nutrition Examination Survey study.⁵ These data represent a first look at the cross-sectional genetics relevant to nutrient metabolism, inflammation, drug metabolism, DNA repair, blood pressure, and oxidative stress. In another large-scale study, the International Warfarin Pharmacogenetics Consortium published a pooled analysis of 5,000 patients to propose both a clinical warfarin dosing algorithm and one that used genetics information.⁶ They showed the superiority of the pharmacogenetic algorithm over the clinical algorithm and demonstrated that patients receiving high doses and those receiving low doses can be prospectively identified.

Gene patenting remains a contentious issue. Salzberg *et al.* empowered individuals to test for mutations in *BRCA1* and *BRCA2* “at home” in potential violation of Myriad’s patents.⁷ The authors argued that this was fundamentally akin to measuring one’s own blood pressure at home and contended that the patents should not have been awarded and that no private entity should have intellectual property rights over naturally occurring gene sequences.

BIOINFORMATICS SIGNATURES IN MOLECULAR DIAGNOSTICS AND PROGNOSTICS

The increasing ability of biologists to make high-throughput measurements of DNA has been accompanied by progress in measuring RNA expression, protein expression, small molecules, and metabolites. The prospect of using these data as markers for clinical

syndromes is tantalizing. The most basic analyses include characterizing cellular populations on the basis of these measurements and clustering them on the basis of similar profiles. There are also efforts to create diagnostic classifiers that can identify subclasses of disease with different prognoses or therapeutic sensitivities. Early work in this area focused on expression data: hematologic malignancies that appeared morphologically similar had different molecular signatures.⁸ Moreover, these signatures correlated with clinical outcomes, suggesting that molecular profiles could provide information on disease risk, disease prognosis, and therapeutic response. The use of nongenetic biomarkers is important because these are better correlated with the relevant phenotypes of interest than is morphology, although perhaps more expensive to measure. The associations between several complex bioinformatic signatures and important disease states have already been recognized, such as those used in diagnostic platforms for staging breast cancer on the basis of gene expression.⁹ The translational bioinformatics literature includes both methodologic advances in the detection and evaluation of biomarkers and the application of these methods to yield disease insights. Recent work focuses on large-scale discovery of biomarkers, sequencing for rare Mendelian disorders, and mining for novel phenotypic consequences of genetic variations, among other applications.

The Wellcome Trust Consortium published a landmark paper in which 14,000 cases of seven common diseases were compared with 3,000 shared controls in a GWAS analysis using 500,000 SNPs.¹⁰ They found 24 independent associations for the seven diseases, and made the data available for the development and refinement of additional methods for GWAS analysis. This study provides tens of potential biomarkers for each of the diseases studied.

Rare genetic diseases are excellent targets for new sequencing technologies. Ng *et al.* showed that sequencing the exons of four affected individuals in three kindreds allowed them to identify the causal gene in Miller syndrome (*DHODH*, which encodes an enzyme for pyrimidine synthesis).¹¹ This was one of the first reports of the analysis of exome data to identify causal variants for rare Mendelian disorders. In a similar study, Roach *et al.* performed whole-genome sequencing of a family quartet to identify this gene and three others.¹²

Denny *et al.* reversed traditional GWAS studies and carried out a prototype of a “phenome-wide association study” or PheWAS in which they sought 776 diseases (phenotypes) associated with five SNPs that were already known to be associated with at least one other disease.¹³ Using this approach, instead of using a single phenotype of interest and one million SNPs, they used one SNP of interest and 776 phenotypes to perform a similar association study to find new gene–disease linkages.

Yet another “*WAS” was the one introduced by Patel *et al.*, namely, the environment-wide association study (EWAS).¹⁴ They used publicly available data from the comprehensive National Health and Nutrition Examination Survey data set to find associations between 266 environmental variables and hyperglycemia. They found significant associations between hyperglycemia and exposure to several environmental agents including the pesticide

heptachlor epoxide and polychlorinated biphenyls, whereas β -carotenes were shown to have a protective effect.

Model systems are always problematic. Although they are necessary for progress, their relevance to human disease must be considered. McGary *et al.* tackled the problem of finding good analogs of human phenotypes in model organisms to facilitate easier study.¹⁵ They created a method for comparing orthologous gene subnetworks and mapping them between organisms. In addition, they reported a yeast model for human angiogenesis, a worm model for breast cancer, and a mouse model of autism.

Other work has identified protein modulators of transcription factor activity in human B cells, using a novel method based on mutual information.¹⁶ Researchers compared genetic variation profiles across six autoimmune diseases and found that there are clusters, and that alleles that are associated with risk for some diseases are protective for others.¹⁷ They also showed that it is possible that GWAS-based links to common variants are actually the result of chance associations of the common variant with rare variants and that the latter are the actual causative variants.¹⁸ This important observation opens up the possibility that many current GWAS “hits” are not pointing to the relevant genomic regions and novel biomarkers, and that we need to move toward full-genome sequencing and new association methods.

COMPUTATIONAL ANALYSIS OF TEXT

Our collective knowledge of biomedicine is contained within the published primary literature. PubMed contains more than 20 million references (most with abstracts available free of charge), and there is no sign that scientists will stop reporting the results of their work in written “natural language” (as opposed to formal, computer-readable language). Therefore, it is important to be able to extract information from natural text and represent it formally in databases that allow automated search, indexing, and inference. Recent progress in natural-language processing has produced algorithms that can reliably determine parts of speech (e.g., nouns, verbs, prepositions, and adjectives) and recognize complex biological entities. Gene names are notoriously difficult to identify because they often involve several otherwise common words strung together (e.g., “sonic hedgehog” is a gene involved in cell signaling). Natural-language processing applications can, however, establish the link between molecular entities and clinical entities and can identify connections between disparate literatures across diverse subdisciplines. Even more difficult than trying to analyze published text is the effort to analyze the health-care-worker notes in EMRs. These are not typically subject to editing, and often do not include full sentences but instead consist of phrases and shorthand abbreviations. Nonetheless, the ability to connect rich descriptions of patient phenotypes with controlled measurements of lab values, imaging studies, and coded diagnoses offers the opportunity to better understand the clinical responses to diseases and therapies. Ultimately, the availability of tissue samples within biobanks will also allow for connections to be made between these descriptions and the underlying molecular players. Exciting work has been published on the analysis of US Food and Drug Administration (FDA) data and the use of text to interpret candidate gene sets, create links between drugs and their targets, and predict drug–drug interactions.

The FDA makes the text of package inserts publicly available. Kuhn *et al.* showed that FDA labels can be downloaded and analyzed to create a valuable database of side effects.¹⁹ They associated 1,450 side effects with 888 drugs to create the SIDER database of on-label side effects.

The biological background knowledge contained in text can be used to guide the interpretation of experimental data. Raychaudhuri *et al.* showed that literature mining could be used to create a set of candidate genes for Crohn's disease and schizophrenia, so that the search for mutations could be focused on relevant biological pathways, thereby increasing interpretability and viability of hits.²⁰

The task of indexing large amounts of text requires automated systems for content-based annotation. Shah *et al.* have created a program that analyzes a text-based resource (e.g., PubMed, patent literature, or FDA labels), annotates individual entries using controlled vocabularies, indexes them, and then links the resources together.²¹

Yildirim *et al.* showed that they could mine FDA data to build associations between drugs and targets.²² In fact, they created a network connecting multiple drugs to multiple targets, and thereby defined the overall landscape of the (current) pharmacologic space. They observed the proliferation of “me too” drugs, and they showed that new drugs are more diverse, in the sense that they hit newer targets.

PubMed contains many reported interactions between genes and drugs, and some of these interactions may imply a mechanism for drug–drug interactions. Percha *et al.* showed that the gene–drug relationships contained within PubMed can be automatically extracted and subsequently used to build a classifier to predict drug–drug interactions.²³ The use of high-fidelity text extraction leads not only to predicted interactions, but also to predicted mechanisms for these interactions.

NETWORKED OR SYSTEMS MEDICINE

The field of systems biology focuses on understanding the behavior of entire biological systems, not just individual components. Systems biology methods range from qualitative network diagrams that associate molecular and phenotypic entities to formal quantitative models based on differential equations. These network views enable the study of emergent phenomena that become apparent only in the context of a comprehensive model of the interactions within a cell and between cells in tissues. Systems medicine takes the ideas of systems biology, applies them to human biology, and focuses on characterizing disease states at the molecular level as well as on therapeutic interventions. Systems pharmacology, in turn, takes the same approach toward the more specific task of understanding drug response, moving away from a one-drug-one-target model to one that is more compatible with the idea that “the network is the target” (the “network” referred to being the network of molecules that interact with one another and with drugs).

The main challenges of network-based or systems medicine are to create models that are useful and predictive, and to validate them. Many of the ideas of system identification from fields of engineering are relevant to these pursuits, and recent work has focused on networks

for human metabolism, cancer, and stem cells. Combining “top down” use of text and “bottom up” use of genomic information, Duarte *et al.* totally reconstructed a human metabolic network based on literature and genome.²⁴ They discovered missing elements of human metabolism and created a starting point for a numerical simulation of human cellular metabolism.

There is much ongoing discussion about whether drug targets tend to be central or peripheral in the network of protein interactions. Goh *et al.* built a network of human diseases and the associated genes from the Online Mendelian Inheritance in Man resource.²⁵ They found that cancer genes are more “central” in the network, as are genes that are essential for life. Most drug targets are not central.

The regulation of stem cell differentiation is a critical challenge for all of biology but has special implications for our ability to differentiate cells for pharmacologic testing. Xu *et al.* reported building a complete computational embryonic stem cell regulatory network by aggregating information from multiple databases, publications, and software packages.²⁶ The result was a large network of transcriptional control of embryonic stem cells, allowing modeling of the effects of introducing different transcription factors in an attempt to encourage differentiation or de-differentiation.

PUBLIC HEALTH AND SOCIETAL APPLICATIONS OF INFORMATICS

There is growing interest in the use of public health records and other population-based clinical information resources (such as those made available by the FDA) for connecting basic biology to medicine. These efforts include detecting and understanding the mechanisms of drug effects and drug interactions. They also include surveillance of the population for signs of communicable diseases such as influenza. The newest set of data sources are those emerging from computer-based social networks in which patients or providers report signs, symptoms, and medical issues. Critical to these efforts is the creation of an infrastructure to capture and share these data. Search engine usage statistics can indicate changes in population health, as people affected by disease or adverse drug effects go to the Internet to learn about the potential causes of their symptoms. Bioinformatics also has application in preventing and detecting bioterrorism activities.

The availability of inexpensive techniques for characterizing food and pathogens using DNA sequencing offers opportunities for testing the provenance of food sources and sources of infection on the basis of genetic similarity and determination of source and ancestry. Trifonov *et al.* analyzed the sequences of key influenza genes and created an evolutionary map of the recombination events leading to the current H1N1 strain.²⁷ They showed that aggressive sampling of species may allow anticipation of future novel strains of influenza.

As our understanding of genetic factors increases, the paucity of information about environmental influences limits our predictive capabilities. Liu *et al.* used 3,342 Medical Entity Subject Headings associated with 3,159 diseases, and linked them to 1,100 genes and 1,034 diseases to create a joint clustering for better understanding the influence of environmental factors on genetic diseases.²⁸ They called this the “etiome” (the collection of all etiologies).

The informatics for integrating biology and the bedside system (i2b2) for translational research²⁹ connects to clinical data warehouses and provides support in both cohort finding and in linking clinical data to molecular and cellular data for the purposes of discovery. It has been installed in more than 50 hospitals and other institutions nationwide.

Most Internet content is text-based and is not marked up sufficiently to allow computers to understand the contents or make inferences about it. A new Web standard is needed to facilitate computational understanding of Web content. Ruttenberg *et al.* provide a vision of a more “semantic” web in which knowledge and relationships are explicitly encoded between entities of interest (such as genes, drugs, diseases, symptoms, patients, and biomarkers).³⁰ They described some of the emerging standards and the promise these standards hold of a more useful data integration infrastructure. Dudley and Butte provided a review of the opportunities and challenges for cloud computing in translational bioinformatics.³¹ The most exciting possibility is that of capturing full copies (or “images”) of computing environments in virtual form so that they can be transferred and run in multiple locations. This provides the potential for publications in translational bioinformatics to be associated with such images, so that readers can literally boot up the entire software and data environment of the authors, and attempt to replicate or extend the results.

PRIVACY OF GENETIC INFORMATION, ETHICS, AND RE-IDENTIFICATION

Translational bioinformatics raises issues of privacy and ethics because it can involve the use of individual patient records. Genome data can be linked to EMRs containing high-fidelity phenotypic measurements, and this may pose a risk of identifying patients (with rare disease, or unusual combinations of findings) whose anonymity has been guaranteed as part of informed consent.

Homer *et al.* showed that a single individual can be identified in a DNA mixture made up of 1,000 individuals, if a high-throughput measurement of that individual’s SNP genotypes is available.³² Previously, it was thought that a mixture of 1,000 genomes would not allow the determination of whether a single individual was part of that mix, but the authors show a statistical method for evaluating this, and show that individuals can be detected with surprisingly high levels of confidence. Nyholt *et al.* showed that, even though the ApoE status of Jim Watson’s genome was redacted, it is trivial to write an algorithm to impute the ApoE alleles on the basis of linkage disequilibrium.³³ The authors did not release actual details of the alleles but made the point that genetic information is hard to hide once it has been partially disclosed. Christley *et al.* published an algorithm for compressing a genome from 3.4 GB of initial data to 4.1 MB fully compressed,³⁴ small enough for a routine e-mail attachment!

Mining EMRs in the pursuit of molecular connections can expose these records to potential abuse or unintentional release because the research environment is often less secure than the clinical care delivery environment. Because of these challenges, there are many debates about whether it is better for patients to (by default) opt in to genomic and data-mining studies or opt out. The two strategies may produce different rates of participation, and there

are considerable technical challenges to implementing each of them. Finally, no matter how consent is obtained, there are important issues of whether research findings should be reported back to study participants, and the technical means for doing this. This issue is accentuated in the case of whole-genome data.³⁵ Vanderbilt University is building an opt-out de-identified biobank.³⁶ The use of opt-out (vs. opt-in consent) leads, not surprisingly, to very high rates of accrual. Importantly, the deployment of aggressive education and outreach also leads to a low rate of participant dissatisfaction.

Conflicting interpretations may be medically reasonable, but they may lead to loss of consumer or physician confidence. Ng *et al.* compared two direct-to-consumer genetic testing companies and found that, although the agreement as to the raw data regarding genotypes was good, the interpretation varied significantly.³⁷ For seven diseases, <50% of the predictions agreed. These observations raise the issue of whether genome interpretations should be regulated by FDA or other oversight bodies.

SYSTEMS PHARMACOLOGY

The opportunities for pharmacology are exciting because drugs are both molecular entities and clinical entities, thereby presenting a very natural means of joining information resources relating these two worlds. The obvious importance of understanding drug action at both the molecular and clinical levels has led to the accumulation of many valuable data resources that associate drugs with clinical concepts (Table 3). Recent work has demonstrated some surprising capabilities.

The side effects of drugs can be used to link drugs with similar genetic targets. Campillos *et al.* showed that the similarity of side-effect profiles of two drugs can be used to impute their targets.³⁸ They created a drug–drug network of 746 drugs and 1,018 relationships, using similarities in side-effect profiles to link drugs. They predicted 261 relationships between chemically dissimilar molecules, and tested 20 of these for binding; 13 of the binding assays were positive. Off-target binding of drugs can also be interrogated using systems approaches. Chang *et al.* showed that three-dimensional structural protein data can be combined with systems modeling methods to understand the side effects of drugs caused by off-target binding.³⁹ In particular, the mechanism through which torcetrapib causes hypertension was dissected using a combined systems–molecular modeling paradigm. Systems pharmacology can sometimes lead to counterintuitive treatment strategies. Torella *et al.* showed that nonlinear dose effects in multidrug antibiotic regimens can be understood through mathematical modeling of the antagonism and synergies associated with different doses.⁴⁰ Surprisingly, the optimal strategy is not always to maximize synergy; mild antagonism may be advantageous.

The drug-binding preferences of targets can be compared to suggest potential repurposing opportunities for drugs, using only the associations mined using informatics. Candidate genes involved in modulating the responses to drugs can be predicted using a network of genetic interactions. Indeed, the cellular gene expression changes created by exposure to a drug can be used to predict new indications for that drug, as well as likely side effects. It is clear that traditional experimental characterization of drug response should be augmented

with computational analysis. These analyses create focused hypotheses that are less expensive to test than large-scale screens. Some of these analyses contribute to our capabilities in pharmacogenomics, because the approach depends on a molecular understanding of drug response. Keiser *et al.* predicted new targets for old drugs by comparing the list of compounds known to bind targets⁴¹; that is, if two targets are associated with a similar list of drugs (based on a statistical analysis of expected vs. observed overlap) then drugs that bind one target are likely to bind the other. This provides another methodologic opportunity for drug repurposing. Kinnings *et al.* showed that similarities in binding sites can be used to predict new uses for old drugs, termed “repurposing.”⁴² They found that a drug used in the treatment of Parkinson’s disease (Comtan) binds a tuberculosis protein and inhibits the progress of tuberculosis. This has opened up the field of structure-based repurposing with several subsequent similar applications. Suthram *et al.* reviewed 4,600 coexpressed functional modules in humans and found that drug targets are not equally distributed in these modules.⁴³ Moreover, many diseases involve overlapping modules, and therefore there are opportunities for repurposing drugs (or developing new ones) based on the presence of a highly “drug-gable” module. Hansen *et al.* showed that a protein–protein interaction network can be used to create a method for ranking proteins in the genome according to the probability that they will modulate the response to a particular drug.⁴⁴ The method uses known protein–drug interactions (both direct and genetic) and then trains a classifier for ranking the genes most likely to be involved in the pharmacodynamics or pharmacokinetics of an input drug.

The future offers the potential of translational bioinformatics to enable more efficient discovery, development, and regulation of new therapeutics.

Each of these areas within translational bioinformatics is relevant to a better understanding of drug action and augmenting drug discovery. Because drugs are central both to molecular physiology and to clinical care, translational bioinformatics promises to impact our understanding of drug response from multiple perspectives:

- The ability to measure whole-genome sequences and relate genetic variation to drug-response phenotypes enables the emergence of pharmacogenomic understanding. This has the potential to drive not only (in the near term) optimal use of existing drugs but also (in the longer term) the development of drugs targeted to particular individuals.
- The ability to associate clinical and molecular biomarkers with drug response, efficacy, and side effects should potentiate the development of a deeper understanding of how drugs affect multiple physiological systems and how we can more intelligently “read out” the detailed physiologic response to a therapeutic, thereby monitoring individual response.
- The published literature is a rich source of information about drug uses, drug effects and adverse events, and drug mechanism and metabolism. Our ability to extract this information systematically from published literature (including the patent literature, case-law literature, and FDA documentation) should yield a rich network of information about each drug.

- A systems understanding of drug response will allow us to move away from a one-drug-one-target model of pharmacology to a more comprehensive view, namely, that the effects of a drug extend to the entire physiology of the cell, organ, and organism. The emerging discipline of systems pharmacology should allow the identification of new uses for existing drugs through rational repurposing, and the design of new therapeutic strategies that take advantage of multiple drugs to synergistically interfere with disease pathways.
- The rich databases from population sources, including the Centers for Disease Control and Prevention, FDA, hospital EMRs, and the patients themselves through Internet health sites and social networking sites will allow unprecedented opportunities for surveillance of drug response. It will also provide new models for drug clinical trials, taking advantage of new technologies for gathering and aggregating data on patient response to treatment.
- This is an exciting time for translational bioinformatics. The fruits of basic biology and the increasing instrumentation of the health-care system provide rich data that require interpretation by a cohort of informatics professionals who understand both biology and medicine, and who are trained in both discovery and application. Translational bioinformatics is not a panacea, and its practitioners must work with colleagues in basic biology and clinical medicine to make maximal impact. But translational bioinformatics is the discipline that will curate, annotate, aggregate, and analyze the rich biomedical data sets that will emerge in the coming years.

References

1. Ashley EA, et al. Clinical assessment incorporating a personal genome. *Lancet*. 2010; 375:1525–1535. [PubMed: 20435227]
2. Ritchie MD, et al. Robust replication of genotype-phenotype associations across multiple diseases in an electronic medical record. *Am. J. Hum. Genet.* 2010; 86:560–572. [PubMed: 20362271]
3. Eriksson N, et al. Web-based, participant-driven studies yield novel genetic associations for common traits. *PLoS Genet.* 2010; 6:e1000993. [PubMed: 20585627]
4. Iossifov I, Zheng T, Baron M, Gilliam TC, Rzhetsky A. Genetic-linkage mapping of complex hereditary disorders to a whole-genome molecular-interaction network. *Genome Res.* 2008; 18:1150–1162. [PubMed: 18417725]
5. Chang MH, et al. Prevalence in the United States of selected candidate gene variants: Third National Health and Nutrition Examination Survey, 1991–1994. *Am. J. Epidemiol.* 2009; 169:54–66. [PubMed: 18936436]
6. Klein TE, et al. Estimation of the warfarin dose with clinical and pharmacogenetic data. *N. Engl. J. Med.* 2009; 360:753–764. [PubMed: 19228618]
7. Salzberg SL, Pertea M. Do-it-yourself genetic testing. *Genome Biol.* 2010; 11:404. [PubMed: 20932271]
8. Alizadeh AA, et al. Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature.* 2000; 403:503–511. [PubMed: 10676951]
9. Paik S. Is gene array testing to be considered routine now? *Breast.* 2011; 20(suppl. 3):S87–S91. [PubMed: 22015300]
10. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature.* 2007; 447:661–678. [PubMed: 17554300]
11. Ng SB, et al. Exome sequencing identifies the cause of a mendelian disorder. *Nat. Genet.* 2010; 42:30–35. [PubMed: 19915526]

12. Roach JC, et al. Analysis of genetic inheritance in a family quartet by whole-genome sequencing. *Science*. 2010; 328:636–639. [PubMed: 20220176]
13. Denny JC, et al. PheWAS: demonstrating the feasibility of a phenome-wide scan to discover gene-disease associations. *Bioinformatics*. 2010; 26:1205–1210. [PubMed: 20335276]
14. Patel CJ, Bhattacharya J, Butte AJ. An Environment-Wide Association Study (EWAS) on type 2 diabetes mellitus. *PLoS ONE*. 2010; 5:e10746. [PubMed: 20505766]
15. McGary KL, Park TJ, Woods JO, Cha HJ, Wallingford JB, Marcotte EM. Systematic discovery of nonobvious human disease models through orthologous phenotypes. *Proc. Natl. Acad. Sci. U.S.A.* 2010; 107:6544–6549. [PubMed: 20308572]
16. Wang K, et al. Genome-wide identification of post-translational modulators of transcription factor activity in human B cells. *Nat. Biotechnol.* 2009; 27:829–839. [PubMed: 19741643]
17. Sirota M, Schaub MA, Batzoglu S, Robinson WH, Butte AJ. Autoimmune disease classification by inverse association with SNP alleles. *PLoS Genet.* 2009; 5:e1000792. [PubMed: 20041220]
18. Dickson SP, Wang K, Krantz I, Hakonarson H, Goldstein DB. Rare variants create synthetic genome-wide associations. *PLoS Biol.* 2010; 8:e1000294. [PubMed: 20126254]
19. Kuhn M, Campillos M, Letunic I, Jensen LJ, Bork P. A side effect resource to capture phenotypic effects of drugs. *Mol. Syst. Biol.* 2010; 6:343. [PubMed: 20087340]
20. Raychaudhuri S, et al. Identifying relationships among genomic disease regions: predicting genes at pathogenic SNP associations and rare deletions. *PLoS Genet.* 2009; 5:e1000534. [PubMed: 19557189]
21. Shah NH, Jonquet C, Chiang AP, Butte AJ, Chen R, Musen MA. Ontology-driven indexing of public datasets for translational bioinformatics. *BMC Bioinformatics*. 2009; 10(suppl. 2):S1. [PubMed: 19208184]
22. Yildirim MA, Goh KI, Cusick ME, Barabási AL, Vidal M. Drug-target network. *Nat. Biotechnol.* 2007; 25:1119–1126. [PubMed: 17921997]
23. Percha B, Garten Y, Altman RB. Discovery and explanation of drug-drug interactions via text mining. *Pac. Symp. Biocomput.* 2012:410–421. [PubMed: 22174296]
24. Duarte NC, et al. Global reconstruction of the human metabolic network based on genomic and bibliomic data. *Proc. Natl. Acad. Sci. U.S.A.* 2007; 104:1777–1782. [PubMed: 17267599]
25. Goh KI, Cusick ME, Valle D, Childs B, Vidal M, Barabási AL. The human disease network. *Proc. Natl. Acad. Sci. USA.* 2007; 104:8685–8690. [PubMed: 17502601]
26. Xu H, Schaniel C, Lemischka IR, Ma'ayan A. Toward a complete in silico, multi-layered embryonic stem cell regulatory network. *Wiley Interdiscip. Rev. Syst. Biol. Med.* 2010; 2:708–733. [PubMed: 20890967]
27. Trifonov V, Khiabani H, Rabadan R. Geographic dependence, surveillance, and origins of the 2009 influenza A (H1N1) virus. *N. Engl. J. Med.* 2009; 361:115–119. [PubMed: 19474418]
28. Liu YI, Wise PH, Butte AJ. The “etiome”: identification and clustering of human disease etiological factors. *BMC Bioinformatics*. 2009; 10(suppl. 2):S14. [PubMed: 19208189]
29. Murphy SN, et al. Serving the enterprise and beyond with informatics for integrating biology and the bedside (i2b2). *J. Am. Med. Inform. Assoc.* 2010; 17:124–130. [PubMed: 20190053]
30. Rutenber A, et al. Advancing translational research with the Semantic Web. *BMC Bioinformatics*. 2007; 8(suppl. 3):S2. [PubMed: 17493285]
31. Dudley JT, Pouliot Y, Chen R, Morgan AA, Butte AJ. Translational bioinformatics in the cloud: an affordable alternative. *Genome Med.* 2010; 2:51. [PubMed: 20691073]
32. Homer N, et al. Resolving individuals contributing trace amounts of DNA to highly complex mixtures using high-density SNP genotyping microarrays. *PLoS Genet.* 2008; 4:e1000167. [PubMed: 18769715]
33. Nyholt DR, Yu CE, Visscher PM. On Jim Watson's APOE status: genetic information is hard to hide. *Eur. J. Hum. Genet.* 2009; 17:147–149. [PubMed: 18941475]
34. Christley S, Lu Y, Li C, Xie X. Human genomes as email attachments. *Bioinformatics*. 2009; 25:274–275. [PubMed: 18996942]
35. Ormond KE, et al. Challenges in the clinical application of whole-genome sequencing. *Lancet*. 2010; 375:1749–1751. [PubMed: 20434765]

36. Pulley J, Clayton E, Bernard GR, Roden DM, Masys DR. Principles of human subjects protections applied in an opt-out, de-identified biobank. *Clin. Transl. Sci.* 2010; 3:42–48. [PubMed: 20443953]
37. Ng PC, Murray SS, Levy S, Venter JC. An agenda for personalized medicine. *Nature.* 2009; 461:724–726. [PubMed: 19812653]
38. Campillos M, Kuhn M, Gavin AC, Jensen LJ, Bork P. Drug target identification using side-effect similarity. *Science.* 2008; 321:263–266. [PubMed: 18621671]
39. Chang RL, Xie L, Xie L, Bourne PE, Palsson BØ. Drug off-target effects predicted using structural analysis in the context of a metabolic network model. *PLoS Comput. Biol.* 2010; 6:e1000938. [PubMed: 20957118]
40. Torella JP, Chait R, Kishony R. Optimal drug synergy in antimicrobial treatments. *PLoS Comput. Biol.* 2010; 6:e1000796. [PubMed: 20532210]
41. Keiser MJ, et al. Predicting new molecular targets for known drugs. *Nature.* 2009; 462:175–181. [PubMed: 19881490]
42. Kinnings SL, Liu N, Buchmeier N, Tonge PJ, Xie L, Bourne PE. Drug discovery using chemical systems biology: repositioning the safe medicine Comtan to treat multi-drug and extensively drug resistant tuberculosis. *PLoS Comput. Biol.* 2009; 5:e1000423. [PubMed: 19578428]
43. 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.* 2010; 6:e1000662. [PubMed: 20140234]
44. Hansen NT, Brunak S, Altman RB. Generating genome-scale candidate gene lists for pharmacogenomics. *Clin. Pharmacol. Ther.* 2009; 86:183–189. [PubMed: 19369935]

Table 1**Key clinical data sources**

Standard terminologies for diseases: Unified Medical Language System, the Medical Entities Subject Headings, the Systematized Nomenclature of Medicine (SNOMED), and others.
Standard terminologies for adverse events: The MedDRA terminology, SNOMED, and others.
Standard terminologies for drugs: the RXNorm Resource and the Veterans Affairs National Drug File–Reference Terminology.
Longitudinal electronic medical records at individual research centers; these records have been de-identified to allow discovery research.
US Food and Drug Administration drug labels and Adverse Events Reporting System.
Standard identifiers for laboratory tests: Logical Observation Identifiers Names and Codes and standards for passing health information between systems, allowing interoperability through Health Level 7.
The Anatomical Therapeutic Chemical classification system associates substances with the organ or system upon which they act, including therapeutic, pharmacologic, and chemical properties.
Too many others to enumerate, all providing valuable clinical information, taxonomies, or terminologies.

Table 2**Key bioinformatics data sources**

Genbank, DNA sequence data, all species.
The UCSC Golden Path Human Genome Browser and the Ensembl Human Genome Browser: access to human genome sequence information and many functional annotations.
Human Genome Nomenclature Committee: standard gene-naming conventions.
PubMed: published biomedical literature.
Gene Expression Omnibus: gene expression data sets.
Protein Data Bank: three-dimensional protein, RNA, and DNA structures.
HumanCyc and KEGG pathways: databases of human metabolism.
Reactome: database of biological pathways.
Many other useful biological databases, particularly large high-throughput experimental data sets that provide information about molecular interactions.

Table 3

Resources that link the molecular and the clinical levels

PubChem: a library of drugs and small molecules.
ChEMBL: database of bioactive drug-like small molecules and their biological binding properties.
DrugBank: a database of drugs and small molecules associated with key properties (particularly targets) and linked to other data resources.
PharmGKB: a database of human genes known to be involved in drug response.
Online Mendelian Inheritance in Man: a database of genetic disorders and their associated genetic variations.
FDA drug labels: information about the molecular mechanisms, side effects, pharmacology, and indications for drugs, among many other characteristics.

FDA, US Food and Drug Administration.