Genome Medicine

REVIEW

Identifying disease mutations in genomic medicine settings: current challenges and how to accelerate progress

Gholson J Lyon*[1,2] and Kai Wang*[2,3]

Abstract

The pace of exome and genome sequencing is accelerating, with the identification of many new disease-causing mutations in research settings, and it is likely that whole exome or genome sequencing could have a major impact in the clinical arena in the relatively near future. However, the human genomics community is currently facing several challenges, including phenotyping, sample collection, sequencing strategies, bioinformatics analysis, biological validation of variant function, clinical interpretation and validity of variant data, and delivery of genomic information to various constituents. Here we review these challenges and summarize the bottlenecks for the clinical application of exome and genome sequencing, and we discuss ways for moving the field forward. In particular, we urge the need for clinical-grade sample collection, high-quality sequencing data acquisition, digitalized phenotyping, rigorous generation of variant calls, and comprehensive functional annotation of variants. Additionally, we suggest that a 'networking of science' model that encourages much more collaboration and online sharing of medical history, genomic data and biological knowledge, including among research participants and consumers/patients, will help establish causation and penetrance for disease causal variants and genes. As we enter this new era of genomic medicine, we envision that consumer-driven and consumer-oriented efforts will take center stage, thus allowing insights from the human genome project to translate directly back into individualized medicine.

A new era in genomic medicine

Since the completion of the initial sequencing of the human genome [1,2] there has been more than a decade of refinement and optimization of whole genome sequencing [3,4]. In 2007, when it was still too expensive to sequence whole genomes in any large quantity, it was first demonstrated that one can capture and sequence the protein coding exons from individual human genomes [5,6]. This was followed by the first in-depth analysis of an 'exome', defined as 'the set of exons in a genome' [7]. The next year saw the sequencing of 12 human exomes [8], followed a year later by the use of exome sequencing to determine the genetic basis for cases of Bartter syndrome [9], Miller syndrome [10] and Kabuki syndrome [11]. By 2010, with the cost of sequencing plummeting due to the rapid development of newer and better technologies, the question arose concerning when and whether it would be better and more cost-effective to go straight to whole genome sequencing (WGS). This was demonstrated by the simultaneous detection of the mutation underlying Miller syndrome by WGS of a family consisting of two affected siblings and their parents with the added bonus of being able to calculate a human intergenerational mutation rate of approximately 1.1×10^{-8} per position per haploid genome [12]. Many researchers have previously reviewed the success and promises of whole exome or genome sequencing in research settings [13-19]. Here, we describe the ongoing developments and challenges for the clinical application of whole exome or genome sequencing, and we discuss strategies to move the field forward.

There are ongoing and rapid developments in sequencing technology (Boxes 1 and 2). We consider whether a centralized model for whole exome or genome sequencing might take advantage of economies of scale and increased efficiency brought about by sequencing in a central location. Such a model of centralized WGS could be a much needed 'disruptive innovation' [20,21], if implemented well. When the company Amazon first started, not many people would have predicted that it would supplant many physical bookstores, but that is

*Correspondence: GholsonJLyon@gmail.com; kaichop@gmail.com
[1] Cold Spring Harbor Laboratory, New York, NY 11797, USA
[2] Institute for Genomic Medicine, Utah Foundation for Biomedical Research (UFBR), Salt Lake City, UT 84106, USA
[3] Zilkha Neurogenetic Institute, Department of Psychiatry and Preventive Medicine, University of Southern California, Los Angeles, CA 90089, USA

BioMed Central

Box 1. Progress in the application of genome sequencing

Researchers are currently sequencing thousands of exome and whole genomes, and the reagent costs for sequencing a whole genome may be as low as US$1,000 in the next couple of years. However, the sequencing technologies and methods used vary widely among researchers, and it is very much an open question how long it will take to achieve what is being called a 'clinical-grade' genome. We hasten to add that even the term 'whole genome sequencing' is misleading, as there are parts of the genome that are not readily amenable to sequencing, particularly with the current short-read sequencing technologies [131]. A newly published long fragment read technology may overcome many of these issues [132]. However, consistent and reproducible interpretation of the available genomic data and translation of the genetic findings into meaningful clinical action will be the more difficult and potentially most expensive part. A new generation of sequencing platforms that performs high-throughput sequencing (HTS) has now made it possible for individual laboratories to generate enormous amounts of DNA sequence data [133]. The inclusion of transcriptome sequencing, chromatin-immunoprecipitation (ChIP) sequencing and epigenetic analyses at unprecedented resolution has enabled the detection of subtle variations such as alternative splicing of individual exons, or base-pair-level binding preferences. There is no doubt that the ongoing avalanche of sequencing data will advance functional genomics studies [134,135], human disease studies [136], population genetics studies [137], metagenomics [138], clinical diagnosis [139], as well as other areas of biomedical importance. HTS data will also provide the foundation for new strategies in systems biology and individualized medicine. The miniaturization of sequencers with the advent of Illumina's MiSeq and Life Technologies' Ion Torrent is making HTS accessible to many more customers in clinical diagnostics, although incorporating these devices into clinical workflows will require substantive testing and validation. The development of newer technologies, including nanopore sequencing [140], could bring about substantially lowered cost, increased convenience, and a potentially simplified bioinformatics pipeline with longer sequence reads. A major drawback of these technologies pertains to possibly higher error rates.

indeed what has occurred. Similarly, a company that can implement an effective centralized sequencing facility and return of genomic data via the internet or using a secure cloud computing architecture could capture a portion of consumer- and hospital-oriented WGS, perhaps augmenting the efforts of localized and academic-based sequencing centers. The industrialization of WGS could also raise the quality standards. An example of industrialization relates to the current manufacturing of oligonucleotide primers used all over the world for polymerase chain reaction (PCR): nowadays most researchers order primers from centralized companies rather than synthesizing primers at local laboratories, mainly because these companies have achieved higher quality at a reduced price. Some WGS is indeed already being performed in central locations by at least two companies: Illumina (San Diego, California, USA) and Complete Genomics (Mountain View, California, USA). The key question will be whether clinicians and hospitals will be willing to send out DNA samples to a centralized location, rather than setting up the sequencing machines and bioinformatics resources locally. A compromise might be along the lines of what has occurred with the disruptive technology magnetic resonance imaging (MRI), which has been deployed in many hospitals but also at stand-alone MRI diagnostic centers throughout the USA.

Over the past 3 years, there has been a period of primarily exome sequencing while awaiting a further decrease in cost for WGS [22]. There have now been, as of July 2012, 747 publications involving exome studies (according to PubMed searching with the term 'exome'). So far, the main achievements of exon capture and high-throughput sequencing in genomic medicine have been the identification of the causes of many clinically characterized Mendelian disorders (that is, single gene disorders). Furthermore, exome sequencing has been applied to the study of a multigenerational pedigree [23] and has also been used in molecular diagnostics (for example, to diagnose neonatal diabetes [24] and an X-linked inhibitor of apoptosis deficiency, with the latter prompting an allogeneic hematopoietic progenitor cell transplantation with promising results [25]). In 2011, X-chromosome exon capture and sequencing was used to determine the genetic basis of a previously undescribed and idiopathic disorder, later named Ogden syndrome, which was shown to result from a defect in the amino-terminal acetylation of proteins [26]. Since that time, many other idiopathic disorders have been identified and their genetic basis determined via exome sequencing [27-30]. Many groups are also applying exome sequencing to the study of complex diseases or traits such as height, hypertension, diabetes and autism, resulting in the identification of rare variants that might play a role in human diseases [31-34].

As the cost of capturing and sequencing exomes has decreased, it has become easier to identify the genetic causes of very rare Mendelian diseases. The major caveat here is that the causative mutation must be present in the currently annotated exome, and we do not have a clear idea of how many diseases will be caused by mutations outside the exome, including in non-coding regulatory regions [35]. Informal polling of many human geneticists suggests that exome sequencing projects currently identify a truly causative variant in only about 10% to 50% of cases, although this estimate is very rough given

Box 2. Ethical issues and practical issues with clinical genome sequencing

In 2011, we discussed how we pursued and handled an unrelated finding emerging from exome sequencing [38], and one of us discussed the ethical, clinical and practical implications of exome and genome sequencing [141,142]. The argument has also been presented that 'there is nothing incidental about unrelated findings', and we consider the term 'incidental' to be misleading because use of this term trivializes, at least in the mind of the public, the amount of work that ought to go into figuring out causality for unrelated, unanticipated or secondary results (all of which are more suitable terms for such a finding) [143]. We are fast approaching a period in which thousands of human genomes will be sequenced, and we advocate for at least the initial germline genome of each person to be sequenced in a clinical grade manner, which includes adequate sample collection, tracking and sequencing in Clinical Laboratory Improvements Amendment (CLIA)-certified laboratories in the USA, so that researchers can return these results to research participants. Already, the company Illumina (San Diego, California, USA) has a CLIA-certified whole genome sequencing (WGS) pipeline (Box 1), and several companies, including 23andMe (Mountain View, California, USA) and Ambry Genetics (Aliso Viejo, California, USA), offer CLIA-certified whole exomes. Several academic institutions are also rapidly moving toward the implementation of CLIA-certified exome and whole genome pipelines in their facilities. We fully expect that the barrier to the clinical-grade sequencing of germline genomes, preferably from blood-derived DNA, will be substantially lower in 12 months' time, thus allowing a huge increase in such sequencing to occur. Several academic medical centers are initiating WGS projects, and we expect that this trend will increase as hospital leaders, executives and insurance companies realize that a germline genome is far cheaper than the current practice of single gene diagnostic odysseys, which can sometimes cost up to US$100,000 and stretch over several years [144].

The current CLIA-certified WGS at Illumina is performed with an average sequencing depth of more than 30×, with more than 95% of all calls made at a depth of greater than 10 reads or more. One could also increase coverage and accuracy by sequencing the same sample on two separate platforms (for example, Illumina and Complete Genomics) and using only shared variants as determined by one variant calling platform (for example, Genome Analysis Toolkit, GATK) [145]. However, this may currently be impractical due to storage and economic considerations, although we expect that this approach could become much more plausible in the future. As an alternative, for now, one could minimize cost but increase accuracy by only using variants called by two or more variant-calling approaches on one set of sequencing data. Additionally, there are different expectations and tolerances to false positives in a research setting versus in a clinical setting, and there is therefore a need to apply different filtering strategies in different settings. There will be constant improvements with assembly and variant-calling algorithms, so that the overall accuracy of WGS will always be improving [146].

that most researchers do not publish their results (or lack thereof) when exome sequencing fails to identify a causative variant. This estimate is also crucially dependent on one's definition and threshold for proof of causation.

Challenges in the application of genome sequencing to identify causal mutations

Recently published whole exome and whole genome sequencing studies have taught us a great deal about the challenges and obstacles of finding disease causal mutations. In this section, we review the recent studies and discuss their implications for the implementation of genomic medicine in clinical settings. Some of these challenges are summarized in Table 1.

Investigating Mendelian diseases in research or clinical settings

Although there are successes in using genome or exome sequencing on single patients or small nuclear families to identify new causal disease mutations, these should be considered rare exceptions given that it is incredibly difficult to prove causality for any mutation (that is, that it is both necessary and sufficient to be the genetic cause) with only one affected individual. In fact, some have suggested that a prerequisite for new disease gene identification should include finding more than one mutation in one gene in more than one pedigree [36]. In the current fragmented healthcare system, one can imagine that it can be incredibly difficult to find a second pedigree for very rare idiopathic conditions. In our own study on Ogden syndrome, we were very fortunate during peer review of our manuscript to find a second pedigree with exactly the same missense mutation, but we have yet to find an additional mutation in this same gene in a third family [26]. Most exome sequencing studies of idiopathic Mendelian diseases would benefit from a substantial collection of cases, or at least multigenerational pedigrees with two or more affected cases, especially when the disease has a dominant inheritance pattern. It is possible that behind every success story in finding new genes for Mendelian diseases, there are several unpublished failures. On the other hand, if the causal gene has already been identified for a Mendelian disease, then it is relatively straightforward to perform whole exome or genome sequencing on patients to confirm the genetic diagnosis of the disease and find potential new causal variants in the known gene. In other words, larger sample sizes are required in research settings for finding causal mutations in new genes, whereas genetic interpretation is feasible for individual patients with known Mendelian diseases in clinical settings with access to databases of known disease-causing variants [37].

Adequate sample selection for any exome or genome sequencing study is important, given that the bioinformatics

Table 1. Considerations and challenges for the identification of disease causal mutations

Considerations		Challenges	Solutions
Mutation detection	Platform selection	Different sequencing platforms have variable error rates	Increased sequencing coverage for platforms with high error rates
	Sequencing target selection	Exome sequencing may miss regulatory variants that are disease causal	Use whole genome sequencing when budget is not a concern, or when diseases other than well-studied classical Mendelian diseases are encountered
	Variant generation	Genotype calling algorithms differ from each other and have specific limitations	Use multiple alignment and variant calling algorithms and look for concordant calls. Use local assembly to improve indel calls
	Variant annotation	Multiple gene models and multiple function prediction algorithms are available	Perform comprehensive set of annotations and make informed decisions; use probabilistic model for ranking genes/variants
	Variant validation	Predicted disease causal mutations may be false positives	Secondary validation by Sanger sequencing or capture-based sequencing on specific genes/regions
Type of mutations	Coding and splice variants	Many prediction algorithms are available	Evaluate all prediction algorithms under different settings. Develop consensus approaches for combining evidence from multiple algorithms
	Untranslated region, synonymous and non-coding variants	Little information on known causal variants in databases such as HGMD	Improved bioinformatics predictions using multiple sources of information (ENCODE data, multispecies conservation, RNA structure, and so on)
Specific application areas	Somatic mutations in cancer	Tissues selected for sequencing may not harbor large fractions of cells with causal mutations due to heterogeneity; variant calling is complicated by stromal contamination; current databases on allele frequencies do not apply to somatic mutations; current function prediction algorithms focus on loss-of-function mutations	Sample several tissue locations for sequencing; utilize algorithms specifically designed for tumor with consideration for heterogeneity; use somatic mutation databases such as COSMIC; develop function prediction algorithms specifically for gain-of-function mutations in cancer-related genes/pathways
	Non-invasive fetal sequencing	Variants from fetal and maternal genomes need to be teased apart; severe consequences when variants are incorrectly detected and predicted to be highly pathogenic	Much increased sequence depth and more sophisticated statistical approaches that best leverage prior information for inferring fetal alleles; far more stringent criteria to predict pathogenic variants
Inheritance pattern	Inherited from affected parents	Rare/private mutations may be neutral	Evaluate extended pedigrees and 'clans' to assess the potential role of private variants
	De novo mutations from unaffected parents	Every individual is expected to carry three *de novo* mutations, including about one amino acid altering mutation per newborn	Detailed functional analysis of the impacted genes
Biological validation	Known disease causal genes	Difficult to conclude causality when a mutation is found in a well-known disease causal gene	Examine public databases such as locus-specific databases
	Previously characterized genes not known to cause the disease of interest	Relate known molecular function to phenotype of interest	Evaluate loss of function by biochemical assays where available
	Genes without known function	Difficult to design functional follow-up assays	Evaluate gene expression data. Use model organisms to recapitulate the phenotype of interest
Statistical validation	Rare diseases	Limited power to declare association	Sequence candidate genes in unrelated patients to identify additional causal variants
	Idiopathic diseases	Lack of additional unrelated patients	Comprehensive functional follow-up of the biospecimens from patients to prove causality
	Mendelian diseases or traits	Finding rare, unrelated individuals with same phenotype and same mutation to help prove causality	Networking of science through online databases can help find similarly affected people with same phenotype and mutation
Type of phenotypes	Mendelian forms of complex diseases or traits	Several major-effect mutations may work together to cause disease	Statistical models of combined effects (additive and epistatic) of multiple variants within each individual
	Complex diseases or traits	Many variants may contribute to disease risk, each with small effect sizes	Refrain from making predictions unless prior evidence suggested that such predictive models are of practical utility (for example, receiver operating characteristic >0.8)

HGMD, Human Gene Mutation Database.

analysis is aided tremendously by the appropriate selection of accurately phenotyped 'normal' or 'unaffected' individuals in a pedigree. For us, the exome sequencing of a quartet with dominantly inherited symptoms of attention-deficit hyperactivity disorder was insufficient to prove causality for any of the rare variants that we identified [38]. In contrast, for Ogden syndrome, sequencing a proband with a new idiopathic disease, presumed carrier mother, presumed carrier grandmother, unaffected brother and unaffected uncle in a pedigree allowed the identification of a potentially causative mutation, followed by confirmation with segregation analysis in affected and unaffected members of the family [26], illustrating the importance of good pedigree collections. Having access to clearly unaffected members of the family also allows one to eliminate sources of systematic error in the sequencing data by eliminating from consideration those variants also found in unaffected members [39].

Appropriate tissue selection

Whole genome or exome sequencing studies typically use DNA samples derived from peripheral blood. However, several recent studies demonstrated the presence of hematopoietic mosaicism [40,41]. The extent of such hematopoietic mosaicism among the population is not well studied, and its influence on variant calls remains largely unknown. When blood samples are not available, other tissues may be used, with buccal samples being the most common due to the non-invasive sampling and the commercial availability of kits for high-quality DNA extraction. It should be noted that buccal samples may be heavily contaminated by bacterial and food content, and these facts should be considered in bioinformatics analysis and data interpretation. Additionally, other sources of DNA can also be assayed under specific settings, such as skin or hair samples in forensic settings, or bones or teeth in archeological settings, all with their own caveats for choosing sequencing and analysis strategies.

For studies on somatic mutations in cancer, a tumor sample is typically used together with a germline sample (adjacent normal tissue or peripheral blood). It should be noted that adjacent normal tissue may still contain initiating mutations that are important for subsequent cancer progression. In some cases, tumor samples are stored in formalin-fixed paraffin-embedded conditions. They can be still used in whole-genome or exome sequencing with help from commercial kits, but one should expect lower data quality, and should carefully select the sequencing platform and library construction methods, given the fragmented nature of these archival DNA samples. Additionally, some non-cancerous diseases of specific tissues may involve somatic mutations, as somatic copy number variations (CNVs) can be identified in differentiated tissues within the same human subjects [42]. If neurological diseases such as epilepsy and schizophrenia are related to somatic mutations at specific brain regions, the evaluation of tissues at lesion sites may offer biological insights that are unattainable from peripheral blood. We discuss more about somatic mosaicism below.

Determining the precise phenotypic characteristics of diseases

Over 5,000 confirmed or suspected Mendelian phenotypes have already been documented in the Online Mendelian Inheritance in Man (OMIM) database, and each of them is represented by a set of specific phenotypic features or diagnostic criteria (although admittedly there are many non-specific diseases documented there). There is also a catalog of human genetic mutations in the Human Genome Mutation Database. In clinical settings, however, the exact disease diagnosis is sometimes difficult to make due to variable expressivity of disease variants, the phenotypic similarity of some diseases and the presence of modifier genes, even for classical Mendelian diseases. It is likely that a precise characterization of each patient is best accomplished by correlating genomic information and longitudinal phenotypic information with each other, thus enabling more accurate diagnoses. In light of this, we envision that a digitalized, longitudinal and more comprehensive description of phenotypes will be especially important for accurate identification of disease genes in clinical genomic sequencing.

Many community efforts are being undertaken to accurately assess phenotypes, including the use of standardized vocabularies and/or specialized diagnostic tests to refine the precise phenotypic characteristics of diseases. In the world of electronic medical records, there are many ongoing projects to develop more precise ontologies, including a Unified Medical Language System and a Systematized Nomenclature of Medicine Clinical Terms [43]. The International Statistical Classification of Diseases and Related Health Problems 10th Revision (ICD-10) is a medical classification list developed by the World Health Organization for the coding of diseases, signs and symptoms, abnormal findings, complaints, social circumstances, and external causes of injury or diseases [44]. This classification allows more than 14,400 different codes for specific phenotypes and permits the digitalized description of phenotypic features or suspected disease traits. The Human Phenotype Ontology (HPO) project has also been proposed to address this challenge [45]. The project provides a standardized vocabulary of phenotypic abnormalities encountered in human disease. It has over 8,000 terms representing individual phenotypic anomalies and has annotated all clinical entries in OMIM with the terms of the HPO. Within specific disease areas, clinicians are also developing specialized tests that better define phenotypic characteristics, especially for hard-to-diagnose diseases such as psychiatric disorders that do not depend on biochemical measurements or biopsies. For example, the Research Domain Criteria have been proposed as new ways of classifying psychopathology based on dimensions of observable behavior and neurobiological measures, thus reducing reliance on the semantics of traditionally defined disorders (such as autism, schizophrenia and bipolar disorder).

Collectively, adoption of these standardized and digitalized diagnostic criteria will help shape the phenotypic characterization of individual patients, identify specific sets of candidate diseases that may fit the phenotypic descriptions, and help to better identify and prioritize candidate genes for diseases in clinical settings. On the other hand, more refined phenotypic assessments come with the price of fewer available samples and less power to detect disease variants and genes. In practice, investigators often have to work with different factors, including being more inclusive in study design or conduct, and applying varying thresholds on phenotype during data analysis, to increase the odds of identifying disease causal genes and variants. For example, in the Kabuki syndrome study [11], the authors assessed the severity of each patient and identified the causal mutation only in the set of patients with more severe phenotypes.

The presence of loss of function mutations

Loss of function (LoF) variants can be classified as those that completely ablate or decrease protein function, and include nonsense mutations, splice site-disrupting mutations, insertions or deletions that disrupt a transcript's reading frame, or deletions encompassing a large part of protein-coding sequence. A recent study showed that genomes from healthy human subjects contain many LoF variants, suggesting unexpected redundancy in the human genome [46]. This well-designed study showed that each genome typically contains approximately 100 genuine LoF variants with about 20 genes completely inactivated and the others partially inactivated [46]. Therefore, the presence of a gene with two LoF variants in a personal genome does not automatically mean that the person will have a disease phenotype.

The establishment of databases of LoF variants, together with careful evaluations of the known functional evidence in the context of the phenotypes of interest, will be required to determine if a gene with LoF mutations is truly causal, at least within particular genetic backgrounds.

The complexity of disease and modes of inheritance

There are approximately 6 billion nucleotides of DNA in every cell of the human body, and there are about 50 to 75 trillion cells in each human body. This genetic complexity is further complicated by somatic mosaicism, epigenetic changes and other possible phenomena (such as heterosis, otherwise known as hybrid vigor, determined by non-mutually exclusive mechanisms like dominance complementation, overdominance and epistasis [47]). Many severe Mendelian diseases are caused by LoF variants. For example, in Kabuki syndrome, a dominantly inherited multiple malformation disorder characterized by a distinctive facial appearance, cardiac anomalies, skeletal abnormalities, and mild to moderate intellectual disability, among 32 disease mutations found in 53 families, 27 (20 nonsense mutations and 7 indels) resulted in a premature stop codon [11,48]. However, it is worth noting that many Mendelian diseases are also caused by missense mutations, with an example being Ogden syndrome, caused by a single amino acid change from a serine to a proline.

Although many Mendelian diseases can be classified as dominant, recessive, inherited or *de novo*, some Mendelian diseases or the Mendelian categories of complex diseases typically do not fall within clear-cut definitions. Instead, Mendelian forms of complex diseases can have genetic heterogeneity, including allelic heterogeneity and also heterogeneity in the mode of inheritance. For example, exome sequencing of 41 families with hypertension and electrolyte abnormalities identified two causal genes, encoding kelch-like 3 (KLHL3) and cullin 3 (CUL3) proteins [33]. However, *KLHL3* mutations were either recessive or dominant, whereas *CUL3* mutations were dominant and predominantly *de novo*. Therefore, caution should be taken when interpreting personal genomes with Mendelian forms of complex diseases, as the mode of inheritance may not be known *a priori*, and there are the combined complexities of locus heterogeneity, mixed models of transmission and *de novo* mutation. In addition, the presence of modifier genes can also have dramatic effects [49]. Somatic mosaicism is an increasingly recognized phenomenon in some diseases [50-53]. Proteus syndrome, which is characterized by the overgrowth of skin, connective tissue, brain and other tissues, was shown to be caused by a somatic activating mutation in the oncogene *AKT1*, encoding an enzyme involved in processes such as cell proliferation and apoptosis, proving that somatic mosaicism and activation of the PI3K-AKT (phosphoinositide 3-kinase - serine-threonine protein kinase) pathway underlies overgrowth and tumor susceptibility in this disorder [53]. Another group showed that ichthyosis with confetti, a severe, sporadic skin disease in humans, is associated with thousands of revertant clones of normal skin that arose from spontaneous loss via mitotic recombination of a region on chromosome 17q containing disease-causing mutations in the gene encoding keratin 10 (*KRT10*). They further suggested that either the revertant stem cell clones are under strong positive selection and/or that the rate of mitotic recombination is elevated in individuals with this disorder [54]. To complicate matters further, another group showed that retrotransposition of long interspersed nuclear elements-1 (LINE-1 or L1s), which are abundant retrotransposons that comprise approximately 20% of mammalian genomes, during brain development can have an impact on gene expression and neuronal function, thereby increasing brain-specific genetic mosaicism. They further showed that L1 retrotransposition can be controlled in a tissue-specific manner and that disease-related genetic mutations can influence the frequency of neuronal L1 retrotransposition [55].

In summary, the complexity of diseases and their mode of inheritance may help us prioritize variants (such as focusing on nonsense variants that arise *de novo* for a specific disease). On the other hand, investigators should be aware of the possibility of heterogeneous patterns of disease between families when performing data analysis.

The need for biological validation of causality for previously uncharacterized variants

Many novel variants will be discovered from exome and genome sequencing. In some cases, the causality is self-evident if the variant leads to severe loss of function and is located in a gene known to cause the phenotype of interest. In many other cases, especially for uncharacterized diseases and/or isolated cases, biological validation is necessary to prove the causality of the disease variants. Unfortunately, biological validation takes a substantial amount of time and effort, which can easily exceed the time taken for the original sequencing experiments. For example, a *de novo* mutation in the gene encoding sodium channel SCN8 was suspected to cause infantile epileptic encephalopathy in a single family, and the researchers analyzed biophysical properties of the mutant by current-clamp analysis in hippocampal neurons transfected with the mutant gene [56]. We identified a mutation in an amino-terminal acetyltransferase gene *NAA10* as a potential disease causal gene for Ogden syndrome, but the establishment of causality came after *in vitro* studies of the mutated protein, showing compromised amino-terminal acetylation. It was only much later that further confirmation came from a second unrelated family with the same phenotype and mutation [26]. Similarly, mutations in the heat shock protein co-chaperone BCL2-associated athanogene 3 (*BAG3*) were suspected to be causal for dilated cardiomyopathy in one family, yet to truly establish the functional relevance, the researchers performed knockdown of *BAG3* in a zebrafish model and recapitulated dilated cardiomyopathy and heart failure [57]. Biological validation in cell culture or various model organisms such as zebrafish and mice can therefore be extremely important to establish causality for uncharacterized variants and genes, particularly when mutations are only identified in one family.

Some previously identified disease causal variants may be false positives

Next-generation sequencing on large human populations could now enable the establishment of databases documenting allele frequencies of genetic variants, and such databases can provide a platform to separate very rare variants (potentially causal for Mendelian diseases) from low-frequency polymorphisms. Many previously reported ‘private’ causal variants have now been found to be present with low allele frequencies in apparently healthy human populations and are thus unlikely to be truly causal or at least are modified by genetic background effects. For example, among 197 previously published rare variants reported as causative of dilated cardiomyopathy, 33 were also present in the National Heart, Lung, and Blood Institute-Exome Sequencing Project (NHLBI-ESP) database (composed of many subjects from a well-phenotyped population), raising the possibility that some of the variants identified as disease-causing in sporadic dilated cardiomyopathy are either false positives or low penetrance alleles in human populations [58]. This example underscores the importance of public databases on allele frequencies from large collections of samples, such as the 1000 Genomes Project and the NHLBI-ESP projects, in helping researchers and clinicians decide on the clinical relevance of specific variants in personal genomes.

A leading example in human genetics concerning genetic background effects involves the concept of tri-allelism, which has been advocated to explain variable penetrance in Bardet-Biedl syndrome, a phenotype defined by the association of retinitis pigmentosa, obesity, polydactyly, hypogenitalism, renal disease and cognitive impairment [59-61]. This model of tri-allelism invokes a third allele in a separate gene as being sometimes involved in the variable expressivity of the phenotype, but this has been challenged by others [62-66], and at least one group maintains that all individuals that they have studied with two autosomal recessive mutations in the same gene have 100% penetrance, but with variable expressivity (that is, one individual might have retinitis pigmentosa only, whereas another individual might have the full-blown symptoms of Bardet-Biedl syndrome [62]). We consider that careful, longitudinal phenotyping of individuals with Mendelian diseases will be necessary for determining the true effects of modifier loci. This is one major reason why we support a ‘networking of science’ approach (see below), including consumer and patient engagement with various online phenotyping tools (for example, PatientsLikeMe).

Bioinformatics challenges in interpreting personal genomes in clinical settings

There is a growing gap between the generation of massively parallel sequencing output and the ability to process, analyze and interpret the resulting data. New users of sequencing technologies are left to navigate a bewildering maze of base calling, alignment, assembly and analysis tools. Many software tools developed for sequencing data are not sufficiently robust and can only work on one type of data generated from one type of sequencing experiment, limiting the ability to obtain critical biological insights. Additionally, many of the academic software tools are not well maintained or documented, perhaps due to the lack of support, funding or motivation after publication of the methodology or software tools. Bridging these gaps is essential, or the coveted US$1,000 genome will come with a US$20,000 or US$100,000 analysis price tag [67,68]. In fact, the challenging nature of bioinformatics prevents many biologists from embracing the new sequencing technologies because of the difficulties involved in analyzing the vast quantities of resulting data, which we refer to as the 'Genomic Deluge'.

There has been growing awareness in the genomics community of the need for novel informatics solutions, including better methods for aligning sequencing reads, variant identification, genotype calling and association tests, in order to take advantage of the massive amounts of sequencing data. In fact, dozens of short read alignment software programs are available now with different functionalities [69], as well as several single nucleotide variant and CNV calling algorithms [70]. However, in addition to identifying variants accurately, interpreting the functional impacts for large amounts of sequencing data is important to pinpoint potential disease causal genes and mutations.

Over the past 2 years, several variant annotation pipelines have been developed by many different groups [26,71-78]. In Table 2, we have summarized some current software tools that are capable of annotating genetic variants from high-throughput sequencing data. Many of these software tools include the ability to report *in silico* protein prediction scores, but they also include many more functionalities. We recognize that the community as a whole has performed extensive research on developing bioinformatic solutions for predicting the functional importance of certain classes of variants such as non-synonymous SNPs. Despite the availability of many such algorithms, different algorithms use different information and each has its own strengths and weaknesses (Table 1). Several groups have worked on improving the various scoring systems for judging the possible extent of the deleterious effects of a mutation [79-92]. However, predictions from different algorithms typically do not agree well: for example, only 5% of deleterious predictions were found to be shared between well-known algorithms (SIFT [82], Polyphen2 [93], LRT [94]) on a data set of three human genomes [94]. One group used data from the pilot3 study of the 1000 Genomes Project, available through Genetic Analysis Workshop 17, to compare (in blinded fashion) the results of four programs (SIFT, PolyPhen, MAPP and VarioWatch) used to predict the functional relevance of variants in 101 genes. The results modestly overlapped in the range of 59.4% to 71.4%, with only 3.5% of variants classified as deleterious and 10.9% as not damaging across all four programs [83]. It has been suggested that investigators should use predictions from multiple algorithms instead of relying on a single one [94]. As a preliminary step to facilitate the process, dbNSFP (database for non-synonymous SNPs' functional predictions) was developed. It compiles prediction scores from four popular algorithms (SIFT, Polyphen2, LRT and MutationTaster [92]), along with a conservation score (PhyloP) and other related information, for every potential non-synonymous variant in the human genome. However, these annotations are far from comprehensive, and we expect that more accurate predictions may be generated using additional information. Furthermore, other classes of variants, including synonymous, UTR, splicing and intergenic variants, are much less well studied, yet recent studies highlight their potential functional significance in human disease [95-98]. Therefore, the lack of comprehensive and powerful approaches for functional prediction is impeding the progress to study newly discovered variants, especially non-coding variants, for their involvement in human disease.

With functional annotation of variants in hand, there are two general strategies for inferring candidate genes: a probabilistic scoring approach and a stepwise filtering approach (Figure 1). The former approach is conceptually more sophisticated and less likely to miss truly causal variants, but the latter approach is currently more easily interpretable by researchers and clinicians without bioinformatics skills, and is therefore more widely used. Various bioinformatics tools, including those in Table 2, offer the ability to filter variants based on user-specified criteria and help with the identification of disease genes.

Disease-specific assignment of functional importance

Although all current software tools predict whether a variant is likely to be deleterious (for example, adversely affects protein or genome functionality), a more important question is whether the variant causes or modifies a certain disease of interest. For example, many genes encoding olfactory receptors are deleted in any human genome, but they can be safely ignored when analyzing Mendelian diseases. Therefore, we urge that disease- or application-specific treatment of genetic

Table 2. A list of open-access bioinformatics software tools or web servers that can perform batch annotation of genetic variants from whole-exome/genome sequencing data*

Tool	URL	Description	Features	Limitations
ANNOVAR	[http://www.openbioinformatics.org/annovar/]	A software tool written in Perl to perform gene-based, region-based and filter-based annotation	Rapid and up-to-date annotations for multiple species; thousands of annotation types are supported	Requires format conversion for VCF files; command line interface cannot be accessed by many biologists
AnnTools	[http://anntools.sourceforge.net/]	A software tool written in Python to annotate SNVs, indels and CNVs	Fast information retrieval by MySQL database engine; output in VCF format for easy downstream processing	Only supports human genome build 37; does not annotate variant effect on coding sequence
Mu2a	[http://code.google.com/p/mu2a/]	A Java web application for variant annotation	Web interface for users with limited bioinformatics expertise; output in Excel and text formats	Does not allow annotation of indels or CNVs
SeattleSeq	[http://snp.gs.washington.edu/SeattleSeqAnnotation/]	A web server that provides annotation on known and novel SNPs	Multiple input formats are supported; users can customize annotation tasks	Limited annotation on indels or CNVs
Sequence Variant Analyzer	[http://www.svaproject.org/]	A graphical Java software tool to annotate, visualize and organize variants	Intuitive graphical user interface; ability to prioritize candidate genes from multiple patients	Functionality is not very customizable; depends on ENSEMBL database for annotations
snpEff	[http://snpeff.sourceforge.net]	A command-line software tool to calculate the effects of variants on known genes such as amino acid changes	Rapid annotation on multiple species and genome builds; supports multiple codon table	Only supports gene-based annotation
TREAT	[http://ndc.mayo.edu/mayo/research/biostat/stand-alone-packages.cfm]	A command-line software tool with rich integration of publicly available and in-house developed annotations	An Amazon Cloud Image is available for users with limited bioinformatics infrastructure; offers a complete set of pipelines to process FASTQ files and generates annotation outputs	Only supports ENSEMBL gene definition and with limited sets of annotations
VAAST	[http://www.yandell-lab.org/software/vaast.html]	A command-line software tool implementing a probabilistic disease-gene finder to rank all genes	Prioritize candidate genes for Mendelian and complex diseases	Main focus is disease gene finding with limited set of annotations
VARIANT	[http://variant.bioinfo.cipf.es]	A Java web application for variant annotation and visualization	Intuitive interface with integrated genome viewer	Highly specific requirement for internet browser; slow performance
VarSifter	[http://research.nhgri.nih.gov/software/VarSifter/]	A graphical Java program to display, sort, filter and sift variation data	Nice graphical user interface; allows interaction with Integrative Genomics Viewer	Main focus is variant filtering and visualization with limited functionality in variant annotation
VAT	[http://vat.gersteinlab.org/]	A web application to annotate a list of variants with respect to genes or user-specified intervals	Application can also be deployed locally; can generate image for genes to visualize variant effects	Requires multiple other packages to work; only supports gene-based annotation by GENCODE
wANNOVAR	[http://wannovar.usc.edu/]	A web server to annotate user-supplied list of whole genome or whole exome variants with a set of pre-defined annotation tasks	Easy-to-use interface for users with limited bioinformatics skills	Limited set of annotation types are available

*Tools that are only commercially available (such as CLC Bio, Omicia, Golden Helix, DNANexus and Ingenuity) or are designed for a specific type of variant (such as SIFT server and PolyPhen server) are not listed here. CNV, copy number variation; SNP, single nucleotide polymorphism; SNV, single nucleotide variation; VCF, variant call format.

variants is needed, beyond simple variant annotation. There are several important reasons to pursue this approach. Firstly, conventional protein functional prediction algorithms only provide a binary prediction on whether a variant is deleterious or tolerated. However, a defect in protein function does not necessarily mean that a specific phenotype will be affected and investigators often have to search for clues on the specific disease classes (for example, cancer, immunological or cardiovascular traits) that the variant may influence. With this information in hand, biologists can design experiments to test the functionality of the variants in the context of

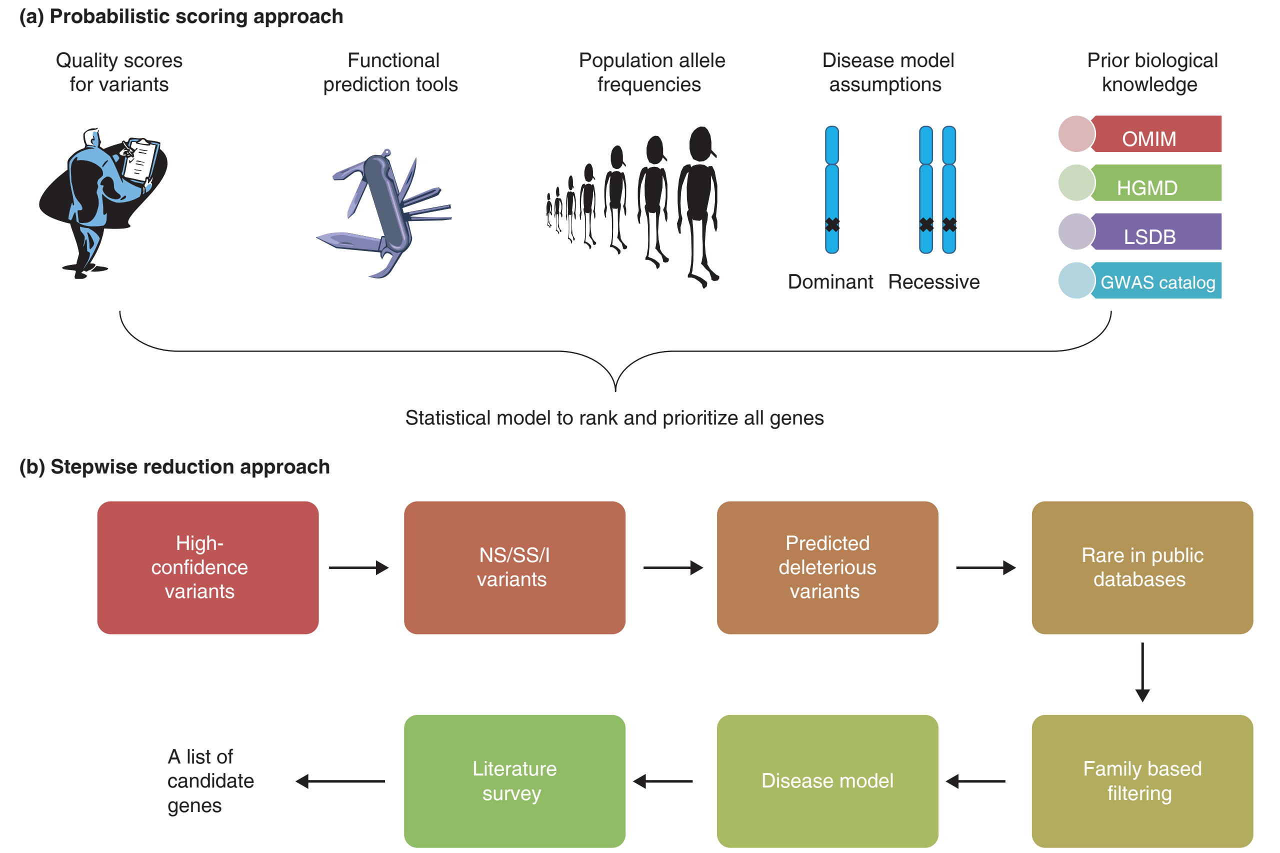
(a) Probabilistic scoring approach
Quality scores for variants
Functional prediction tools
Population allele frequencies
Disease model assumptions
Prior biological knowledge
OMIM
HGMD
LSDB
GWAS catalog
Dominant
Recessive
Statistical model to rank and prioritize all genes
(b) Stepwise reduction approach
High-confidence variants
NS/SS/I variants
Predicted deleterious variants
Rare in public databases
Family based filtering
Disease model
Literature survey
A list of candidate genes

Figure 1. Two approaches for prioritizing disease causal genes from whole-genome or exome sequencing data. (a) The probabilistic scoring approach collects relevant information from multiple data sources, and compiles a statistical model that ranks all genes in the genome by their likelihood of being disease causal. **(b)** The stepwise reduction approach removes variants that are unlikely to be disease causal based on a series of filtering criteria, until a small set of candidate genes is found. The first approach may be more effective and rigorous, yet the second approach may be easier for non-specialists to understand and interpret. GWAS, genome-wide association study; HGMD, Human Gene Mutation Database; I, indel; LSDB, locus-specific database; NS, non-synonymous; OMIM, Online Mendelian Inheritance in Man; SS, splice acceptor or donor site.

in vivo or *in vitro* models for the specific diseases of interest. Secondly, it is common practice to use prior knowledge of particular rare variants when trying to understand their clinical significance. For example, since the cholesteryl ester transfer protein gene, *CETP*, is associated with cardiovascular diseases [99], then any rare variants within *CETP* could more likely contribute to cardiovascular diseases, compared with other diseases. However, this is not necessarily the case. Indeed, genes can pleiotropically influence risk for multiple diseases. In fact, a single gene may influence risk for multiple diseases (for example, apolipoprotein E and heart and Alzheimer's diseases). Thirdly, for intergenic non-coding variants, large-scale projects such as ENCODE [100] provide ample amounts of tissue-specific information for assigning functional relevance. Although some enhancers and promoters are 'housekeeping' genomic elements that are active in most cell types, many enhancers, promoters, or repressors function in a tissue-specific or disease-specific manner, and the functional relevance of intergenic variants has to be assigned accordingly. For example, active enhancers in prostate cell lines are potentially more likely to contribute to prostate cancer formation and progression than enhancers in other cell lines; indeed, the 8q24 genome-wide association study (GWAS) hit sits inside an enhancer, and this same GWAS target has already been associated with several different cancers [101-104]. Such a region has strong existing evidence for involvement in cancer and is potentially less relevant for other types of diseases, and this information could be incorporated into predicting variant function.

Approaches for accelerating data collection and analysis

We have discussed various practical challenges and considerations for identifying disease-relevant mutations in genomic medicine settings, and we have highlighted some key differences compared to research settings. We have discussed ongoing community efforts to address these challenges, including clinical-grade sample collection, digitalized phenotyping, acquisition of high-quality sequencing data, rigorous generation of variant calls, comprehensive functional annotation of variants, and biological validation of variant function. In this section, we propose that additional crucial steps are needed to move the field forward and accelerate the clinical adoption of whole exome or genome sequencing. In particular, we strongly support the collaborative 'networking of science' model, which involves community efforts to identify disease mutations. We also discuss how putting more focus on families affected by specific diseases and 'clan genomics' (analysis of individuals sharing very recent ancestry) can help to improve the success of genomic medicine.

The 'networking of science' model

The advent of the internet and the digital age, including the explosion of social media, is enabling an incredible increase in the amount of networking worldwide. The increased ability to collaborate on a massive scale has resulted in projects such as the HapMap project [105-107], the 1000 Genomes Project [108] and the ENCODE project [109], just to name a few. As presented recently [110], this 'networking of science' is also giving rise to much more citizen science and online patient advocacy groups, conducting their own research with the online sharing of their own data, with an example being PatientsLikeMe [111,112]. We consider that a 'networking of science' model that encourages much more collaboration and online sharing of genomic data [113], including among the research participants and patients, will help to truly unleash the power of genome sequencing in clinical applications. This model should comprise open-access databases on the internet, including, at the very least, anonymized phenotype and genomic information. The Human Variome Project and others are working on standards for genome and variant annotation [114,115], and it will remain important to foster international collaborations between academic and commercial centers.

Networks form the basis of the scientific community, contributing to career development and the dissemination of scientific knowledge among peer groups. However, networks do not have to rely on persons whom a scientist already knows, and networks may indeed accelerate collaborative research with anonymous peers, citizen scientists, research participants and patients. In the genomic medicine setting, where each individual's expertise and knowledge may only focus on one part of the genome (such as for specific genes or diseases), network-based collaborative investigation may become especially important. One major obstacle to the implementation of 'networked genomic medicine' will be privacy concerns, as there are many cultural, political and legal obstacles to the sharing of genetic data, only some of which have been partially alleviated in the USA with the Genetic Information Nondiscrimination Act of 2008. There are also major problems with the entire electronic medical record system in the USA, most recently articulated in a succinct commentary [116]. It is perhaps naive to expect that these obstacles can be overcome within the next 20 years, and it may very well be the case that there might be a 50-year time horizon on the secure implementation of clinical genomics and individualized medicine. We certainly hope that every newborn will have the vast majority of their genome sequenced and digitally available by the year 2062.

Online sharing of genomic data, phenotype data and biological knowledge

Traditionally, online sharing of genomic data has been restricted to healthy human subjects (such as the 1000 Genomes Project) due to potential regulatory concerns. Recently, online sharing of patients' data, including medical history, diagnosis and genome sequence information, has been increasingly recognized as important for advancing collaborative scientific efforts [111,112]. For example, Complete Genomics has now shared genome sequence information for four anonymous patients with breast cancer online [117]. The San Antonio 1000 Cancer Genome Project is planning to share medical history as well as complete DNA sequencing for 1,000 cancer patients online [118]. Additionally, many individuals are brave enough to post their own genome data in blogs or personal websites, letting the whole world gain unrestricted access [119,120]. These types of online genomic data sharing for disease populations or healthy subjects will undoubtedly arouse interest from experts in different scientific areas, and allow them to apply their expertise to find better ways to examine genomic information.

In addition to sharing of genotype or phenotype data, we expect that collaborative online annotation of disease genes and disease variants may supplement current bioinformatics solutions. The possibility to build a gene-function wiki was proposed in 2006 [121] and several such wiki implementations were established several years later [122,123]. Given that different researchers or clinicians have different areas of expertise, it is possible that collaborative annotation of disease-relevant variants can accelerate the pace of variant prioritization with respect to specific disease phenotypes (much like the 'literature review' procedure in Figure 1 performed by the community). Frameworks of collaborative variant annotation already exist, including OMIM and GeneTalk [124], though OMIM is currently limited only to a small group of experts. It is foreseeable that the collective brainpower from the worldwide community of scientists and citizens, through online annotation of disease variants, may significantly facilitate the annotation of disease causal mutations in clinical settings.

One can imagine that in the not too distant future, each person's omics information (for example, genome, methylome, transcriptome and proteome), or at least whole genome DNA variants, will be sequenced at birth and automatically deposited onto a secure website, perhaps in a centralized location with massive online storage space. In such a setting, each individual will be able to review pertinent genetic findings at any time, including any rare variants shown to increase the risk of any diseases by at least ten-fold or greater within the 'clan' or population substructure of that particular individual [125,126]. Such participant-centric initiatives might also help to empower individuals with their own genomic data [127]. This will be further enabled by adequate ancestry and family history tracking, and one can already see that this is possible with social networking (for example, Twitter, Facebook or similar incarnations) and the current efforts of companies like 23andMe and Ancestry.com.

Lastly, there is an increasing movement for genomic data and samples to belong to the person from which they are derived, and this concept of 'portable legal consent' is being championed by Sage Bionetworks [113]. We acknowledge the possible security concerns that people have to face, if they choose to release genomic data online, just as people currently have to maintain the security of their social security numbers and other identifiable information. We consider that it might be impossible to maintain complete de-identification of genomic DNA in the face of potential 'genomic hackers', but the benefits far outweigh potential risks, at least for families with devastating diseases. Increased security is one possible reason to support (or even require) a more centralized and industrialized whole germline genome clinical sequencing effort, although we fully realize that this will likely not occur in the absence of regulation at the federal level.

The focus on diseases in families and 'clans'

In research settings, a research participant's questionnaire is assessed and then samples of blood or saliva are obtained, with the samples kept in a de-identified manner among those of many other study participants. In comparison, in genomic medicine settings, we emphasize that it is extremely important to put more focus and effort on families with diseases. Ultimately, the families, rather than researchers or research funders, are those who will benefit most from clinical genomic sequencing efforts, and they have a strong desire to find disease-relevant mutations and advance scientific discovery regarding these mutations. A second reason to focus on families is that the ability to build up personal relationships with families will help to obtain further support from these families for research efforts. This becomes especially important in light of the need to perform biological validation of genetic findings, as this will likely require the assessment of additional biospecimens, including blood, skin fibroblasts, urine, or any other tissues that might be of interest for investigating the role of genetic variants. Furthermore, the patients' families may help to recruit additional family members, including remote members who share very recent family history, for further genetic research, and this might help to separate truly disease causal variants from private but neutral variants in the family.

This idea can be extended to the concept of 'clan genomics' [125,126], which posits that there are unique combinations of rare variants in recent family lineages, playing a causative role in disease. In other words, recently arisen rare variants shared by a clan are more likely to be disease causal, compared with older common variants shared by a population or ethnicity group, particularly as such variants are too new to have been culled by natural selection. Therefore, for a specific patient, it is more important to consider the recent 'genetic history' of the patient's extended pedigree or clan, rather than their overall ethnic background. Genome sequencing on clans of subjects with elevated disease risks, such as on the Micronesian island of Kosrae [128] and Old Order Amish pedigree [129], may offer more important biological insights for identifying major-effect disease genes and variants.

Conclusions

We have reviewed here the explosive growth of exome and genome sequencing, highlighting the progress and challenges of applying these technologies to find and catalog disease mutations in genomic medicine settings. At some point, we will reach a tipping point in which it will be more cost-effective to sequence a whole genome rather than a collection of candidate genes or even large candidate gene panels or exomes. With the ongoing efforts to further reduce the cost of sequencing and improve the quality and bioinformatic analysis of sequencing data, the challenge in proving genetic causality will become more salient. In addition to improved bioinformatic and experimental approaches, this can be facilitated by a 'networking of science' model that requires collaborative efforts, and also by focusing on families in large pedigrees or the 'clan genomics' of subjects sharing recent genetic history. Forging strong ties with and among families will also enable access to other tissues to study newly discovered loci with many emerging technologies. As we enter this new period of individualized medicine [130], we expect that consumer-driven and consumer-oriented efforts will allow translation of the human genome project directly back to each individual.

Abbreviations

CNV, copy number variant; GWAS, genome-wide association study; HPO, Human Phenotype Ontology; HTS, high-throughput sequencing; LoF, loss of function; MRI, magnetic resonance imaging; NHLBI-ESP, National Heart, Lung, and Blood Institute-Exome Sequencing Project; OMIM, Online Mendelian Inheritance in Man; PCR, polymerase chain reaction; UTR, untranslated region; WGS, whole genome sequencing.

Competing interests

The authors declare that they have no competing interests.

Acknowledgements

We apologize to those whose work we could not cite due to reference limitations. We thank Peter Krawitz, Degui Zhi, Shane McCarthy, Jason O'Rawe, Edwin Nyambi, Reid Robison and three anonymous reviewers for helpful comments and suggestions for the manuscript. GJL is supported by funds from the Stanley Institute for Cognitive Genomics at Cold Spring Harbor Laboratory. KW is supported by NIH grant number HG006465.

Published: 26 July 2012

References

1. Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, Dewar K, Doyle M, FitzHugh W, Funke R, Gage D, Harris K, Heaford A, Howland J, Kann L, Lehoczky J, LeVine R, McEwan P, McKernan K, Meldrim J, Mesirov JP, Miranda C, Morris W, Naylor J, Raymond C, Rosetti M, Santos R, Sheridan A, Sougnez C, *et al*.: Initial sequencing and analysis of the human genome. *Nature* 2001, **409:**860-921.
2. Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, Smith HO, Yandell M, Evans CA, Holt RA, Gocayne JD, Amanatides P, Ballew RM, Huson DH, Wortman JR, Zhang Q, Kodira CD, Zheng XH, Chen L, Skupski M, Subramanian G, Thomas PD, Zhang J, Gabor Miklos GL, Nelson C, Broder S, Clark AG, Nadeau J, McKusick VA, Zinder N, *et al*.: The sequence of the human genome. *Science* 2001, **291:**1304-1351.
3. Venter JC: Genome-sequencing anniversary. The human genome at 10: successes and challenges. *Science* 2011, **331:**546-547.
4. Lander ES: Initial impact of the sequencing of the human genome. *Nature* 2011, **470:**187-197.
5. Hodges E, Xuan Z, Balija V, Kramer M, Molla MN, Smith SW, Middle CM, Rodesch MJ, Albert TJ, Hannon GJ, McCombie WR: Genome-wide in situ exon capture for selective resequencing. *Nat Genet* 2007, **39:**1522-1527.
6. Albert TJ, Molla MN, Muzny DM, Nazareth L, Wheeler D, Song X, Richmond TA, Middle CM, Rodesch MJ, Packard CJ, Weinstock GM, Gibbs RA: Direct selection of human genomic loci by microarray hybridization. *Nat Methods* 2007, **4:**903-905.
7. Ng PC, Levy S, Huang J, Stockwell TB, Walenz BP, Li K, Axelrod N, Busam DA, Strausberg RL, Venter JC: Genetic variation in an individual human exome. *PLoS Genet* 2008, **4:**e1000160.
8. Ng SB, Turner EH, Robertson PD, Flygare SD, Bigham AW, Lee C, Shaffer T, Wong M, Bhattacharjee A, Eichler EE, Bamshad M, Nickerson DA, Shendure J: Targeted capture and massively parallel sequencing of 12 human exomes. *Nature* 2009, **461:**272-276.
9. Choi M, Scholl UI, Ji W, Liu T, Tikhonova IR, Zumbo P, Nayir A, Bakkaloglu A, Ozen S, Sanjad S: Genetic diagnosis by whole exome capture and massively parallel DNA sequencing. *Proc Natl Acad Sci U S A* 2009, **106:**19096-19101.
10. Ng SB, Buckingham KJ, Lee C, Bigham AW, Tabor HK, Dent KM, Huff CD, Shannon PT, Jabs EW, Nickerson DA, Shendure J, Bamshad MJ: Exome sequencing identifies the cause of a mendelian disorder. *Nat Genet* 2010, **42:**30-35.
11. Ng SB, Bigham AW, Buckingham KJ, Hannibal MC, McMillin MJ, Gildersleeve HI, Beck AE, Tabor HK, Cooper GM, Mefford HC, Lee C, Turner EH, Smith JD, Rieder MJ, Yoshiura K, Matsumoto N, Ohta T, Niikawa N, Nickerson DA, Bamshad MJ, Shendure J: Exome sequencing identifies MLL2 mutations as a cause of Kabuki syndrome. *Nat Genet* 2010, **42:**790-793.
12. Roach JC, Glusman G, Smit AF, Huff CD, Hubley R, Shannon PT, Rowen L, Pant KP, Goodman N, Bamshad M, Shendure J, Drmanac R, Jorde LB, Hood L, Galas DJ: Analysis of genetic inheritance in a family quartet by whole-genome sequencing. *Science* 2010, **328:**636-639.
13. Gilissen C, Hoischen A, Brunner HG, Veltman JA: Disease gene identification strategies for exome sequencing. *Eur J Hum Genet* 2012, **20:**490-4907.
14. Biesecker LG, Shianna KV, Mullikin JC: Exome sequencing: the expert view. *Genome Biol* 2011, **12:**128.
15. Shendure J: Next-generation human genetics. *Genome Biol* 2011, **12:**408.
16. Oetting WS: Exome and genome analysis as a tool for disease identification and treatment: The 2011 human genome variation society scientific meeting. *Hum Mutat* 2012, **33:**586-590.
17. Ku CS, Cooper DN, Polychronakos C, Naidoo N, Wu M, Soong R: Exome sequencing: dual role as a discovery and diagnostic tool. *Ann Neurol* 2012, **71:**5-14.
18. Singleton AB: Exome sequencing: a transformative technology. *Lancet Neurol* 2011, **10:**942-946.
19. Cooper GM, Shendure J: Needles in stacks of needles: finding disease-causal variants in a wealth of genomic data. *Nat Rev Genet* 2011, **12:**628-640.
20. Christensen CM, Grossman JH, Hwang J: *The Innovator's Prescription: a Disruptive Solution for Health Care.* New York: McGraw-Hill; 2009.
21. Drmanac R: The advent of personal genome sequencing. *Genet Med* 2011, **13:**188-190.
22. Teer JK, Mullikin JC: Exome sequencing: the sweet spot before whole genomes. *Hum Mol Genet* 2010, **19:**R145-151.
23. Hedges DJ, Burges D, Powell E, Almonte C, Huang J, Young S, Boese B, Schmidt M, Pericak-Vance MA, Martin E, Zhang X, Harkins TT, Züchner S: Exome sequencing of a multigenerational human pedigree. *PLoS One* 2009, **4:**e8232.
24. Bonnefond A, Durand E, Sand O, De Graeve F, Gallina S, Busiah K, Lobbens S, Simon A, Bellanné-Chantelot C, Létourneau L, Scharfmann R, Delplanque J, Sladek R, Polak M, Vaxillaire M, Froguel P: Molecular diagnosis of neonatal diabetes mellitus using next-generation sequencing of the whole exome. *PLoS One* 2010, **5:**e13630.
25. Worthey EA, Mayer AN, Syverson GD, Helbling D, Bonacci BB, Decker B, Serpe JM, Dasu T, Tschannen MR, Veith RL, Basehore MJ, Broeckel U, Tomita-Mitchell A, Arca MJ, Casper JT, Margolis DA, Bick DP, Hessner MJ, Routes JM, Verbsky JW, Jacob HJ, Dimmock DP: Making a definitive diagnosis: successful clinical application of whole exome sequencing in a child with intractable inflammatory bowel disease. *Genet Med* 2011, **13:**255-262.
26. Rope AF, Wang K, Evjenth R, Xing J, Johnston JJ, Swensen JJ, Johnson WE, Moore B, Huff CD, Bird LM, Carey JC, Opitz JM, Stevens CA, Jiang T, Schank C, Fain HD, Robison R, Dalley B, Chin S, South ST, Pysher TJ, Jorde LB, Hakonarson H, Lillehaug JR, Biesecker LG, Yandell M, Arnesen T, Lyon GJ: Using VAAST to identify an X-linked disorder resulting in lethality in male infants due to N-terminal acetyltransferase deficiency. *Am J Hum Genet* 2011, **89:**28-43.
27. van Bon BW, Gilissen C, Grange DK, Hennekam RC, Kayserili H, Engels H, Reutter H, Ostergaard JR, Morava E, Tsiakas K, Isidor B, Le Merrer M, Eser M, Wieskamp N, de Vries P, Steehouwer M, Veltman JA, Robertson SP, Brunner HG, de Vries BB, Hoischen A: Cantu syndrome is caused by mutations in ABCC9. *Am J Hum Genet* 2012, **90:**1094-1101.
28. LaRusch J, Barmada MM, Solomon S, Whitcomb DC: Whole exome sequencing identifies multiple, complex etiologies in an idiopathic hereditary pancreatitis kindred. *JOP* 2012, 13:258-262.
29. Kirwan M, Walne AJ, Plagnol V, Velangi M, Ho A, Hossain U, Vulliamy T, Dokal I: Exome sequencing identifies autosomal-dominant SRP72 mutations associated with familial aplasia and myelodysplasia. *Am J Hum Genet* 2012, **90:**888-892.
30. Austin ED, Ma L, Leduc C, Berman Rosenzweig E, Borczuk A, Phillips JA, 3rd, Palomero T, Sumazin P, Kim HR, Talati MH, West J, Loyd JE, Chung WK: Whole exome sequencing to identify a novel gene (caveolin-1) associated with human pulmonary arterial hypertension. *Circ Cardiovasc Genet* 2012, **5:**336-343.
31. Sanders SS: Whole-exome sequencing: a powerful technique for identifying novel genes of complex disorders. *Clin Genet* 2011, **79:**132-133.
32. Tennessen JA, Bigham AW, O'Connor TD, Fu W, Kenny EE, Gravel S, McGee S, Do R, Liu X, Jun G, Kang HM, Jordan D, Leal SM, Gabriel S, Rieder MJ, Abecasis G, Altshuler D, Nickerson DA, Boerwinkle E, Sunyaev S, Bustamante CD, Bamshad MJ, Akey JM; Broad GO; Seattle GO; NHLBI Exome Sequencing Project: Evolution and functional impact of rare coding variation from deep sequencing of human exomes. *Science* 2012, **337:**64-69.
33. Boyden LM, Choi M, Choate KA, Nelson-Williams CJ, Farhi A, Toka HR, Tikhonova IR, Bjornson R, Mane SM, Colussi G, Lebel M, Gordon RD,

Semmekrot BA, Poujol A, Välimäki MJ, De Ferrari ME, Sanjad SA, Gutkin M, Karet FE, Tucci JR, Stockigt JR, Keppler-Noreuil KM, Porter CC, Anand SK, Whiteford ML, Davis ID, Dewar SB, Bettinelli A, Fadrowski JJ, Belsha CW: **Mutations in kelch-like 3 and cullin 3 cause hypertension and electrolyte abnormalities.** *Nature* 2012, **482:**98-102.

34. Iossifov I, Ronemus M, Levy D, Wang Z, Hakker I, Rosenbaum J, Yamrom B, Lee YH, Narzisi G, Leotta A, Kendall J, Grabowska E, Ma B, Marks S, Rodgers L, Stepansky A, Troge J, Andrews P, Bekritsky M, Pradhan K, Ghiban E, Kramer M, Parla J, Demeter R, Fulton LL, Fulton RS, Magrini VJ, Ye K, Darnell JC, Darnell RB, *et al.*: **De novo gene disruptions in children on the autistic spectrum.** *Neuron* 2012, **74:**285-299.
35. Cartault F, Munier P, Benko E, Desguerre I, Hanein S, Boddaert N, Bandiera S, Vellayoudom J, Krejbich-Trotot P, Bintner M, Hoarau JJ, Girard M, Génin E, de Lonlay P, Fourmaintraux A, Naville M, Rodriguez D, Feingold J, Renouil M, Munnich A, Westhof E, Fähling M, Lyonnet S, Henrion-Caude A: **Mutation in a primate-conserved retrotransposon reveals a noncoding RNA as a mediator of infantile encephalopathy.** *Proc Natl Acad Sci U S A* 2012, **109:**4980-4985.
36. **Full spectrum genetics.** *Nat Genet* 2012, **44:**1.
37. Need AC, Shashi V, Hitomi Y, Schoch K, Shianna KV, McDonald MT, Meisler MH, Goldstein DB: **Clinical application of exome sequencing in undiagnosed genetic conditions.** *J Med Genet* 2012, **49:**353-361.
38. Lyon GJ, Jiang T, Van Wijk R, Wang W, Bodily PM, Xing J, Tian L, Robison RJ, Clement M, Lin Y, Zhang P, Liu Y, Moore B, Glessner JT, Elia J, Reimherr F, van Solinge WW, Yandell M, Hakonarson H, Wang J, Johnson WE, Wei Z, Wang K: **Exome sequencing and unrelated findings in the context of complex disease research: ethical and clinical implications.** *Discov Med* 2011, **12:**41-55.
39. Meacham F, Boffelli D, Dhahbi J, Martin DI, Singer M, Pachter L: **Identification and correction of systematic error in high-throughput sequence data.** *BMC Bioinformatics* 2011, **12:**451.
40. Jacobs KB, Yeager M, Zhou W, Wacholder S, Wang Z, Rodriguez-Santiago B, Hutchinson A, Deng X, Liu C, Horner MJ, Cullen M, Epstein CG, Burdett L, Dean MC, Chatterjee N, Sampson J, Chung CC, Kovaks J, Gapstur SM, Stevens VL, Teras LT, Gaudet MM, Albanes D, Weinstein SJ, Virtamo J, Taylor PR, Freedman ND, Abnet CC, Goldstein AM, Hu N: **Detectable clonal mosaicism and its relationship to aging and cancer.** *Nat Genet* 2012, **44:**651-658.
41. Laurie CC, Laurie CA, Rice K, Doheny KF, Zelnick LR, McHugh CP, Ling H, Hetrick KN, Pugh EW, Amos C, Wei Q, Wang LE, Lee JE, Barnes KC, Hansel NN, Mathias R, Daley D, Beaty TH, Scott AF, Ruczinski I, Scharpf RB, Bierut LJ, Hartz SM, Landi MT, Freedman ND, Goldin LR, Ginsburg D, Li J, Desch KC, Strom SS, *et al.*: **Detectable clonal mosaicism from birth to old age and its relationship to cancer.** *Nat Genet* 2012, **44:**642-650.
42. Piotrowski A, Bruder CE, Andersson R, Diaz de Stahl T, Menzel U, Sandgren J, Poplawski A, von Tell D, Crasto C, Bogdan A, Bartoszewski R, Bebok Z, Krzyzanowski M, Jankowski Z, Partridge EC, Komorowski J, Dumanski JP: **Somatic mosaicism for copy number variation in differentiated human tissues.** *Hum Mutat* 2008, **29:**1118-1124.
43. Pathak J, Wang J, Kashyap S, Basford M, Li R, Masys DR, Chute CG: **Mapping clinical phenotype data elements to standardized metadata repositories and controlled terminologies: the eMERGE Network experience.** *J Am Med Inform Assoc* 2011, **18:**376-386.
44. **World Health Organization: International Classification of Diseases.** [http://www.who.int/classifications/icd/en/]
45. Robinson PN, Kohler S, Bauer S, Seelow D, Horn D, Mundlos S: **The Human Phenotype Ontology: a tool for annotating and analyzing human hereditary disease.** *Am J Hum Genet* 2008, **83:**610-615.
46. MacArthur DG, Balasubramanian S, Frankish A, Huang N, Morris J, Walter K, Jostins L, Habegger L, Pickrell JK, Montgomery SB, Albers CA, Zhang ZD, Conrad DF, Lunter G, Zheng H, Ayub Q, DePristo MA, Banks E, Hu M, Handsaker RE, Rosenfeld JA, Fromer M, Jin M, Mu XJ, Khurana E, Ye K, Kay M, Saunders GI, Suner MM, Hunt T, *et al.*: **A systematic survey of loss-of-function variants in human protein-coding genes.** *Science* 2012, **335:**823-828.
47. Ackermann RR, Rogers J, Cheverud JM: **Identifying the morphological signatures of hybridization in primate and human evolution.** *J Hum Evol* 2006, **51:**632-645.
48. Hannibal MC, Buckingham KJ, Ng SB, Ming JE, Beck AE, McMillin MJ, Gildersleeve HI, Bigham AW, Tabor HK, Mefford HC, Cook J, Yoshiura K, Matsumoto T, Matsumoto N, Miyake N, Tonoki H, Naritomi K, Kaname T, Nagai T, Ohashi H, Kurosawa K, Hou JW, Ohta T, Liang D, Sudo A, Morris CA, Banka S, Black GC, Clayton-Smith J, Nickerson DA, *et al.*: **Spectrum of MLL2 (ALR) mutations in 110 cases of Kabuki syndrome.** *Am J Med Genet A* 2011, **155A:**1511-1516.
49. Hamilton BA, Yu BD: **Modifier genes and the plasticity of genetic networks in mice.** *PLoS Genet* 2012, **8:**e1002644.
50. Jongmans MC, Verwiel ET, Heijdra Y, Vulliamy T, Kamping EJ, Hehir-Kwa JY, Bongers EM, Pfundt R, van Emst L, van Leeuwen FN, van Gassen KL, Geurts van Kessel A, Dokal I, Hoogerbrugge N, Ligtenberg MJ, Kuiper RP: **Revertant somatic mosaicism by mitotic recombination in dyskeratosis congenita.** *Am J Hum Genet* 2012, **90:**426-433.
51. Steinbusch C, van Roozendaal K, Tserpelis D, Smeets E, Kranenburg-de Koning T, de Waal K, Zweier C, Rauch A, Hennekam R, Blok M, Schrander-Stumpel C: **Somatic mosaicism in a mother of two children with Pitt-Hopkins syndrome.** *Clin Genet* 2012. doi: 10.1111/j.1399-0004.2012.01857.x.
52. Yamada M, Okura Y, Suzuki Y, Fukumura S, Miyazaki T, Ikeda H, Takezaki S, Kawamura N, Kobayashi I, Ariga T: **Somatic mosaicism in two unrelated patients with X-linked chronic granulomatous disease characterized by the presence of a small population of normal cells.** *Gene* 2012, **497:**110-115.
53. Lindhurst MJ, Sapp JC, Teer JK, Johnston JJ, Finn EM, Peters K, Turner J, Cannons JL, Bick D, Blakemore L, Blumhorst C, Brockmann K, Calder P, Cherman N, Deardorff MA, Everman DB, Golas G, Greenstein RM, Kato BM, Keppler-Noreuil KM, Kuznetsov SA, Miyamoto RT, Newman K, Ng D, O'Brien K, Rothenberg S, Schwartzentruber DJ, Singhal V, Tirabosco R, Upton J, *et al.*: **A mosaic activating mutation in AKT1 associated with the Proteus syndrome.** *N Engl J Med* 2011, **365:**611-619.
54. Choate KA, Lu Y, Zhou J, Choi M, Elias PM, Farhi A, Nelson-Williams C, Crumrine D, Williams ML, Nopper AJ, Bree A, Milstone LM, Lifton RP: **Mitotic recombination in patients with ichthyosis causes reversion of dominant mutations in KRT10.** *Science* 2010, **330:**94-97.
55. Coufal NG, Garcia-Perez JL, Peng GE, Marchetto MC, Muotri AR, Mu Y, Carson CT, Macia A, Moran JV, Gage FH: **Ataxia telangiectasia mutated (ATM) modulates long interspersed element-1 (L1) retrotransposition in human neural stem cells.** *Proc Natl Acad Sci U S A* 2011, **108:**20382-20387.
56. Veeramah KR, O'Brien JE, Meisler MH, Cheng X, Dib-Hajj SD, Waxman SG, Talwar D, Girirajan S, Eichler EE, Restifo LL, Erickson RP, Hammer MF: **De novo pathogenic SCN8A mutation identified by whole-genome sequencing of a family quartet affected by infantile epileptic encephalopathy and SUDEP.** *Am J Hum Genet* 2012, **90:**502-510.
57. Norton N, Li D, Rieder MJ, Siegfried JD, Rampersaud E, Zuchner S, Mangos S, Gonzalez-Quintana J, Wang L, McGee S, Reiser J, Martin E, Nickerson DA, Hershberger RE: **Genome-wide studies of copy number variation and exome sequencing identify rare variants in BAG3 as a cause of dilated cardiomyopathy.** *Am J Hum Genet* 2011, **88:**273-282.
58. Norton N, Robertson PD, Rieder MJ, Zuchner S, Rampersaud E, Martin E, Li D, Nickerson DA, Hershberger RE: **Evaluating pathogenicity of rare variants from dilated cardiomyopathy in the exome era.** *Circ Cardiovasc Genet*2012, **5:**167-174.
59. Eichers ER, Lewis RA, Katsanis N, Lupski JR: **Triallelic inheritance: a bridge between Mendelian and multifactorial traits.** *Ann Med* 2004, **36:**262-272.
60. Katsanis N, Ansley SJ, Badano JL, Eichers ER, Lewis RA, Hoskins BE, Scambler PJ, Davidson WS, Beales PL, Lupski JR: **Triallelic inheritance in Bardet-Biedl syndrome, a Mendelian recessive disorder.** *Science* 2001, **293:**2256-2259.
61. Davis EE, Katsanis N: **The ciliopathies: a transitional model into systems biology of human genetic disease.** *Curr Opin Genet Devel* 2012, **22:**290-303.
62. Abu-Safieh L, Al-Anazi S, Al-Abdi L, Hashem M, Alkuraya H, Alamr M, Sirelkhatim MO, Al-Hassnan Z, Alkuraya B, Mohamed JY, Al-Salem A, Alrashed M, Faqeih E, Softah A, Al-Hashem A, Wali S, Rahbeeni Z, Alsayed M, Khan AO, Al-Gazali L, Taschner PE, Al-Hazzaa S, Alkuraya FS: **In search of triallelism in Bardet-Biedl syndrome.** *Eur J Hum Genet* 2012, **20:**420-427.
63. Smaoui N, Chaabouni M, Sergeev YV, Kallel H, Li S, Mahfoudh N, Maazoul F, Kammoun H, Gandoura N, Bouaziz A, Nouiri E, M'Rad R, Chaabouni H, Hejtmancik JF: **Screening of the eight BBS genes in Tunisian families: no evidence of triallelism.** *Invest Ophthalmol Vis Sci* 2006, **47:**3487-3495.
64. Laurier V, Stoetzel C, Muller J, Thibault C, Corbani S, Jalkh N, Salem N, Chouery E, Poch O, Licaire S, Licaire S, Danse JM, Amati-Bonneau P, Bonneau D, Mégarbané A, Mandel JL, Dollfus H: **Pitfalls of homozygosity mapping: an extended consanguineous Bardet-Biedl syndrome family with two mutant genes (BBS2, BBS10), three mutations, but no triallelism.** *Eur J Hum Genet* 2006, **14:**1195-1203.
65. Nakane T, Biesecker LG: **No evidence for triallelic inheritance of MKKS/BBS loci in Amish Mckusick-Kaufman syndrome.** *Am J Med Genet A* 2005, **138:**32-34.

66. Mykytyn K, Nishimura DY, Searby CC, Beck G, Bugge K, Haines HL, Cornier AS, Cox GF, Fulton AB, Carmi R, Iannaccone A, Jacobson SG, Weleber RG, Wright AF, Riise R, Hennekam RC, Lüleci G, Berker-Karauzum S, Biesecker LG, Stone EM, Sheffield VC: **Evaluation of complex inheritance involving the most common Bardet-Biedl syndrome locus (BBS1).** *Am J Med Genet A* 2003, **72:**429-437.
67. McPherson JD: **Next-generation gap.** *Nat Methods* **2009, 6:**S2-5.
68. Schuster SC, Miller W, Ratan A, Tomsho LP, Giardine B, Kasson LR, Harris RS, Petersen DC, Zhao F, Qi J, Alkan C, Kidd JM, Sun Y, Drautz DI, Bouffard P, Muzny DM, Reid JG, Nazareth LV, Wang Q, Burhans R, Riemer C, Wittekindt NE, Moorjani P, Tindall EA, Danko CG, Teo WS, Buboltz AM, Zhang Z, Ma Q, Oosthuysen A, *et al.*: **Complete Khoisan and Bantu genomes from southern Africa.** *Nature* 2010, **463:**943-947.
69. Trapnell C, Salzberg SL: **How to map billions of short reads onto genomes.** *Nat Biotechnol* 2009, **27:**455-457.
70. Dalca AV, Brudno M: **Genome variation discovery with high-throughput sequencing data.** *Brief Bioinform* 2010, **11:**3-14.
71. Teer JK, Green ED, Mullikin JC, Biesecker LG: **VarSifter: Visualizing and analyzing exome-scale sequence variation data on a desktop computer.** *Bioinformatics* 2012, **28:**599-600.
72. Asmann YW, Middha S, Hossain A, Baheti S, Li Y, Chai HS, Sun Z, Duffy PH, Hadad AA, Nair A, Liu X, Zhang Y, Klee EW, Kalari KR, Kocher JP: **TREAT: a bioinformatics tool for variant annotations and visualizations in targeted and exome sequencing data.** *Bioinformatics* 2012, **28:**277-278.
73. Stitziel NO, Kiezun A, Sunyaev S: **Computational and statistical approaches to analyzing variants identified by exome sequencing.** *Genome Biol* 2011, **12:**227.
74. Pippucci T, Benelli M, Magi A, Martelli PL, Magini P, Torricelli F, Casadio R, Seri M, Romeo G: **EX-HOM (EXome HOMozygosity): a proof of principle.** *Hum Hered* 2011, **72:**45-53.
75. Yandell M, Huff C, Hu H, Singleton M, Moore B, Xing J, Jorde LB, Reese MG: **A probabilistic disease-gene finder for personal genomes.** *Genome Res* 2011, **21:**1529-1542.
76. Ionita-Laza I, Makarov V, Yoon S, Raby B, Buxbaum J, Nicolae DL, Lin X: **Finding disease variants in Mendelian disorders by using sequence data: methods and applications.** *Am J Hum Genet* 2011, **89:**701-712.
77. Ji HP: **Improving bioinformatic pipelines for exome variant calling.** *Genome Med* 2012, **4:**7.
78. Li MX, Gui HS, Kwan JS, Bao SY, Sham PC: **A comprehensive framework for prioritizing variants in exome sequencing studies of Mendelian diseases.** *Nucleic Acids Res* 2012, **40:**e53.
79. Lopes MC, Joyce C, Ritchie GR, John SL, Cunningham F, Asimit J, Zeggini E: **A combined functional annotation score for non-synonymous variants.** *Hum Hered* 2012, **73:**47-51.
80. Hu J, Ng PC: **Predicting the effects of frameshifting indels.** *Genome Biol* 2012, **13:**R9.
81. Ng PC, Henikoff S: **Predicting the effects of amino acid substitutions on protein function.** *Annu Rev Genomics Hum Genet* 2006, **7:**61-80.
82. Kumar P, Henikoff S, Ng PC: **Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm.** *Nat Protoc* 2009, **4:**1073-1081.
83. Jaffe A, Wojcik G, Chu A, Golozar A, Maroo A, Duggal P, Klein AP: **Identification of functional genetic variation in exome sequence analysis.** *BMC Proc* 2011, **5(Suppl 9):**S13.
84. Wei P, Liu X, Fu YX: **Incorporating predicted functions of nonsynonymous variants into gene-based analysis of exome sequencing data: a comparative study.** *BMC Proc* 2011, **5(Suppl 9):**S20.
85. Sethumadhavan R, Doss CG, Rajasekaran R: **In silico searching for disease-associated functional DNA variants.** *Methods Mol Biol* 2011, **760:**239-250.
86. Zou M, Baitei EY, Alzahrani AS, Parhar RS, Al-Mohanna FA, Meyer BF, Shi Y: **Mutation prediction by PolyPhen or functional assay, a detailed comparison of CYP27B1 missense mutations.** *Endocrine* 2011, **40:**14-20.
87. Hicks S, Wheeler DA, Plon SE, Kimmel M: **Prediction of missense mutation functionality depends on both the algorithm and sequence alignment employed.** *Hum Mutat* 2011, **32:**661-668.
88. Thusberg J, Olatubosun A, Vihinen M: **Performance of mutation pathogenicity prediction methods on missense variants.** *Hum Mutat* 2011, **32:**358-368.
89. Flanagan SE, Patch AM, Ellard S: **Using SIFT and PolyPhen to predict loss-of-function and gain-of-function mutations.** *Genet Test Mol Biomarkers* 2010, **14:**533-537.
90. Wei Q, Wang L, Wang Q, Kruger WD, Dunbrack RL Jr: **Testing computational prediction of missense mutation phenotypes: functional characterization of 204 mutations of human cystathionine beta synthase.** *Proteins* 2010, **78:**2058-2074.
91. Liu X, Jian X, Boerwinkle E: **dbNSFP: a lightweight database of human nonsynonymous SNPs and their functional predictions.** *Hum Mutat* 2011, **32:**894-899.
92. Schwarz JM, Rodelsperger C, Schuelke M, Seelow D: **MutationTaster evaluates disease-causing potential of sequence alterations.** *Nat Methods* 2010, **7:**575-576.
93. Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P, Kondrashov AS, Sunyaev SR: **A method and server for predicting damaging missense mutations.** *Nat Methods* 2010, **7:**248-249.
94. Chun S, Fay JC: **Identification of deleterious mutations within three human genomes.** *Genome Res* 2009, **19:**1553-1561.
95. Wang J, Zhang J, Li K, Zhao W, Cui Q: **SpliceDisease database: linking RNA splicing and disease.** *Nucleic Acids Res* 2012, **40:**D1055-1059.
96. Sauna ZE, Kimchi-Sarfaty C: **Understanding the contribution of synonymous mutations to human disease.** *Nat Rev Genet* 2011, **12:**683-691.
97. Esteller M: **Non-coding RNAs in human disease.** *Nat Rev Genet* 2011, **12:**861-874.
98. Klopocki E, Mundlos S: **Copy-number variations, noncoding sequences, and human phenotypes.** *Annu Rev Genomics Hum Genet* 2011, **12:**53-72.
99. Zhong S, Sharp DS, Grove JS, Bruce C, Yano K, Curb JD, Tall AR: **Increased coronary heart disease in Japanese-American men with mutation in the cholesteryl ester transfer protein gene despite increased HDL levels.** *J Clin Invest* 1996, **97:**2917-2923.
100. Myers RM, Stamatoyannopoulos J, Snyder M, Dunham I, Hardison RC, Bernstein BE, Gingeras TR, Kent WJ, Birney E, Wold B, Crawford GE: **A user's guide to the encyclopedia of DNA elements (ENCODE).** *PLoS Biol* 2011, **9:**e1001046.
101. Al Olama AA, Kote-Jarai Z, Giles GG, Guy M, Morrison J, Severi G, Leongamornlert DA, Tymrakiewicz M, Jhavar S, Saunders E, Hopper JL, Southey MC, Muir KR, English DR, Dearnaley DP, Ardern-Jones AT, Hall AL, O'Brien LT, Wilkinson RA, Sawyer E, Lophatananon A; UK Genetic Prostate Cancer Study Collaborators/British Association of Urological Surgeons' Section of Oncology; UK Prostate testing for cancer and Treatment study (ProtecT Study) Collaborators, Horwich A, Huddart RA, Khoo VS, Parker CC, Woodhouse CJ, Thompson A, Christmas T, Ogden C, Cooper C, Donovan JL, Hamdy FC, Neal DE, Eeles RA, Easton DF: **Multiple loci on 8q24 associated with prostate cancer susceptibility.** *Nat Genet* 2009, **41:**1058-1060.
102. Enciso-Mora V, Broderick P, Ma Y, Jarrett RF, Hjalgrim H, Hemminki K, van den Berg A, Olver B, Lloyd A, Dobbins SE, Lightfoot T, van Leeuwen FE, Försti A, Diepstra A, Broeks A, Vijayakrishnan J, Shield L, Lake A, Montgomery D, Roman E, Engert A, von Strandmann EP, Reiners KS, Nolte IM, Smedby KE, Adami HO, Russell NS, Glimelius B, Hamilton-Dutoit S, de Bruin M, *et al.*: **A genome-wide association study of Hodgkin's lymphoma identifies new susceptibility loci at 2p16.1 (REL), 8q24.21 and 10p14 (GATA3).** *Nat Genet* 2010, **42:**1126-1130.
103. Kiemeney LA, Thorlacius S, Sulem P, Geller F, Aben KK, Stacey SN, Gudmundsson J, Jakobsdottir M, Bergthorsson JT, Sigurdsson A, Blondal T, Witjes JA, Vermeulen SH, Hulsbergen-van de Kaa CA, Swinkels DW, Ploeg M, Cornel EB, Vergunst H, Thorgeirsson TE, Gudbjartsson D, Gudjonsson SA, Thorleifsson G, Kristinsson KT, Mouy M, Snorradottir S, Placidi D, Campagna M, Arici C, Koppova K, Gurzau E, *et al.*: **Sequence variant on 8q24 confers susceptibility to urinary bladder cancer.** *Nat Genet* 2008, **40:**1307-1312.
104. Zanke BW, Greenwood CM, Rangrej J, Kustra R, Tenesa A, Farrington SM, Prendergast J, Olschwang S, Chiang T, Crowdy E, Ferretti V, Laflamme P, Sundararajan S, Roumy S, Olivier JF, Robidoux F, Sladek R, Montpetit A, Campbell P, Bezieau S, O'Shea AM, Zogopoulos G, Cotterchio M, Newcomb P, McLaughlin J, Younghusband B, Green R, Green J, Porteous ME, Campbell H, *et al.*: **Genome-wide association scan identifies a colorectal cancer susceptibility locus on chromosome 8q24.** *Nat Genet* 2007, **39:**989-994.
105. **The International HapMap Project.** *Nature* 2003, **426:**789-796.
106. International HapMap Consortium, Frazer KA, Ballinger DG, Cox DR, Hinds DA, Stuve LL, Gibbs RA, Belmont JW, Boudreau A, Hardenbol P, Leal SM, Pasternak S, Wheeler DA, Willis TD, Yu F, Yang H, Zeng C, Gao Y, Hu H, Hu W, Li C, Lin W, Liu S, Pan H, Tang X, Wang J, Wang W, Yu J, Zhang B, Zhang Q, Zhao H, *et al.*: **A second generation human haplotype map of over 3.1 million SNPs.** *Nature* 2007, **449:**851-861.
107. International HapMap 3 Consortium, Altshuler DM, Gibbs RA, Peltonen L,

Altshuler DM, Gibbs RA, Peltonen L, Dermitzakis E, Schaffner SF, Yu F, Peltonen L, Dermitzakis E, Bonnen PE, Altshuler DM, Gibbs RA, de Bakker PI, Deloukas P, Gabriel SB, Gwilliam R, Hunt S, Inouye M, Jia X, Palotie A, Parkin M, Whittaker P, Yu F, Chang K, Hawes A, Lewis LR, Ren Y, Wheeler D, *et al.*: Integrating common and rare genetic variation in diverse human populations. *Nature* 2010, **467**:52-58.

108. A map of human genome variation from population-scale sequencing. *Nature* 2010, **467**:1061-1073.
109. ENCODE Project Consortium, Birney E, Stamatoyannopoulos JA, Dutta A, Guigó R, Gingeras TR, Margulies EH, Weng Z, Snyder M, Dermitzakis ET, Thurman RE, Kuehn MS, Taylor CM, Neph S, Koch CM, Asthana S, Malhotra A, Adzhubei I, Greenbaum JA, Andrews RM, Flicek P, Boyle PJ, Cao H, Carter NP, Clelland GK, Davis S, Day N, Dhami P, Dillon SC, Dorschner MO, Fiegler H, *et al.*: Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature* 2007, **447**:799-816.
110. Nielsen MA: *Reinventing Discovery: The New Era of Networked Science.* Princeton, NJ: Princeton University Press; 2012.
111. Wicks P, Vaughan TE, Massagli MP, Heywood J: Accelerated clinical discovery using self-reported patient data collected online and a patient-matching algorithm. *Nat Biotechnol* 2011, **29**:411-414.
112. Brownstein CA, Brownstein JS, Williams DS, 3rd, Wicks P, Heywood JA: The power of social networking in medicine. *Nat Biotechnol* 2009**, 27**:888-890.
113. Derry JM, Mangravite LM, Suver C, Furia MD, Henderson D, Schildwachter X, Bot B, Izant J, Sieberts SK, Kellen MR, Friend SH: Developing predictive molecular maps of human disease through community-based modeling. *Nat Genet* 2012, **44**:127-130.
114. Patrinos GP, Smith TD, Howard H, Al-Mulla F, Chouchane L, Hadjisavvas A, Hamed SA, Li XT, Marafie M, Ramesar RS, Ramos FJ, de Ravel T, El-Ruby MO, Shrestha TR, Sobrido MJ, Tadmouri G, Witsch-Baumgartner M, Zilfalil BA, Auerbach AD, Carpenter K, Cutting GR, Dung VC, Grody W, Hasler J, Jorde L, Kaput J, Macek M, Matsubara Y, Padilla C, Robinson H, *et al.*: Human variome project country nodes: Documenting genetic information within a country. *Hum Mutat* 2012. doi: 10.1002/humu.22147.
115. Smith TD, Robinson HM, Cotton RG: The Human Variome Project Beijing meeting. *J Med Genet* 2012, **49**:284-289.
116. Mandl KD, Kohane IS: Escaping the EHR trap - the future of health IT. *N Engl J Med* 2012, **366**:2240-2242.
117. Complete Genomics: Cancer Data Set [http://www.completegenomics.com/public-data/cancer-data/]
118. 1000 Cancer Genome Project [http://cancergenome.mooreworks.net/cms/san-antonio.aspx]
119. Jung's Biology Blog [http://jchoigt.wordpress.com/2012/07/02/a-first-look-at-my-exome-variants-from-23andme/]
120. illumina [http://investor.illumina.com/phoenix.zhtml?c=121127&p=irol-newsArticle&ID=1401402&highlight]
121. Wang K: Gene-function wiki would let biologists pool worldwide resources. *Nature* 2006, **439**:534.
122. Hoffmann R: A wiki for the life sciences where authorship matters. *Nat Genet* 2008, **40**:1047-1051.
123. Huss JW, 3rd, Orozco C, Goodale J, Wu C, Batalov S, Vickers TJ, Valafar F, Su AI: A gene wiki for community annotation of gene function. *PLoS Biol* 2008, **6**:e175.
124. GeneTalk [http://www.gene-talk.de]
125. Lupski JR, Belmont JW, Boerwinkle E, Gibbs RA: Clan genomics and the complex architecture of human disease. *Cell* 2011, **147**:32-43.
126. Bittles AH, Black ML: Evolution in health and medicine Sackler colloquium: Consanguinity, human evolution, and complex diseases. *Proc Natl Acad Sci U S A* 2010, **107(Suppl 1)**:1779-1786.
127. Kaye J, Curren L, Anderson N, Edwards K, Fullerton SM, Kanellopoulou N, Lund D, Macarthur DG, Mascalzoni D, Shepherd J, Taylor PL, Terry SF, Winter SF: From patients to partners: participant-centric initiatives in biomedical research. *Nat Rev Genet* 2012, **13**:371-376.
128. Kenny EE, Gusev A, Riegel K, Lutjohann D, Lowe JK, Salit J, Maller JB, Stoffel M, Daly MJ, Altshuler DM, Friedman JM, Breslow JL, Pe'er I, Sehayek E: Systematic haplotype analysis resolves a complex plasma plant sterol locus on the Micronesian Island of Kosrae. *Proc Natl Acad Sci U S A* 2009, **106**:13886-13891.
129. Yang S, Wang K, Gregory B, Berrettini W, Wang LS, Hakonarson H, Bucan M: Genomic landscape of a three-generation pedigree segregating affective disorder. *PLoS One* 2009, **4**:e4474.
130. Topol EJ: *The Creative Destruction Of Medicine: How The Digital Revolution Will Create Better Health Care.* New York: Basic Books; 2012.
131. Lee H, Schatz MC: Genomic Dark Matter: The reliability of short read mapping illustrated by the Genome Mappability Score. *Bioinformatics* 2012 [Epub ahead of print].
132. Peters BA, Kermani BG, Sparks AB, Alferov O, Hong P, Alexeev A, Jiang Y, Dahl F, Tang YT, Haas J, Robasky K, Zaranek AW, Lee JH, Ball MP, Peterson JE, Perazich H, Yeung G, Liu J, Chen L, Kennemer MI, Pothuraju K, Konvicka K, Tsoupko-Sitnikov M, Pant KP, Ebert JC, Nilsen GB, Baccash J, Halpern AL, Church GM, Drmanac R: Accurate whole-genome sequencing and haplotyping from 10 to 20 human cells. *Nature* 2012, **487**:190-195.
133. Metzker ML: Sequencing technologies - the next generation. *Nat Rev Genet* 2010, **11**:31-46.
134. Werner T: Next generation sequencing in functional genomics. *Brief Bioinform* 2010, **11**:499-511.
135. Morozova O, Marra MA: Applications of next-generation sequencing technologies in functional genomics. *Genomics* 2008, **92**:255-264.
136. Cirulli ET, Goldstein DB: Uncovering the roles of rare variants in common disease through whole-genome sequencing. *Nat Rev Genet* 2010, **11**:415-425.
137. Durbin RM, Abecasis GR, Altshuler DL, Auton A, Brooks LD, Gibbs RA, Hurles ME, McVean GA: A map of human genome variation from population-scale sequencing. *Nature* 2010, **467**:1061-1073.
138. Pallen MJ, Loman NJ, Penn CW: High-throughput sequencing and clinical microbiology: progress, opportunities and challenges. *Curr Opin Microbiol* 2010, **13**:625-631.
139. Raymond FL, Whittaker J, Jenkins L, Lench N, Chitty LS: Molecular prenatal diagnosis: the impact of modern technologies. *Prenat Diagn* 2010, **30**:674-681.
140. Wanunu M: Nanopores: A journey towards DNA sequencing. *Phys Life Rev* 2012, **9**:125-158.
141. Lyon GJ: Personalized medicine: Bring clinical standards to human-genetics research. *Nature* 2012, **482**:300-301.
142. Genomes Unzipped. Guest post: Time to bring human genome sequencing into the clinic. [http://www.genomesunzipped.org/2012/02/guest-post-time-to-bring-human-genome-sequencing-into-the-clinic.php]
143. Lyon GJ: There is nothing "incidental" about unrelated findings. *Per Med* 2012, **9**:163-166.
144. Mroch AR, Flanagan JD, Stein QP: Solving the puzzle: case examples of array comparative genomic hybridization as a tool to end the diagnostic odyssey. *Curr Probl Pediatr Adolesc Health Care* 2012, **42**:74-78.
145. Lam HY, Clark MJ, Chen R, Natsoulis G, O'Huallachain M, Dewey FE, Habegger L, Ashley EA, Gerstein MB, Butte AJ, Ji HP, Snyder M: Performance comparison of whole-genome sequencing platforms. *Nat Biotechnol* 2012, **30**:78-82.
146. Li H: Exploring single-sample SNP and INDEL calling with whole-genome de novo assembly. *Bioinformatics* 2012, **28**:1838-1844.

doi:10.1186/gm359
Cite this article as: Lyon GJ, Wang K: Identifying disease mutations in genomic medicine settings: current challenges and how to accelerate progress. *Genome Medicine* 2012, **4**:58.