Whole-Genome Sequencing in Personalized Therapeutics

P Cordero[1] and EA Ashley[2]

Eleven years since the initial drafts of the human genome were published, we have begun to see the first examples of the application of whole-genome sequencing to personalized diagnosis and therapeutics. The exponential decline in sequencing costs and the constant improvement in these technologies promise to further advance the use of a patient's full genetic profile in the clinic. However, realizing the potential benefit of such sequencing will require a concerted effort by science, medicine, law, and management. In this review, we discuss current approaches to decoding the 6 billion–letter genetic code of a whole genome in a clinical context, give current examples of translating this information into therapy-guiding knowledge, and list the challenges that will need to be surmounted before these powerful data can be fully exploited to forward the goals of personalized medicine.

INTRODUCTION

In 1990, after decades of considerable advancements in nucleic acids research sparked by the birth of recombinant DNA techniques, the US Department of Energy announced a bold and ambitious project with a goal of sequencing, in nucleotide resolution, the entirety of the human genome. The resulting cascade of events that eventually spread the use of genomics in the biological and medical sciences is well known. Researchers had earlier realized that, with sufficient effort, they could uncover the code that holds the key to an organism's physical traits. It did not take the more technologically minded part of the scientific community long to create novel methodologies to bring about vast increases in the throughput of genomic sequencing. This technological race began its exponential acceleration when public and private entities started competing to be the first to announce the first draft of the human genome.[1,2] Even after this historic milestone was reached, the ambition of making a whole-genome sequence available on an individual basis for clinical purposes pushed the sequencing industry to further accelerate and streamline the technology, thereby driving down the costs. Other genotyping technologies were significantly overhauled in the process: chip capture–based profiling of the most common single-nucleotide polymorphisms (SNPs) became powerful and inexpensive enough to jumpstart the personalized genomics industry and enable the first genome-wide association studies (GWAS). As a result of these mid-throughput methods, companies such as 23andMe, Navigenics, and others have begun to navigate the challenging terrain comprising the ethical, logistic, and legal implications of affordable personal genomics data.[3]

GWAS have provided the first glimpse of the power of large-scale genetic studies and the statistical and interpretational challenges that arise therein—characteristics that contrast with those of traditional linkage studies.[4,5] GWAS have revealed associations between traits and common single-nucleotide genotypes in certain populations. In the area of pharmacogenomics, large-scale efforts have pinpointed loci important for a multitude of genotype–drug relationships, for example, anticoagulant, antiplatelet, and interferon-α-enhanced antiviral drug response.[6–9] Such studies have the potential to establish criteria for personalized dosing. Drug-related adverse events have also been used, albeit less frequently, as the target trait in GWAS, yielding associations between adverse events and commonly used drugs. Examples include simvastatin-induced myopathy,[10] flucloxacillin-induced liver injury,[11] and thalidomide-related neuropathy.[12]

The results of GWAS are both encouraging and frustrating. Some of the genotype–phenotype relationships discovered in pharmacogenomics have proven strong enough that genotyping has been proposed as a recommended practice to the US Food and Drug Administration, thereby advancing the ideal of personalized therapeutics.[13] However, the heritability explained by genetic markers for most traits has been found surprisingly low, a dilemma now known as the "missing heritability problem."[14] Even while the less expensive genotyping technologies have helped disseminate the genome-wide paradigm, the price and turnaround time of high-throughput sequencing methods continue to fall exponentially. As a result, whole-exome and whole-genome studies

[1]Program in Biomedical Informatics, Stanford University, Stanford, California, USA; [2]Center for Inherited Cardiovascular Disease, Division of Cardiovascular Medicine, Stanford University, Stanford, California, USA. Correspondence: EA Ashley (euan@stanford.edu)

have recently gained traction as a viable strategy for genotyping in the future. Whole-exome studies, a less expensive alternative to whole-genome methods, have in recent years proved useful in determining the genetic causes of Mendelian diseases in small cohorts; they have solved genetic mysteries such as those represented by Miller, Kabuki, and Schinzel–Giedon syndromes,[15–17] amyotrophic lateral sclerosis,[18] and congenital cortical malformations.[19] Whole exomes have also been used to analyze genetic inheritance in families, in unprecedented resolution.[20]

Furthermore, we are seeing successful applications of whole-genome sequencing in uncovering the genetic culprits in Charcot–Marie–Tooth neuropathy[21] and dopa (3,4-dihydroxyphenylalanine)–responsive dystonia[22]; indeed, the latter study even provided enough information for the development of a personalized therapy. Even in cases where there is no adverse phenotype, clinical assessment of a whole genome has been shown to provide useful pharmacogenomic information and indications regarding risk of disease.[23,24] Whole-genome inheritance patterns and statistics have also shed light on the perils of using a "generic" reference sequence and stimulated the creation of novel methods for combining multiple disease risks and assessing synonymous coding SNPs.[24]

The constant improvement in sequencing technologies makes it likely that a patient's whole genome will be incorporated into the clinical profile. When integrated with the electronic health records and analyzed using tools for automatic genome analysis, the full genetic information of patients will provide a strong foundation for the development of drugs and therapeutics tailored to specific genetic profiles. It will also enable hypothesis-free, large-scale population studies with enough power to reliably discern loci of interest with single-nucleotide resolution.

In the following sections, we review the range of human genetic variation and its impact on genomic-guided therapies, describe current examples of translating whole-genome data into personalized diagnostics and therapeutics, and pinpoint challenges faced by the several disciplines involved in whole-genome clinical analysis.

ANATOMY OF HUMAN GENETIC VARIATION

The development of personalized therapies based on a patient's complete genetic information is both empowered and limited by our understanding of human genetic variation. After the release of the first two human reference sequences,[1,2] large-scale studies of the SNPs and structural rearrangement underlying human genetic variation were initiated.[25–27] The wealth of information that these studies and their resulting consortia have amassed has vastly increased our understanding of the plasticity of the human genome and has become the foundation of modern human genetics. Below we present the range of human genetic variation, give examples of genetic biomarkers for drug and disease phenotypes, and discuss their role in whole-genome clinical assessment.

Single-nucleotide variants

Single-nucleotide variants (SNVs) occurring at a single site have been the main class of variants studied since sequencing efforts began. In the human genome there are approximately 11 million SNPs with observed major allele frequencies of at least 1%.[28] These variants occur in a highly correlated manner, forming linkage disequilibrium (LD) blocks of SNPs that are likely to be inherited in tandem and that vary widely among population ancestries.[26] The presence of these LD blocks allows for choosing a reduced set of SNPs that can serve as tag variants for large regions of the genome. LD blocks are the foundation of haplotype structure, the genetic profile that is commonly used to determine genotype–phenotype associations from GWAS.

Through linkage studies and GWAS, a multitude of SNVs have been associated with a wide range of traits: diseases such as inflammatory bowel disease, type 2 diabetes, breast cancer, multiple sclerosis, macular degeneration, and myocardial infarction,[29–32] as well as pharmacodynamic/pharmacokinetic relationships such as those seen in Coumadin anticoagulant dosing[6] and responses to clopidogrel[9] and methotrexate.[33] These associations are recorded in publicly available databases such as the database of Genotypes and Phenotypes (dbGaP),[34] the Human Gene Mutation Database (HGMD),[35] and the Online Mendelian Inheritance in Man (OMIM)[36] database (**Table 1**) and form the cornerstone for genetic diagnosis and subsequent personalization of treatment.

Structural variants

In a broad sense, any genetic variation that is not a base change at one site is considered a structural variant. Small insertions and deletions—indels—have sizes ranging from 1 to 10,000 nucleotides and are estimated to occur with a frequency of 1 million per genome.[37] Indels occurring in coding regions can cause a reading frame shift or create a premature stop codon, possibly affecting a gene's protein products; these indels are labeled non-sense mutations. Contrary to biological intuition, nonsense mutations do not always have a deleterious effect, although they do occur less frequently than their less disruptive counterparts.[38]

Larger genome rearrangements include large regions that are duplicated in consecutive positions (known as copy number variants or CNVs) or inverted (inversions). In contrast with indels, large structural variants, especially CNVs, have been more commonly studied from a disease association perspective. CNVs are found in more than 20% of the human genome and have been linked to autoimmune diseases such as systemic lupus erythematosus as well as brain disorders such as autism and schizophrenia.[39–41] CNVs have also been found to be present in abnormally elevated numbers in the genomes of several types of cancer cells.[42] Similarly, inversions have been found to be more common than initially expected, occurring with a frequency of ~50 to 60 per genome and spanning regions with an average of 500,000 bases; they include the well-known genetic causes of hemophilia A and Hunter syndrome.[43,44]

Studies in the past decade have focused on the effects of single-nucleotide variants in drug metabolism and response. However, there has been a steady shift toward studying the impact of structural variations in these phenotypes as next-generation sequencing—the best method for detecting them—becomes increasingly more cost-effective. Pharmacogenetic studies have revealed

Table 1 Publicly available tools and databases for various tasks of genetic variant annotation and prioritization

Category	Database/tool/project	Description	URL
Genetic variant data sources	dbSNP[68]	Comprehensive, curated SNP and short indel database	http://www.ncbi.nlm.nih.gov/projects/SNP
	DbVar[69]	Comprehensive, curated database for structural variants	http://www.ncbi.nlm.nih.gov/dbvar
	DGV[70]	Human structural variants from samples with no phenotype	http://projects.tcag.ca/variation
Functional characterization of genomic elements	ENCODE[71]	High-throughput functional characterization of DNA elements, including noncoding regions	http://www.genome.gov/10005107
	SIFT[72], PolyPhen[73]	Prioritization of nonsynonymous SNPs	http://sift.jcvi.org, http://genetics.bwh.harvard.edu/pph2
Public gene–trait associations	dbGaP[34]	Comprehensive listing of genotype-to-phenotype mappings	http://www.ncbi.nlm.nih.gov/gap
	EGA[74]	Genotype–phenotype experiment archive	http://www.ebi.ac.uk/ega
Disease-associated mutations	HGMD[35]	Database for human disease mutations	http://www.hgmd.org
	OMIM[36]	Mendelian disease gene associations	http://www.ncbi.nlm.nih.gov/omim
	SwissVar[76]	Variant catalog of the UniProt knowledge bases	http://swissvar.expasy.org
	GAD[77]	NCBI source for genotype–disease associations	http://geneticassociationdb.nih.gov
	GWAS catalog from NHGRI[78]	SNP-phenotype associations found by GWAS	http://www.genome.gov/gwastudies
Whole-genome repositories	Complete genomics public genomes[79]	Complete genomics for 69 genomes from multiple ancestries (includes samples from the NHGRI and NIGMS repositories)	http://www.completegenomics.com/sequence-data/download-data
	1,000 Genomes[80]	Expanding resource currently housing three low-coverage whole genomes of multiple ancestries	http://www.1000genomes.org
Ancestry-focused variant data sources	HapMap[26]	Haplo-block mapping for diverse populations	http://www.hapmap.org
	HGDP[27]	SNP profiles of samples from several endogenous populations	http://hagsc.org/hgdp
Pharmacogenomic associations and data sources	PharmGKB[56]	Variant–pharmacokinetic/pharmacodynamic trait associations and gene–drug interactions	http://www.pharmgkb.org
	DrugBank[81]	Drug-target database with biochemical properties	http://drugbank.ca

dbGaP, database of Genotypes and Phenotypes; DGV, Database of Genomic Variants; EGA, European Genome–Phenome Archive; ENCODE, Encyclopedia of DNA Elements; GAD, Genetic Association Database; GWAS, genome-wide association studies; HGDP, Human Genome Diversity Project; HGMD, Human Gene Mutation Database; NCBI, National Center for Biotechnology Information; NHGRI, National Human Genome Research Institute; NIGMS, National Institute of General Medical Sciences; OMIM, Online Mendelian Inheritance in Man database; SIFT, tool that Sorts Intolerant From Tolerant amino-acid substitutions; SNP, single-nucleotide polymorphism.

CNVs that are possibly involved in sulfotransferase-related drug metabolism[45] and an indel that affects codeine-to-morphine metabolism.[46] The latter induces metabolism enhancement by generating an open reading frame in the CYP2D7 pseudogene, resulting in a functional protein that metabolizes codeine more efficiently as compared with the wild-type gene. This is a prime example of a mechanism that affects therapeutic efficacy and is present only in structural variants.

Rare variants

Although most attention has been focused on common variants, it now seems likely that the majority of genetic variation is, in fact, rare. Rare SNVs are defined as those having a major allele frequency of <1%; rare structural variants are more difficult to categorize because our catalog of structural variation in humans is still small. Prior to GWAS, it was widely believed that common variants were a main driver in the heritability of common traits and diseases. This view began to change when GWAS failed to detect common SNPs that strongly controlled the heritability of the most common traits and diseases, instead finding only associations with modest effects.[14] This missing heritability problem is still unresolved and remains a central question among the genetics community. A tentative hypothesis, known as the "rare-variant hypothesis," proposes that the compounded genetic effects of rare variants act as the main controllers of common traits.[47] Recent studies have started to accumulate evidence that supports this hypothesis, revealing that rare variants found to be associated with a particular trait tend to explain a larger percentage of the heritability than common variants do. An excellent example of this trend is the discovery of a rare variant in *MYH6* that was found to be highly associated with the risk for sick sinus syndrome.[48] This variant was identified by a combination of SNP-chip genotyping and whole-genome sequencing and is associated with the syndrome with a surprising odds ratio of 12.53, which is an effect size ~10-fold higher than that for most GWAS variants. A more recent study has shown that rare variants in SLCO1B1 have greater effects on methotrexate clearance than common variants do, thereby suggesting that therapeutics-associated traits may also be affected by rare variants.[49] These large effects on drug hypersensitivity

have attracted further efforts to uncover the mechanisms through which genetic variations in *OATP1B1* and *OATP1B3* may increase the risks of drug toxicity.[50]

Noncoding, synonymous variants, repeat regions, and pseudogenes

Although we have a framework for the interpretation of many of the categories of genetic variation discussed above, a gap exists in our understanding of noncoding and synonymous variations. Our lack of understanding does not, however, imply unimportance. For example, the majority of SNVs in GWAS that are significant at the genome-wide level are present in noncoding regions (and a significant minority are present in haplotype blocks that do not contain any gene).[51] Furthermore, a good number of synonymous SNVs have been implicated in biological processes and diseases.[52–55]

Certain areas of the genome pose a technical challenge to identification even when next-generation sequencing approaches are employed. Repeat regions and pseudogenes are good examples. Many genes that are of importance in pharmacogenomics have associated pseudogenes. With short-read sequencing approaches, assembly algorithms cannot indicate where a given short read should be placed. Long-read technologies will help greatly in tackling these challenges; however, in the meantime it is important to be aware of these limitations when analyzing next-generation sequencing data. As mentioned previously, structural variations have been shown to produce functional versions of otherwise nonfunctional pseudogenes, and these functional versions can even outperform the wild-type homolog with respect to efficacy in drug metabolism.[46] The findings suggest that these genomic elements cannot be ignored.

VARIANT PRIORITIZATION AND META-ANALYSIS FOR CLINICAL INTERPRETATION

The vast amounts of data available from whole-genome sequencing represent a challenge in interpretation, often requiring automated methods for annotation and prioritization of the variants (see **Figure 1**). Resources that house curated relationships between genotypes, disorders, and pharmacogenomic traits (**Table 1**) are essential for this task. Of particular interest to personalized therapeutics is the PharmGKB database,[56] a curated database of more than 2,000 genes linked to drug metabolism or implicated in drug response. Once the relationship between a variant and a particular trait has been found through the annotation process, a score that encapsulates the strength of the dependency, as supported by external evidence, can be calculated. There is an ongoing debate over how to integrate the usually noisy evidence extracted from the literature; most clinical interpretations simply report the list of variants found to be associated with a trait of interest, without venturing into estimating the total risk. In our recent whole-genome clinical interpretation efforts, we have proposed the use of likelihood ratios as an approach to integrating effect size.[23,57] Given that multiple studies may support a particular association, their combined likelihood ratio can be calculated using straightforward Bayesian methods.

Regardless of the methods used to combine genotype–phenotype association evidence, and even if statistical assumptions of test/study independence hold, several caveats should be kept in mind when performing this task:

- The quality of the study must be taken into account: some would discount all candidate gene studies performed before the GWAS era.
- The ancestry of the study population is key, and biases in the population samples must be accounted for.
- Combined likelihoods can lose their effect size as more studies are used in the meta-analysis, usually producing odds ratios that are much lower than those of studies supporting typical diagnoses or laboratory tests.[57] It is difficult to discern whether this phenomenon, which may confound risk interpretation, is due to a regression-to-the-mean process or to an incorrect assumption in the procedure for combining the likelihood ratios.
- There are a great number of conditional dependencies between a multitude traits and diseases, precluding an accurate, independent estimate of risk for each trait. The conditional dependencies may vary depending on the traits that are involved, and the complicated nature of the relationships precludes the use of standard statistical calculations to estimate the final risk. However, the directionality of the relationship can be calculated, giving an indication of the trait interactions.[23]

Major environmental factors that could outweigh the genetic effects must also be considered. The interplay among environment, genes, and disease is still poorly understood, mainly because of the vast number of possible gene–environment interactions. Even though environmental associations can be incorporated in the meta-analysis in much the same way as genotype–trait studies, the inclusion of the patient's actual environment into the equation is not trivial and can be done only in a qualitative fashion. Furthermore, environmental association studies are relatively rare as compared with their genetic counterparts; this is a trend that may be altered with environment-wide association studies.[58]

CLINICAL ASSESSMENT INCORPORATING PERSONAL GENOMES

Whole-genome sequencing is beginning to emerge as a clinical tool in certain controlled circumstances. The methodology has already been used to elucidate possible causes of a Mendelian disease: Lupski *et al.* conducted a study that found associations between Charcot–Marie–Tooth disease and several of the compound alleles in *SH3TC2*. Using the whole-genome sequence of an affected individual, the authors validated the relevant variants in family members using lower-throughput methods. In the realm of personalized therapeutics, whole-genome sequencing was recently used in a pair of 14-year-old fraternal twins diagnosed with dopa (3,4-dihydroxyphenylalanine)–responsive dystonia (DRD).[22] The analysis provided further insights into DRD and revealed causative compound heterozygous mutations in the

SPR gene that encodes for sepiapterin reductase. The findings motivated the physicians to prescribe 5-hydroxytryptophan, a serotonin precursor, alongside the L-dopa dopamine precursor commonly used to treat DRD. This recommended treatment improved the clinical outcomes for both twins, further reinforcing the fact that genome sequencing has potential in clinical therapeutics.

Cancer therapeutics

The promise offered by whole-genome sequencing in patients with cancer is the opportunity to move beyond historical organ- and microscopy-based diagnostic approaches to molecular diagnosis, changing the diseases' classification and suggesting tailored interventions. In a recent study, the whole genomes and whole transcriptomes of tumor biopsies were combined with targeted whole-exome sequencing of tumors and normal DNA in four patients with metastatic cancers.[59] One patient in

the study, with metastatic colorectal cancer, was found to have somatic point mutations in the *NRAS*, *TP53*, *AURKA*, *FAS*, and *MYH1*1 genes as well as an amplification and overexpression of cyclin-dependent kinase 8 (CDK8); another patient with malignant melanoma had point mutations in *HRAS* and a structural rearrangement affecting CDKN2C. Interestingly, both resulting mutational landscapes provided evidence of common affected pathways that could be targeted by a combination of therapies: CDK8 amplification could be targeted by CDK inhibitors, whereas the Ras-associated mutations could be treated with MEK (mitogen-activated or extracellular signal-regulated protein kinase) and PI3K (phosphatidylinositol 3-kinase) inhibitors. Unfortunately, because neither of these patients was eligible for enrollment in the trials that were actively testing these compounds, the therapeutic hypotheses could not be tested.

Another exposition of the use of whole-genome sequencing in personalized oncology and treatment was published in a study involving a patient with acute promyeolocytic leukemia.[60] This patient was diagnosed with acute myeloid leukemia and received all-trans-retinoic acid (ATRA) chemotherapy treatment for this disease until routine fluorescence *in vitro* hybridization tests revealed patterns of cryptic fusions that were not consistent with acute myeloid leukemia. The diagnostic conundrum of whether the patient had acute promyeolocytic leukemia or acute myeloid leukemia was solved with the detection, through whole-genome sequencing, of previously unseen breakpoints that resulted in a cryptic fusion oncogene consistent with non-ATRA-resistant acute promyeolocytic leukemia types. Consequently, ATRA was re-prescribed and had a positive effect. In the treatment of cancer, the power of whole-genome sequencing lies in its enhanced ability both to classify the disease and to personalize the therapy.

Preventive medicine

The noninvasive nature of whole-genome sequencing also allows for it to be applied to preventive medicine. Lo *et al.* presented low-coverage whole-genome sequencing of a fetus through sequencing maternal cell-free plasma DNA. In this study, both parents were carriers of mutations in the *HBB* gene that results in the autosomal recessive disorder β-thalassemia. Interrogation of the trio's DNA samples revealed that the fetus had inherited only the paternal *HBB* mutations, making him a carrier of the
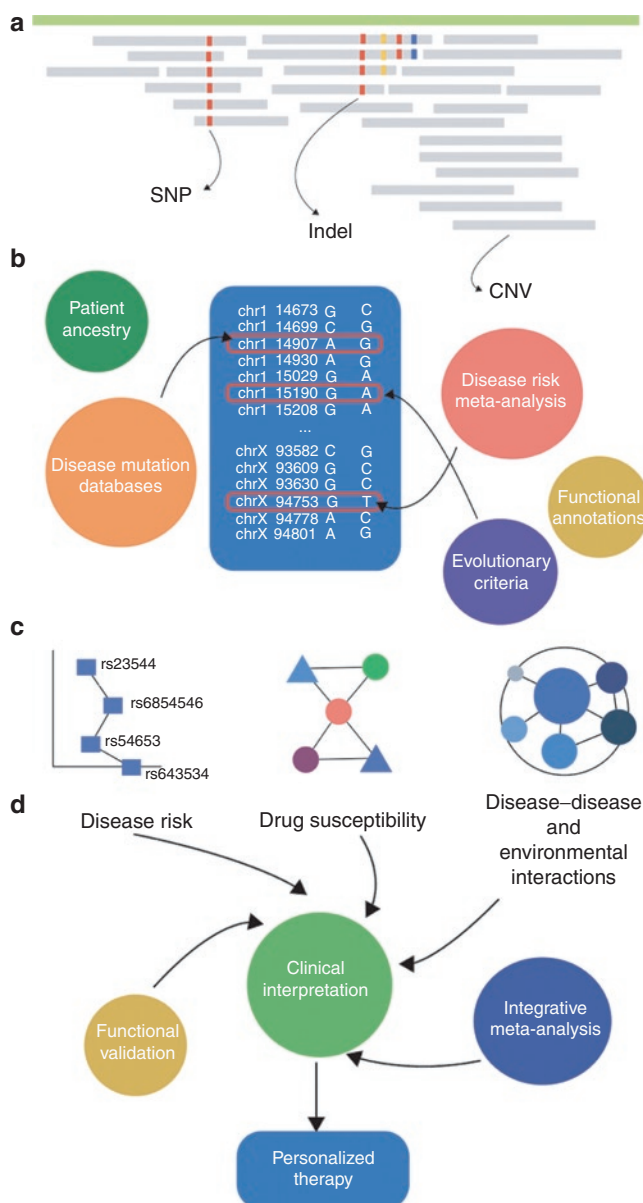

Figure 1 Bird's-eye view of the pipeline that transforms a patient's whole-genome information into personalized therapeutics. (**a**) A whole genome is usually received in the form of aligned reads with perhaps some tentative genotype calls. These calls must be validated and supplemented with structural variant calls, such as indels, inversions, and CNVs. (**b**) The resulting calls are then annotated and prioritized using meta-analyses to calculate disease risk, patient ancestry to accurately measure genetic variation characteristics, and computational methods based on evolutionary criteria and functional annotations to assess the effects of rare and personal variants, as well as variants falling in noncoding loci. (**c**) Disease risk, pharmacogenomic particularities, and disease–disease/disease–environment interactions are delivered in the final analysis and (**d**) interpreted clinically, informed by further, integrative meta-analysis, and orthogonal and functional validation to produce personalized therapies and recommendations. CNV, copy number variant; SNP, single-nucleotide polymorphism.

disease but not affected by it. In other settings, whole-genome sequencing can provide unparalleled power in risk assessment. We recently performed whole-genome clinical interpretations in five individuals. The first, whom we met in our clinic at the Stanford Center for Inherited Cardiovascular disease, was the fifth individual to undergo whole-genome sequencing. Given the individual's family history of sudden death and coronary artery disease, our assessment focused on cardiovascular risk and pharmacogenomics advice.[23] In a follow-up study, we refined methods for whole-genome clinical assessment in a family of four, developing a novel major-allele human reference sequence, building tools for computational long-range phasing, and revealing an inherited predisposition to blood clot formation and response to blood-thinning medications.[24]

As genome-wide studies are extended beyond common variants to include rare variants, risk scores will inevitably improve and become part of the clinical information given to patients. Several groups are exploring patient response to genetic findings presented both by direct-to-consumer genetic testing[61] and by research groups targeting the reporting of specific disease risks such as those for type 2 diabetes[62] and cardiovascular disease (J. Knowles *et al.*, personal communication). In direct-to-consumer clients, who usually have no associated phenotype, there has been no evidence of significant behavioral change after post-test risks are reported; however, this may not hold true for individuals who receive reports with a limited clinical scope (as in the type 2 diabetes and cardiovascular disease example studies). Behavioral studies are required to investigate the changes in lifestyle patterns in patients who have been exposed to knowledge of their estimated genetic risks for disease. Data from such studies will be crucial for understanding how the knowledge of personal risk affects therapeutic outcome.

PERSPECTIVES AND FUTURE CHALLENGES

Eleven years after the first drafts of the human genome were published, we have finally arrived at a point where we can start making use of the complete genetic information of an individual to personalize disease diagnosis and therapeutics. The acceleration of sequencing technology has allowed high-throughput sequencing to reach the clinical realm unexpectedly early. Whole-genome sequencing presents enormous potential but also significant challenges. The promise of personalized medicine can come to pass only through concerted efforts between multiple groups in a wide range of disciplines (see **Figure 2**):

- *Statistical and population genetics:* the goal of future efforts in statistical genetics will be to provide methods that can deliver high-confidence genotype–phenotype associations by routinely, or even automatically, scanning patients' genomic information. Current GWAS data already present formidable statistical challenges given the number of loci they interrogate; these difficulties will be exacerbated when 3 billion loci per person need to be handled. A knowledge of the true LD structure of each particular population or individual will help enormously, but this will require a uniform and rapid diffusion of sequencing technology.

- *Health-care administration and information technology*: the electronic medical record will be key in linking clinical observables to genetic data. Standardized and robust systems will facilitate studies that use genomic data and test the storage and delivery of genetic information. The manner of interaction of these systems is likely to lead to a shift in the ownership of clinical data, from the physician to the patient. Policies will have to be designed to allow rapid interinstitutional sharing of genetic data, channeled either through the patients themselves or through the use of open electronic health record platforms.

- *Medical sciences*: apart from the tremendous efforts that will be needed to fully interpret a whole genome, linking genetic information with phenotypic traits requires that the phenotyping itself be properly carried out. Adequate quantification of disease states that go beyond simple physical descriptions and aggregate measurements will be critical in making whole-genome sequencing clinically relevant. Interplay between phenotypes will also have to be taken into consideration; perhaps the concept would have to be redefined altogether and move toward a consideration of trait networks rather than of lone traits.

- *High-throughput experimentation*: now that we can measure and digitize the entire genetic material of an individual, additional molecular phenotyping will be necessary to track down genetic effects in the genotype–phenotype chain and to discover relevant biomarkers for further personalization of diagnoses and therapeutics. High-throughput experimentation technologies that give us insight into the transcriptomics, proteomics, metabolomics, and other biological aspects of an individual will have to mature further before they can be used in the clinic.[63] Our discovery that higher-order phenotypes can be affected through epigenetic events such as DNA methylation is a great example of the usefulness of these technologies. The genomic methylation landscape and other forms of genomic organization have been tremendously useful in characterizing several cancers,[64] and other forms of epigenetic flow may be at the core of other pathologies and drug–response phenotypes.[65]

- *Biomedical informatics*: the bioinformatics community will be challenged with the scope of population whole-genome data. The current limitations of storage technology, computing capacity, and technical ability of health-care staff will have to be overcome in the short term with automated, open-source pipelines for genomic clinical interpretation. Confident meta-analysis of these data will require tools and databases that are updated constantly with new findings from the scientific literature.

In summary, the translation of the diagnostic power of genome sequencing into therapeutics will require strategies to handle huge amounts of biological and medical data and present significant results in intuitive and clinically meaningful ways. We look forward to continuous, incremental, and ultimately transformative changes in clinical research and practice motivated by the
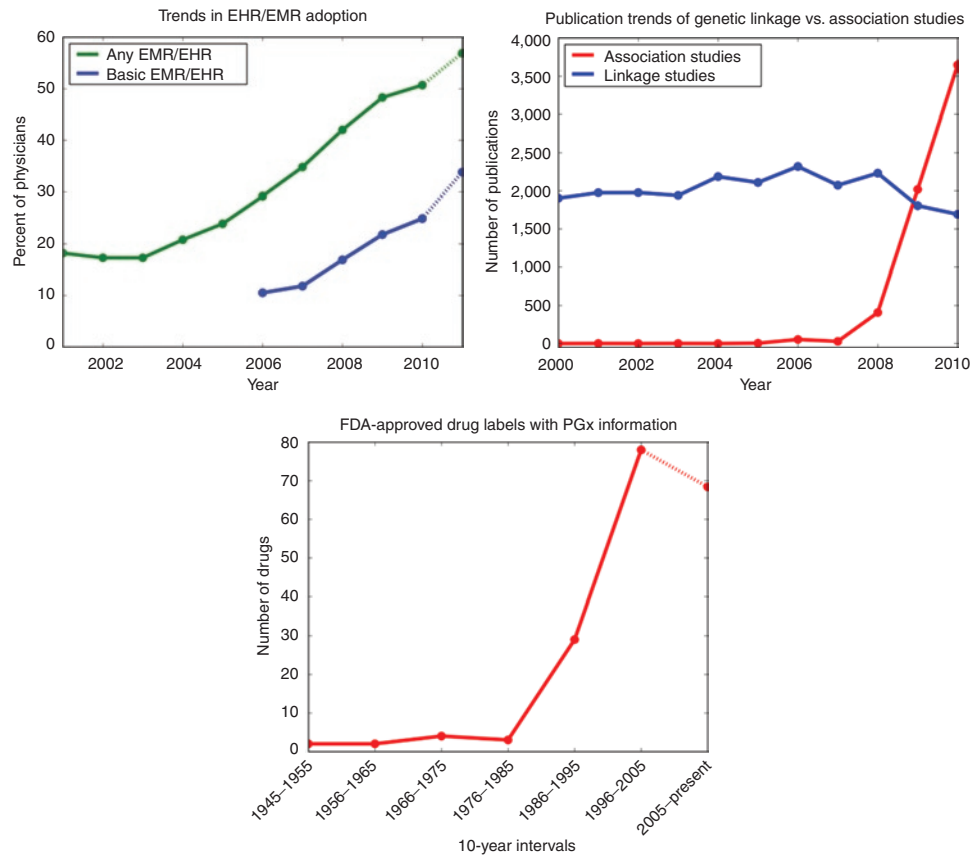
Figure 2 Setting the stage for genomics-enabled personalized therapeutics. Top left: adoption of electronic records by physicians in the clinical setting from 2000 to 2009 (2010–2011 are estimated values; data obtained from the CDC/NHCS National Ambulatory Medical Care Survey[66]). EHRs will greatly facilitate the clinical actionability of patient genetic data by storing genomic information and linking it to clinically relevant phenotypes. Top right: trends in genetic studies using linkage and association methodologies. The recent fall in the costs of mid- and high-throughput genotyping technologies have shifted genetic studies from low-resolution, family-based linkage studies to high-resolution, population-based association studies. This latter approach will be further enhanced when large-scale whole-genome studies are enabled by inexpensive sequencing technology. Bottom: number of FDA-approved drugs with pharmacogenomic label information per decade (data from 1945 to 2005 taken from Frueh *et al.*;[67] data from 2006 to the present inferred from publicly available information and online FDA reports). Although there is a promising trend toward incorporating pharmacogenomic knowledge into drug labels as additional information, the advent of personalized medicine will challenge current regulatory processes with respect to pharmaceuticals and clinical trials. CDC, Centers for Disease Control and Prevention; EHR, electronic health record; EMR, electronic medical record; FDA, US Food and Drug Administration; PGX, pharmacogenomics.

multidisciplinary advancements and analyses of whole-genome sequencing. In the next few years, the utility of a patient's complete genetic profile will become increasingly apparent, to the benefit of many.

CONFLICT OF INTEREST
E.A.A. owns stock in Personalis. P.C. declared no conflict of interest.

1. Venter, J.C. *et al.* The sequence of the human genome. *Science* **291**, 1304–1351 (2001).
2. Lander, E.S. *et al.*; International Human Genome Sequencing Consortium. Initial sequencing and analysis of the human genome. *Nature* **409**, 860–921 (2001).
3. Lenzer, J. & Brownlee, S. Knowing me, knowing you. *BMJ* **336**, 858–860 (2008).
4. Chee-Seng, K., En Yun, L., Yudi, P. & Kee-Seng, C. Genome-wide association studies: the success, failure and future. *Wiley Online Library* (2001). doi:10.1002/9780470015902.a0021995
5. Witte, J.S. Genome-wide association studies and beyond. *Annu. Rev. Public Health* **31**, 9–20 4 p following 20 (2010).
6. Cooper, G.M. *et al.* A genome-wide scan for common genetic variants with a large influence on warfarin maintenance dose. *Blood* **112**, 1022–1027 (2008).
7. Tanaka, Y. *et al.* Genome-wide association of IL28B with response to pegylated interferon-alpha and ribavirin therapy for chronic hepatitis C. *Nat. Genet.* **41**, 1105–1109 (2009).
8. Suppiah, V. *et al.* IL28B is associated with response to chronic hepatitis C interferon-alpha and ribavirin therapy. *Nat. Genet.* **41**, 1100–1104 (2009).
9. Shuldiner, A.R. *et al.* Association of cytochrome P450 2C19 genotype with the antiplatelet effect and clinical efficacy of clopidogrel therapy. *JAMA* **302**, 849–857 (2009).
10. Link, E. *et al.*; SEARCH Collaborative Group. SLCO1B1 variants and statin-induced myopathy–a genomewide study. *N. Engl. J. Med.* **359**, 789–799 (2008).
11. Daly, A.K. *et al.*; DILIGEN Study; International SAE Consortium. HLA-B*5701 genotype is a major determinant of drug-induced liver injury due to flucloxacillin. *Nat. Genet.* **41**, 816–819 (2009).
12. Johnson, D.C. *et al.* Genetic factors underlying the risk of thalidomide-related neuropathy in patients with multiple myeloma. *J. Clin. Oncol.* **29**, 797–804 (2011).
13. Daly, A.K. Pharmacogenomics of anticoagulants: steps toward personal dosage. *Genome Med.* **1**, 10 (2009).
14. Manolio, T.A. *et al.* Finding the missing heritability of complex diseases. *Nature* **461**, 747–753 (2009).
15. Ng, S.B. *et al.* Exome sequencing identifies the cause of a Mendelian disorder. *Nat. Genet.* **42**, 30–35 (2010).
16. Ng, S.B. *et al.* Exome sequencing identifies MLL2 mutations as a cause of Kabuki syndrome. *Nat. Genet.* **42**, 790–793 (2010).

17. Hoischen, A. *et al*. De novo mutations of SETBP1 cause Schinzel-Giedion syndrome. *Nat. Genet.* **42**, 483–485 (2010).

18. Johnson, J.O. *et al*.; ITALSGEN Consortium. Exome sequencing reveals VCP mutations as a cause of familial ALS. *Neuron* **68**, 857–864 (2010).

19. Bilgüvar, K. *et al*. Whole-exome sequencing identifies recessive WDR62 mutations in severe brain malformations. *Nature* **467**, 207–210 (2010).

20. Roach, J.C. *et al*. Analysis of genetic inheritance in a family quartet by whole-genome sequencing. *Science* **328**, 636–639 (2010).

21. Lupski, J.R. *et al*. Whole-genome sequencing in a patient with Charcot-Marie-Tooth neuropathy. *N. Engl. J. Med.* **362**, 1181–1191 (2010).

22. Bainbridge, M.N. *et al*. Whole-genome sequencing for optimized patient management. *Sci. Transl. Med.* **3**, 87re3 (2011).

23. Ashley, E.A. *et al*. Clinical assessment incorporating a personal genome. *Lancet* **375**, 1525–1535 (2010).

24. Dewey, F.E. *et al*. Phased whole-genome genetic risk in a family quartet using a major allele reference sequence. *PLoS Genet.* **7**, e1002280 (2011).

25. Iafrate, A.J. *et al*. Detection of large-scale variation in the human genome. *Nat. Genet.* **36**, 949–951 (2004).

26. Tanaka, T. The International HapMap Project. *Nature* **426**, 789–796 (2003).

27. Cann, H.M. *et al*. A human genome diversity cell line panel. *Science* **296**, 261–262 (2002).

28. Frazer, K.A., Murray, S.S., Schork, N.J. & Topol, E.J. Human genetic variation and its contribution to complex traits. *Nat. Rev. Genet.* **10**, 241–251 (2009).

29. Wellcome Trust Case Control Consortium. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* **447**, 661–678 (2007).

30. Sladek, R. *et al*. A genome-wide association study identifies novel risk loci for type 2 diabetes. *Nature* **445**, 881–885 (2007).

31. Duerr, R.H. *et al*. A genome-wide association study identifies IL23R as an inflammatory bowel disease gene. *Science* **314**, 1461–1463 (2006).

32. Easton, D.F. *et al*.; SEARCH collaborators; kConFab; AOCS Management Group. Genome-wide association study identifies novel breast cancer susceptibility loci. *Nature* **447**, 1087–1093 (2007).

33. Hinks, A. *et al*. Association of the 5-aminoimidazole-4-carboxamide ribonucleotide transformylase gene with response to methotrexate in juvenile idiopathic arthritis. *Ann. Rheum. Dis.* **70**, 1395–1400 (2011).

34. Mailman, M.D. *et al*. The NCBI dbGaP database of genotypes and phenotypes. *Nat. Genet.* **39**, 1181–1186 (2007).

35. Cooper, D.N. & Krawczak, M. Human Gene Mutation Database. *Hum. Genet.* **98**, 629 (1996).

36. Hamosh, A., Scott, A.F., Amberger, J., Valle, D. & McKusick, V.A. Online Mendelian Inheritance in Man. *Hum. Mutat.* **15**, 811–812 (2000).

37. Mullaney, J.M., Mills, R.E., Pittard, W.S. & Devine, S.E. Small insertions and deletions (INDELs) in human genomes. *Hum. Mol. Genet.* **19**, R131–R136 (2010).

38. Zia, A. & Moses, A.M. Ranking insertion, deletion and nonsense mutations based on their effect on genetic information. *BMC Bioinformatics* **12**, 299 (2011).

39. Willcocks, L.C. *et al*. Copy number of FCGR3B, which is associated with systemic lupus erythematosus, correlates with protein expression and immune complex uptake. *J. Exp. Med.* **205**, 1573–1582 (2008).

40. Marshall, C.R. *et al*. Structural variation of chromosomes in autism spectrum disorder. *Am. J. Hum. Genet.* **82**, 477–488 (2008).

41. Rare chromosomal deletions and duplications increase risk of schizophrenia. *Nature* **455**, 237–241 (2008).

42. Cappuzzo, F. *et al*. Epidermal growth factor receptor gene and protein and gefitinib sensitivity in non-small-cell lung cancer. *J. Natl. Cancer Inst.* **97**, 643–655 (2005).

43. Antonarakis, S.E. *et al*. Factor VIII gene inversions in severe hemophilia A: results of an international consortium study. *Blood* **86**, 2206–2212 (1995).

44. Bondeson, M.-L. *et al*. Inversion of the IDS gene resulting from recombination with IDS-related sequences in a common cause of the Hunter syndrome. *Hum. Mol. Genet.* **4**, 615–621 (1995).

45. Hebbring, S.J., Moyer, A.M. & Weinshilboum, R.M. Sulfotransferase gene copy number variation: pharmacogenetics and function. *Cytogenet. Genome Res.* **123**, 205–210 (2008).

46. Pai, H.V., Kommaddi, R.P., Chinta, S.J., Mori, T., Boyd, M.R. & Ravindranath, V. A frameshift mutation and alternate splicing in human brain generate a functional form of the pseudogene cytochrome P4502D7 that demethylates codeine to morphine. *J. Biol. Chem.* **279**, 27383–27389 (2004).

47. Dickson, S.P., Wang, K., Krantz, I., Hakonarson, H. & Goldstein, D.B. Rare variants create synthetic genome-wide associations. *PLoS Biol.* **8**, e1000294 (2010).

48. Holm, H. *et al*. A rare variant in MYH6 is associated with high risk of sick sinus syndrome. *Nat. Genet.* **43**, 316–320 (2011).

49. Ramsey, L.B. *et al*. Rare versus common variants in pharmacogenetics: SLCO1B1 variation and methotrexate disposition. *Genome Res.* **22**, 1–8 (2012).

50. van de Steeg, E. *et al*. Complete OATP1B1 and OATP1B3 deficiency causes human Rotor syndrome by interrupting conjugated bilirubin reuptake into the liver. *J. Clin. Invest.* **122**, 519–528 (2012).

51. McCarthy, M.I. *et al*. Genome-wide association studies for complex traits: consensus, uncertainty and challenges. *Nat. Rev. Genet.* **9**, 356–369 (2008).

52. Carlini, D.B. & Genut, J.E. Synonymous SNPs provide evidence for selective constraint on human exonic splicing enhancers. *J. Mol. Evol.* **62**, 89–98 (2006).

53. Chen, R., Davydov, E.V., Sirota, M. & Butte, A.J. Non-synonymous and synonymous coding SNPs show similar likelihood and effect size of human disease association. *PLoS ONE* **5**, e13574 (2010).

54. Pasmant, E., Sabbagh, A., Vidaud, M. & Bièche, I. ANRIL, a long, noncoding RNA, is an unexpected major hotspot in GWAS. *FASEB J.* **25**, 444–448 (2011).

55. Ørom, U.A. & Shiekhattar, R. Long non-coding RNAs and enhancers. *Curr. Opin. Genet. Dev.* **21**, 194–198 (2011).

56. McDonagh, E.M., Whirl-Carrillo, M., Garten, Y., Altman, R.B. & Klein, T.E. From pharmacogenomic knowledge acquisition to clinical applications: the PharmGKB as a clinical pharmacogenomic biomarker resource. *Biomark. Med.* **5**, 795–806 (2011).

57. Morgan, A.A., Chen, R. & Butte, A.J. Likelihood ratios for genome medicine. *Genome Med.* **2**, 30 (2010).

58. Patel, C.J., Bhattacharya, J. & Butte, A.J. An Environment-Wide Association Study (EWAS) on type 2 diabetes mellitus. *PLoS ONE* **5**, e10746 (2010).

59. Roychowdhury, S. *et al*. Personalized oncology through integrative high-throughput sequencing: a pilot study. *Sci. Transl. Med.* **3**, 111ra121 (2011).

60. Welch, J.S. *et al*. Use of whole-genome sequencing to diagnose a cryptic fusion oncogene. *JAMA* **305**, 1577–1584 (2011).

61. Bloss, C.S., Schork, N.J. & Topol, E.J. Effect of direct-to-consumer genomewide profiling to assess disease risk. *N. Engl. J. Med.* **364**, 524–534 (2011).

62. Grant, R.W. *et al*. Design of a randomized trial of diabetes genetic risk testing to motivate behavior change: the Genetic Counseling/lifestyle Change (GC/LC) Study for Diabetes Prevention. *Clin. Trials* **8**, 609–615 (2011).

63. Chen, R. *et al*. Personal omics profiling reveals dynamic molecular and medical phenotypes. *Cell* **148**, 1293–1307 (2012).

64. Tsai, H.C. & Baylin, S.B. Cancer epigenetics: linking basic biology to clinical medicine. *Cell Res.* **21**, 502–517 (2011).

65. Johnstone, S.E. & Baylin, S.B. Stress and the epigenetic landscape: a link to the pathobiology of human diseases? *Nat. Rev. Genet.* **11**, 806–812 (2010).

66. National Center for Health Statistics. Electronic health record systems and intent to apply for meaningful use incentives among office-based physician practices: United States 2001-2011. <http://www.cdc.gov/nchs/data/databriefs/db79.htm> (2011).

67. Frueh, F.W. *et al*. Pharmacogenomic biomarker information in drug labels approved by the United States food and drug administration: prevalence of related drug use. *Pharmacotherapy* **28**, 992–998 (2008).

68. Day, I.N. dbSNP in the detail and copy number complexities. *Hum. Mutat.* **31**, 2–4 (2010).

69. Church, D.M. *et al*. Public data archives for genomic structural variation. *Nat. Genet.* **42**, 813–814 (2010).

70. Zhang, J., Feuk, L., Duggan, G.E., Khaja, R. & Scherer, S.W. Development of bioinformatics resources for display and analysis of copy number and other structural variants in the human genome. *Cytogenet. Genome Res.* **115**, 205–214 (2006).

71. Birney, E. *et al*.; ENCODE Project Consortium; NISC Comparative Sequencing Program; Baylor College of Medicine Human Genome Sequencing Center; Washington University Genome Sequencing Center; Broad Institute; Children's Hospital Oakland Research Institute. Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature* **447**, 799–816 (2007).

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

73. Adzhubei, I.A. *et al*. A method and server for predicting damaging missense mutations. *Nat. Methods* **7**, 248–249 (2010).

74. Leinonen, R. *et al*. The European Nucleotide Archive. *Nucleic Acids Res.* **39**, D28–D31 (2011).

75. Stenson, P.D. *et al*. The Human Gene Mutation Database: 2008 update. *Genome Med.* **1**, 13 (2009).

76. Mottaz, A., David, F.P., Veuthey, A.L. & Yip, Y.L. Easy retrieval of single amino-acid polymorphisms and phenotype information using SwissVar. *Bioinformatics* **26**, 851–852 (2010).

77. Becker, K.G., Barnes, K.C., Bright, T.J. & Wang, S.A. The genetic association database. *Nat. Genet.* **36**, 431–432 (2004).

78. Hindorff, L.A. *et al*. Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proc. Natl. Acad. Sci. USA* **106**, 9362–9367 (2009).

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

80. Durbin, R.M. *et al*. A map of human genome variation from population-scale sequencing. *Nature* **467**, 1061–1073 (2010).

81. Knox, C. *et al*. DrugBank 3.0: a comprehensive resource for 'omics' research on drugs. *Nucleic Acids Res.* **39**, D1035–D1041 (2011).