Translational Bioinformatics

25

Jessica D. Tenenbaum, Nigam H. Shah,
and Russ B. Altman

After reading this chapter, you should know the answers to these questions:

- How does translational bioinformatics differ from the more general field of bioinformatics?
- What do T1 and T2 refer to in the context of translational research?
- What is a biomarker, and why is it important in medicine?
- What is personalized medicine, and how does it differ from traditional medical practice?
- What is the difference between pharmacokinetics and pharmacodynamics?
- What changes are needed from a clinical IT perspective to support personalized medicine?
- What is the difference between statistical significance and clinical significance?
- How are genomic data being used today in research, clinical care, and consumer health?
- What are some legal and ethical issues surrounding direct-to-consumer genetic testing?
- How are ontologies useful in translational bioinformatics?

J.D. Tenenbaum, PhD (✉)
Duke Translational Medicine Institute,
Duke University, 2424 Erwin Rd,
Durham 27705, NC, USA
e-mail: jessie.tenenbaum@duke.edu

N.H. Shah, MBBS, PhD
Department of Medicine, Stanford University,
1265 Welch Road, X-229, Stanford 94305, CA, USA
e-mail: nigam@stanford.edu

R.B. Altman, MD, PhD
Departments of Bioengineering,
Genetics and Medicine, Stanford University,
Clark Center, 318 Campus Drive, S242,
Stanford 94305, CA, USA
e-mail: russ.altman@stanford.edu

25.1 What Is Translational Bioinformatics?

The preceding chapter described the field of bioinformatics, or the study of how information from biological systems is represented and analyzed. **Translational Bioinformatics** (**TBI**) is bioinformatics applied to human health and disease. It uses and extends the concepts and methods from bioinformatics to facilitate the practice of **translational medicine**, i.e. the translation of biological ("bench") discoveries into actual impact on clinical care ("bedside") and ultimately on population health (Fig. 25.1). The American Medical Informatics Association (AMIA) defines Translational Bioinformatics as "the development of storage, analytic, and interpretive methods to optimize the transformation of increasingly voluminous biomedical data, and genomic data, into proactive, predictive, preventive, and participatory health."[1] Those latter terms are often grouped together along with the more common descriptive of **personalized medicine**. The realization of personalized medicine will require methods for standards-based data

[1] http://www.amia.org/applications-informatics/translational-bioinformatics (Accessed 30/11/2012).

E.H. Shortliffe, J.J. Cimino (eds.), *Biomedical Informatics*,
DOI 10.1007/978-1-4471-4474-8_25,

storage and retrieval, novel methods for analysis and interpretation, and aid to the clinician in decision support. In AMIA's description of TBI, they further state that it includes "…the evolution of clinical informatics methodology to encompass biological observations." That is, clinical informatics approaches will need to expand in scope to incorporate **omics** data (genomics, transcriptomics, metabolomics, and proteomics data) as it increasingly pertains to clinical care. In this chapter, we will discuss the three components of AMIA's definition—the novel methods, the voluminous data, and the personalized approach to health that TBI enables. We conclude with a discussion of challenges and future directions for the field.

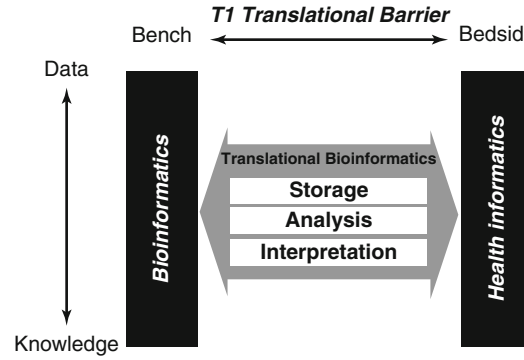

Fig. 25.1 TBI bridges the gap between bioinformatics on the "bench" side of the T1 barrier and health informatics on the "bedside" end of the spectrum. Novel methods for storage, analysis, and interpretation span the spectrum from data to knowledge (Adapted with permission from (Sarkar et al. 2011a, b))

25.1.1 Differences from "Traditional" Bioinformatics

TBI differs from the larger field of bioinformatics in a number of key ways. As described above, the focus of TBI is human health. As such, the discipline centers primarily, though not exclusively, around human data. This fact has a number of implications from an informatics perspective. First, one encounters a range of data management, regulatory, and privacy issues that do not arise in handling data from mice, yeast, *Escherichia coli*, or other model organisms. Laws such as the Health Information Portability and Accountability Act (HIPAA)[2] (see Chap. 10) dictate how patient data must be handled and safeguarded to protect patient privacy. Title 21 of the Code of Federal Regulations Part 11 (21 CFR part 11)[3] mandates how data must be managed if they are to be included as part of a submission to the Food and Drug Administration. In addition, institutional review boards (IRBs) typically require measures to ensure subject confidentiality before they will approve a research protocol. Making complete datasets publically accessible for a mouse experiment is good scientific citizenship. Making the same type of data accessible for a human study would be a serious violation of privacy and confidentiality.

Another difference is that while experimental perturbation through small molecule agonists or antagonists, siRNA, or knock-out genes are straightforward and common in yeast or *E. coli*, such approaches would be neither feasible nor ethical in human subjects. This has significant implications for data generation and collection in translational research. **Phase I clinical trials** are the notable exception to this rule, but they are performed only on ostensibly therapeutic agents. They also require a number of preliminary steps, are very expensive, and are performed in a very small number of subjects. Other factors that differentiate research with human subjects include genetic and environmental heterogeneity, which can be controlled in model organisms. Instead, much translational data from human beings comes from *in vitro* experiments on cell lines and observational inquiries regarding factors such as genotype, environmental factors, and outcomes. With so much inherent noise, very large sample sizes are typically required for new discoveries. Novel approaches to data integration, mining, and reuse are thus particularly important in translational research.

[2] http://aspe.hhs.gov/admnsimp/pl104191.htm (Accessed 30/11/2012).

[3] http://www.accessdata.fda.gov/scripts/cdrh/cfdocs/cfcfr/cfrsearch.cfm?cfrpart=11 (Accessed 30/11/2012).

25.1.2 A Few Words About Omics and Hamburgers

The word "genome" was coined in 1930 by a German Botanist named Hans Winkler, as a combination of the words "gene" and "chromosome."[4] The suffix "-ome" has subsequently come to be used in the context of biology to mean the totality of a thing, e.g. the *proteome* represents all proteins, and *proteomics*, the study of all proteins. Some use the term "genomics" to refer to information from genomes *and* their derivatives, i.e. RNA, proteins, and metabolites. Others use the more general term "omics" to refer to these different totalities. Here we shall adopt the more general *omics* neologism.

Omics technologies, then, are frequently referred to as "high-throughput" methods. This usage is not inaccurate in that many of these methods can be done in batch, for example in 384-well plates. However, focusing on the ability to run these assays *en masse* largely misses the point—the breadth or *bandwidth* of the technologies. A genetic assay measures one gene; a genomic assay measures thousands of genes. As an analogy, consider the food industry: McDonald's sells large quantities of food, quickly. In that respect, it is a high-throughput operation. But the breadth of their offerings is fairly narrow. A grocery store, on the other hand, sells thousands of types of food items, nearly every kind available. It is both high throughput (assuming cashiers are both plentiful and competent) and *high-bandwidth* (Butte 2009). This concept of bandwidth, the ability to observe thousands of types of molecules in a given assay, is the key attribute to these omic technologies.

25.2 The Rise of Translational Bioinformatics

25.2.1 Promise of the Human Genome Project

In January of 2000, two different groups announced that they had fully sequenced the human genome (see Chap. 24). The public project, published in *Nature*, was based on multiple individuals (Lander et al. 2001). The other genome, published in *Science*, was a private venture, performed on the DNA of biologist and entrepreneur Craig Venter (Venter et al. 2001). The vision for the human genome was that once all the genes were identified, they could be assigned functional annotations, and we would thus be able to understand what goes wrong when human beings succumb to disease. Additionally, this knowledge would help us to understand exactly which pathways and molecules needed to be targeted in order to prevent or cure disease. Of course, biological reality is not quite so straight forward. To begin with, the "central dogma" of biology (Crick 1970)—DNA is transcribed into mRNA, which is then translated into protein—is overly simplistic. Variations in regulatory regions can affect when the gene is turned on, and to what degree. Most genes have a number of different splice variants, producing a number of different proteins. In addition, proteins undergo post-translational modifications, which impact their structure and function. Finally, additional complexity is added through **epigenetics**, or heritable traits that are not coded for through DNA sequencing alone. An example is methylation of the DNA molecule, which has been shown to affect transcription (Cedar 1988). Despite this, the sequencing of the complete human genome marked a decisive turning point in biomedical research. The parts list had been assembled and researchers could move on to the more interesting aspects of the genome—what each part does, how the parts differ among individuals, and what it all means. The impact this would have on the field of clinical informatics was recognized immediately, reflected in the theme for the 2002 AMIA[5] Annual Symposium: "Bio*Medical Informatics: One Discipline" (Tarczy-Hornoch et al. 2007).

[4] http://dictionary.reference.com/browse/genome (Accessed 30/11/2012).

[5] AMIA is the American Medical Informatics Association, Bethesda, MD: http://www.amia.org

25.2.2 Translational Research and the CTSA Era

In the early 2000s, there was growing acknowledgement that the population at large, and patients in particular, were not reaping the full benefits of the considerable amount of research money being devoted to scientific discovery. It was recognized that researchers do not do a good job translating their discoveries "from bench to bedside," i.e. between biological discoveries in the lab and clinical application of the findings (Lenfant 2003). Two significant roadblocks were identified (Fig. 25.2)—one in translating discoveries into clinical care guidelines (dubbed T1 translation), and the other in translating clinical care guidelines into actual practice (T2 translation) (Sung et al. 2003). In 2004, the National Institutes of Health (NIH) launched the Roadmap for Medical Research, aimed at transforming life science research in the twenty-first century. Biomedical informatics plays a strong role across all three of the major Roadmap themes: New Pathways to Discovery, Research Teams of the Future, and Reengineering the Clinical Research Enterprise (Zerhouni 2006). As part of this Roadmap, the NIH embarked on a major initiative to break down translational barriers. The Clinical and Translational Science Award (CTSA) was a new funding mechanism established through the National Center for Research Resources within the NIH, with 12 awards in 2006 and approximately twelve each subsequent year until 2011, when the CTSA consortium comprised 60 institutions. As indicated in the NIH's request for applications, Biomedical Informatics was considered a key functional component of a Clinical and Translational Science Institution.[6] In fact, the word "informatics" appeared 38 times in the RFA itself, reflecting recognition by the NIH that informatics is crucial to addressing challenges in both T1 and T2 translational research (Butte 2008b).

It was in this context, with newfound attention to translational research, that Butte and Chen coined the term "translational bioinformatics" at the AMIA annual symposium in 2006 in a paper entitled "Finding Disease-Related Genomic Experiments Within an International Repository: First Steps in Translational Bioinformatics" (Butte and Chen 2006). To assist in translating discoveries made using genomic technologies to medicine, they asserted, this emerging discipline "is focusing on the development of analytic, storage, and interpretive methods to optimize the transformation of increasingly voluminous genomic and biological data into diagnostics and therapeutics for the clinician." From there, the field became increasingly recognized by the larger informatics community. AMIA added TBI as one of its key supported domains and in 2008 held its first annual Summit on Translational Bioinformatics. (Although the "annual" qualifier was cautiously withheld until the event had proven to be a success.) Later that year, the *Journal of the American Medical Informatics Association* (JAMIA) published a perspective on TBI's "Coming of Age" that enumerated several reasons why the time was right for TBI to come into its own as a field (Butte 2008b). In 2009, the editors of the *Journal of Biomedical Informatics* published an explicit Editorial to announce a change in the journal's editorial policy to "focus its bioinformatics attention on innovations in the area of *translational bioinformatics*" (Shortliffe et al. 2009).

25.2.3 Personalized Medicine as a Driving Force

Personalized medicine is health care that is based on an individual's unique clinical, genetic, omic, and environmental profile, in addition to his or her specific values and preferences. The focus of personalized medicine is not only on diagnosis and therapeutic course, once a person falls ill. It also helps to guide health management based on individualized risk of disease in the future. Historically, guidelines for a healthy lifestyle have been presented as universal. Clinical care guidelines have been based primarily on macroscopic symptoms and qualities that can be observed in the course of a physical exam or reported by the patient. Personalized medicine aims to take advantage of all available modalities

[6] http://grants.nih.gov/grants/guide/rfa-files/RFA-RM-06-002.html (Accessed 30/11/2012).

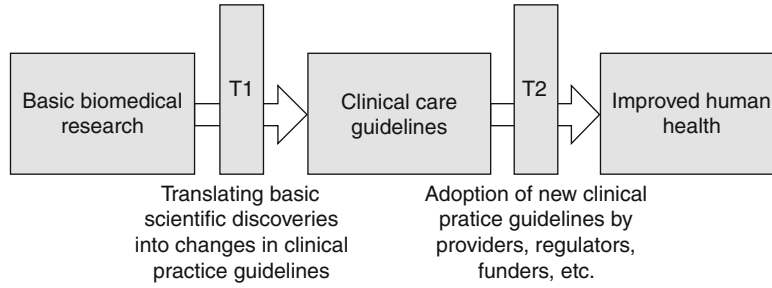

Fig. 25.2 Translational roadblocks along the continuum of biomedical research from scientific discoveries to changes in clinical practice and improvement of human health

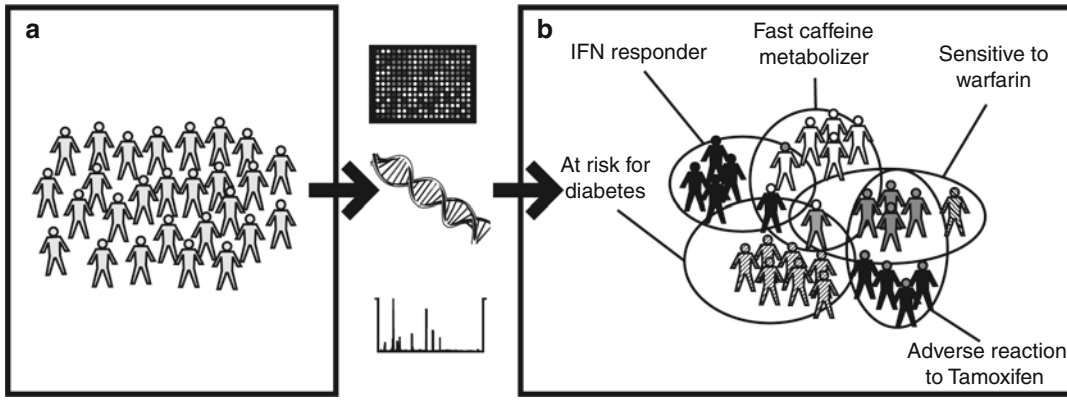

Fig. 25.3 Stratified or genomic medicine. A seemingly homogeneous group of people (**a**) can be divided into subgroups (**b**) based on molecular fingerprints. This stratification can help to guide therapeutic decision

and sources of information regarding the individual. This may entail such non-traditional aspects as genotypic information or other molecular **biomarkers**, but also exposure to chemical cleaning agents, aversion to risk, a high threshold for pain, the impending birth of a first grandchild, etc. Of course, some of these variables have always been a part of clinical care to some extent, together with some degree of intangible qualitative judgment on the part of the clinician. Personalized medicine formalizes this integrated, patient-based approach, and incorporates the omics aspect. In 2004, Lee Hood coined the term "P4 medicine": predictive, personalized, preventive, and participatory (Weston and Hood 2004). Based on an individual's specific risk factors, interventions or changes in lifestyle could be adopted *before* the person falls ill, improving quality of life *and* saving significant costs in health care spending. Armed with this individualized knowledge, patients would be empowered to play an active role in their own health and medical care. Quality medical care has never been one-size-fits-all; personalized medicine acknowledges this fact and seeks to change the practice of clinical care accordingly.

Many would agree that personalized medicine is the lodestar of medical practice; few would argue that we are close to achieving it. The same translational barriers that apply to clinical research more generally apply to personalized medicine in particular. Discoveries in the lab still take years to be incorporated into clinical care guidelines, and personalized clinical care guidelines take years to be widely adopted in practice. One stepping stone to truly personalized medicine that arguably *has* been achieved, and continues to evolve, is **genomic medicine**, also called **stratified medicine** (Trusheim et al. 2007; Ginsburg and Willard 2009). Stratified medicine involves clinical care that is based on specific qualities of an individual, often derived from molecular fingerprints. Subjects may be divided into different classes based on their genetics or some other high dimensional omic-scale biomarker (Fig. 25.3). This classification can then be used to guide clinical decision making.

25.3 Key Concepts for Translational Bioinformatics

As noted in the definition above, TBI involves the development of novel methods for data storage, analysis, and interpretation. In this section we elaborate on these different levels of informatics methodologies which can be framed as falling along a spectrum from data to knowledge (Fig. 25.1). *Data* represent specific values; at the simplest level, they can be reduced to ones and zeros. In the middle of the spectrum is *information*—ascribing new meaning to the data at hand through analysis. Finally, we arrive at *knowledge*—the ability to interpret information in a specific context, and for that interpretation to guide actions and behavior.

25.3.1 Data Storage

Data storage takes place at a number of different levels corresponding to different stages along the translational pipeline. At the "bench" end of bench-to-bedside, there is the need to store massive files of raw data generated through omics technologies (Stein 2010). In the case of genome sequencing, these files can be so large that is has been suggested (though not necessarily concluded) that for easily regenerated samples, it might be more cost-effective to discard the raw data and, if necessary, re-sequence at a later time (Hsi-Yang Fritz et al. 2011). For each raw data file type, one can generally choose among several different processing tools or algorithms. Thus, in addition to the raw data, a researcher or core facility may want to store one or more versions of processed data files, still frequently very large in size. In addition to the actual data, experimental **metadata** are needed in order to understand how the data were generated and how they were processed or analyzed. Annotation facilitates both comprehension and data provenance. Unfortunately, that information is rarely standardized, and frequently stored only in the researcher's head, paper notes, or hard drive. Increasingly, standards and tools such as the Ontology of Biomedical Investigations (OBI) (Brinkman et al. 2010), Minimum Information lists (Taylor et al. 2008) and the Investigation/Study/Assay (ISA) infrastructure (Rocca-Serra et al. 2010) (see Chap. 24), are being developed to address this issue.

In the middle of the translational spectrum, there is an increasing need to store information related to subject consent. As DNA **Biobanks** (described below) become more common, researchers will have greater access to tissue samples of subjects they themselves did not recruit. It will no longer suffice to have consent information stored on a paper form, locked away in a file drawer. Even scanned images of the signed paper forms are insufficient for rapid and accurate information retrieval. Researchers and biobank administrators will need the ability to know to what each participant has consented, and to perform electronic queries to determine consent status on demand. Can John Doe's tissue be used for research beyond the study for which he was enrolled? Can the blood collected as a byproduct of care be used for **Genome-Wide Association Studies** (**GWAS**)? Can Jane Doe be contacted for enrollment in a follow-up study? In parallel with work being done to address issues of ethics and governance for this type of data capture and management, researchers are working to develop tools and terminologies to facilitate research permissions management (Obeid et al. 2010). To date, a clear solution has not emerged.

At the bedside end of the spectrum, clinicians do not have the time, nor usually the training, to analyze the underlying data. They need easy access to a patient's genotype, protein biomarker pattern, or metabolite profile without having to wade through volumes of sequence and biomarker data to learn the results of the test. They may also want to know some type of confidence or quality score for the data provided. HL7's Clinical Genomics Work Group is working to develop an HL7 standard in this area.[7] Just as important as the data itself, clinicians need to know what to do with that information.

[7] http://wiki.hl7.org/index.php?title=CG (Accessed 12/3/2012).

Incorporating omic data into the EHR (Chap. 12) will not improve clinical care without the incorporation of these data types into clinical guidelines and tools for clinical decision support as discussed in Chap. 22 (Hoffman 2007).

25.3.2 Biomarkers

Fundamentally, the newfound ability to analyze and interpret high-throughput molecular datasets is about the discovery of biomarkers. The term **biomarker** has been used for decades, referring to any observation that could be used as an indication of an underlying physiological state. One commonly accepted definition is "a characteristic that is objectively measured and evaluated as an indicator of normal biological processes, pathogenic processes, or pharmacologic responses to a therapeutic intervention" (Atkinson et al. 2001). Exactly what constitutes a biomarker has historically depended in part on what types of observations could be made. Early biomarkers would have included fever, increased respiratory rate, or a rash. As our ability to probe living organisms increased, the domain of biomarkers expanded to the presence or concentration of specific molecules in the blood. For example, increased levels of glucose are indicative of diabetes, and an increase in PSA (prostate-specific antigen) suggests a risk for prostate cancer. Omics-era methodologies give us new types of data to which we can apply novel analytic methods to predict disease and progression. In the genomic era, biomarkers may consist of not just one but many different characteristics, which together give insight into underlying states or processes. Gene expression signatures are a common example of this type of multi-dimensional biomarker.

One important distinction to be made is that of predictive versus mechanistic biomarkers. Predictive biomarkers are essentially correlative markers of a given observation or outcome. They may or may not be causal for that outcome, but they can assist both clinicians and researchers through decision support by predicting outcomes or suggesting new focus areas for research. Mechanistic biomarkers, on the other hand, can help shed light on what is happening at the molecular level that causes, for example, pathology, disease progression, or sensitivity to a given drug. Understanding a mechanism allows researchers to try to modify it through the activation or inhibition of specific molecules or pathways.

25.3.2.1 Predictive Biomarkers for Clinical Use

Predictive biomarkers can facilitate decision making in a number of ways. A biomarker indicating poor prognosis might suggest a more aggressive course of therapy than if that biomarker were not present. A signature indicating that lifestyle changes are likely to offer significant benefit to a patient could provide the motivation needed to follow through. For example, a signature indicating that weight loss is likely to improve insulin resistance could identify individuals for whom an intensive lifestyle changes is likely to have the most impact. Shah et al. were able to identify a **metabolomic** profile in subjects who had lost weight that, while independent of the amount of weight lost, was correlated with changes in insulin resistance (Shah et al. 2009b). On the flip side, a signature indicating that lifestyle changes alone are unlikely to confer the desired benefits may suggest that pharmaceutical intervention should be considered as well. Even if a biomarker is in no way actionable *yet*, it can be useful for biomedical research. As an example, osteoarthritis is a debilitating disease that is treated primarily through palliative measures to alleviate symptoms, but for which no disease-modifying therapeutic agents exist. One reason for this is the time and cost required to carry out a clinical trial. Gold standard radiographic methods for observing structural disease progression lack sensitivity and work best when the degeneration of a joint can be observed over time. Studies must therefore be carried out over of a period of several years. In addition, without knowledge of which subjects are likely to progress, studies must enroll large numbers of participants in order to be significantly powered (see Chap. 11). Identifying biomarkers to predict progression

would enable cohort enrichment for individuals in whom disease progression is more likely, thus cutting the total number of subjects required and hence the cost of the trial (Kraus et al. 2011).

Biomarkers that are not clinically actionable may be personally actionable. As an example, relapsing-remitting multiple sclerosis (RRMS) is a form of multiple sclerosis in which the patient experiences exacerbations or relapses of neurologic symptoms, followed by periods of partial or complete recovery. If a test could be developed to enable RRMS patients to know in advance if relapses were likely to occur within an upcoming span of weeks or months, it could enable them to make more informed personal or professional decisions, such as when to plan a vacation, or whether to take a new job (Gregory 2011). Another example is the LRRK2 mutation that confers a significantly increased lifetime risk for Parkinson's disease. After learning of his carrier status for this mutation, Google co-founder Sergey Brin contributed tens of millions of dollars to Parkinson's research (Goetz 2010). A unique situation, to be sure, but it would be hard to argue that knowledge of a biomarker in this case was not actionable.

One major area for biomarker use is that of pharmacogenomics, described in Sect. 25.4.5 below. In many cases, a therapeutic gold standard exists, but only a fraction of patients respond to the given therapy. Knowing in advance who is likely not to respond to therapy, or who needs a higher or lower dose than the standard guidelines suggest, can be useful for tailoring therapeutic interventions. Interestingly, while the success of genetic biomarker discovery for common disease has been limited, genotypic biomarkers for response to drugs may be more promising because these variations would not have been selected against through evolution (Cirulli and Goldstein 2010). This may explain why, among published GWAS finding to date, the pharmacogenetic associations tend to have much higher odds ratios than those of genes associated with common diseases.

25.3.2.2 Molecular Mechanism for Therapeutic Targeting

Biomarkers may also be used for elucidation of disease mechanism which can then enable therapeutic targeting toward a specific molecule or pathway. Comparative analysis of high dimensional molecular signatures in patients versus healthy volunteers, tumors versus normal tissue, responders versus non-responders, etc., can reveal a set of molecules that are differentially expressed among these groups. One can then study those specific molecules more closely, or the pathways in which those molecules are involved, for example through gene ontology (GO) enrichment (see Sect. 25.4.2.2) or analysis using a curated pathway database such as Ingenuity's IPA, or Thomson Reuter's MetaCore (Chan et al. 2007). These types of tools also help to address a major challenge with pattern detection in high throughput data. Particularly in human data sets where differences are observational and not perturbed, it can be difficult, if not impossible, to know what is causal and what is simply correlated. Systems biology, described in Chap. 24, attempts to address this.

25.4 Computational Approaches to Biomarker Discovery

25.4.1 Classification and Prediction

One of the most common uses of biomarkers is to categorize samples or patients: cancerous samples versus normal tissues, good versus poor prognosis, bacterial versus viral infection. There are a number of ways to approach this problem, all of which fall under the heading of **supervised learning**. Fundamentally, supervised learning entails taking a set of inputs and corresponding outputs to try to learn a model that will enable one to predict output when faced with a previously unseen input. One is trying to predict one value, the *dependent variable*, based on some number of other values, also called *features* (in computer science), *independent variables* (in statistics), or *risk factors* (in clinical practice). If the dependent variable is categorical, typically one is actually predicting the probability of belonging to one class or the other. For example, one might want to predict whether a person will have a heart attack based on age, race, gender, weight, and

cholesterol level. Or, in the context of TBI, one might want to predict the likelihood of a heart attack based on gene expression. Note that this latter approach is useful only if the gene expression signature increases the predictive capabilities beyond that offered by the clinical variables, which are typically easier to collect. Algorithmic approaches to classification and prediction are described in Chap. 24.

25.4.1.1 Clinical Relevance Versus Statistical Significance

Statistical significance is a measure of certainty that a test will give the right answer. Clinical relevance is a measure of how valuable this information is in guiding clinical care. It incorporates not only statistics, but also efficacy, safety, and cost. A test may be able to predict with 90 % accuracy whether, for example, a patient is likely to respond better to a treatment with unpleasant side effects over another, more innocuous therapy. However, if those side effects would significantly lower the patient's quality of life, then the test, while statistically significant, may not be clinically relevant. Similarly, if the cost of a false negative is very high, for example if a test predicts with 90 % accuracy that a patient will survive without a given intervention, that intervention will likely still be administered. On the other hand, if a test predicts with 90 %, or even 100 %, accuracy that a patient is likely to live 1 year longer with a given intervention but the intervention costs $1 million, this highly significant test is still not likely to affect clinical care.

Similarly, incorporation of molecular data or improvement upon an analytic method may increase a test's accuracy to a *statistically* significant degree while still not affecting clinical practice. Accuracy of a test is often measured by the area under the **receiver operating characteristic** (**ROC**) curve (AUC) (see Chap. 3), or the **C statistic**. The ideal ROC curve goes straight up the y-axis at $x = 0$, and then straight across the x-axis at $y = 1$, giving an AUC of 1. The more accurate a test, the closer it comes to that perfect path.

Figure 25.4 shows hypothetical ROC curves for two tests. Test 2 is a more accurate test in that it has a statistically significant higher C statistic, but as with the examples above, that may not change any clinical decisions. In light of this fact,

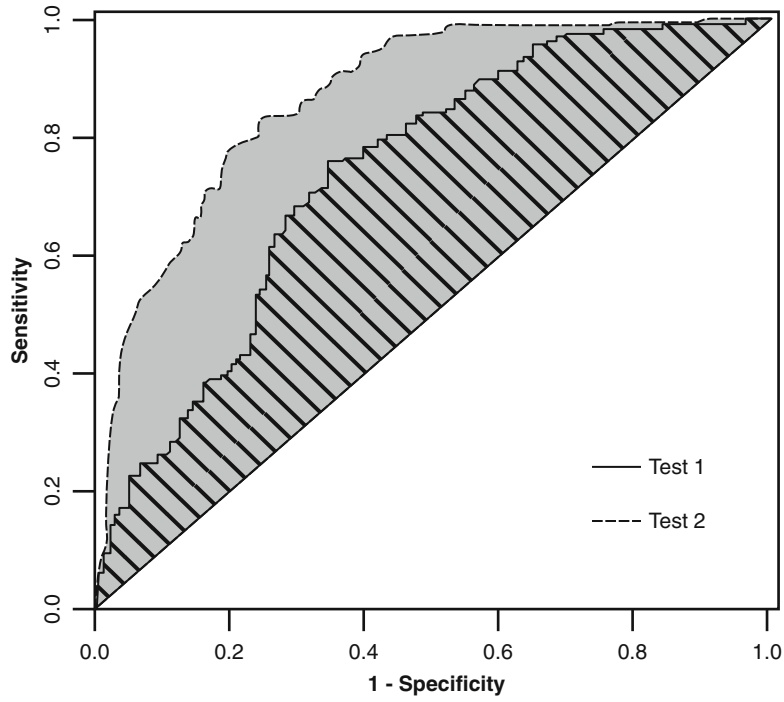

Fig. 25.4 A comparison between two Receiver Operator Characteristic (ROC) curves. The area under the curve (AUC) or C statistic is higher for Test 2 (*gray*) than for Test 1 (*diagonal lines*) to a statistically significant degree, but this increased accuracy does not necessarily imply clinical relevance

Table 25.1 Hypothetical reclassification of disease risk between two prognostic tests

	Number of individuals (actual rate)			
		Predicted 5-year risk for test 2		
Predicted 5 year risk for Test 1		**0–5 %**	**5–20 %**	**> 20 %**
	0–5 %	300 (3 %)	20 (2 %)	0
	5–20 %	30 (3 %)	300 (11 %)	40 (37 %)
	> 20 %	0	10 (35 %)	300 (42 %)

some have proposed that a better measure is needed for judging the incremental value of novel biomarkers and analytical approaches (Pencina et al. 2008). Alternative methods include **net reclassification improvement** (**NRI**), a measure of the net fraction of reclassifications made in the correct direction using the given biomarker or method over a method without the designated improvement (Steyerberg et al. 2011). This concept is illustrated in Table 25.1. Rows represent the risk level predicted by the hypothetical Test 1 for 1,000 subjects, columns represent the risk level predicted by Test 2. Values along the diagonal were predicted to have the same risk by both tests. Subjects in the black cells (30 + 40 = 70) were correctly reclassified by Test 2 (i.e., the actual rate in parentheses matches the appropriate risk category). Subjects in the light gray cells (10 + 20 = 30) were reclassified incorrectly. The resulting net reclassification improvement is (70–30)/1,000, or 4 %.

25.4.1.2 Biomarkers for Drug Repurposing

One very promising area for use of biomarkers is in **drug repurposing**, or drug repositioning. That is, identifying existing drugs that may be useful for indications other than those for which they were initially approved. Doing so allows circumvention of early clinical trials for toxicity as those have already been performed. A number of different approaches have typically been used to identify candidates for repositioning. In some cases, overlapping symptoms may suggest a potential match between one disease area and another. In other cases, empirical observation of unexpected positive effects may suggest alternative uses for a given drug. With omic-scale biomarker discovery, it is possible to use underlying molecular pathway signatures to suggest new uses for existing drugs.

One of the prominent early examples of this approach came from the Broad Institute in the form of the "Connectivity Map," a resource intended to enable researchers to identify functional connections between drugs, genes, and diseases (Lamb 2007). The general idea was to identify a gene expression signature in a state of interest, e.g. a disease, and then compare that signature to the gene expression patterns observed upon exposure to a number of different compounds. Correlated signatures suggested pathways that were similarly perturbed between a disease state and an intervention. More importantly, anti-correlated signatures suggested potential utility for a given compound in trying to reverse the underlying molecular mechanisms of a given disease. A similar approach was used by Sirota et al. to identify the anti-ulcer drug cimetidine as a candidate agent to treat lung adenocarcinoma. They were then able to validate this alternate use *in vivo* using an animal model of the disease (Sirota et al. 2011). Their approach is illustrated in Fig. 25.5.

25.4.2 Ontologies for Translational Research

In order to apply computational methods for biomarker discovery, one needs a consistent way to refer to diseases, drugs, devices, etc. Several ontologies exist in the biomedical domain, many under active development, that provide the necessary terms for creating consistent annotations—preferably in an automated manner—for the various datasets that are at the core of conducting research in TBI. One primary need in TBI is to identify and refer unambiguously to disease using one or more disease ontologies. We use the term disease **ontology** to refer to artifacts—terminologies and vocabularies as well as true ontologies—that can provide a hierarchy of parent–child terms for disease conditions. Disease-specific and other clinically-oriented ontologies are discussed in detailed in Chap. 7.

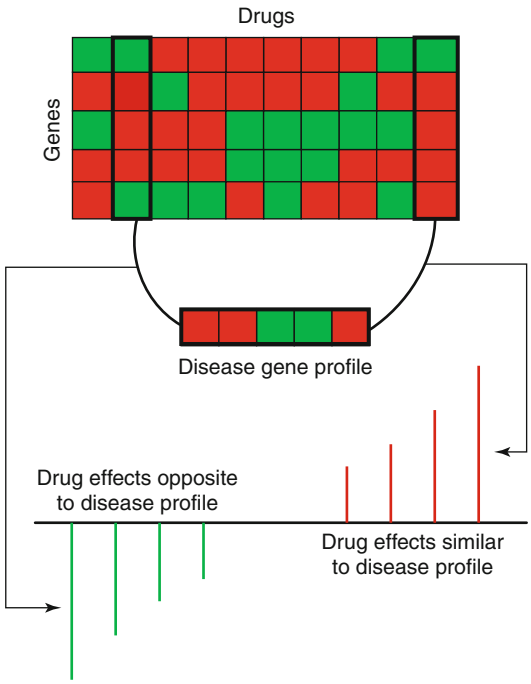

Fig. 25.5 A computational approach to candidate selection for drug repurposing. Sirota et al. first generated genomic signatures representing both diseases and drug exposure. For each disease signature, they compared it to the panel of drug signatures and assigned a drug-disease score based on profile similarity. Drugs whose pattern were most significantly *dis*similar to the disease state were ranked as lead candidates to treat the disease of interest

The Ontology for Biomedical Investigations (OBI) was developed as a collaboration among a number of experimental communities around the world in order to represent common aspects of biological and clinical investigations. It includes broadly applicable terms such as *assay*, as well as more specific terms, such as *transcription profiling by array assay*. It is particularly useful for annotation of experimental metadata, for example to record that a *protein expression profiling assay* was performed on a *blood specimen* (Brinkman et al. 2010).

25.4.2.1 Ontology-Related Resources for Translational Scientists

The use of ontology-based analyses for TBI, especially disease and drug ontologies as well as analyses using multiple ontologies, is a recent development and the adoption and use of ontologies is likely to accelerate. Several resources are available for researchers who wish to use ontologies in making sense of large scale datasets. The UMLS, or Unified Medical Language System (see Chaps. 2 and 7), is a set of files and software that brings together many health and biomedical vocabularies and standards to enable interoperability among computer systems. One primary use of the UMLS is linking health information, medical terms, drug names, and billing codes across different computer systems, for example linking terms and codes among doctors' offices, pharmacies, and insurance companies. The UMLS has many other uses, including search engine retrieval, data mining, public health statistics reporting, and terminology research. The UMLS consists of three components, called the Knowledge Sources. These are: (1) The Metathesaurus, which provides terms and codes from many vocabularies, including RxNorm and SNOMED CT; (2) The Semantic Network, which provides broad categories (semantic types) such as 'Neoplastic process', 'Pharmacological Substance' and the relationships (semantic relations) among the categories; and (3) The SPECIALIST Lexicon and Lexical Tools, which provide natural language processing tools (see Chap. 8).

The National Library of Medicine provides Web based services called The UMLS Terminology Services (UTS), which provide comprehensive access to components of the Unified Medical Language System. UTS allows users to search and display content from the UMLS Metathesaurus, the Semantic Network and SNOMED CT. Users can also download content such as the UMLS knowledge sources, RxNorm monthly updates and SNOMED CT. UTS also provides an API for accessing the UMLS Metathesaurus content over the Web. In the field of TBI, the UMLS is a relatively underutilized resource, but that is changing quickly with the increase in the variety of access options (Aronson 2001; Bodenreider 2004; Aronson et al. 2008; Shah and Musen 2008; Aronson and Lang 2010; Mork et al. 2010) and heightened dissemination efforts by the National Library of Medicine.

The National Center for Biomedical Ontology maintains a repository of biomedical ontologies called BioPortal (Musen et al. 2011) which provides access through both Web pages and Web Services to more than 270 biomedical ontologies and controlled terminologies. Users go to the BioPortal Web site to browse biomedical ontologies and to search for specific ontologies relevant to their work. Using tools such as the Ontology Recommender, scientists who are unsure which of the dozens of ontologies in BioPortal provide the best coverage for capturing the entities in a particular application area can choose the right ontology. The NCBO Ontology Recommender Service (Jonquet et al. 2010) takes as input representative textual data relevant to a domain of interest and returns as output an ordered list of ontologies that would be most appropriate for annotating the corresponding text. By browsing ontologies on BioPortal and using tools such as the ontology recommender, a cancer biologist may find, for example, that although the Gene Ontology offers some terms for annotating her experimental data related to cell division, there are more precise terms in the NCIt. She may discover that the Foundational Model of Anatomy ontology provides terms for consistently naming body parts from which the experimental specimens were obtained, or that the National Drug File - Reference Terminology (NDF-RT) provides the properties of the drugs used in generating the experimental data. BioPortal allows users to navigate ontologies using a tree browser. Users also can visualize ontologies in BioPortal using special views that offer cognitive support for understanding the complexities of large ontologies (Fig. 25.6).

To provide the relationships between terms in two *different* ontologies, BioPortal provides mappings between the terms (Ghazvinian et al. 2009). The mappings can inform the user that the term *Melanoma* in the NCI Thesaurus is related to the term *Malignant*, *Melanoma* in SNOMEDCT and to *Melanoma* in the Human Disease Ontology. These mappings allow users to compare the use of related terms in different ontologies and to analyze how whole ontologies compare with one another (Ghazvinian et al. 2011). In addition to the curated mappings from the UMLS metathesaurus, Bioportal enables algorithmic and user-generated mappings as well.

25.4.2.2 Enrichment Analysis

Enrichment analysis is a statistical method to determine whether, for a set of items, a given concept or value is statistically over-represented compared to what one would expect at random. For example, informatics-related terms are over-represented in this book compared to what one would expect to find in a random sampling of words from all textbooks. The canonical example of enrichment analysis involves a list of genes differentially expressed in some condition. To determine the biological meaning of such a list, the usual solution is to perform enrichment analysis with the GO (Gene Ontology), which provides terms for consistent naming of the cellular component (CC) of gene products, the molecular functions (MF) they carry out, and the biological processes (BP) in which they participate. Several curation projects use the GO terms to annotate gene products from multiple organisms with terms from the three branches (CC, MF, BP) of the GO (Camon et al. 2003). These annotations form the basis for enrichment analysis in which we can aggregate the annotating GO concepts for each gene in this list, and arrive at a profile of the biological processes or mechanisms affected by the condition under study. This approach does have certain limitations, for example incomplete annotations for a number of genes, lack of conditional independence between annotations, and lack of a systematic mechanism to compensate for differing levels of depths in different branches of the ontology hierarchy (Khatri and Draghici 2005; Rhee et al. 2008). Despite this, such analysis is widely popular in the bioinformatics community and has resulted in over 100 tools listed on the GO website[8] and over 7,000 citations to the landmark paper on the Gene Ontology (Ashburner et al. 2000).

Disease and drug ontologies can be used to perform enrichment analysis in a manner simi-

[8] http://www.geneontology.org/GO.tools.shtml#alphabet (Accessed 12/3/2012).

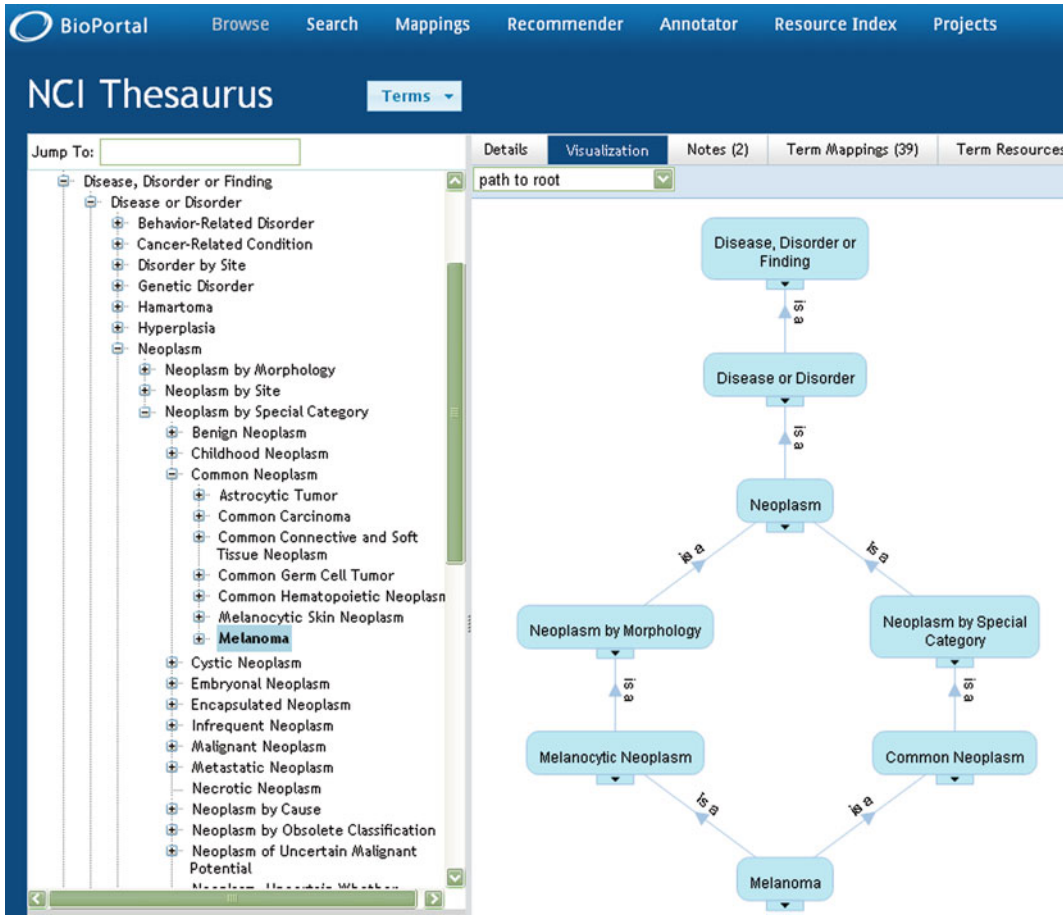

Fig. 25.6 A portion of the National Cancer Institute's thesaurus. The left pane shows a standard tree view for the term 'Melanoma'. The right pane shows a visualization that provides additional context by showing the parent classes of melanoma, all the way to the root node of 'Disease, Disorder or Finding'. The navigation bar just above the graphical visualization provides access to additional information, such as mappings which provide hooks into other disease ontologies that contain the concept Melanoma

lar to GO-based analyses for gene expression data (Subramanian et al. 2005; LePendu et al. 2011a). Just as scientists can ask *Which biological process is over-represented in my set of interesting genes or proteins*, we can also ask *Which disease* (*or class of diseases*) *is over-represented in my set of interesting genes or proteins*? For example, by annotating known protein mutations with disease terms, Mort et al. were able to identify a class of diseases—blood coagulation disorders—that were associated with a lower than expected rate of amino acid substitutions at O-linked glycosylation sites (Mort et al. 2010).

25.4.3 Natural Language Processing for Information Extraction

Ontologies are also useful in the context of extracting information from a body of text. In-depth methods for natural language processing are discussed in Chap. 8. Here we describe some applications in the context of translational research.

25.4.3.1 Mining Electronic Health Records

Researchers have shown that is possible to profile patient cohorts from EHRs using a variety of

ontologies including SNOMED CT, MedDRA, and RxNorm (LePendu et al. 2011b). For example, LePendu et al. developed methods to annotate clinical text and methods for the mining of the resulting annotations to compute the risk of having a myocardial infarction on taking Vioxx (rofecoxib) for Rheumatoid arthritis. Their preliminary results show that it is possible to apply annotation analysis methods for detecting drug safety signals using electronic medical records (LePendu et al. 2011b).

Mining EHR data has also been proposed as a solution to the challenge of the large number of subjects that are needed for genome wide association studies (GWAS). Patients are increasingly able to consent to, or in some cases to opt out of, allowing excess biospecimens taken in the course of clinical care to be used in a de-identified fashion for genomic testing. Even for relatively strong genetic effects, GWAS requires thousands of individuals for sufficient statistical power (Chap. 11). For weaker effects, tens of thousands of subjects are likely to be needed. Although the cost of genotyping continues to decrease, recruitment and sample collection for these large numbers is both costly and labor-intensive. Leveraging the health care system and EHRs for subject recruitment offers a potential approach to circumvent this problem. Ritchie et al. demonstrated the feasibility of this approach by using EHR data and an associated biobank to replicate a number of previously discovered genotype-phenotype associations (Ritchie et al. 2010).

One major initiative in this area is the eMERGE (Electronic Medical Records and Genomics) Network, whose initial aim was to demonstrate that data captured through routine clinical care are sufficient to identify cases and controls accurately for GWAS (Thorisson et al. 2005). The eMERGE consortium includes five institutions with DNA repositories and associated electronic medical record systems. For each site, ontology-based data extraction and natural language processing algorithms are applied to the EHR in order to determine phenotypes such as dementia, cataracts, peripheral artery disease, type 2 diabetes, and cardiac conduction defects. This analysis is performed in a high-throughput, scalable fashion with results compared to a manually curated gold standard in order to determine positive and negative predictive values for cases and controls for the phenotypes in question (Kho et al. 2011). The consortium is also looking at cross-institutional algorithm application, ethical, legal, and social issues around DNA biobanks, and the potential for future incorporation of GWAS findings into clinical care.

These types of EMR-associated biobank resources enable a number of other approaches to data mining. For example, Denny et al. used BioVU at Vanderbilt University to perform what they called a "PheWAS," or a systematic, high-throughput **phenome-wide association scan** (Denny et al. 2010). Instead of measuring whole genomes across thousands of patients in order to find a gene associated with a phenotype in question, they measured only five alleles across thousands of patients and performed enrichment analysis for various diseases based on ICD9 codes. They then were able to reproduce known associations between those genes and certain diagnoses, *and* to generate new hypotheses for associations between these genes and other diagnoses that were statistically enriched for a given genotype. The ability to connect, at a molecular level, diseases that were not previously associated can have implications for therapeutic intervention.

25.4.3.2 Dataset Annotations

In addition to EHRs, public repositories for omics-scale datasets have been a heretofore underutilized resource for data mining. Upon submission, these datasets are typically annotated using only free-text descriptions. To address the lack of annotations, researchers have demonstrated that translational analyses are enabled by automatically annotating tissue and gene microarray datasets with ontology terms (Shah et al. 2009a). Researchers have also mapped the text annotations for records in databases such as the Stanford Tissue Microarray Database to terms from the NCI thesaurus to enable integrated analysis and query (Shah et al. 2007). Such automated annotation approaches have been generalized to create systems that process the

free text metadata of diverse database elements such as gene expression data sets, descriptions of radiology images, clinical-trial reports, and PubMed article abstracts to annotate and index them with concepts from appropriate ontologies (Jonquet et al. 2011).

25.4.4 Network Analysis

Biology lends itself in various ways to modeling through networks or **graphs**. The term "graph" simply refers to a set of *nodes* or circles connected by a set of *edges* or lines. Typically, a node represents a molecular entity, and an edge represents some form of relationship between those molecular entities. This relationship may be a physical interaction (e.g., binds to), an influence (e.g., activates), or a similarity (e.g., is co-expressed with), among other possibilities. One frequently sees graphical models of gene regulatory networks, protein-protein interactions, and signaling cascades. The set of all of these sorts of physical interactions has been referred to as the *interactome* (Barabasi et al. 2011). Studying this **interactome** and its properties from a graph theory perspective enables useful insights regarding gene modules and pathways, and how these are disrupted in disease.

A number of researchers have attempted to develop gene association networks using gene expression data either alone or together with other sources of network data such as protein-protein interactions. The general idea is that co-expressed genes are likely to interact with each other or participate in the same pathway. But of course, correlation does not equal causality, and to be useful from a translational perspective, it is important to know the directionality of the influence between two molecules. Consider two genes, X and Y, whose expression is correlated (See Fig. 25.7). One can conclude that the genes interact in some way, whether directly or indirectly (i.e., through another molecule). However, without additional knowledge of any sort, we cannot know whether X influences Y (Fig. 25.7a), or Y influences X (Fig. 25.7b) or they share a third causal gene, Z (Fig. 25.7c). Which model

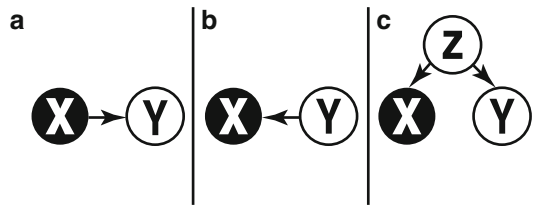

Fig. 25.7 Three possible causal relationships between two co-expressed genes. (**a**) Gene X affects Y. (**b**) Gene Y affects X. (**c**) Both X and Yare affected by a third causal gene Z

represents the true underlying relationship is important to know because if Y is involved in poor outcome, then targeting X will help to alleviate this condition in the first model, but not in the second or third.

One way to determine the actual underlying relationship, used frequently in model systems, is to actively perturb a specific variable in the system. If the other molecule changes accordingly, then we know that the perturbed variable was causal. This is the approach frequently used in a systems biology approach (see Chap. 24). Unfortunately, this is much harder to do in human beings than in *E. coli* or yeast. One clever approach to determination of causality in human biological networks is to integrate gene expression with genotypic information, in which case DNA sequence can be assumed to be the independent variable. If differential gene expression is correlated with differential genotype, one can conclude that the genotype caused the gene expression pattern and not the other way around. This is the basis for the approach taken by Eric Schadt et al. to develop **probabilistic causal networks** which can then be used to identify key drivers of disease (Zhu et al. 2008).

Network analysis in translational research need not be confined to concrete objects such as molecules. The Human Disease Network is a graphical model where nodes represent both known disease genes and disorders, linked by known associations between a given gene and disease (Goh et al. 2007). Figure 25.8 shows the "diseaseome" bipartite network, as well as the Human Disease Network, which connects diseases based on common genes, the and Disease

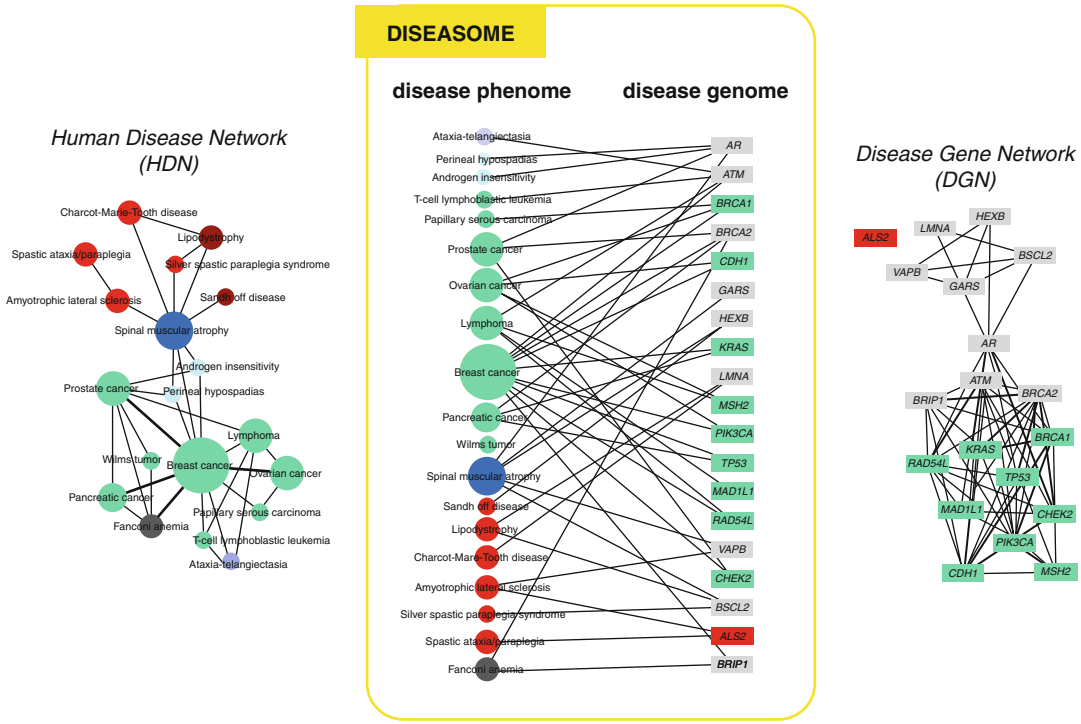

Fig. 25.8 The Human Disease Network. The *middle panel* shows a small subset of the bi-partite gene-disease network based on OMIM (Online Mendelian Inheritance in Man) gene-disease relationships. The Human Disease Network on the *left* shows diseases as nodes, with connections representing common related genes. The Disease Gene Network on the *right* depicts genes as node, with connections indicating that they have been implicated in one or more of the same disorders (From Goh et al. (2007), ©2007 National Academy of Sciences, U.S.A.)

Gene Network, connecting genes based on diseases in common. Combining these disparate data types enables a graph theoretic approach to study the genetic basis for disease. Using this framework, one can analyze similarity between genes based not on co-expression or GO term annotation but based on the pathologies in which a gene is known to be involved. Such similarities could easily go undetected through gene expression analysis if, for example, the different diseases are caused by over-activation or inhibition of the gene respectively. A disease-gene network also enables the comparison of diseases not traditionally studied together, based on common underlying molecular mechanisms. Identifying disease similarities based on gene expression requires that one analyze expression data from those two diseases together in the first place, making it more difficult to discover novel, previously unsuspected relationships.

Building upon the Human Disease Network, Yildirim et al. created a network of drug-gene target interactions, thus enabling an additional layer of analysis regarding similarity between different drugs based on targeted genes, and between target molecules based on the drugs that target them (Yildirim et al. 2007). This type of network can be used as the basis for a number of different observations, including trends in drug development over time. For example, analysis of the structure of the graph revealed significant clustering of drug-gene interactions, suggesting a significant "me too" pattern to drug development (see Fig. 25.9). Inclusion of drugs still under investigation, i.e., not yet FDA approved at the time of analysis, demonstrated that the breadth of drug targets is expanding, suggesting a trend toward target diversity. Incorporating the cellular component of target proteins showed that the distribution of cellular location for target proteins,

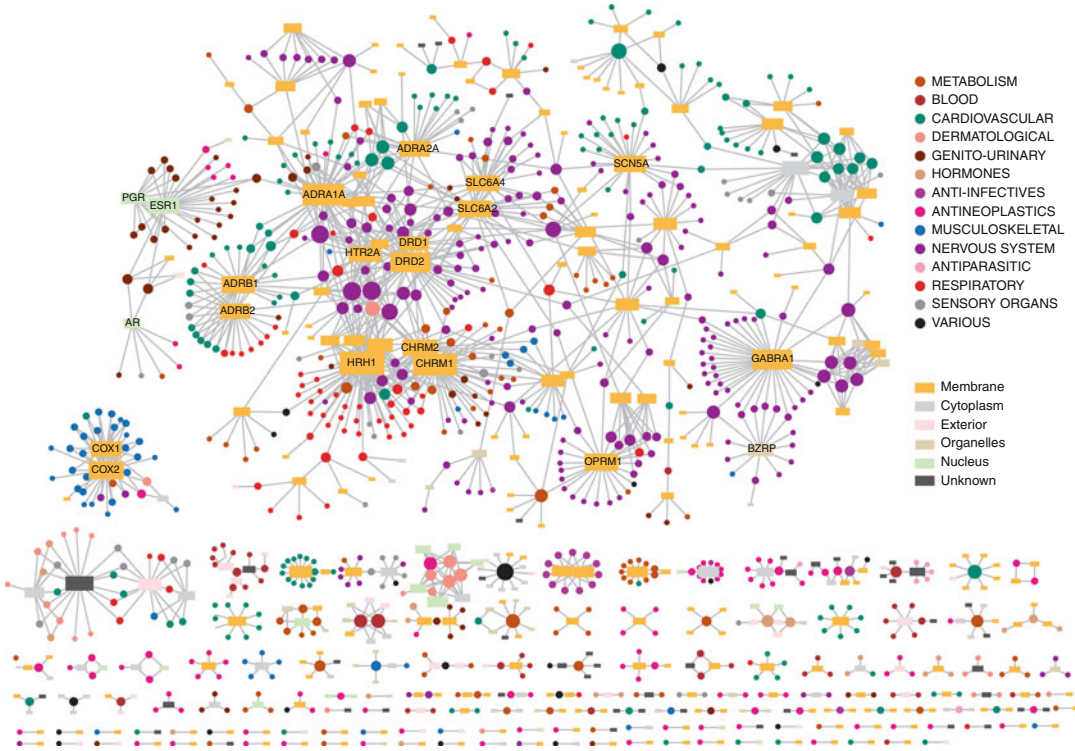

Fig. 25.9 Yildirim et al.'s Drug-Target network. Circles represent FDA-approved drugs and rectangles represent target proteins. Diseases are color-coded by anatomical system and protein targets according to their cellular location. Clusters of drugs associated with one target reflect the pharmaceutical industry's tendency to develop 'follow-on' drugs (Used with permission, Albert- Laszlo Barabasi)

previously nearly two-thirds membrane-associated, is becoming more diverse, better matching the known distribution for disease proteins. Finally, this group incorporated protein-protein interaction data to facilitate the study of network properties of drug target gene products. They looked for the shortest path between drug target genes and known disease genes for the disorder that drug was intended to treat and found that this number appears to be decreasing over time, suggesting that drugs are moving from a palliative approach (i.e. treating the symptoms and not the cause) to rational drug design (Yildirim et al. 2007).

25.4.5 Pharmacogenomics

Pharmacogenomics is the study of how genes and genetic variation influence drug response. Drug response is a phenotype that can change in the context of genetic variation. The primary challenges for pharmacogenomics are to (1) identify the key genes that influence drug response, and (2) understand how specific variations in these genes modulate drug response. The term **pharmacogenetics** generally refers to drug-gene relationships that are dominated by a single gene, whereas the more general term refers to drug responses that result from a combination of interacting gene products. In this section, we use the word "gene" loosely to refer not only to the DNA coding regions for proteins and RNAs but also the protein and RNA products themselves. In many cases, the gene-drug relationship is really a relationship between the drug and the gene's protein product.

Pharmacogenomics is a prototypical TBI activity because it involves both clinical entities such as drugs, diseases, symptoms as well as

molecular entities such as genes, proteins, DNA, RNA, small molecules and cellular processes. Because drug response is the key phenotype of interest, it is useful to review the basis for drug response. When a drug is administered, there are two distinct genetic "programs" that are relevant. The first is the **pharmacokinetic program** or PK, which describes the absorption, distribution, metabolism and excretion of the drug in the body. Genes implement this program (they encode transporter molecules that move the drug across membranes and the liver enzymes that transform the drug and prepare it for elimination via the kidney or liver) and variation in these genes can lead to a different blood level of drugs or a different timing of these levels. The second is the **pharmacodynamics program** or PD, which describes how the drug works, its protein target, and the mechanism by which it impacts cellular physiology in order to alleviate or cure disease. Genes are clearly also involved in this program (they encode the drug's primary targets, and the other proteins that interact with these targets to create the cellular response to the drug), and variation in these genes can lead to a different response to the drug. In short, PK is "what the body does to the drug", and PD is "what the drug does to the body." The goal of pharmacogenomics is to understand, for every drug, the key PK and PD genes, and which variation impacts their response. This will allow us to realize the vision of using the genome to choose drugs based on maximizing their likely efficacy and minimizing their likely toxicity.

25.4.5.1 Key Entities and Associated Data Resources

The key computational entities in pharmacogenomics are genes, drugs, and drug-related phenotypes (indications and effects, including side effects). There exist good informatics resources for all of these:

1. Genes. These are typically specified using the Human Genome Nomenclature Committee (HGNC) standard (Seal et al. 2011). They are typically situated within the genome as a series of exons that are spliced together to create a mature mRNA transcript that is then translated into a protein. This basic concept is made more complex because the strategy for splicing the exons may be variable (alternative splicing) thereby leading to several proteins, the RNA transcript may be degraded before it is translated, and the proteins may be modified after they are created. There are many resources on the web for gene information, and many aggregators of this information. These create a remarkably powerful network of associations that can be used creatively to make new associations. For example, Fig. 25.10 shows the links on a resource called PharmGKB (Pharmacogenomics Knowledge Base) for the drug VKORC1 and includes links to:
   - Entrez Gene: summarizes the sequence, variations, homologs across species
   - OMIM: provides information about rare genetic diseases involving this gene
   - UniProt: provides mapping information to relate this gene to its protein products
   - GeneCards: provides aggregated information about function, tissue localization, expression levels compound, literature references and more
2. Drugs and small molecules. The RxNorm standard for specifying drugs is a terminology standard (Parrish et al. 2006). DrugBank provides information about drugs, their targets, pharmacology, uses, and many other characteristics (Knox et al. 2011). There are only around 2,000 approved drugs in the United States, and so this list is relatively short. The list of small molecules that are not drugs is much larger and includes the contents of PubChem (Wang et al. 2009b, 2010), an NIH-built resource with basic information about the structure, function and literature on small molecules. The Zinc Database (Irwin and Shoichet 2005) lists 13 million commercially available compounds that can be purchased for use in research. Much drug information is contained within the "package insert" that is included in most drug packaging. This is information created by the drug company, but approved by the FDA. The FDA makes these available on a drug information

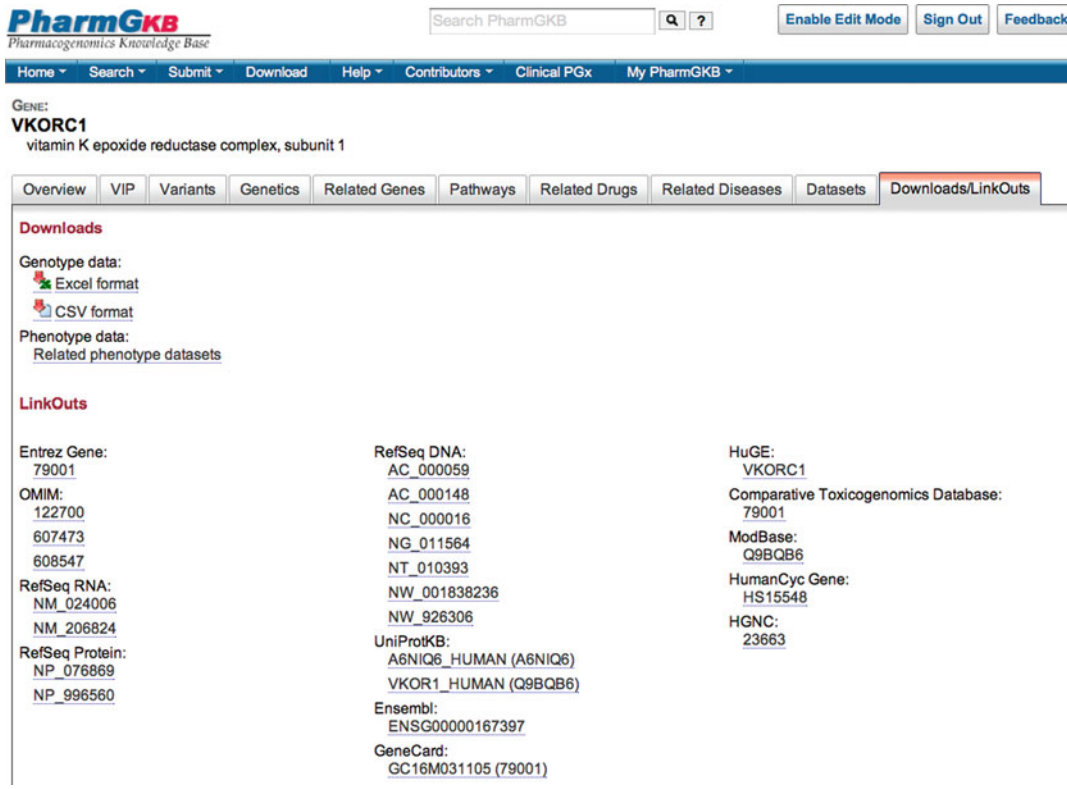

Fig. 25.10 PharmGKB gene pages are organized by tabs and the "Downloads/LinkOuts" tab shown here has links to many other sites with valuable information about human genes (Copyright PharmGKB, used with permission from PharmGKB and Stanford University)

site called DailyMed. For patients, the National Library of Medicine's MedlinePlus resource provides basic drug information as well.

3. Drug indications and drug effects. Drugs are used to treat particular diseases, and so controlled terminologies of drug indications and drug effects are useful for computational efforts. At the organism level, indications and effects are often diseases (diabetes is an indication) or side effects (hyperglycemia is a side effect). The UMLS and MeSH terminologies are often used to characterize such disease phenotypes (Bodenreider 2004). Of course, other disease terminologies such as SNOMED are also useful (Spackman et al. 1997a, b). For side effects, there are specialized terminologies, including the MedDRA terminology used by the FDA in adverse event reporting (MedDRA replaced a previous terminology called COSTART), and the WHOART (World Health Organization Adverse Reactions Terminology) dictionary for adverse reactions (Brown et al. 1999; Alecu et al. 2006). The SIDER database (Kuhn et al. 2010) provides information mined from drug package inserts about drug indications and side effects. The Anatomic Therapeutic Chemical classification (ATC) from the World Health Organization provides a high level classification of drugs organized hierarchically by the anatomical location of target, the therapeutic category, the pharmacological subgroup, chemical subgroup and precise chemical substance (Miller and Britt 1995).
4. Pharmacological properties of drugs. There are resources on the web that provide molecular level assay data related to small molecules, including many drugs. The ChEMBLdb resource provides the ability to find targets,

binding affinities, inhibition concentrations and information about other drug-oriented assays (Overington 2009). BindingDB also provides binding affinities for small molecules and proteins (Liu et al. 2007).

5. Population-based data on drug effects. The FDA maintains information about all reports of adverse events in the FDA Adverse Event Reporting System (FDA AERS). These reports include demographic information, indications for treatment, drugs administered, side effects experienced and a summary of clinical outcomes. They are freely available at the FDA website. The Canadian equivalent system, also making data freely available, is available through the Health Canada website. These data are very noisy and have many confounding variables, but nonetheless can be useful for discovering "signals" suggesting dangerous side effects or drug-drug interactions (Tatonetti et al. 2011a).
6. Genomic data resources. Fundamental to understanding complex gene-drug interactions are high-throughput genomic measurements that provide information about these interactions. The primary sources of genomic data include:
   - Information about genetic variation. These are available through the HapMap projects which catalog variation over a wide variety of ethnic populations, in order to define the occurrence and frequency of common genetic variations (Rusk 2010). The 1,000 Genomes project is taking HapMap further to categorize the occurrence of more rare variations (changes in single DNA bases, as well as insertions/deletions, segmental duplications, and larger scale inversions and translocations) (Via et al. 2010). There are also resources about copy number variations.[9] dbSNP—**d**ata**b**ase of **S**ingle **N**ucleotide **P**olymorphisms is a publicly available catalog of genome variation. Contents primarily represent single nucleotide substitutions, but also include a small number of other types of variation, for example microsatellite repeats and small insertions and deletions (Homerova et al. 2002). The PharmGKB resource specifically annotates genetic variations relevant to drug response (Altman 2007).
   - Gene expression information. The Gene Expression Omnibus contains an extremely large and diverse collection of high throughput gene expression experiments which allow one to evaluate whether a disease (or drug exposure) leads to up- or down-regulation of gene expression (Edgar et al. 2002). Particularly useful examples of gene expression for drug response are the Connectivity Map data set in which gene expression in response to 164 drugs was measured (Lamb et al. 2006). Similarly, the NCI 60 is a set of 60 cancer cell lines that have been exposed to hundreds of drugs in order to determine their sensitivity (Ross et al. 2000). Other efforts have looked at genetic variations that correlate with gene expression in order to associate these genomic regions with the function of the correlated genes (Gamazon et al. 2010; Nicolae et al. 2010).
   - Gene associations. The Genetic Association Database (Becker et al. 2004) provides curated information about the results of genetic association studies, including those studies that relate genetic variation to variation in drug response. The Human Genome Mutation Database (HGMD) also provides this information in a highly curated form (Stenson et al. 2009). dbGaP—**d**ata**b**ase of **G**enotypes **a**nd **P**henotypes is a resource to archive and distribute information about the interaction between genotype and phenotype (Mailman et al. 2007). The PharmGKB resource is devoted entirely to providing information about associations between human genetic variation and drug response phenotypes (Altman 2007).
   - Genetic pathways. Understanding drug action requires understanding the pathways and networks of drug action and drug

[9] http://humanparalogy.gs.washington.edu/structuralvariation/general/intro.html (Accessed 12/3/2012).

metabolism. The PharmGKB provides curated drug pathways for both drug action and drug metabolism, and has links to relevant external pathways created by the National Cancer Institute Pathway Interaction Database (Schaefer et al. 2009), Reactome (Joshi-Tope et al. 2005), and others.

25.4.5.2 TBI Applications in Pharmacogenomics

The network of data described above is a rich potential source of hypotheses about how genes combine to create drug response, as well as for predicting the particular consequences of genetic variation. This is still a new field, and there remain many opportunities for innovative use of these data. We highlight a few here to illustrate how integration of data can lead to novel discoveries.

GWAS to Discover Drug Response Genes

The most straightforward way to associate genes with drug response is to perform a genome-wide association study (GWAS) in which two groups are compared. One group (cases) has a drug response of interest (e.g., an adverse event in response to the drug or a particularly good response to it) and the other group (controls) does not have the drug response of interest. It is critical to ensure that the phenotype or response is carefully defined and measured. With each group, DNA is collected and typically 500,000 or 1,000,000 **SNPs** (single nucleotide polymorphisms) are measured using microarray technology. Then, for each SNP, an association is measured between the genotypes in cases and controls and the response of interest using a simple statistical test such as the chi-squared test. The SNPs that are most highly associated may represent regions of the genome that are involved in the response. These must be carefully vetted statistically, as there are many potential confounding variables. For example, it is important that the cases and controls are drawn from populations with similar ethnic origin, and that the significance remains after correcting for multiple testing. When one tests 500,000 or 1,000,000 hypotheses, adjustments such as Bonferoni correction (see Chap. 24) must be made in order to take into account the chance that an association is spurious. If the result is real, then the SNP may be used to identify nearby genes in the region that may be important for the drug response. For example, Shuldiner and colleagues were interested in the ability of the drug clopidogrel to protect patients from cardiovascular events. They found that a polymorphism RS12777823 was associated with a high-likelihood of having a cardiovascular event. They noted that this SNP was very close to the metabolizing enzyme CYP2C19, and in particular the "risk" allele for this SNP co-occurred with the CYP2C19*2 variant. Thus, they showed that CYP2C19 is important for the desired effect of clopidogrel, and found a variation of this gene that predicted poor response to the drug in affected patients (Shuldiner et al. 2009).

Mining the FDA AERS to Find Drug-drug Interactions

The FDA Adverse Events database associates multiple drugs with multiple diseases as indications as well as side effects. This database shows promise as a way to find new associations between single drugs and their side effects, as well as multiple drugs and their side effects. As mentioned above, the SIDER database is a "top down" database of side effects derived from the package label of drugs. Another approach to getting good lists of side effects is a more data-driven approach. One way to do this is to look for patterns of side effects associated with certain types of drugs using machine learning. For example, one may analyze the side effects of drugs that alter glucose in order to create a signature of the "typical" profile of side effects associated with a glucose-altering drug. Then, one can search a database of side effects (such as FDA AERS) for other drugs that match this profile. This was done by Tatonetti et al., who created a profile for glucose-altering drugs and found a set of 10 side effects either enriched or deficient (compared to background) in these drugs: hyperglycemia, diarrhea, hypoglycemia, and pain were higher than others, and paresthesia, nausea, pyrexia,

abdominal pain, and anorexia were less likely than others (Tatonetti et al. 2011b). Using this pattern, more than 93 % of drugs that are known to alter glucose could be recovered. More interestingly, however, this pattern could be applied to patients on pairs of drugs to search for pairs that altered glucose. A highly correlated combination was the antidepressant paroxetine and the cholesterol medication pravastatin (Tatonetti et al. 2011b). In subsequent validation in three independent EHR systems, large increases in glucose were observed in patients on these two drugs, and in mouse studies of these two drugs, glucose was substantially increased. Thus, the adverse event patterns could be used to create patterns and detect new signals, not specifically reported in the database, but implied by the pattern of other side effects observed.

Mining the Literature to Build a Database of Gene-Drug Associations

Biomedical text can also be an important source of high quality information about the relationships among genes, drugs, and diseases (Garten et al. 2010). High fidelity natural language processing techniques (Chap. 8) can be used to extract information about gene-drug interactions. In some cases, the association between genes and drugs can be inferred simply by their co-occurrence in sentences (Garten and Altman 2009). In these cases, however, there can be many false positives due to sentences in which genes and drugs are mentioned, but are not actually interacting. A more precise method is based on careful parsing of sentences to find subjects and objects that are genes and drugs, and which are related by verbs that connect them (e.g. "CYP2D metabolizes codeine" has CYP2D as the subject, codeine as the object, and the verb "metabolizes" establishes their relationship) (Coulet et al. 2010). The rate of false positives is reduced in this case because more strict criteria are applied before claiming a relationship. These high quality interactions can be chained together to infer new knowledge. For example, drug-drug interactions often occur because two drugs share a common metabolizing gene and that gene becomes saturated in the presence of both drugs, and cannot adequately metabolize both of them. Thus, the observations that "CYP2D metabolizes codeine" and "CYP2D metabolizes metoprolol" might be combined to infer that codeine and metoprolol have a potential drug-drug interaction. There are a large number of similar inferences that could be drawn about the relationships between genes, drugs and diseases given a high quality database of pairwise interactions drawn from the published literature. Of course, some pairwise interactions may be incorrect, and so evidence for interactions should be combined from several sources (including EMR validation, for example) and once predictions are made, they should be embraced only with skepticism.

Using Drug-Target Interactions to Predict New Ones

Another way to find new uses for old drugs is to predict interactions between drugs and new potential targets. Many drugs are designed to interact with a single target based on a detailed understanding of disease pathology. Once the drugs are administered, however, they may not bind only the original target, but they may unexpectedly have effects based on their binding to other targets. Most commonly, these "off target" effects are considered side effects and are avoided. In some cases, however, the "off target" effect may be beneficial in the setting of some other disease. Thus, both for explaining the molecular basis of side effects and for finding new molecular evidence for beneficial novel effects, it is useful to connect drugs to proteins. One way to do this is to build computational and visualization methods for docking a 3D representation of a small molecule into the 3D structure of a target protein. This can be very successful, and has led to the hypothesis that a Parkinson's disease drug may treat tuberculosis (Kinnings et al. 2009)! In that case, the 3D structure of a tuberculosis protein had a pocket that appeared to have high binding potential to a known Parkinson's disease drug, and thus the hypothesis arose that the Parkinson's drug might inhibit TB growth. These structure-based methods are powerful but limited because we have the 3D structure of only a subset of human proteins. Another approach,

therefore, is based on looking for similarities in the list of drugs that have been shown experimentally to bind a protein. In this case, all that is needed are data from chemical assays showing which drugs bind which proteins. These are routinely collected in large screening experiments, and are available at the ChEMBL resource (Heikamp and Bajorath 2011), for example. Given two proteins with two lists of interacting drugs, we can compare the list of drugs to look for commonalities. If there are many commonalities between protein A and protein B, then one might conclude that the drugs that bind protein A may also bind protein B. This was the approach taken in the Similarity Ensemble Approach (SEA) where the list of drugs binding two proteins are compared using a measure of chemical similarity (Keiser et al. 2009). When the chemicals on the two lists are statistically similar (more than would be expected by chance), then the SEA method predicts cross-binding of ligands for the two structures. When this was applied to a large set of proteins, the authors found that the antidepressant fluoxetine (Prozac) had high potential binding to the beta-adrenergic receptor, and this was found experimentally to block the beta-1 receptor—demonstrating that Prozac is a type of beta-blocker!

Identifying Drug Targets Using Side-Effect Similarity

A critical goal in pharmacogenomics is to associate drugs with their target proteins (and thus their coding genes) in order to know where to look for variation that may affect drug response. Determining drug targets can involve a difficult and lengthy experimental program. Thus, it would be very useful to have computational methods for determining targets. One way to do this is to associate drugs to their side effects, and to look for side effect profiles that are similar across drugs. If one drug has a known target, and if another drug has a similar pattern of side effects, then the two drugs may share that target. This is based on the assumption that side effects arise from a few common mechanisms, and so genes involved in this mechanism may be targeted by multiple drugs or drug classes. In one study, Campillos et al. showed that they could create 1,018 drug-drug relationships based on shared side effects (Campillos et al. 2008). The side effects were taken from the SIDER database, and the drugs came from a list of 746 marketed drugs. Twenty of these drug-drug relationships were tested experimentally, and 13 of them were shown to bind common targets. Thus, a relatively straightforward association of drugs based on side effects allowed the definition of molecular targets. In related work, Hansen et al. showed that genes could be ranked by their likelihood of interacting with a drug based on looking at the degree of similarity between chemical structure and indications-of-use between the query drugs, and small molecules known to interact with the gene products and their close protein interaction neighbors (Hansen et al. 2009).

The examples we have discussed have several common features: they deal with the basic objects of diseases, drugs, and disease or adverse-event phenotypes; they integrate at least two sources of data to establish new relationships between these basic objects; and they connect clinical entities (drugs and diseases or adverse events) to molecular entities. Such examples represent only a small subset of the types of questions that can be asked with these valuable datasets. The key technical challenges are typically (1) finding adequate gold standards (Chap. 2) to evaluate the success of methods before applying them for novel discoveries; (2) understanding the sources of error and bias so that predictions are as reliable as possible; (3) designing careful statistical tests to ensure that the scoring and estimates of significance are accurate and useful (minimizing false positives, in particular); and (4) identifying and engaging experimental collaborators who can, when appropriate, test the predictions that are made in human or model systems. Recently, it has become clear that despite their shortcomings, EHRs can be extremely useful for initial validation of hypotheses about connections between drugs and adverse events (Tatonetti et al. 2011a). Gene-drug associations are typically tested in model systems with genes altered in order to reduce or eliminate their normal function, or by looking for covariation in human subjects.

25.4.5.3 Challenges for Pharmacogenomics

Target Expansion: Molecules to Networks

The emerging field of systems pharmacology is abandoning the view of "one drug, one target" and moving instead toward a view that "the network is the target." That is, the larger network of interacting genes is targeted by a drug at several points, and thus the systemic effects of drugs needs to be evaluated in order to understand better the molecular underpinnings of drug response. The challenges to systems pharmacology are similar to the challenges to the more general systems biology: defining the network topology and key players, creating ways to measure parameters, modeling nonlinear responses, and understanding how variation in the basic molecular players impacts the resulting phenotype—in this case drug response phenotypes.

Rare Variants

As whole genome sequencing increasingly provides data about rare variants, the paradigm of looking for common genetic variation that explains variation in drug response will need to be modified. There may be cases when variation in drug response is explained by multiple rare variants rather than one or a few common variants. This is particularly challenging because there will often be insufficient statistics to evaluate rare variants. In some cases, huge population-based studies may provide enough samples, but in other cases even these large cohorts will not have sufficient examples of any rare variant to allow statistical validation. In those cases, we will have to rely on computational techniques to assess the significance of very rare variations.

Computational Methods to Leverage Stem Cell-Based Model Systems

The rise in the use of stem cells will create opportunities for combining direct measurements of cellular response to drugs with systems models of response, whole genome variation, and epigenetic information. As we perfect methods for creating induced pluripotent stem cells and differentiating them into the target tissues, we will be in a position to measure the response to drugs directly on these cells, with identical genetic and perhaps epigenetic backgrounds. Computational methods for analyzing these responses and relating them to the expected response in the patients from whom these cells are derived will be a major challenge in the years ahead.

25.5 Basepairs to Bedside

Although the sequenced human genome has not been a panacea for human disease, it has enabled the beginnings of a new approach to human health and to the practice of **P4** (**personalized**, **predictive**, **preventive and participatory**) **medicine** (Hood and Friend 2011). As the price of genomic sequencing falls and as our knowledge regarding the meaning of genomic variation increases, genotypic data is poised to become a standard component of a person's medical record. In this section we describe the translational path of genomics, from sequencing in the lab to clinical relevance for individuals.

25.5.1 Whole Genome Sequencing

25.5.1.1 Technologic Advances

The DNA-probe approach to genotyping, described in Chap. 24, may be compared to looking for one's car keys under the proverbial street lamp. That is, the technology shines the light on a certain portion of the genetic landscape, and that is where we look. Which SNPs are included on a chip is determined in large part by which SNPs have been detected in the past, for example through the HapMap project (Kang et al. 2006). A number of new associations have been found in this way, whether because the SNPs themselves were of interest, or due to genetic linkage—the tendency for alleles located close to one another on a chromosome to be inherited together. However, more recent findings have demonstrated that the concept of "common disease-common variant" is flawed (Zhu et al. 2011). Indeed, there has been some disappointment in the extent to which GWAS has been able to explain common diseases with known genetic components

(Manolio et al. 2009). Whole genome or, in some cases, whole exome sequencing allows researchers to identify rare variants (i.e., those with a minor allele frequency of <1 %) that account for genetic disease. Advances in sequencing technologies (e.g., "**nextgen**" **sequencing**, see Chap. 24) and the corresponding decrease in cost, make genome-scale sequencing increasingly feasible in translational research and even in clinical care.

25.5.1.2 Whole Genome Versus Exome

Even with recent advances in genome sequencing technology, the cost to sequence a full genome at a rate of coverage that enables the identification of novel SNPs is still significant, on the order of thousands of dollars. However, recall that only about 1 % of the genome actually codes for proteins and 85 % of known disease-causing mutations with large effects occur in proteins (Choi et al. 2009). One way to further decrease the time and cost of sequencing is to look at only those stretches that code for actual proteins. This can be justified because most variants known to underlie Mendelian disorders disrupt protein-coding sequences. Of course, this approach will miss causal variations if they exist in the other 99 % of the genome. Moreover, a recent cluster of publications from the ENCODE (**Enc**yclopedia **o**f **D**NA **E**lements) Consortium asserts assignment of biochemical function for 80 % of the genome (Dunham et al. 2012). Some additional components, such as regulatory regions or splice acceptor and donor sites may be included as well to increase sensitivity without incurring significant additional cost. Exome sequencing in a small number of individuals has been used to identify the causal variant for rare diseases such as Miller's syndrome, a multiple malformation disorder (Ng et al. 2010), and Proteus syndrome, a disorder causing the overgrowth of tissues and organs, thought to have afflicted the 19th century Englishman known as *The Elephant Man* (Lindhurst et al. 2011).

25.5.1.3 Publicly Available Resources

The genomics community has generally been ahead of the curve with respect to data sharing. According to the *Policy for Sharing of Data Obtained in NIH Supported or Conducted Genome-Wide Association Studies* (*GWAS*),[10] genotypic data must be deposited to the NIH database of Genotypes and Phenotypes (dbGaP). Genomic variation data is also available through a number of other online resources—see Sect. 24.4.5 and (Sherry et al. 2001; WTCCC 2007; Altshuler et al. 2010a). In 2008, researchers demonstrated that the presence of a single genome within a complex mixture of DNA samples could be ascertained (Homer et al. 2008). This caused both NIH and the Wellcome Trust[11] to limit access not only to individual genomes, but to aggregate genomic information as well. (Note that the ability to determine the presence or absence of an individual's DNA in a heterogeneous sample presupposes the availability of detailed genomic information about the individual in question.) These actions prompted responses that ranged from "too little, too late" to "a heavy-handed bureaucratic response to a practically minimal risk that will unnecessarily inhibit scientific research" (Church et al. 2009). Current NIH policy allows investigators to submit a data access request to be reviewed by an NIH Data Access Committee. Access to data is granted once a Data Use Certification is co-signed both by the investigator and the appropriate official[s] at the investigator's affiliated institution.[12] Additional online genomic resources include TCGA (**T**he **C**ancer **G**enome **A**tlas) and WTCCC (**W**ellcome **T**rust **C**ase **C**ontrol **C**onsortium). TCGA is a joint effort of the National Cancer Institute (NCI) and the National Human Genome Research Institute (NHGRI) to accelerate understanding of the molecular basis for cancer through application of genomic technologies, including genome sequencing.[13] The WTCCC, established in 2005, comprises 50 research groups across the UK who have performed a series of genome-wide

[10] http://grants.nih.gov/grants/guide/notice-files/NOT-OD-07-088.html (Accessed 12/6/2012).

[11] http://www.wellcome.ac.uk/ (Accessed 12/6/2012).

[12] http://grants.nih.gov/grants/guide/notice-files/NOT-OD-07-088.html (Accessed 12/6/2012).

[13] http://cancergenome.nih.gov/ (Accessed 12/6/2012).

association studies and made the data available through application to a Consortium Data Access Committee.[14]

25.5.2 Here Are Some Human Beings

Promising though genomic medicine may be, much remains to be worked out technically, scientifically, and from the ELSI (ethical, legal, and social implications) perspective (see Chap. 10). Some notable pilot projects have been embarked upon in order to catalyze progress in all of these areas. Craig Venter was the first person to have his complete genome published in 2007. Since then, a number of human genomes have been sequenced, and some of those have been made available in the public domain. The question becomes: now what? What can any given individual learn from his or her complete genomic sequence? What does an individual *want* to learn, or not want to learn, as the case may be? The only reliable way to answer these questions is with empirical input.

George Church is a pioneer in genomic sequencing, inventor of the Polonator sequencer, and founder of personal genome sequencing company Knome. In 2005, Church started the Personal Genome Project (PGP), ultimately aiming to sequence 100,000 individuals in order to advance understanding of how genes contribute, along with environment, to human traits. The project "hopes to make personal genome sequencing more affordable, accessible, and useful for humankind."[15] A vanguard of ten volunteers—the PGP-10—were selected to have their genomes sequenced. This endeavor differs from other projects in one crucial way: in addition to making the sequence data publicly available, complete phenotypic data, including personal and health information, family history, and even name and photographs would be shared as well. This was a departure for the type of projects the NIH typically funds and supports. Generally, informed consent includes information on how the research team plans to secure privacy and confidentiality for the subject. In this case, sharing of personal data was part of the protocol itself. The first set of integrated data from this group was made available in October 2008.

Making this type of data both publicly available and personally identifiable was stepping out into socio-scientific *terra incognita*, generating some worry that it could affect health care, employment, insurance, and more. In 2008, the Genetic Information Nondiscrimination Act (GINA) was signed into law, but its scope is limited to employment and health care insurance. It does not address life, disability, or long term care insurance (Hudson et al. 2008). Though rare, there are a few notorious examples of lawsuits where employers performed genetic and health-related testing on employees without their consent (Angrist 2010), and though unlikely, the PGP warns prospective participants that their DNA could be artificially synthesized and planted at a crime scene (Lunshof et al. 2010). As much as the PGP has pushed the boundaries and helped to advance the technology, data management, and clinical issues involved with personal genomes, and will continue to do so, it also serves as a weather balloon from the ELSI perspective, generating empirical data on sociological atmosphere, ethical pressures, and legal winds of change (see Chap. 10 for additional discussion on these points).

Misha Angrist, a bioethicist at Duke University, is PGP Participant #4. As documented in his book *Here Is a Human Being*: *At the Dawn of Personal Genomics*, the early sequencing was slow going, the technology took time to work out the kinks, and the preliminary results were underwhelming even to the individuals who had been sequenced (Angrist 2010). The infrastructure is not yet in place to empower someone with his complete genomic profile to do much with that information. Angrist describes his own attempts to make use of tools for genomic interpretation—SNPedia,[16] Sequence Variant Analyzer (Ge et al.

[14] https://www.wtccc.org.uk/index.shtml (Accessed 12/6/2012).

[15] http://www.personalgenomes.org/ (Accessed 12/6/2012).

[16] http://www.snpedia.com/index.php/SNPedia (Accessed 12/6/2012).

2011), and the Church lab's open source Trait-o-Matic[17]—which he compares to the dial-up days of the internet. Out of all of the variants carried by the PGP10, only one was deemed serious. Steven Pinker carried a mutation for MYL2, which had been shown in some cases to cause hypertrophic cardiomyopathy (Angrist 2010). More importantly, however, the project has created a publicly available, integrated resource for genomic, environmental, and trait (GET) data (Lunshof et al. 2010) and an empirical test bed for tackling the ELSI issues brought to bear by such a resource.

As another proof of concept, collaborators at Stanford and Harvard did a complete sequencing, analysis, and genetic counseling for a 40-year-old male with family history of sudden death from cardiac arrest (Ashley et al. 2010). The goal was to determine how whole genome sequencing would translate to clinical application. The patient was found to have increased risk for myocardial infarction, type 2 diabetes, and some cancers. While most of the findings were not actionable, the patient had both increased risk for cardiovascular disease and genetic disposition to benefit from the use of statins and aspirin. Despite this, just over a year after publication, the patient maintained that he had "not been convinced that statins or aspirin would have enough beneficial effect relative to their risks," and had not therefore changed his pharmaceutical behavior (Quake 2011).

Just over a year after the Quake profile, the same group published their findings from performing whole exome sequencing on the first healthy nuclear family (Dewey et al. 2011). They generated an ethnically concordant reference sequence (i.e. a reference sequence based on a European population, reflecting the European background of the family in question), which enabled increased accuracy for rare mutations. Findings included high resolution inference of sites of recombination (i.e., where the parents' chromosomes "cross over" during meiosis), and a novel approach to HLA (Human Leukocyte Antigen) typing—important for risk in a number of diseases, particularly autoimmune disorders. For the family in question, they were able to determine that the father had passed down to his daughter a mutation for Factor V Leiden that poses increased risk for blood clotting. This is actionable information for women as the Factor V mutation is a contraindication for estrogen-based birth control pills (Singer 2011), and inherited thrombophilia is a known risk factor for pregnancy outcomes (Tenenbaum et al. 2012). (Note that Factor V mutations are also included in chip-based genotyping services, so whole genome sequencing was not the key enabling technology in this case.)

One key item reported in the paper by Ashley et al. was the fact that, in the absence of a centrally curated resource of all rare and disease-associated variants, the authors spent *hundreds of hours* reviewing databases. Moreover, the work was a collaborative effort among a number of highly trained experts in clinical genetics, genetic counseling, bioinformatics, internal medicine, pharmacogenomics, etc. (Ormond et al. 2010). Clearly new tools, automation, and infrastructure, as well as a whole new paradigm in genetic counseling, will be required to make incorporation of genomic data into health care feasible for the population at large.

25.5.2.1 Sequencing Early in Life

One crucial complication in the search for genomic explanations for any given disease or phenotype is the impact of environmental interactions. Over time, every person on earth is exposed to environmental factors that may differ based not only on a factory that disposes of industrial waste near a drinking water supply or the traffic on the street they grew up on, but also by the foods they eat, the climates in which they live, and the infections they have harbored. Those external variables, hard to control for and sometimes even to know, can have major effects on the downstream products and activities of one's genomic fingerprint. Early in life, however, those effects are less pronounced. Of course, the impact of the *in-utero* environment on the well-being of the developing fetus is well established. But a

[17] https://github.com/xwu/trait-o-matic/wiki (Accessed 12/19/2012).

genetic defect is much more likely to be the cause in a newborn with an unidentified disease than in an adult patient who has undergone a lifetime of environmental insults. In this vein, a number of initiatives have been established across the US to offer clinical sequencing services for young patients, including programs at Children's Hospital of Philadelphia, Duke University, Partners Health care, and the Baylor College of Medicine, and the Medical College of Wisconsin. More controversial on paper, and not yet being performed in practice, is pre-natal genome sequencing. Ethicists are just beginning to explore the potential implications of this possible direction (Donley et al. 2012).

Addressing the time and resources needed to perform genome interpretation, one striking success story was achieved at Children's Mercy Hospitals and Clinics in Kansas City, MO (Saunders et al. 2012). Investigators used an Illumina HiSeq 2500 machine and an internally-developed automated analysis pipeline to perform whole-genome sequencing and make a differential diagnosis for genetic disorders in under 50 h. The diagnoses in question are among the ~3,500 known monogenetic disorders that have been characterized. In this case, WGS is not being used to identify novel, previously unknown mutations. Rather, it is shortening to just over 2 days the family's agonizing waiting time that would traditionally take 4–6 weeks as a battery of tests were performed sequentially.

We offer one final example in which genome sequencing was used as a last resort in a medical odyssey to identify the cause of a mysterious bowel condition in a 4-year-old boy named Nicholas Volker (Worthey et al. 2011). Having ruled out every diagnosis they could conceive of, doctors resorted to exome sequencing, leading to the identification of 16,124 mutations, of which 1,527 were novel. A causal mutation was discovered in the gene XIAP. This gene was already known to play a role in XLP, or X-linked lymphoproliferative syndrome and retrospective review showed that colitis had been observed in 2 XLP patients in the past. Based on these findings, a cord blood transplant was performed, and 2 years later, Nic's intestinal issues have not returned. News coverage of this story by the Milwaukee Journal Sentinel was awarded a Pulitzer Prize for explanatory reporting.[18]

25.5.3 Direct to Consumer Genetics

In the wake of the human genome project and the commoditization of genotypic data, a number of companies were founded to provide consumers with their own genetic information directly. These direct-to-consumer (DTC) genomic companies began making the services broadly available when deCODE genetics launched the deCODEme service in November 2007, followed a few days later by 23andMe, with the slogan "Genetics just got personal. Don't worry. We're here to help" (Davies 2008). Navigenics was launched the following spring. These companies offered consumers the opportunity to provide a saliva specimen or buccal swab through the mail, and in exchange to receive genotypic information for a range of known genetic markers. Different companies emphasized different aspects of genetic testing. Navigenics focused on known disease risk markers, while 23andMe was much broader, including disease markers but also ancestry information and "recreational" genetic information, for example earwax type and the ability to smell a distinct odor in urine after eating asparagus. Navigenics offered free genetic counseling as part of their service, while 23andMe and deCODEme provided referrals to genetic counselors. A study of concordance between these three services found >99.6 % agreement among them, but in some cases the predicted relative risks differed in magnitude or even direction (Imai et al. 2011). This disagreement is likely due to differences in the specific SNPs and the reference population used to calculate risk.

From the companies' perspectives, their customers offer a rich resource of genomic data for potential research and data mining. 23andMe created a research initiative called 23andWe through which they enlist customers as "collaborators,

[18] http://www.jsonline.com/news/milwaukee/120091754.html (Accessed 12/21/12).

advisers and participants."[19] They invite users to fill out questionnaires and then use the phenotypic information to perform genome-wide analysis studies. This approach enabled researchers at the company to replicate a number of known associations, and to discover a number of novel associations, recreational though they may be, for curly hair, freckling, sunlight-induced sneezing, and the ability to smell a metabolite in urine after eating asparagus (Sobradillo et al. 2011). deCODE, purchased by Amgen in 2012, boasts a large number of medically significant genetic discoveries to have come out of their volunteer registry of 140,000 Icelanders, more than half of the adult population of the country.[20] Navigenics was purchased by Life Technologies in 2012 and no longer offers their Health Compass genetic testing service.

25.5.3.1 Ethical, Legal, and Social Issues (ELSI)

Unknown Unknowns

In a companion article to the Quake profile, it was asserted that consent for a process in which the risks of knowledge gained are not wholly understood is more complex than for simple genetic testing. People have trouble interpreting probabilities. Patients must be advised that they may find out things they did not want to know about. The eminent scientist James Watson made a point of requesting that his ApoE status be redacted from the release of his full genome because he did not want to know if he was at risk. His grandmother had died of Alzheimer's at 83, and he did not want to worry that every subsequent memory lapse marked the onset of dementia (Angrist 2010). Statistics predict that any given patient will find out he is a carrier for *some* lethal autosomal recessive disease. Illness aside, the average global non-paternity rate has, astoundingly, been estimated to be as high as 10 % (Olson 2007). All of this information could also have implications for the patient's children, present or future, and for other family members. Patients, this group concluded, must have access to trained professionals to provide answers to their questions, where answers exist. This will be difficult, lengthy, and expensive, but not to do it would undermine the consent process (Ormond et al. 2010).

ELSI and the Genomic Consumer

Direct to consumer genetic testing, and pursuit of genomic medicine more generally, raise a number of ethical, legal, and social issues (see also Chap. 10). Some worry that people are ill-equipped to process the results of these tests. But it is not clear that a paternalistic approach is a better alternative; there was a time when it was considered acceptable for a doctor not to disclose a cancer diagnosis to the patient himself (Novack et al. 1979). In addition, new discoveries are being made all the time—what are the obligations to follow up if something new (and dire? and actionable?) is discovered about a given subject? Other questions include whether enough is known for the results to be of any practical use, whether the service should be provided outside of the context of a relationship with a clinical caregiver, and whether results could have detrimental effects on a person's ability to secure health insurance. Some states have banned the services, others have made stipulations requiring clinician involvement and **CLIA certification** for the labs that handle the samples and process the results.[21]

Although knowing the "parts list" for the human genome is an important step, much remains to be understood about how genes factor into human health and disease. For most diseases, the environment plays as much, if not more, of a role as a person's DNA. Aside from some notable, deterministic exceptions such as Huntington's disease, most known risk alleles confer fairly low odds ratios unto themselves (see Chap. 3), making an individual, for example, approximately 1.1

[19] http://spittoon.23andme.com/2008/12/17/turning-research-participants-into-research-partners/ (Accessed 12/6/2012).

[20] http://www.decode.com/research/ (Accessed 12/6/2012).

[21] http://www.genomeweb.com/dxpgx/will-other-states-follow-ny-calif-taking-dtc-genetic-testing-firms (Accessed 12/6/2012).

times as likely as the average individual to develop a given condition. Even when ratios are as high as, say, twofold, it is of dubious actual utility to know that based on one's genotype, the odds of being diagnosed with Crohn's disease went from 0.5 in 100 to 1 in 100.

For certain disease markers, such as Alzheimer's or BRCA1 and BRCA2, it was, and largely still is, unknown what impact negative results might have on a customer's mental and emotion well-being. Some studies have shown that while a person experiences negative emotions immediately in the wake of learning the bad news, over a time period of months there is no significant difference in anxiety, depression, or test-related distress (Green et al. 2009). In any case, DTC genetics companies' websites must provide the ability to view sensitive results while protecting the customer from stumbling on these findings unintentionally. 23andMe, as an example, has spent considerable resources on the design of a user-friendly interface through which to present an individual's "health reports," or their individual genotype for markers that have been characterized through reliable, established research methods. Along with a text explanation, these health reports give a graphical depiction of a person's relative risk. Figure 25.11 shows one such graphic for an individual's risk of venous thromboembolism based on three specific clotting markers as well as the individual's sex and ethnicity. For sensitive results such as BRCA1 and 2, and markers for Alzheimer's and Parkinson's disease, the information is initially "locked." Users must explicitly click through an additional screen to confirm that they truly want to know genotype and relative risk for that trait.

Rulings and Regulations

From a regulatory perspective, it was not (and to some degree is still not) clear whether these services qualify as medical devices as defined by the FDA, and are therefore subject to regulation by the Agency. The DTC testing landscape is still rapidly evolving, and will continue to do so for the foreseeable future. Logistically, a prospective customer typically registers on the DTC company's website and a sample collection kit is sent in the mail, though 23andMe's kit is also available for purchase through Amazon.com. In May 2010, Pathway Genomics and Walgreens announced a plan to sell these kits in Walgreens drugstores, but the FDA sent a letter to Pathway Genomics indicating their belief that the company's genomic report qualified as medical device (Bradley et al. 2011) and as such required FDA approval. Plans to sell the saliva collection container in brick-and-mortar stores were put on hold until the regulatory issues could be resolved, or at least addressed.

Another high profile legal issue is the case of *Assoc. for Molecular Pathology v. Myriad Genetics, Inc.*, et al., regarding Myriad's patent on the BRCA1 and BRCA2 genes, which are included in 23 and Me's offerings,[22] and more generally whether genes should be patentable at all. In 2011, a federal appeals court overturned a lower court in the case of and found that genes can, in fact, be patented (Pollack 2011). This ruling was upheld in a court of appeals in 2012, however, in 2013, the Supreme Court partially overturned that ruling and found that isolated genomic DNA (gDNA) is not patent-eligible, but cDNA is. Disappointingly, this ruling did not do much to reduce ambiguity around these issues.

Clinical Training

Another group that is affected by DTC genetic testing is primary care physicians. Most doctors have only a basic level of training in genetics, and are ill-equipped to answer in-depth questions from patients who bring to an appointment printouts of their results from these services (Frueh and Gurwitz 2004). More knowledge is required, in addition to training and tools, before family care providers, internists, and even specialists, are prepared to incorporate genomic information into their clinical practice (Chan and Ginsburg 2011; Ormond et al. 2010).

[22] https://www.23andme.com/health/BRCA-Cancer/ (Accessed 12/19/12).

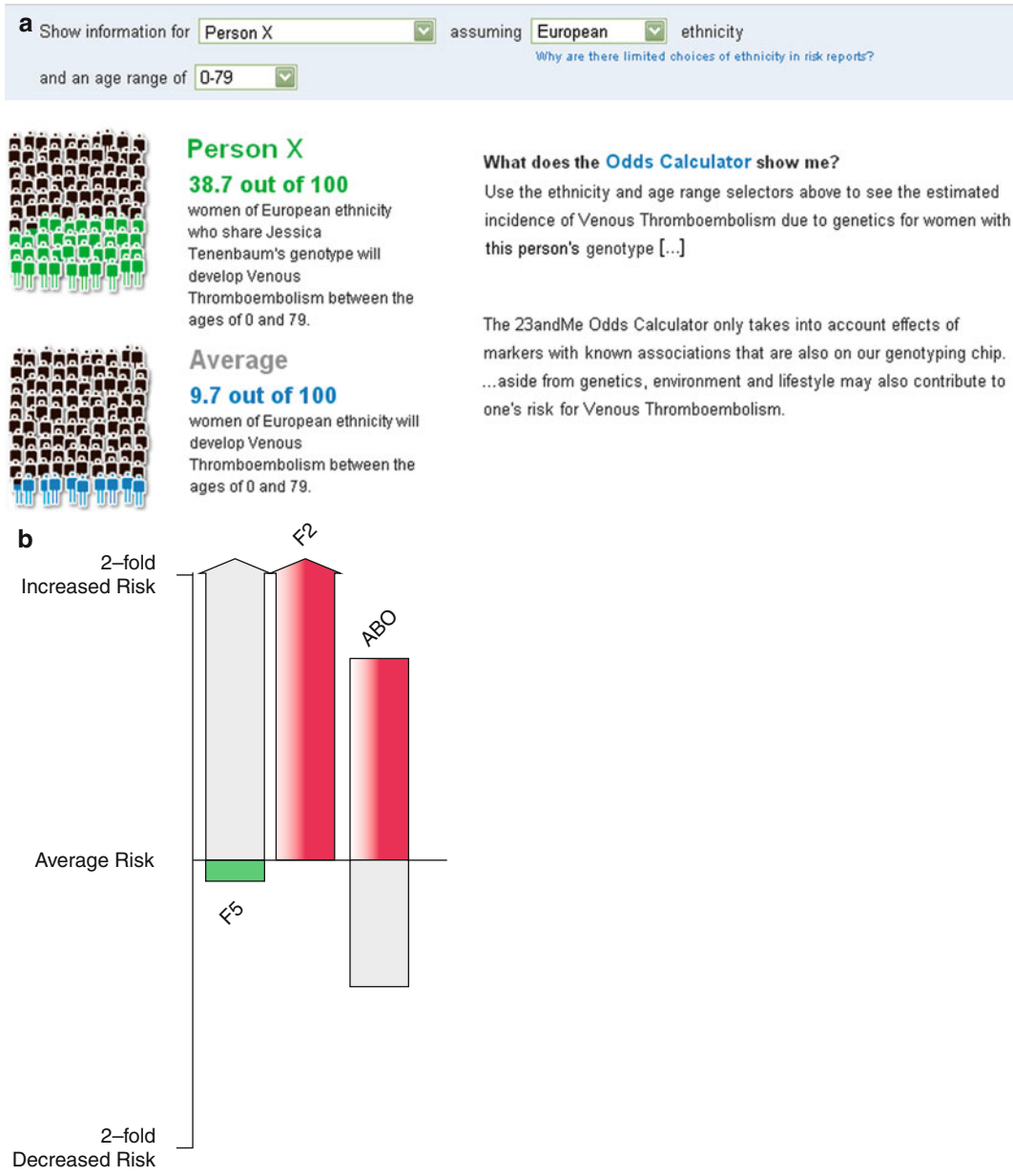

Fig. 25.11 23and Me's graphical representation of relative risk (© 23andMe, Inc. 2007–2012. All rights reserved; distributed pursuant to a Limited License from 23andMe.) (**a**) *Colored figures* represent the number of people on average out of 100 who are likely to develop venous thromboembolism over the course of a lifetime. *Green figures* represent the individual's personal reported risk; *blue figures* represent the average risk for females of European descent. (Accompanying text has been shortened for clarity.) (**b**) The individual's relative risk for each of three reported markers: factor V, factor 2, and ABO. (Specific values are displayed on the website when the user hovers the mouse over the *colored bars*)

25.6 Challenges and Future Directions

TBI as a discipline is in an exciting and dynamic phase—so much so that a number of items included in this chapter were announced or published during its writing and it is inevitable that new developments will have occurred by the time it is being read. Though challenges remain, the field is poised to become an increasingly crucial element of biomedical research and clinical practice. We conclude this chapter with a discussion of future directions and key challenges for this burgeoning discipline.

25.6.1 Expansion of Data Types

Genomic data are already being used to guide clinical care. Genomic data themselves are relatively straightforward in that an individual's genome is relatively static, and through the intrinsic physical properties of ribonucleic acids and the transcriptional process, DNA and RNA are relatively easy to capture, observe, and quantify. Proteins and metabolites are more challenging in this regard. Proteomic and metabolomic methodologies have primarily centered around isotopic labeling, but more recent approaches enable unbiased label-free identification and even quantification (Du et al. 2008; Wishart 2011). Identification of metabolites associated with disease has already enabled enzymatic drug targeting in diabetes, obesity, cardiovascular disease, and cancer, among other conditions (Chan and Ginsburg 2011). We expect that as proteomics and metabolomics standards and technologies continue to mature, they will play an increasingly significant role in translational research and practice.

The role of epigenetics needs to be understood more fully. It is clear that the environment can induce changes in the packaging and labeling of DNA. These environmental cues can include lifetime exposures to toxins, viruses, bacteria and nutritional compounds as well as drug exposures. Understanding the ways in which these epigenetic modifications affect phenotype is in its infancy, and so we must understand how to measure these effects, and then compute with them. The human microbiome is also an active area for translational research. **Metagenomics**, or genomics across organisms derived from an environmental sample, may be applied to the hundreds of bacterial species that make up the gut flora of every human being (Bruls and Weissenbach 2011). Finally, as standards are developed and clinicians and researchers see the value to be gained from structured data collection through studies such as The National Children's Study (Landrigan et al. 2006), structured environmental data is likely to be increasingly available to complete the picture for gene-environment interactions (Schwartz and Collins 2007).

25.6.2 Changes for Medical Training, Practice, and Support

Clinicians will need enhanced training in genetics and other areas described above. Curricular components relating to genetics, pharmacogenomics, statistics, and data standards will be increasingly important. Expertise in these fields will also need to be supplemented by an expanded workforce of genetic counselors. Increasingly, therapies will require accompanying diagnostic tests. As the opportunities for use of genomic data in clinical care continue to advance, it will become increasingly important to incorporate this information into both the electronic health record and into machine readable clinical care guidelines for clinical decision support. This in turn will require new standards to capture genomic findings, and new decision support tools to enable clinicians to incorporate this ever-increasing amount of information into their therapeutic decision making processes (Hoffman 2007). Clearly a number of standards exist in this space. The key will be in educating prospective users and in enforcing adoption. This applies to the full translational spectrum, from annotation of experimentally generated datasets to a common format for the exchange of clinically relevant omic data between EHR systems.

25.7 Conclusions

As the cost of data generation and storage continues to decrease, and the methods for data analysis and interpretation continue to advance, TBI is poised to be a key enabler of the vision for P4 medicine. One can imagine a day when every newborn has his or her genome sequenced and this information becomes a part of the medical record, much as blood type is recorded today. The biggest challenges to achieving this vision are likely not to be technical ones, but rather ethical, legal, and economic in nature (Schadt 2012). Society must strike a balance between privacy protection and facilitating progress in biomedical research. Legal issues will need to be worked out around direct-to-consumer genetic testing, preventing genetic discrimination, gene patenting, and many other such issues. Return on investment will need to be established through economic analysis combined with comparative effectiveness research (see Chaps. 11 and 26). Ultimately, someone will have to pay for these accompanying diagnostic tests. Major change is unlikely until an organization like the Center for Medicare and Medicaid Services (CMS) changes its policies. For example, CMS coverage for the genetic test to guide warfarin dosing is currently conditional upon it being ordered as part of a research protocol (Meckley and Neumann 2010). TBI will continue to play a key role in transforming these types of scientific discoveries into improvements in human health.

Suggested Readings

Altman, R. B., & Miller, K. S. (2011). 2010 translational bioinformatics year in review. *Journal of the American Medical Informatics Association, 18*, 358–366. This article summarizes Dr. Altman's third annual "year in review" presentation delivered at the 2010 AMIA Joint Summits on Translational Science in San Francisco.

Angrist, M. (2010). *Here is a human being: at the dawn of personal genomics*. New York: Harper. This text is written by one of the Personal Genome Project's first subjects, describing the project, the cohort, and the experience. It also gives a good overview of the background of the project and a number of ethical, legal, and social issues that it raises.

Capriotti, E., Nehrt, N. L., Kann, M. G., & Bromberg, Y. (2012 July). Bioinformatics for personal genome interpretation. *Briefings in Bioinformatics, 13*(4), 495–512. The authors of this review summarize key databases and bioinformatics tools that have been developed in recent years to aid in the intepretation of genomic variance. Resources covered include databases of variants, genotype/phenotype annotation databases, tools for gene prioritization and tools for interpretation of single nucleotide variants.

Davies, K. (2010). *The $1000 genome: The revolution in DNA sequencing and the new era of personalized medicine*. New York: Free Press. This text, written by the editor of BioIT World magazine, documents the characters, events, and issues in the race to achieve the $1000 Genome.

Hastie, T., Tibshirani, R., & Friedman, J. H. (2009). *The elements of statistical learning : data mining, inference, and prediction*. New York: Springer. A useful primer on the statistical concepts underlying machine learning approaches to biomarker discovery.

Kann, M.G, & Lewitter, F., (Eds.). (2012). *Translational bioinformatics*. PLOS Computational Biology Collections eBook. This eBook represents both the first "textbook" devoted entirely to TBI, and the first online, open access textbook from PLOS. In addition to many of the topics covered in this chapter, the collection includes chapters on related topics such as cancer genome analysis, micribiome analysis, structural variation, and protein interactions in disease.

Masys, D. R., Jarvik, G. P., Abernethy, N. F., Anderson, N. R., Papanicolaou, G. J., Paltoo, D. N., Hoffman, M. A., Kohane, I. S., & Levy, H. P. (2012). Technical desiderata for the integration of genomic data into electronic health records. *Journal of Biomedical Informatics, 45*(3), 419–422. The authors describe the characteristics of biomolecular data that differentiate it from other EHR data, enumerate a set of technical desiderata for management of biomolecular data in clinical settings (e.g., separation of molecular data observations from clinical interpretation, lossless data compression, support for readability by both humans and machines), and propose a technical approach to its representation.

Payne P.R.O., Sarkar, I.N., Embi, P.J., & Kahn M. (2011, December). *Journal of Biomedical Informatics*, *44*(Suppl 1), S1–S108. This supplement to JBI's 44th volume highlights the top papers from the 2011 AMIA Joint Summits on Translational Science. An editorial by N. Sarkar, 2011 TBI scientific program committee chair, gives an overview of the conference contents, and provides context for the papers selected to appear in this issue.

Sarkar, I. N., & Payne, P. R. O. (2011, December). The joint summits on translational science: crossing the translational chasm. *Journal of Biomedical Informatics, 44*(Suppl 1), S1–2. This editorial discusses the spectrum of biomedical informatics, from

biology to medicine, in the context of the NIH Roadmap and the Clinical and Translational Science Award program. It gives the history of the AMIA Joint Summits on Translational Science, and explains the emergence of TBI and CRI as disciplines unto themselves, intended to address the same issues that motivated those initiatives- namely translating scientific discoveries into meaningful changes in health care delivery.

Sarkar, I. N., Butte, A. J., Lussier, Y. A., Tarczy-Hornoch, P., & Ohno-Machado, L. (2011). Translational bioinformatics: linking knowledge across biological and clinical realms. *Journal of the American Medical Informatics Association, 18*, 354–357. The authors present the field of TBI in the context of successes from bioinformatics and health informatics.

Questions for Discussion

1. Should DTC genetic testing be regulated by the FDA?
2. Should genes be patentable?
3. Are there sufficient legal protections in place to prevent discrimination based on genomic information? If not, what regulations are needed?
4. Are we headed toward full disclosure of genomic information?
5. What are some reasons a researcher might not want to share research data? Should they be required to share? If so, under what circumstances (e.g., 6 months after first publication)?
6. For novel analyses applied to complex, high-dimensional datasets, should there be new guidelines in place to prevent reporting erroneous results through user error or data fraud? Why or why not?
7. What are the major barriers to incorporating the benefits of personalized medicine fully into standard practice?