Diagnostic Applications of High-Throughput DNA Sequencing

Scott D. Boyd

Department of Pathology, Stanford University, Stanford, California 94305;
email: sboyd1@stanford.edu

Annu. Rev. Pathol. Mech. Dis. 2013. 8:381–410

First published online as a Review in Advance on November 1, 2012

The *Annual Review of Pathology: Mechanisms of Disease* is online at pathol.annualreviews.org

This article's doi:
10.1146/annurev-pathol-020712-164026

Keywords

genome, exome, Mendelian, cancer, transplant, prenatal

Abstract

Advances in DNA sequencing technology have allowed comprehensive investigation of the genetics of human beings and human diseases. Insights from sequencing the genomes, exomes, or transcriptomes of healthy and diseased cells in patients are already enabling improved diagnostic classification, prognostication, and therapy selection for many diseases. Understanding the data obtained using new high-throughput DNA sequencing methods, choices made in sequencing strategies, and common challenges in data analysis and genotype-phenotype correlation is essential if pathologists, geneticists, and clinicians are to interpret the growing scientific literature in this area. This review highlights some of the major results and discoveries stemming from high-throughput DNA sequencing research in our understanding of Mendelian genetic disorders, hematologic cancer biology, infectious diseases, the immune system, transplant biology, and prenatal diagnostics. Transition of new DNA sequencing methodologies to the clinical laboratory is under way and is likely to have a major impact on all areas of medicine.

INTRODUCTION

We are living at a remarkable time, when the genetic information accumulated in the history of life on Earth has rapidly become available for study, classification, and correlation with the biological activities of cells, tissues, and whole organisms. Recent advances in DNA sequencing technology mean that pathologists, geneticists, and members of other medical disciplines are witnessing a flood of new genetic data that are providing insights into disease predisposition, pathogenesis, prognostication, and therapeutic strategies. In addition to readily interpreted results, this torrent of high-throughput sequencing (HTS) data contains other findings whose meaning is not immediately apparent, and will not be useful until correlated with additional research. It also seems likely that large numbers of sequence variants will resist interpretation, and remain a kind of "background noise" of no definite significance in each patient's genomic record. Although high-throughput DNA sequencing may help predict which diseases a particular healthy individual will eventually develop, genomic information will probably find its most immediate application in the context of patients with signs and symptoms that cause them to seek medical attention. In this setting, the interpretation of genetic results can be informed by a patient's history, a physical examination, conventional laboratory tests, histologic evaluation, and imaging studies that measure the manifestations of gene function in cell and organ biology.

This review begins with a brief description of high-throughput DNA sequencing technologies, choices of templates for sequencing, and common questions that arise in analysis and interpretation of HTS data. The remaining sections survey recent diagnostic advances fueled by HTS and consider their impact on our understanding of Mendelian genetic disorders, somatic genetic changes in hematologic cancers, detection and classification of infectious organisms, monitoring of the immune system, organ transplantation, and prenatal testing. Several recent review articles have addressed many of these areas with a different emphasis or greater depth and are recommended for further reading (1–5). Although the logistical and educational challenges posed by the inclusion of large-scale genetic data into the practice of diagnostic medicine are formidable, the potential benefits to patients are great, and the magnitude of the impact of these new data on our understanding of human diseases may be comparable to the impact that the techniques of light microscopy or microbiology had when they were first employed.

DNA SEQUENCING METHODS

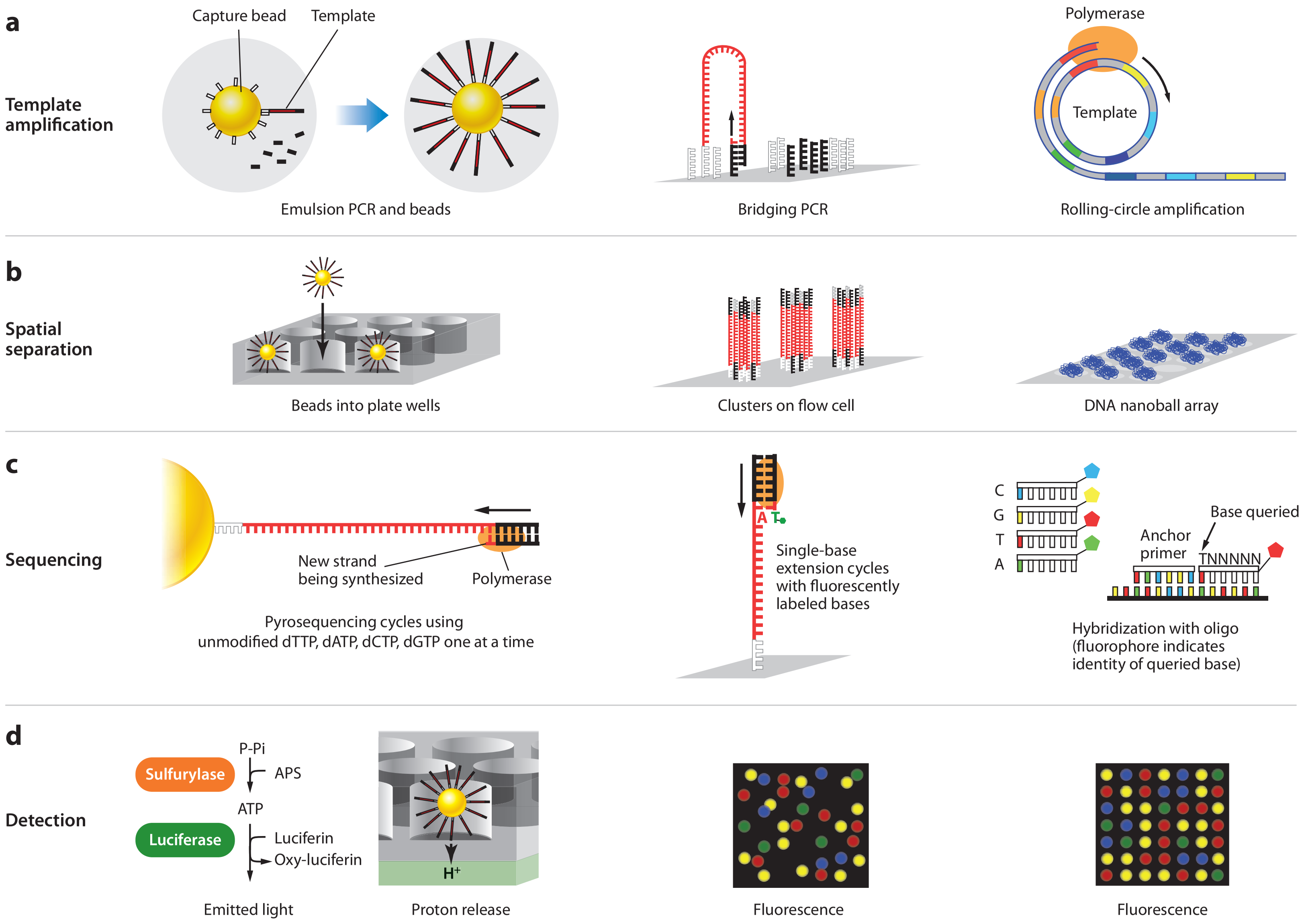
Soon after the historic publication of the Sanger-sequenced Human Genome Project's draft results in 2001 and the "finished" euchromatin sequence in 2004, several new DNA sequencing technologies were described in the literature; these technologies have been commercialized and commodified at an astonishing pace (**Figure 1**) (6–10). The details of these methods have been reviewed elsewhere (1), but most use a solid flow-cell surface or beads in an emulsion of aqueous droplets in oil to spatially segregate individual DNA template molecules so that they can be amplified in situ, then sequenced with simultaneous data acquisition in parallel from millions of templates via optical or electronic detection. Determination of the template DNA sequence is performed (*a*) by synthesis of the complementary strand in a manner that permits identification of the nucleotides being added as synthesis proceeds (i.e., sequencing by synthesis) or (*b*) by repeated cycles of hybridization with oligonucleotide probes that identify the base at a particular position in the template molecule (**Figure 1**). The Illumina HiSeq^{TM} systems, which can currently determine billions of paired 100-bp reads in a single run, have become the dominant platform for many applications in genome centers; they use in situ polymerase chain reaction (PCR) on a flow-cell surface for template amplification, followed by sequencing by synthesis using reversibly 3′-blocked fluorescently labeled nucleotides (11). The Roche 454 platform uses emulsion PCR for template amplification and then employs pyrosequencing (detection of liberated pyrophosphate upon nucleotide incorporation) to generate approximately one million ∼450-bp reads. This platform is popular for applications wherein long read lengths are critical, such as sequencing DNA from mixed microbial populations or immunoglobulin gene rearrangements (8, 12, 13). Various other systems have joined the commercial competition, including (*a*) the Ion Torrent platform, which begins with an emulsion PCR template preparation step similar to that of the Roche 454 technology but uses solid-state electronic detection of protons to detect base incorporation; (*b*) several methods based on true single-molecule sequencing, such as those developed by Helicos BioSciences and Pacific Biosciences; and (*c*) Applied Biosystems's SOLiD^{TM} platform and Complete Genomics's platform, which use hybridization and ligation-based sequencing (14–18). Complete Genomics employs a unique strategy in not selling instruments and kits but rather focusing solely on performing human genome sequencing at a central laboratory facility as a service for academic and commercial clients.

It remains to be seen whether some of these technologies will quickly become outmoded or whether each will occupy a distinct ecological niche for different sequencing applications. The competition among manufacturers has led to remarkable increases in sequencing capacity and decreases in sequencing costs for researchers. A human genome sequence at 30× coverage (in which each base, on average, is present in 30 distinct sequence reads) can now be obtained for less than US$5,000. Comparisons between this cost with that typically charged for other laboratory tests in CLIA (Clinical Laboratory Improvement Amendments)-certified laboratories are somewhat misleading because obtaining the data from a 30× coverage human genome is only the starting point for an involved process of data analysis and further validation, particularly if clinically important decisions will be based on the sequence data. Nonetheless, the current cost of sequencing is low enough that the primary challenges in incorporating genome-scale sequencing for patient care are now those of data reliability and interpretability, as well as regulatory, reimbursement, and intellectual-property considerations. Each DNA sequencing technology has characteristic error modes and rates, and all of the new methods have a higher per-base error rate within a single sequence read than modern Sanger sequencing does; however, many HTS applications seek to overcome this obstacle by obtaining many independent sequence reads for each base to be determined. In the immediate future, interpretation of the complex data derived from HTS experiments will benefit from time-honored strategies in biological research: (*a*) obtaining replicate data sets (preferably with biological replicates) from multiple individuals with a given clinical disorder, as well as repeat sampling from a given individual; (*b*) performing technical replicates that repeat the same sequencing library preparation and sequencing protocol on a sample; and finally (*c*) increasing sequencing depth (19). Such replication strategies may be easiest to implement in discovery-phase research activities when disease-related mutations are being sought, but they should also be considered for clinical laboratory protocols generating data used to guide clinical decision making. For example, generating two independent libraries from a patient's genomic DNA and sequencing could help to detect false positive or false negative results arising from experimental variation in library preparation.

GENETIC MATERIAL FOR DIAGNOSIS

Whole-Genome Sequencing

The choice of which nucleic acids are sequenced and which upstream sample-preparation methods are used determines what type of genetic or epigenetic data is obtained from a patient sample. Genomic DNA is particularly stable and amenable to sequencing. Most


a
Template amplification
Capture bead
Template
Emulsion PCR and beads
Bridging PCR
Polymerase
Template
Rolling-circle amplification
b
Spatial separation
Beads into plate wells
Clusters on flow cell
DNA nanoball array
c
Sequencing
New strand being synthesized
Polymerase
Pyrosequencing cycles using unmodified dTTP, dATP, dCTP, dGTP one at a time
A
T
Single-base extension cycles with fluorescently labeled bases
C
G
T
A
Anchor primer
Base queried
TNNNNN
Hybridization with oligo (fluorophore indicates identity of queried base)
d
Detection
Sulfurylase
P-Pi
APS
ATP
Luciferase
Luciferin
Oxy-luciferin
Emitted light
H+
Proton release
Fluorescence
Fluorescence

HTS research has used genomic DNA derived from fresh or frozen tissue or cell samples. In contrast, formalin fixation/paraffin embedding is still the mainstay of sample preparation for histologic examination and immunostaining for clinical diagnosis. To enable broad application of HTS to patient tissue specimens, particularly in cancer diagnosis, changes in routine laboratory protocols may be required to ensure storage of frozen tissue that represents the patient's lesion. Initial studies with low-coverage genome sequencing of frozen versus formalin-fixed tissue DNA samples provide some reason to hope that fixed tissues could be a source of usable template, although as outlined below, the challenges posed by false-positive and false-negative results in genome-scale data sets may best be addressed by approaches that avoid any additional sources of error (20). HTS studies of chromatin (genomic DNA with associated proteins and noncoding RNA) extend the scope of genome sequence analysis and are instructional for evaluating the effects of nucleosome positioning; nucleosome modifications on DNA function; and associations between DNA and transcription factors, the RNA polymerase complex, and noncoding RNAs. Such studies are essential for mechanistic understanding of the readout of genetic information by the cell. Many of these approaches may be less amenable to routine diagnostic use because of their experimental complexity (21). Studies of DNA that is selected or chemically marked on the basis of direct covalent modifications, such as methylation, have underscored the effects of those epigenetic marks on gene regulation,

←

Figure 1

Components of high-throughput DNA sequencing technologies. Several of the most widely used high-throughput sequencing methodologies are shown. (*a*,*b*) Amplification and spatial separation. Most platforms include a step in which templates in the form of single DNA molecules are amplified and spatially separated from one another following initial library preparation, which adds sequencer-specific linker or primer sequences to the ends of the template molecules. (*Left*) In the Roche 454 and Ion Torrent platforms, template separation and amplification are carried out by first diluting the template molecules and generating an emulsion of aqueous droplets in oil, with an average of less than one template molecule per aqueous droplet, then performing polymerase chain reaction (PCR) in the emulsion in the presence of capture beads to attach the amplified template to a solid surface. (*Center*) In the Illumina platform, template separation and amplification occur through the addition of diluted template to a flow cell that captures the template and permits so-called bridging PCR amplification through the use of primers attached to the flow-cell surface. (*Right*) In the Complete Genomics platform, sequence tags derived from the template molecule (*colored segments separated by gray linker sequences*) in a closed circular DNA template molecule are amplified by a bacteriophage rolling-circle polymerase; they are then isolated spatially on an array surface as DNA nanoballs. (*c*,*d*) Sequencing and detection. (*Left*) The Roche 454 and Ion Torrent platforms use a sequencing-by-synthesis approach, in which repeating cycles of extension of the DNA strand complementary to the template are carried out in reaction mixtures containing only one of the deoxyribonucleotide triphosphates (dATP, dCTP, dGTP, or dTTP). Detection of incorporation of one or more nucleotides is indicated in the Roche 454 platform by pyrosequencing employing coupled enzymatic reactions to drive the generation of photons of light by the luciferase enzyme, using energy originally derived from the pyrophosphate released following the incorporation of each new nucleotide into the synthesized DNA strand. In the Ion Torrent platform, miniaturized ion sensors in the sequencing plate detect the incorporation of nucleotides into the growing template by measuring the release of protons that occurs as each nucleotide is added. (*Center*) The Illumina platform carries out cycles of single-base extension sequencing by synthesis, in which each nucleotide is labeled with a distinct fluorophore that reveals the identity of the incorporated nucleotide and prevents further extension of the template. Once the flow cell is imaged, the fluorescent labels are cleaved off, and the templates are ready for the next cycle of extension and imaging. (*Right*) In the Complete Genomics platform and in other methods using hybridization and ligation for sequencing, the identity of the base at a particular position in the unknown sequence tag derived from the template is interrogated using oligonucleotide probes that are degenerate in sequence at all positions apart from the positions that are being queried. The identity of the queried base is indicated by the color of the fluorescent label of the oligonucleotide. Hybridization of these partially degenerate probes next to an anchor probe complementary to the linker sequences is followed by ligation under conditions wherein the efficiency of ligation depends on base-pairing between the queried nucleotide in the template and the specified nucleotide in the probe. Following imaging of the flow cell, the ligated primers are washed away and a new set of oligonucleotide probes is added to interrogate the next position in the unknown sequence tag. The components and procedures for each of the methods depicted are schematic and simplified for purposes of illustration. Several alternative methods of high-throughput DNA sequencing, including methods that do not require template amplification or that use different combinations of the amplification, spatial isolation, and sequencing steps, have also been reported in the literature and commercialized.

with relevance for cancer biology and inherited disorders (22).

Exome Sequencing

The cost of whole-genome sequencing (WGS) has been steadily decreasing, but "genome partitioning" methods for physically isolating selected regions of the genome on the basis of their sequence have become popular as less expensive alternative strategies. Notable among these approaches are exome methods for isolating the protein coding ~1% of the genome. Hybridization-based methods using oligonucleotides that are complementary to the exons of the human genome, either in solution or attached to a solid substrate, are the most common approach; other methods, such as those that employ connector inversion probes that hybridize flanking particular genomic regions and are extended by genome-templated synthesis, then sequenced, are also used for this purpose (**Figure 2**) (23–26).

Studies of Mendelian genetic disorders and cancer genomes have particularly benefited from this methodology, as described below. With continually declining WGS costs, exome sequencing may become less popular, but for applications in which particularly deep coverage of protein-coding regions of the genome is desirable, such as in tests for low-frequency gene mutations that are present in some subclones of a tumor sample, the approach may retain an advantage. The commercial availability of customized arrays of capture probes to enable resequencing of genome areas of interest makes the genome partitioning approach relatively flexible for the validation of sequence variants detected in WGS (27).

Amplicon Sequencing

The HTS application that is most readily compatible with current clinical laboratory work flows and regulatory frameworks is amplicon sequencing. In this technique, PCR is used to amplify one or a panel of regions of the genome from genomic DNA or a complementary DNA (cDNA) template, and then the amplicons are sequenced with HTS (**Figure 2**). Such tests are conceptually similar to the Sanger sequencing work flow of laboratory sequencing tests; an important difference is that HTS methods generate many distinct sequences from each amplicon that represent the sequences of individual templates, rather than a consensus sequence trace in which variant sequence positions appear as mixed peaks. HTS of amplicon panels will probably represent an intermediate step in the development of clinical genome sequencing, to validate and confirm sequence variants detected by WGS or exome sequencing. Limitations of amplicon panels include their poor scalability and the tendency of some variant sequences to be amplified at lower efficiency and be lost from analysis. The number of sequence variants that have been studied well enough to form the basis for clinical management decisions is still small enough (in cancer diagnosis, approximately 50–100 genes have a clear therapeutic or prognostic implication) that amplicon panels can encompass these actionable variants. Investigators have demonstrated the robustness of this method for detection of chronic myeloid leukemia–associated somatic mutations and for the resequencing of large genes such as *BRCA1* and *BRCA2* (28, 29).

Other applications in which the selectivity of amplicon sequencing is critical are those in which a great depth of sequencing over a limited region is most desirable, as in detection of viral sequence variants that contain mutations conferring resistance to antiviral medications (30). Amplicon sequencing is also well suited to other niche applications in specialized regions of the genome, such as the rearranged immunoglobulin and T cell receptor (TCR) loci in lymphocytes, and may provide biomarkers of immune-mediated disease activity or adaptive immune system health (12, 31).

Transcriptome Sequencing

Measurements of RNA transcripts by use of microarray hybridization have been performed

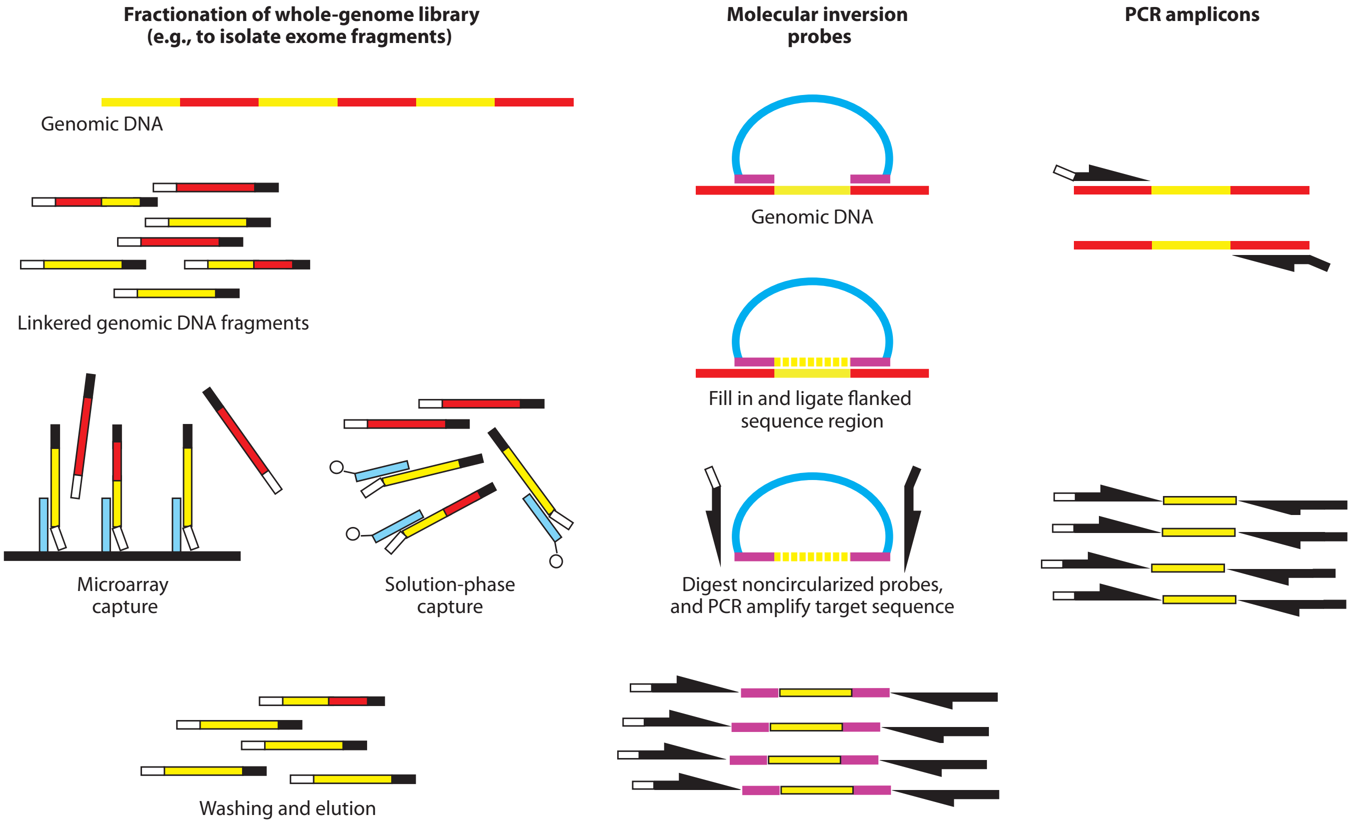

Fractionation of whole-genome library (e.g., to isolate exome fragments)
Genomic DNA
Linkered genomic DNA fragments
Microarray capture
Solution-phase capture
Washing and elution
Molecular inversion probes
Genomic DNA
Fill in and ligate flanked sequence region
Digest noncircularized probes, and PCR amplify target sequence
PCR amplicons

Figure 2

Methods of enrichment of genomic regions for sequencing. (*Left*) Physical enrichment of genomic DNA regions of interest, such as protein-coding exons of the genome, appears in yellow. Genomic DNA is fragmented, typically by sonication; then linker elements (*black and white*) are added to the ends of the fragments by ligation. The linked genomic DNA is exposed to probe sequences that are complementary to the DNA regions of interest, either on a solid array surface or in solution phase with probes that can later be isolated via a biotin tag that will be captured by streptavidin bound to beads, or analogous methods. The DNA of interest hybridizes to the capture probes, and unbound DNA is washed away. The population of DNA molecules enriched for the sequences of interest is then eluted from the array or beads and is sequenced. (*Center*) Molecular inversion probes and other, similar reagents contain sequences that hybridize 5′ and 3′ of a region of sequence to be interrogated. The probe is then extended with a polymerase to make a copy of the unknown sequence between the probe ends, and ligated to form a closed circle. After exonuclease digestion of unligated probes, polymerase chain reaction (PCR) amplification of the unknown sequence region and high-throughput sequencing are performed. (*Right*) PCR amplification can be used to amplify sequence regions of interest for high-throughput DNA sequencing.

for more than 15 years (32). HTS of cDNA has been widely adopted for research applications because it combines transcript counting and detection of transcript variants, including spliceform variants, pathogenic fusion transcripts derived from chromosomal rearrangements, and other mutations (33–36). The use of HTS to profile the expression of cellular genes is even being explored in formalin-fixed, paraffin-embedded samples through the use of the 3SEQ method, in which expressed gene-sequence tags from oligodeoxythymidine templated cDNA, or cDNA generated using random hexamer priming of RNA fragments possessing poly-A-tails, are characterized by HTS (37). Overall, the greater biological variability and chemical lability of RNA compared with those of DNA, and the introduction of additional sequencing errors that arise from the low fidelity of reverse transcriptase, are disadvantages of the use of RNA as a template for diagnostic sequencing but may be outweighed by the additional information provided by this method.

GENETIC VARIATION IN HUMANS AND HUMAN DISEASES

In analyses of data from HTS applications, key questions recur. One of the most important questions is how to deal with rare sequence variants (be they in the patient's germ-line genome, the genome of the patient's cancer, or the genome of an infectious agent) that may be important for an individual patient's disease pathogenesis, prognosis, or therapeutic response, but whose significance is poorly understood because it has not been observed in enough people or studied under the conditions of a controlled clinical trial. Statistical distributions with many rare members have been referred to as having long tails, and the collective set of sequence variants in human genomes, cancer genomes, and immunoglobulin or TCR gene rearrangements are medically relevant examples of such distributions. The set of sequence variants of unknown significance and the expected artifactual false-positive results from sequencing errors that are present in any human genome sequence has been termed the incidentalome by Kohane et al. (38, 39); these variants will undoubtedly pose a significant challenge for the responsible interpretation of genome-sequence information in clinical settings.

Before evaluating genomic DNA sequence variants that may be related to human diseases, one must understand the genetic variation in healthy human populations (i.e., humans who have not yet developed the diseases that they will eventually die from). The accumulation of our current understanding of this topic has been recently reviewed (2, 3). The assembled haploid genome in the Human Genome Project was a consensus patchwork of sequences derived from a pool of anonymized donor samples (6). The task of uncovering sequence variants and haplotypes (the blocks of sequence-containing variants that are inherited together because meiotic recombination has not separated them) in different human populations was undertaken by the International HapMap Project (40). This effort built on earlier research compiling the dbSNP database of polymorphic sites in human DNA and used a combination of microarray single-nucleotide polymorphism (SNP) typing and amplicon sequencing of samples from parent-child trios to determine the haplotypes for common sequence variants that have a minor allele frequency (MAF) of at least 5% in at least one of the ethnic groups studied (40–42). The variants and haplotype blocks defined through this method formed the basis for numerous genome-wide association studies (GWAS) that aimed to identify genomic regions linked to disease phenotypes, under the hypothesis that common diseases with a significant heritable component may be related to common sequence variants, even if these variants individually have weak effects (43). Although some GWAS have met with success, the overall explanatory power of these experiments has been modest; the identified variants typically contribute only a few percent of the suspected heritability. Proposed explanations for this missing-heritability discrepancy have included a strong effect of rare variant sequences that are not detected in GWAS, overestimation of the heritability of diseases, and very weak contributions of numerous common variants toward disease risk. Although some mechanistic clues to disease pathogenesis have been provided by genes that contain common sequence variants, in most cases to date GWAS results have not had a major impact on disease diagnosis, classification, prognosis, or therapeutic choices for individual patients.

More recently, the pilot stage of the 1,000 Genomes Project has sought to discover lower-frequency sequence variants, with the goal of detecting 95% of variants with a MAF of at least 1%, and surveying rarer variants, particularly in exons. The initial pilot phase of the project used several different HTS methods to generate shallow (2–4×) WGS data from 179 individuals from 3 ethnic groups (Han Chinese, Japanese, and Nigerian Yoruban); deep (40×) WGS data from parent-child trios of European and Yoruban origin; and deep exome sequencing data from 697 individuals from 7 population groups (44). These data, combined with the results from a series of WGS data sets (including those of Craig Venter and James Watson, African individuals from Khoisan and Bantu groups in Southern Africa and three populations from the Kalahari Desert, a Han Chinese individual, two Korean individuals, and more recently a Gujarati South Asian individual), among others, have provided a good initial survey of the major features of human genomic diversity (3, 11, 44–52). Also, although the existence of segmental duplication regions (long, duplicated stretches of near-identical sequence within a genome) had been appreciated in the Human Genome Project data, subsequent research revealed that copy number variation in the form of amplified, deleted, or rearranged regions of DNA constitutes one of the largest sources of total variant bases among individual human genomes (53–55). Taken together, the data from published individual human genome sequences and the initial results from the 1,000 Genomes Project support earlier estimations of the frequency of variant positions in the human genome (approximately 1 per 1,000 bases), and these data have greatly enhanced our understanding of the types and distribution of sequence variants in human populations. Some of the key findings regarding human genomic diversity are presented in the sidebar titled Summary of Human Genome Diversity Findings That Are Relevant to Medical Testing (**Figure 3**) (3).

This background of expected variation in any human genome is the noise from which the signal of disease-related variants in a given patient must be extracted in a diagnostic setting. Rare (or private) single-nucleotide variants (SNVs) present the greatest challenge, given that these variants continue to be discovered in large numbers as additional genome sequences are completed. Therefore, there will be a long tail to the list of known human genetic variants, and each new patient is likely to have many novel variants, some of which will be predicted to affect gene function. Previous genetic research performed without the benefit of WGS data has provided important lessons that must be heeded in HTS data interpretation (see the sidebar titled Caveats About Genotype-Phenotype Relationships).

One practical implication of the inconsistent penetrance of phenotypes related to a given sequence change is that filtering out variants in an asymptomatic individual, to narrow down the number of variants that could explain that patient's disease, will impede the identification of some causative mutations. A second implication is that high-quality and detailed phenotype data will be very helpful in ensuring that genotype-phenotype correlations in the HTS era will have some chance of detecting associations with alternate or milder phenotypic effects of genetic variants that cause severe disorders in a different genetic background or environmental context. Various good strategies for identifying causative mutations for suspected single-gene disorders have been applied

SUMMARY OF HUMAN GENOME DIVERSITY FINDINGS THAT ARE RELEVANT TO MEDICAL TESTING

A given individual's genome differs from the current reference human genome by an average of approximately three to four million single-nucleotide variants (SNVs). Rare (or private) SNVs in an individual's genome may comprise 200,000 (~5.0–7.5%) of the total SNVs; these variants are particular to that individual, his or her family, or other relatively closely related individuals.

Approximately 20,000 (~0.50–0.75%) of the SNVs in an individual's genome are in exons of genes; of these, approximately half are nonsynonymous variants that affect protein coding. Each person carries an average of 250 to 300 loss-of-function variants in annotated genes in his or her genome, as well as 50–100 variants that have been implicated in genetic disorders.

Most individuals have approximately 1,000 of the more common copy number variants (namely those with a MAF greater than 5%) that range from a few hundred bases to more than 1 Mb and affect a larger number of total bases than do the SNVs in the genome. Many of these copy number variants are related to the most common forms of human retrotransposon elements (sequences originating from transposable elements that mobilize via an RNA intermediate): the LINE (long interspersed) and SINE (short interspersed) elements.

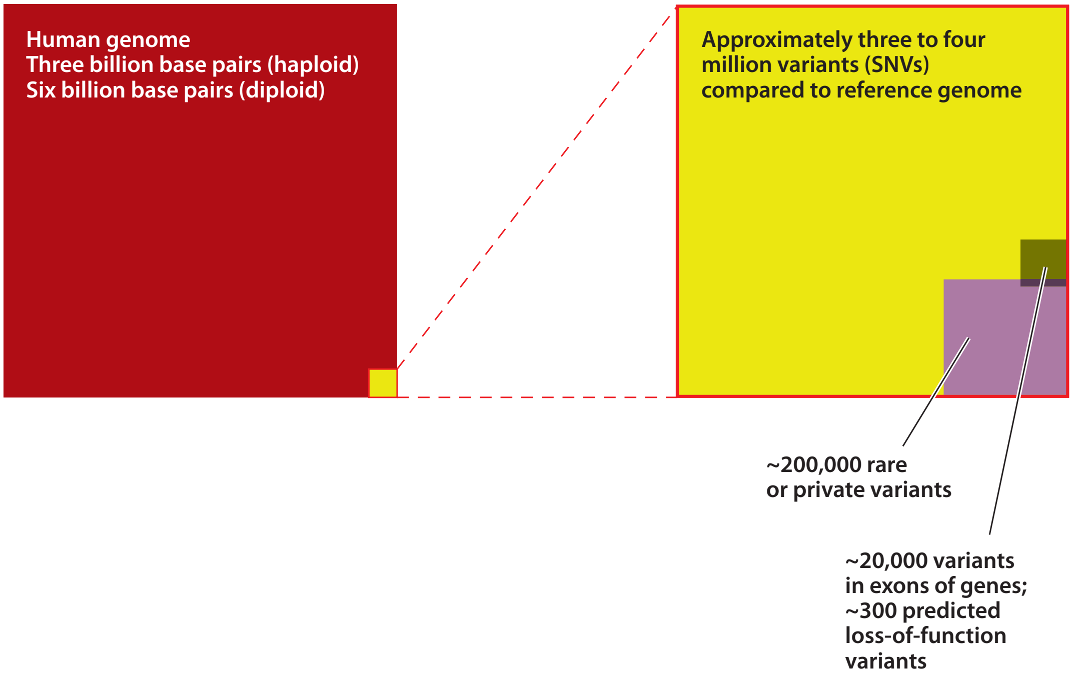


Figure 3

Sequence variants detected in whole-genome sequencing. The colored areas represent the approximate numbers of DNA single-nucleotide variants (SNVs) identified in human whole-genome sequencing experiments. (*Left*) In a diploid human genome of approximately six billion base pairs, approximately three to four million variants are identified upon comparison to a reference genome sequence. The number of variants detected depends on the human population groups from which the new genome sequence and the reference sequence are derived; reference sequences used in much of the published literature are from individuals of European descent. (*Right*) Of the three to four million SNVs, approximately 200,000 are rare (or private) variants that are particular to the individual, his or her family members, and others with relatively recent shared descent. Approximately 20,000 of the SNVs are located in gene exons. Approximately 300 of the SNVs identified in a given individual's genome are loss-of-function variants that are predicted to impair gene function.

in recent exome and WGS studies and have been reviewed elsewhere (2, 3). We discuss some highlights of this research in the section titled Mendelian Genetic Disorders, below.

In HTS studies of human cancers, the key goal has been to identify somatic mutations that are responsible for oncogenesis. Although some germ-line sequence variants in the cancer patient's genome can greatly increase the risk of developing a malignancy, additional mutations distinguishing the cancer from the nonmalignant cells of the body are almost certainly required. The challenge, then, is to identify differences between the patient's cancer genome (or genomes, given that many cancers have genetic heterogeneity) and the genome of the rest of their cells, as well as to distinguish between the causative driver mutations specific to the cancer and the passenger mutations in the cancer that have no functional effect. Just as rare mutations in individuals' genomes represent a challenge for interpretation and correlation with phenotypes, a critical question in cancer genome sequencing is whether the somatic mutations that cause cancers to develop consist mainly of common recurrent mutations in known oncogenes and tumor suppressors, or whether each cancer has its own private spectrum of rare mutations acting in concert, with complex interactions between mutations. The landscape of cancer mutations is beginning to come into focus with the results of several tumor and healthy tissue whole-genome comparisons that have been published during the past 4 years.

Several large collaborative projects that are currently under way, such as the Cancer Genome Atlas and the International Cancer Genome Consortium, are sequencing genomes and exomes from hundreds of individual cases of many of the most common types of cancers to discover the somatic mutations that characterize each distinct malignancy. This research has the potential to discover genetically defined subclasses within current diagnostic entities (see **http://cancergenome.nih.gov** and **http://www.icgc.org**). Over the next several years, these efforts will uncover many new sequence variants that are suspected to be involved in oncogenesis. If the past is any guide to the future, it will be a longer-term project to correlate the new sequence knowledge with sufficiently long follow-up periods to confirm the prognostic importance of individual mutations or combinations of mutations. Genetic insights into tumor biology may also suggest targeted therapies that could be attempted for individual tumors, and there will undoubtedly be anecdotal reports of particular patients who respond to such early versions of personalized therapy. The desire to treat a patient's tumor with a highly customized therapeutic strategy conflicts somewhat with traditional clinical trial research practice, in which the effects of any therapeutic intervention should be tested on sufficient numbers of patients randomized to receive either that particular therapy or an alternative or placebo treatment to enable valid statistical analysis of the benefits of the therapy. New clinical trials with a customized therapy arm, compared with a standard therapy arm, may permit a fair comparison between the personalized and standard therapy approaches but may be challenging to interpret if numerous different treatment approaches are permitted in the personalized therapy category. However, recurrent gene mutations observed in a subset of cases of a particular cancer type should permit clinical trials with enough patients who have the gene mutation, or who lack it, to determine whether the gene mutation is an appropriate guide for therapy selection. Testing for *EGFR* and *ALK* mutations in non-small-cell lung cancer, to decide whether to treat with targeted agents directed at those gene products, may offer a preview of strategies that will become available for a wider range of cancers over time (61–64).

CAVEATS ABOUT GENOTYPE-PHENOTYPE RELATIONSHIPS

Gene variants that are strongly associated with disease in affected individuals may also be present at significant levels in subclinical cases or in completely healthy individuals. For example, homozygous *HFE C282Y* mutations that are strongly associated with the recessive condition hereditary hemochromatosis are occasionally observed in healthy individuals, indicating incomplete penetrance of the phenotype (56–58). Environmental factors can have a major effect on penetrance; consumption of alcohol has been implicated in hemochromatosis (59).

Mutations in the same gene may contribute to different phenotypes that are classified as different diseases in different individuals. A British study of patients who had the sporadic form of porphyria cutanea tarda found that almost half had homozygous *HFE C282Y* mutations but that none had symptoms of hemochromatosis (60).

Even diseases that are classically considered single-gene Mendelian disorders can have strong modifier effects from other loci that affect the phenotype. For example, GWAS of cystic fibrosis revealed that modifier loci at chromosomes 11p13 and 20q13.2 significantly affect the severity of the phenotype of patients with *CFTR* mutations.

RECENT DIAGNOSTIC INSIGHTS FROM HIGH-THROUGHPUT SEQUENCING

Mendelian Genetic Disorders

There may be at least 5,000 to 6,000 single-gene disorders in humans; approximately 4,000 disorders are currently associated with a known genetic change in OMIM (the Online Mendelian Inheritance in Man database) (65). The current pace of causative mutation discovery with HTS suggests that the mutations responsible for many of the remaining disorders will be identified within the next few years. In some cases, such discoveries may prompt a reclassification, or at least a better understanding of clinical syndromes with overlapping features, if a common gene mutation or set of mutations in a pathway can provide a unifying diagnostic category, as has been proposed for muscular dystrophy dystroglycanopathies (65). Exome sequencing and, to a lesser extent, WGS have been popular and successful approaches for Mendelian disorder gene discovery with HTS. The following early examples of exome sequencing mutation identification, followed by additional examples highlighting experimental design and data-analysis strategies, show how the needle of a clinically significant mutation can be extracted from the haystack of complexity in individual genomes.

The first example of human exome sequencing in Mendelian disease studies used microarray-captured exome DNA from 12 individuals, of whom 4 were unrelated patients with a diagnosis of a rare dominantly inherited disorder, Freeman–Sheldon syndrome (FSS) (66). The exomes were sequenced at a 51-fold-average coverage by use of the Illumina platform, and the assay performance for sequence variant detection was estimated from detection of heterozygous base positions for the eight control subjects in the study, who were HapMap patients from whom independent sequence variant data sets were available; these observations showed that approximately 4% of known heterozygous positions were sequenced too shallowly in the exome data for confident detection but that >99% of called heterozygous positions in the exome data were concordant with prior HapMap data. The intersection of the sets of genes containing nonsynonymous coding SNPs, splice-site variants, or coding indels in the four FSS patients identified ~2,000 candidate genes, but when sequence variants present in dbSNP or in the HapMap exomes were filtered out, the only gene remaining was *MYH3*, which had previously been identified as causative for FSS in candidate gene screening studies, thereby validating the exome sequencing and data-analysis approach (67).

The first novel causative gene for a Mendelian disorder identified by exome sequencing came from studies of the recessive Miller syndrome in two affected siblings and two unrelated affected individuals. Similar to the approach used in the FSS study, several gene variants that would alter protein coding in both siblings, as well as in the unrelated affected individuals, and that were not found in dbSNP or HapMap data, were narrowed down to a single candidate: the pyrimidine de novo biosynthesis pathway gene *DHODH*, in which compound heterozygote mutations were found in the affected patients (68). Highlighting the challenge of predicting the functional consequences of gene mutations, a program (PolyPhen) designed to predict whether a sequence variant would be deleterious predicted that one of the *DHODH* variants present in the affected siblings was benign. Another striking result in this study was that the affected siblings had an atypical phenotype involving recurrent lung infections, bronchiectasis, and chronic obstructive pulmonary disease in addition to the classic dysmorphic features of Miller syndrome; they also had mutations in both copies of the *DNAH5* gene, a cause of primary ciliary dyskinesia, which offers a plausible explanation for the additional phenotypic features.

A more convoluted process of analysis led to the identification of *MLL2* as the causative gene for the autosomal dominant disorder Kabuki syndrome (69). In the study reporting this finding, 10 unrelated affected individuals' exomes were sequenced, and the variants were evaluated with a filtering scheme that was similar to that used in the FSS and Miller syndrome studies; no single gene appeared to be mutated in all 10 individuals. In an examination of genes that were mutated in subsets of the patients ranked according to the canonical dysmorphic features and phenotypic severity, *MLL2* emerged as the only candidate gene, and loss-of-function mutations were identified in 7 of the 10 individuals. Additional experimental work using Sanger resequencing of the gene in two of the apparently unmutated cases revealed that they carried indels causing frameshifts that had not been detected in the HTS data set. Follow-up studies confirmed *MLL2* mutations in 26 of 43 additional cases of Kabuki syndrome.

These early studies highlighted the feasibility of exome HTS data analysis for discovery of gene variants that are responsible for inherited genetic disorders. Additional early examples of exome sequencing in genetic disorder diagnosis are summarized below, with an emphasis on the data-analysis strategies used to narrow down the sequence variants that are most likely to cause the patient's disease. Further gene mutations identified in Mendelian disorders with HTS approaches have been reviewed elsewhere (2).

Strategies used in exome data analysis in Mendelian disorder mutation discovery include the following.

1. Identity by descent. Comparison between gene variants found in members of an affected pedigree, particularly the most distantly related affected members, can help to filter out individual private sequence variants that are unrelated to the disease. Exome sequencing of three siblings with the recessive hyperphosphatasia mental retardation syndrome was used to infer inherited haplotypes; the investigators focused on genes in which all three siblings had inherited the same haplotype from each parent. Only two candidate genes with mutations in all siblings remained within the 20% of the exome that met these criteria, and the mannosyl transferase gene, *PIGV*, was subsequently identified as the causative gene (70).
2. Homozygosity by descent. In offspring of consanguineous parents, homozygous genomic regions derived from a common ancestor (i.e., autozygous regions) surround areas where recessive disease-causing homozygous gene mutations may be located. Identification of such regions was used to identify a homozygous mutation in *SERPINF1* from exome sequencing of a single patient with osteogenesis imperfecta, whose parents were second cousins (71). The mutation was conspicuous for being present in the largest contiguous homozygous genomic region in the data. Another example in an offspring of consanguineous parents identified homozygous *STIM1* splicing mutations in a child who died of classic Kaposi's sarcoma, probably due to unchecked HHV-8 infection in the context of the primary immunodeficiency phenotype that had previously been associated with *STIM1* mutation (72).
3. Compound heterozygosity or homozygosity in suspected recessive disorders. Even when only a single patient sample is available, prioritizing genes containing homozygous or compound heterozygous mutations can filter out most of the nonpathogenic private sequence variants from exome data. This strategy proved successful in single-patient studies identifying (*a*) mutation of *WDR35* as the cause of Sensenbrenner syndrome and (*b*) mutation of the peroxisomal enzyme gene *HSD17B4* as the cause of Perrault syndrome (73, 74).
4. Shared mutation in unrelated individuals. When sufficient numbers of unrelated affected patients are available, sequencing exomes and searching for genes sharing mutation in most or all patients can also be effective, as demonstrated by a study that identified *SETB1* as the mutated gene in four unrelated individuals with Schinzel–Giedion syndrome and in another study that identified *ASXL1* as the mutated gene in three unrelated individuals with Bohring–Opitz syndrome (75, 76). These examples also illustrated that this strategy can be effective for discovery of de novo mutations responsible for congenital disorders (75). Data from four unrelated patients with the rare, suspected recessive blood-clotting disorder gray platelet syndrome narrowed down the list of candidate genes from exome sequencing to a single gene, *NBEAL2*, which is related to genes that are involved in cellular granule formation (77). In

a similar strategy, sequencing of three unrelated individuals with a hypomyelinating syndrome characterized by diffuse cerebral hypomyelination and cerebellar atrophy and hypoplasia of the corpus callosum uncovered compound heterozygous mutations in the genes for the two largest subunits of RNA polymerase III, *POLR3A* (in one individual) and *POLR3B* (in the other two individuals). These findings implicated impairment in transcription of small noncoding RNAs, including 5S ribosomal RNA and transfer RNAs, in the pathogenesis of the disorder (78).

5. Candidate gene filtering. Even relatively large numbers of genes containing private sequence variants represent a drastic narrowing of candidates for an inherited disorder, and the list can sometimes be further pared down on the basis of disease physiology. Exome sequencing of a single patient with a suspected mitochondrial respiratory chain complex I disorder identified a compound heterozygous mutation of the mitochondrial gene *ACAD9* (79). Data from model organisms can also greatly aid interpretation of human mutations, as demonstrated by an exome sequencing study of affected members of a family with combined hypolipidemia (a disorder with markedly reduced plasma levels of LDL and HDL cholesterol and triglycerides) (80). Compound heterozygous mutations in the *ANGPTL3* gene were discovered, which is consistent with the hypolipidemia phenotype that is observed in mice when this gene is disrupted (81). The use of candidate gene logic to sort through exome sequence variants is limited by our still-rudimentary understanding of the functions of most genes and proteins in the cell. Surprising connections between phenotypes and pathways have come to light in recent exome sequencing papers; for example, mutations in the gene for the double-stranded DNA-repair protein RAD51 appear to cause the neurological condition congenital mirror movements; mutations in the spliceosome gene *EFTUD2* are responsible for the rare sporadic syndrome mandibulofacial dysostosis with microcephaly; and mutation of the gene for ribosomal protein L21 appears to cause hereditary hypotrichosis simplex, an inherited nonsyndromic hair loss disorder (82–84).

The mutation discovery efforts described above relied on exome sequencing, a popular method due to both its lower cost, compared with that of full genome sequencing, and the fact that the exome and its splice sites represent the 1% of the genome that is most readily associated with known biological activities. As sequencing costs continue to drop, WGS methods are also being applied to inherited-disease studies. The earliest example of this approach studied a pedigree with recessive Charcot–Marie–Tooth neuropathy; the investigators identified compound heterozygous mutations in the *SH3TC2* gene in affected family members, as well as subclinical neuropathy susceptibility phenotypes in family members who inherited a single variant form of the gene (85). As genome sequencing costs continue to decrease, WGS may become the standard method for studies in which evaluation of the exome as well as the rest of the patient's genome may detect mutations that are in regions poorly captured in exome isolation, or in regions of the genome not currently suspected to have critical functions.

Have exome or WGS data yet been used for clinical decision making for inherited disorders? Some anecdotal reports are appearing in the literature. One paper describes a diagnosis of congenital chloride diarrhea on the basis of exome sequencing detection of a homozygous missense mutation in the *SLC26A3* gene in a patient who had not been given a definitive diagnosis, but in whom the renal salt-wasting disease Bartter syndrome had been suspected (86). Clinical follow-up confirmed the diagnosis and showed that the patient's dehydration and electrolyte abnormalities could be attributed to gastrointestinal losses rather than a renal defect. It would be fair to say that traditional laboratory testing and clinical workup combined with, and guided by, the exome sequencing data were the basis for the diagnosis, in the sense that it would not have been sufficient to do the sequencing analysis alone. A more dramatic example, in which a therapeutic choice with a high risk of morbidity and mortality was made on the basis of exome sequencing data, is a report of a boy with an intractable Crohn disease–like illness who was found to have a hemizygous mutation of the X-linked inhibitor of apoptosis gene (*XIAP*), a regulator of inflammatory responses (87). After Sanger sequencing in a clinical laboratory confirmed the mutation, myeloablative conditioning, followed by hematopoietic stem cell transplantation, was carried out because patients with *XIAP* mutation are at risk of death due to hemophagocytic lymphohistiocytosis unless reconstitution of the immune system is achieved. The investigators reported that the patient recovered well and the gastrointestinal disease manifestations also resolved (87).

Other treatment decisions guided by exome or other HTS data are probably being made for additional patients, as genomic data begin to enter the clinical arena. As in the rest of the scientific literature, there is likely to be a publication bias in favor of success stories. In these very early days of applying this new technology to clinical questions, there is an acute need for confirmation of results; for ensuring that full, informed consent is obtained from the patient; and for multidisciplinary evaluation of the diagnostic evidence and treatment options to weigh the risks and benefits of any therapies considered. For patients with life-threatening illnesses and no clear diagnosis from traditional approaches, careful application of HTS-based methods may provide valuable diagnostic clues, but such efforts straddle the boundary between clinical research and clinical care.

Cancer Genetics

Applications of HTS to human cancers during the past 5 years have provided new data for refining diagnostic and prognostic categories for malignancies, suggesting mechanisms of oncogenesis, and indicating potential therapeutic targets. As with studies of Mendelian disorders, a major task in the interpretation of cancer genomes, exomes, or transcriptomes is filtering out sequence variants that are unrelated to the disease, to focus the list of putative pathogenic gene variants so that they can be evaluated further in larger series of cases and characterized in greater detail. Detection of somatic mutations in cancers on a genome-wide scale currently requires comparison with the genome or exome sequence from unaffected tissue from the same patient, so that both common and private SNPs and copy number variants in the patient's germ line can be excluded from the list of potential somatic mutations. Of course, the germ-line DNA of some patients contains gene variants that predispose them to the development of cancers, as in the case of *BRCA1* and *BRCA2* mutation, or rarer syndromes, such as Li–Fraumeni syndrome in patients with *TP53* germ-line mutation, but most HTS analyses focus on identifying the somatic mutations unique to the neoplastic cells. The fastest progress has arguably been made in hematologic malignancies, in part because obtaining high-purity samples of malignant cells is easier than for solid-tissue malignancies, in which tumor cells and nonneoplastic stromal cells may be extensively intermingled. Also, many hematologic malignancies do not have highly aneuploid karyotypes and extensively rearranged genomes that complicate data analysis, as is the case for many carcinomas. In the following survey of recent results in cancer genome sequencing we focus on a sampling of results from hematologic malignancies, but there are equally compelling results from new HTS studies across the whole spectrum of human cancers, as recently reviewed (4, 5).

General Concepts in Cancer Biology Informed by High-Throughput Sequencing Studies

Many of the cancer genome sequencing papers published during the past 3 years have shed new light on familiar cancer biology concepts, but some have revealed entirely unexpected features of cancer cells as well. The most striking of these findings are described below.

1. Somatic mutation frequencies differ between different cancer types. WGS of a non-small-cell lung carcinoma identified more than 300 nonsynonymous mutations in protein-coding genomic regions of the tumor cells, and a similar WGS analysis of plasma cell myeloma found an average of 35 protein-altering point mutations per case and 21 chromosomal rearrangements that altered protein coding (88, 89). Diffuse large B cell lymphomas have an average of 15 somatic point mutations in their exomes, and there are approximately 30 total exomic mutations (point mutations, copy number changes, and translocations) per case (90). Acute myeloid leukemia (AML) genome sequences reported to date contain approximately 10 somatic mutations per case that affect protein-coding genes (91–93).
2. New core oncogenes and tumor suppressors mutated in diverse cancers continue to be discovered. For example, the *BRAF* gene had long been known to be mutated in a limited number of cancer types, such as melanomas, but it has now been identified in a much broader range of tumors, including several hematologic malignancies (94–96).
3. New genetic mechanisms for mutations have been discovered. A prominent example is chromothripsis, a process in which clustered regions of tens to hundreds of pieces from one or a few chromosomes are apparently fragmented and rearranged in a single catastrophic event (97). Studies of patients with plasma cell myeloma indicate that the presence of chromothripsis lesions correlates with (*a*) rapid relapse following treatment and (*b*) decreased survival (98). Notably, similar chromosome-rearrangement patterns have recently been reported in the germ-line DNA of patients with de novo congenital anomalies (99).
4. Cellular pathways that were not previously considered major targets for oncogenesis are mutated in human cancers. Two categories of mutations stand out in recent results from various tumors: those involving members of RNA splicing pathways and those involving central metabolic pathways such as glycolysis and the Krebs cycle metabolism (92, 100, 101). Genes involved in histone modification also loom large in the results from WGS of hematologic malignancies (102–106). The unexpected biology of tumor mutations suggests that we are only beginning to be able to predict which pathways are the most important as mutation targets for particular cancers. On a positive note, these findings may also hint that there are many currently unexplored targets for future drug development.
5. Tumor heterogeneity is becoming more accessible to study via HTS methods. Studies of genome mutations in relapsed cancers, and comparison between them and the molecular lesions observed at initial diagnosis, can reveal the relationships between subclones in the initial diagnostic specimen and the response to therapy. Recent work on relapsed AML indicates that relapsed disease can be associated with the predominant clone at diagnosis gaining additional mutations, or it can represent expansion and further mutation of minor subclones present in the initial diagnostic specimen (107). Similarly, studies of myelodysplastic syndrome (MDS) and subsequent AML in a series of patients have provided evidence of clonal evolution in the course of disease progression (27).
6. New links between viral pathogens and cancers are being discovered. A dramatic example is the discovery by transcriptome sequencing of a new polyoma virus integrated into the genomes of approximately 80% of cases of Merkel

cell carcinoma, a cancer that is observed at increased rates in immunosuppressed patients but is also associated with sun exposure of the skin (108, 109). Clonal integration of the virus appears to be an early event, and expression of viral large T antigen further supports a direct role for this virus in oncogenesis.

Hematologic malignancies. Many of the earliest highlights of cancer genome sequencing derived from studies of hematologic malignancies. The first reported cancer genome sequence was from an AML patient; this research and subsequent studies demonstrated major lessons of mutation discovery and filtering in the WGS era (91, 92). The initial AML genome that was sequenced and compared with a control genome from skin fibroblast DNA from the same individual had a normal karyotype, as determined by conventional cytogenetics. By comparing the skin fibroblast genome and the AML genome with the reference genome, investigators identified approximately 2.65 million well-supported SNVs in the tumor, of which 2.58 million were also present in the skin. Additional filtering removed SNVs observed in the Watson or Venter genome, thereby narrowing the list to 32,000 potential somatic mutations in the tumor. Of these, 181 were in coding regions or splice sites of the AML genome and were predicted to alter gene function. Interestingly, only 8 of the 181 potential functional variants were confirmed on resequencing; most of the rest were false positives from sequencing error or were present in the skin sample. Also, mutations in *FLT3* and *NPM1* (which are known to be recurrently mutated genes in AML, and which were known to be mutated in the case studied) were detected. When the 8 novel and confirmed somatically mutated genes were sequenced in 187 other AML cases, no additional mutations in these genes were found (91). Fortunately, a subsequent WGS study of an additional AML case succeeded in finding novel recurrent mutations: The authors of this study identified somatic mutations in the isocitrate dehydrogenase 1 (*IDH1*) gene, and in the mitochondrial gene *ND4*, that were also mutated in 1 or more of 188 additional AML cases (92). A second isocitrate dehydrogenase gene, *IDH2*, located on chromosome 15 and whose protein product functions in mitochondria, rather than in the cytosol like IDH1, is also recurrently mutated in AML. These findings reinforced the evidence that metabolic disturbances may significantly contribute to AML pathogenesis (110, 111). Further evaluation and deeper sequencing of the AML case in the very first AML genome study eventually revealed that this genome also contained a novel recurrent mutation in the *DNMT3* DNA methyltransferase gene, which was also a target of mutation in 22% of additional de novo AML cases tested (112).

The prognostic importance of novel recurrent mutations in AML has been studied for many of the most common mutations, although only a few years have passed since their discovery. Mutation of *DNMT3* is associated with a worsened prognosis in patients whose leukemia has a normal cytogenetic profile, and in patients with an intermediate risk profile (112). A Phase III clinical trial that randomized 398 AML patients under 60 years old to receive high-dose or standard-dose daunorubicin, and that sequenced 18 recurrently mutated genes in each leukemia, found reduced overall survival in cases with internal tandem duplication in *FLT3* (FLT3-ITD), partial tandem duplication in *MLL* (MLL-PTD), and mutations in *ASXL1* and *PHF6*; in contrast, improved overall survival was observed in cases with *CEBPA* or *IDH2* mutation (113). The same study found that *NPM1* mutation was a positive prognostic factor, but only in the presence of *IDH1* or *IDH2* mutation. High-dose daunorubicin improved overall survival only in patients with mutated *DNMT3A* or *NPM1*, or with translocations of *MLL*, which suggests that these genetic features could be used to stratify patients to different courses of treatment (113).

Studies of relapsed AML have provided evidence that clonal progression is a common feature of disease relapse following treatment (107). WGS of eight patients' initial leukemia, relapsed leukemia, and skin fibroblasts, followed by targeted recapture and deep resequencing of genes containing leukemia somatic mutations, showed that the relapsed AML genomes appeared to originate via accumulation of additional mutations in either the dominant initial leukemic clone or a subclone of the patient's initial disease. In the course of the study, additional novel genes subject to recurrent mutation in AML were discovered (*WAC*, *SMC3*, *DIS3*, *DDX41*, and *DAXX*). Also notable is that many of the new mutations in the relapsed AML samples were transversions, which suggests that cytotoxic chemotherapy agents play a role in generating the new mutations associated with relapse.

New mutations in the AML-precursor malignancy MDS have also recently come to light; this discovery has highlighted frequent mutations in splicing-related genes (114, 115). Mutations in members of a set of the splicing pathway genes *SRSF2*, *U2AF1*, *ZRSR2* and/or *SF3B1* were found in 67 of 193 MDS patients (34.7%) assayed by targeted resequencing. The *SRSF2* and *SF3B1* mutations were the most common; each was observed in more than 10% of cases. Mutation of *SRSF2* also appeared to be associated with poorer prognosis. Results from another MDS exome sequencing study of 29 cases also identified mutations in the splicing factors *U2AF35*, *ZRSR2*, *SRSF2*, and *SF3B1* in 45% to 85% of cases (116). Detailed transcriptome studies that aim to map the effects of spliceosome pathway mutations on other expressed genes will further explore this new direction in myeloid oncogenesis.

Other HTS and comparative genomic hybridization array studies have identified novel recurrent mutations in MDS (and other myeloid neoplasms) that affect histone modification pathways, including *EZH2*, a polycomb group histone H3 and H1 methylase that has probable loss-of-function mutations in up to 6% of MDS cases; *TET2*, a hydroxylase that is thought to be involved in hydroxylation of methylcytosine and that is mutated in approximately 20% of MDS cases; and *ASXL1*, an enhancer of trithorax and polycomb group proteins, that is mutated in up to 15% of MDS cases (105, 106, 117–119). Some of these mutated genes reside in regions of the genome that are frequently deleted in MDS, and they may help account for the recurrence of such deletions. For example, *EZH2* is located at chromosome 7q36 and may be a major reason that recurrent chromosome 7q deletions are observed, whereas chromosome 20q deletions may be related to the presence of *ASXL1* in chromosome 20q.

The mechanism of MDS progression to AML has been studied in detail in seven patients; in this study, the authors sequenced the secondary AML genome, then tested for AML-associated somatic mutations in samples from the preceding MDS, by using customized oligonucleotide arrays for targeted capture of genome regions containing candidate somatic SNVs found in the AML genome (27). The great majority (approximately 85%) of the cells in each MDS patient's bone marrow were clonal prior to progression to AML, and all cases showed an antecedent founding clone in the MDS sample that contained 182 to 660 somatic mutations that were also detected in AML. Progression to AML was associated with the detection of one or more subclones, which harbored dozens to hundreds of new mutations in addition to the earlier mutations observed in the MDS sample; this observation provided evidence that this disease progresses linearly and that this progression is correlated with the accumulation of new mutations. Four new recurrently mutated genes in AML were reported in the study (*CDH23*, *SMC3*, *UMODL1*, and *ZSWIM4*), and each MDS sample had one somatic mutation that was predicted to affect protein coding or an RNA gene; the subsequent AML samples acquired at least one additional mutation that affected protein coding, but most of the somatic mutations used to infer the clonal lineage relationships were outside the exome and of uncertain significance. Researchers have demonstrated the negative prognostic impact of some of the new recurrent mutations detected in MDS, independent of the International Prognostic Scoring System score, by using multivariate analysis. Mutations in *TP53*, *EZH2*, *ETV6*, *RUNX1*, and *ASXL1* showed hazard ratios ranging from 2.48 for *TP53* mutation down to 1.38 for *ASXL1* mutation (120).

In research on B cell malignancies, a particularly striking finding reported in the past year was the presence of an activating *V600E* mutation in *BRAF* in a case of hairy cell leukemia (HCL) studied by exome sequencing; this mutation was subsequently confirmed in 100% of 47 additional cases (95). Notably, other B cell malignancies with some morphologic and immunophenotypic overlap with HCL, such as marginal zone lymphomas, did not show *BRAF* mutation, nor did other peripheral B cell leukemias or lymphomas. Although HCL is typically responsive to current treatment strategies, these findings raise the question of whether BRAF inhibitors could play a role in the treatment of this neoplasm. The set of neoplasms in which *BRAF* mutations have been detected now also includes T lymphoblastic leukemia, multiple myeloma, and even Langerhans cell histiocytosis (88, 96).

The causative mutations for the B cell neoplasm chronic lymphocytic leukemia (CLL) have been elusive. WGS of four cases, and follow-up resequencing in 363 additional cases, identified recurrent mutations in the genes *NOTCH1*, exportin 1 (*XPO1*), myeloid differentiation primary response gene 88 (*MYD88*), and Kelch-like 6 (*KLHL6*) (121). The frequency of these mutations correlated well with the two major prognostic disease categories; cases with mutated immunoglobulin genes and good prognosis had a predominance of mutations in *MYD88* and *KLHL6*, whereas *NOTCH1* and *XPO1* mutations were observed mainly in poorer-prognosis cases with unmutated immunoglobulins. A large study that used WGS and/or exome sequencing in 88 cases of CLL identified recurrent mutations in previously known genes such as *TP53* and *ATM*, as well as in *MYD88* and *NOTCH1*, but it also identified novel mutation targets (*SF3B1*, *ZMYM3*, *MAPK1*, *FBXW7*, and *DDX3X*) (122). The mutation of the spliceosome component *SF3B1*, taken together with similar mutations found in cases of MDS and AML, indicates that splicing pathways are probably broadly relevant to the development of many cancers. The new genomic studies have greatly increased the number of known recurrently mutated genes in CLL, and the modest overlap between the gene sets they identified may indicate that the search for recurrent CLL mutations is not yet complete.

Diffuse large B cell lymphoma (DLBCL) and follicular lymphoma (FL), which have been extensively studied with earlier generations of technology such as microarray measurement of gene expression, have also been the subject of several striking discoveries obtained through HTS. Recurrent mutations in the *EZH2* histone methyltransferase gene were first reported in DLBCL germinal-center B cell category cases (22% of cases) and FL (7% of cases) on the basis of combined WGS and transcriptome sequencing (102). Subsequent work revealed the very high rate of recurrent mutation in the *MLL2* histone methyltransferase gene in DLBCL (32% of cases) and FL (89% of cases), as well as the rate of recurrent mutation in the *MEF2B* histone acetyltransferase complex member (11.4% of DLBCL and 13.4% of FL cases), among 109 genes that have recurrent somatic mutations in the lymphomas (103). An additional, extensive list of recurrent somatic gene mutations in DLBCL, based on WGS studies and targeted resequencing, has recently been published (90).

Plasma cell myeloma is a hematologic malignancy that is among the most refractory to treatment. A large WGS study of 38 cases and control tissues discovered recurrent somatic mutations in the *DIS3* RNA exonuclease gene involved in regulating levels of mRNA in the cell (11% of cases); *FAM46C*, which is thought to be an mRNA stability factor (13% of cases); *XBP1*, a transcription factor involved in plasma cell differentiation and the unfolded protein response (5% of cases); and the *LRRK2* gene, which encodes a kinase of the translation initiation factor 4E–binding protein (8% of cases) (88). These genes appeared to form a theme of RNA processing and protein translation control, functions that could be related to the heavy protein synthesis burden borne by plasma cells as they generate and secrete immunoglobulins. Other recurrent mutations were found in *IRF1* (also known as *MUM1*), a transcription factor involved in B cell and T cell differentiation (5% of cases), and its downstream target *PRDM1* (also known as *BLIMP1*), another transcription factor that is involved in plasma cell differentiation (5% of cases). Of the 38 genomes that were sequenced, 1 had an activating *BRAF* mutation, and 4% of an additional 161 cases also showed *BRAF* mutation. Similar to findings in MDS, AML, and B cell lymphomas, genes involved in histone methylation were significantly enriched for mutation in myeloma genomes; mutations were detected in *MLL*, *MLL2*, *MLL3*, *UTX*, *WHSC1*, and *WHSC1L1* (88). Other observations included mutation in coagulation pathway members and recurrent noncoding mutations in the *BCL7A* gene, a putative oncogene in Burkitt lymphoma. Correlation of these findings with the clinical behavior or therapeutic responses of plasma cell myeloma will be of great interest. As noted above, other researchers have reported that plasma cell myeloma patients whose cancers show evidence of chromothripsis genetic lesions represent a particularly high risk group and are prone to early relapse and poor outcomes (98).

T lymphoblastic leukemia is not as well characterized genetically as B lymphoblastic leukemia; in the former disease, evidence of recurrent translocations or known mutations is more limited. A recent WGS study of 12 cases with an early T precursor phenotype has considerably expanded our knowledge of mutations in this malignancy and has highlighted recurrent activating mutations in genes that regulate cytokine receptor and RAS signaling (two-thirds of cases had mutations in *NRAS*, *KRAS*, *FLT3*, *IL7R*, *JAK3*, *JAK1*, *SH2B3*, or *BRAF*), inactivating lesions that disrupt hematopoietic development (more than half of cases had mutations in *GATA3*, *ETV6*, *RUNX1*, or *IKZF1*), and histone-modifying genes (approximately half of cases had mutations in *EZH2*, *EED*, *SUZ12*, *SETD2*, or *EP300*). Novel recurrent mutations in genes *DNM2*, *ECT2L*, and *RELN* were also detected (94).

Minimal residual disease testing. An additional application of HTS that is relevant to the care of cancer patients is the use of sequencing assays to detect residual disease following treatment. Sequencing of the immunoglobulin V(D)J gene rearrangements of lymphomas or lymphoid leukemias for minimal residual disease (MRD) testing has been validated in pilot studies (12, 123). A more general approach that is applicable to solid tumors involves HTS detection of translocations or fusion gene regions that are characteristic of the patient's cancer, followed by design of real-time PCR assays to detect those rearrangements (124). The use of patient-specific real-time PCR assays could limit the scalability of this approach, but it is likely that HTS could eventually be used as the basis for MRD detection of most cancers.

Infectious Disease

HTS assays that aim to detect and characterize the genomes of microbes have made it much easier to identify novel infectious organisms and to track outbreaks or epidemics of disease. This new era of pathogen identification can be considered to have begun, prior to the availability of HTS methods, with the rapid identification of the SARS virus in 2003, which was achieved through a combination of viral nucleic acid microarray hybridization and traditional viral culture and real-time PCR, followed by sequencing (125). Proof-of-concept experiments using HTS for similar goals have been published; in particular, such experiments have searched for viral pathogens in respiratory, diarrheal, and hemorrhagic infections (126–133). Key factors that should be considered in such experiments, and in adapting these methods for wider diagnostic use, are the quality of the evidence implicating an infectious agent in a particular disease and the need to sample bodily fluids or tissue sites in a sufficient quantity to ensure adequate sensitivity for the detection of pathogens that may be rare, particularly if the pathogens are novel.

The traditional microbiology lab methods for detecting and identifying bacterial pathogens include Gram staining, solid or liquid culture, the use of the live microbes in tests of biochemical activities and antibiotic resistance, and targeted molecular testing. For most common pathogenic bacteria in humans, these methods are effective, but unusual or novel species can prove difficult to characterize. A recent study of a patient suspected to have died from a fatal inhalation anthrax infection shows the power that HTS brings to microbial characterization: The bacterial genome from the pathogen was rapidly sequenced and found to be that of a new strain of *Bacillus cereus*, rather than anthrax (134).

HTS is also being used to identify the source and track the spread of infectious disease outbreaks and epidemics. An example is an analysis of a US epidemic of community-associated methicillin-resistant *Staphylococcus aureus*, which found that most strains in different regions of the United States were very closely related; this finding implicates expansion from a single population rather than convergent evolution of different strains (135). Other HTS epidemiologic insights have come from a study of Haitian cholera outbreaks that traced their probable origin to Bangladesh, as well as from an analysis of a particularly virulent Shiga toxin–producing *Escherichia coli* O104:H4 strain in Germany, which revealed that the enhanced virulence probably arose by horizontal transfer of a prophage carrying genes for Shiga toxin 2, other virulence factors, and antibiotic resistance (136, 137).

Patterns of H1N1 influenza strains in different geographical regions and the eventual dominance of particular strains over time have been studied in the United States, and HTS has also been used to investigate viral populations in immunosuppressed patient populations that may contribute disproportionately to the development of drug-resistant influenza strains (138, 139). In patients who are infected with HIV, deep sequencing of viral subpopulations detects low-frequency mutant viral strains with antiviral resistance–associated sequence changes, and assessment of the role for this type of assay in clinical management is under way (30, 140).

A very large body of data is also being collected as part of the National Institutes of Health–funded Human Microbiome Project, as well as by other investigators, to catalog the microbial flora present in different human anatomic sites and to correlate microbial populations with possible effects of other human phenotypes and diseases, including obesity and cancer (141, 142). The scope of these studies is vast and precludes a full discussion here, but the concept demonstrates how a significant environmental factor can be subjected to HTS analysis for combination with studies of the human host's genome.

Immune System Monitoring

The rearrangement and junctional diversification of genomic DNA segments encoding the immunoglobulin receptors in B cells and the TCR of T cells represent yet another source of genomic complexity that is very relevant to human diseases. The immunoglobulin and TCR proteins are the basis for lymphocyte recognition and targeting of foreign pathogens, but they also contribute to pathogenic targeting of host tissues in the case of autoimmune disorders, and excessive immunoglobulin E–mediated immune responses due to misregulation of lymphocytes specific for benign environmental or food allergens, to cause life-threatening allergic disorders. HTS methods have made the repertoires of immunoglobulin and TCR gene rearrangements in human samples accessible to comprehensive monitoring, and they should significantly improve our understanding of normal and pathogenic immune responses. It may be optimistic to suppose that there will be recurrent immunoglobulin or TCR rearrangements that will be both specific and sensitive markers of particular autoimmune disorders, given the vast diversity of these receptor populations and the numerous target molecules and epitopes they may target, but HTS technologies provide an excellent means of tracking individual B cell or T cell clones that are associated with a patient's disease. Initial research in diagnostic applications in this area has focused on detecting residual disease from treated B cell lymphomas or leukemias, as well as mapping out the basic features of immune repertoires in healthy individuals. We expect that these tools will be applied to autoimmune diseases, transplant biology, allergic disorders, and other immune-mediated disease topics (12, 31, 123, 143–145). Monitoring whether an individual's B cell and T cell populations have normal, healthy diversity may also be useful, for example, during reconstitution of the adaptive immune system following bone marrow transplant, characterization of immunodeficiency disorders, and evaluation of immune system function in the elderly. HTS of immunoglobulin gene rearrangements is already providing new detailed insights into the interactions between viral pathogens such as HIV and the immune response directed against them, and this technology is a powerful new way to evaluate the diversity and persistence of adaptive immune system responses stimulated by vaccine candidates (146, 147).

Transplant Biology

Recently, investigators described an innovative way to monitor solid-organ transplant rejection through HTS of cell-free DNA from the blood (148). Short fragments of DNA are present in the serum or plasma of all individuals, and they can be ligated to linker DNA and sequenced in depth. In organ transplant recipients, the ratio of recipient genomic DNA to graft-derived donor DNA, distinguished by SNPs that are specific to the recipient or the donor, provides a measure of the number of graft cells that are dying and releasing their DNA into the blood. In a pilot study of heart transplant recipients, episodes of acute cellular organ rejection were marked by increases in the proportion of donor-derived DNA in the blood (148). Advantages of this approach over traditional periodic biopsies of the graft tissue are that it is less invasive, may be less affected by sampling errors if lymphocytic infiltrates or regions of cell damage in the graft tissue are unevenly distributed, and could be more sensitive in detecting early stages of graft rejection. In combination with other methods such as monitoring for pathogen nucleic acids, this approach may be able to distinguish between graft rejection and infections or other sources of damage to graft tissue.

Prenatal Diagnostics

Continuing the theme of monitoring foreign DNA in the blood, prenatal testing can be performed with HTS of cell-free DNA to detect fetal trisomies by comparing the ratios of the number of DNA fragments derived from each chromosome (149–151). Remarkably, the detection of trisomies does not require one to distinguish whether individual fragments are of maternal or fetal origin, but rather detects the small increase in representation of the trisomic chromosome even though fetal DNA typically comprises less than 10% of the total cell-free DNA in the mother's blood (152). Recently, investigators clinically validated this approach in a study of 753 pregnant women at high risk for fetal trisomy 21. The technique demonstrated 100% sensitivity and 97.9% specificity of detection of fetal trisomy 21, which yielded a positive predictive value of 96.6% and a negative predictive value of 100% (153). More remarkably, recent studies have provided evidence that noninvasive determination of the fetal whole-genome sequence from maternal plasma cell–free DNA is feasible, when the interpretation is guided by phased maternal genome sequence data, and either nonphased paternal genome sequence or entirely inferred paternal genome sequence (154, 155).

CONCLUSIONS

We are entering a new era in which a large amount of genetic data can be readily made available from a patient's genome sequence, as well as from other applications of HTS. There will probably continue to be a gap between the availability of such data and the ability to comprehensively interpret the results for clinical decision making, but efforts to close the gap with targeted resequencing and validation of key recurrent mutations in independent clinical cohorts and in clinical trials will be vigorously pursued. Ongoing controversies about the best clinical use of laboratory assays—even relatively simple ones such as the prostate-specific antigen test—underscore the challenges that lie ahead in applying much more complicated data from WGS or exome sequencing in clinical settings. The number and design of clinical trials, and the disease and therapeutic topics that are addressed in such trials, have typically been major rate-limiting factors in advancing modern clinical knowledge. This may also be the case for genomic medicine, although the speed with which novel AML- and MDS-associated somatic mutations have been tested for prognostic significance shows that such validation can proceed rapidly (112, 113, 120, 156). The ethical and practical consequences of entirely ignoring data whose use has not been validated by conventional clinical trials, versus running the risk of overinterpreting some of the new information in an effort to help a patient and potentially causing harm, are still a matter of active debate; such considerations should be weighed with reference to the patient's own preferences regarding being an early adopter of new technology and about conservative versus proactive management of his or her health concerns. Private or public entities that provide genomic testing services may be tempted to exaggerate the value of these new testing methods, so it will be in the public interest to ensure responsible regulatory monitoring of testing methods whose results affect patients' health or life choices. A larger societal challenge may be to resolve the tensions between (*a*) the ready availability of patient-specific, fine-grained DNA sequence data of varying degrees of interpretability and (*b*) the increasing regulatory and economic drives in the US healthcare system to make patient care more standardized and less expensive.

New genetic data may refine or revise diagnostic classification schemes in many disease areas, but our ability to predict the eventual consequences of a genetic change is not yet fully developed. Therefore, HTS genetic testing is a powerful addition to the diagnostic armamentarium, but will probably not entirely replace most current diagnostic approaches. Pathologists' interpretations of histologic findings, and efforts to integrate these data with other laboratory and clinical results, are somewhat analogous to the task of interpreting genomic information. A microscope slide can contain numerous artifacts of sample preparation, and the visual interpretation of cell and tissue morphology summarizes a great deal of data, but human cancers show frequent deviations from the ideal entities described in the disease classification schemes at a given period of time. With accumulating data and experience, ongoing clinical trials, and appropriate training, pathologists and other physicians will soon be able to make responsible judgment calls in clinical genomics even with imperfect information, much as such judgments are made in histologic interpretation. Integration of clinical genomic and HTS data with other laboratory testing, histology, imaging studies, and patient history and physical findings will probably be the best check against overinterpretation or erroneous decision making based on a single data source. Synoptic reports, such as those used in hematopathology to integrate morphologic findings, flow cytometry, or immunohistochemical stain results, along with cytogenetic and molecular testing data, offer a good model for the integration of clinical genomic data with other patient data. "Tumor board"–style meetings of clinicians and diagnostic specialists from pathology and radiology, in which genomic testing data are discussed in the context of a patient's entire clinical record, will also be a prudent way to integrate new results into patient care. Inclusion of WGS data and other high-throughput DNA sequencing data sets in the clinical care of patients will require a period of transition, in which clinical decision making will be based only on reliable interpretation of sequence variants of known importance; the currently uninterpretable portions of genetic data sets can be retained, if the patient wishes, for future use as this area of medicine progresses. Although this body of knowledge is already becoming complex and detailed, the remarkable discoveries so far obtained through HTS methods in almost all areas of medicine probably represent a mere glimpse of what lies ahead.

DISCLOSURE STATEMENT

The author is a consultant to Immumetrix, LLC.

LITERATURE CITED

1. Mardis ER. 2011. A decade's perspective on DNA sequencing technology. *Nature* 470:198–203
2. Gilissen C, Hoischen A, Brunner HG, Veltman JA. 2012. Disease gene identification strategies for exome sequencing. *Eur. J. Hum. Genet.* 20:490–97
3. Gonzaga-Jauregui C, Lupski JR, Gibbs RA. 2012. Human genome sequencing in health and disease. *Annu. Rev. Med.* 63:35–61
4. Graubert TA, Mardis ER. 2011. Genomics of acute myeloid leukemia. *Cancer J.* 17:487–91
5. Mardis ER. 2012. Genome sequencing and cancer. *Curr. Opin. Genet. Dev.* 22:245–50
6. Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, et al. 2001. Initial sequencing and analysis of the human genome. *Nature* 409:860–921
7. Int. Hum. Genome Seq. Consort. 2004. Finishing the euchromatic sequence of the human genome. *Nature* 431:931–45
8. Margulies M, Egholm M, Altman WE, Attiya S, Bader JS, et al. 2005. Genome sequencing in microfabricated high-density picolitre reactors. *Nature* 437:376–80
9. Bennett S. 2004. Solexa Ltd. *Pharmacogenomics* 5:433–38
10. Shendure J, Porreca GJ, Reppas NB, Lin X, McCutcheon JP, et al. 2005. Accurate multiplex polony sequencing of an evolved bacterial genome. *Science* 309:1728–32
11. Bentley DR, Balasubramanian S, Swerdlow HP, Smith GP, Milton J, et al. 2008. Accurate whole human genome sequencing using reversible terminator chemistry. *Nature* 456:53–59
12. Boyd SD, Marshall EL, Merker JD, Maniar JM, Zhang LN, et al. 2009. Measurement and clinical monitoring of human lymphocyte clonality by massively parallel VDJ pyrosequencing. *Sci. Transl. Med.* 1:12ra23
13. Wommack KE, Bhavsar J, Ravel J. 2008. Metagenomics: read length matters. *Appl. Environ. Microbiol.* 74:1453–63
14. Rothberg JM, Hinz W, Rearick TM, Schultz J, Mileski W, et al. 2011. An integrated semiconductor device enabling non-optical genome sequencing. *Nature* 475:348–52
15. Eid J, Fehr A, Gray J, Luong K, Lyle J, et al. 2009. Real-time DNA sequencing from single polymerase molecules. *Science* 323:133–38
16. Harris TD, Buzby PR, Babcock H, Beer E, Bowers J, et al. 2008. Single-molecule DNA sequencing of a viral genome. *Science* 320:106–9
17. Drmanac R, Sparks AB, Callow MJ, Halpern AL, Burns NL, et al. 2010. Human genome sequencing using unchained base reads on self-assembling DNA nanoarrays. *Science* 327:78–81
18. Valouev A, Ichikawa J, Tonthat T, Stuart J, Ranade S, et al. 2008. A high-resolution, nucleosome position map of *C. elegans* reveals a lack of universal sequence-dictated positioning. *Genome Res.* 18:1051–63
19. Heinrich V, Stange J, Dickhaus T, Imkeller P, Kruger U, et al. 2011. The allele distribution in next-generation sequencing data sets is accurately described as the result of a stochastic branching process. *Nucleic Acids Res.* 40:2426–31
20. Schweiger MR, Kerick M, Timmermann B, Albrecht MW, Borodina T, et al. 2009. Genome-wide massively parallel sequencing of formaldehyde-fixed paraffin-embedded (FFPE) tumor tissues for copy number and mutation analysis. *PLoS ONE* 4:e5548

21. Robertson G, Hirst M, Bainbridge M, Bilenky M, Zhao Y, et al. 2007. Genome-wide profiles of STAT1 DNA association using chromatin immunoprecipitation and massively parallel sequencing. *Nat. Methods* 4:651–57

22. Harris RA, Wang T, Coarfa C, Nagarajan RP, Hong C, et al. 2010. Comparison of sequencing-based methods to profile DNA methylation and identification of monoallelic epigenetic modifications. *Nat. Biotechnol.* 28:1097–105

23. Albert TJ, Molla MN, Muzny DM, Nazareth L, Wheeler D, et al. 2007. Direct selection of human genomic loci by microarray hybridization. *Nat. Methods* 4:903–5

24. Hodges E, Xuan Z, Balija V, Kramer M, Molla MN, et al. 2007. Genome-wide in situ exon capture for selective resequencing. *Nat. Genet.* 39:1522–27

25. Ng PC, Levy S, Huang J, Stockwell TB, Walenz BP, et al. 2008. Genetic variation in an individual human exome. *PLoS Genet.* 4:e1000160

26. Akhras MS, Unemo M, Thiyagarajan S, Nyren P, Davis RW, et al. 2007. Connector inversion probe technology: a powerful one-primer multiplex DNA amplification system for numerous scientific applications. *PLoS ONE* 2:e915

27. Walter MJ, Shen D, Ding L, Shao J, Koboldt DC, et al. 2012. Clonal architecture of secondary acute myeloid leukemia. *N. Engl. J. Med.* 366:1090–98

28. De Leeneer K, Hellemans J, De Schrijver J, Baetens M, Poppe B, et al. 2011. Massive parallel amplicon sequencing of the breast cancer genes *BRCA1* and *BRCA2*: opportunities, challenges, and limitations. *Hum. Mutat.* 32:335–44

29. Kohlmann A, Klein HU, Weissmann S, Bresolin S, Chaplin T, et al. 2011. The Interlaboratory RObustness of Next-generation sequencing (IRON) study: a deep sequencing investigation of *TET2*, *CBL* and *KRAS* mutations by an international consortium involving 10 laboratories. *Leukemia* 25:1840–48

30. Simen BB, Simons JF, Hullsiek KH, Novak RM, Macarthur RD, et al. 2009. Low-abundance drug-resistant viral variants in chronically HIV-infected, antiretroviral treatment–naïve patients significantly impact treatment outcomes. *J. Infect. Dis.* 199:693–701

31. Robins HS, Campregher PV, Srivastava SK, Wacher A, Turtle CJ, et al. 2009. Comprehensive assessment of T cell receptor β-chain diversity in αβ T cells. *Blood* 114:4099–107

32. Schena M, Shalon D, Davis RW, Brown PO. 1995. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science* 270:467–70

33. Marioni JC, Mason CE, Mane SM, Stephens M, Gilad Y. 2008. RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays. *Genome Res.* 18:1509–17

34. Morin R, Bainbridge M, Fejes A, Hirst M, Krzywinski M, et al. 2008. Profiling the HeLa S3 transcriptome using randomly primed cDNA and massively parallel short-read sequencing. *Biotechniques* 45:81–94

35. Nagalakshmi U, Wang Z, Waern K, Shou C, Raha D, et al. 2008. The transcriptional landscape of the yeast genome defined by RNA sequencing. *Science* 320:1344–49

36. Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B. 2008. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat. Methods* 5:621–28

37. Beck AH, Weng Z, Witten DM, Zhu S, Foley JW, et al. 2010. 3′-end sequencing for expression quantification (3SEQ) from archival tumor samples. *PLoS ONE* 5:e8768

38. Kohane IS, Hsing M, Kong SW. 2012. Taxonomizing, sizing, and overcoming the incidentalome. *Genet. Med.* 14:399–404

39. Kohane IS, Masys DR, Altman RB. 2006. The incidentalome: a threat to genomic medicine. *J. Am. Med. Assoc.* 296:212–15

40. Int. HapMap Consort. 2003. The International HapMap Project. *Nature* 426:789–96

41. Sherry ST, Ward MH, Kholodov M, Baker J, Phan L, et al. 2001. dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res.* 29:308–11

42. Altshuler DM, Gibbs RA, Peltonen L, Dermitzakis E, Schaffner SF, et al. 2010. Integrating common and rare genetic variation in diverse human populations. *Nature* 467:52–58

43. Reich DE, Lander ES. 2001. On the allelic spectrum of human disease. *Trends Genet.* 17:502–10

44. 1,000 Genomes Proj. Consort. 2010. A map of human genome variation from population-scale sequencing. *Nature* 467:1061–73

45. Schuster SC, Miller W, Ratan A, Tomsho LP, Giardine B, et al. 2010. Complete Khoisan and Bantu genomes from southern Africa. *Nature* 463:943–47

46. McKernan KJ, Peckham HE, Costa GL, McLaughlin SF, Fu Y, et al. 2009. Sequence and structural variation in a human genome uncovered by short-read, massively parallel ligation sequencing using two-base encoding. *Genome Res.* 19:1527–41

47. Kim JI, Ju YS, Park H, Kim S, Lee S, et al. 2009. A highly annotated whole-genome sequence of a Korean individual. *Nature* 460:1011–15

48. Ahn SM, Kim TH, Lee S, Kim D, Ghang H, et al. 2009. The first Korean genome sequence and analysis: full genome sequencing for a socio-ethnic group. *Genome Res.* 19:1622–29

49. Wheeler DA, Srinivasan M, Egholm M, Shen Y, Chen L, et al. 2008. The complete genome of an individual by massively parallel DNA sequencing. *Nature* 452:872–76

50. Wang J, Wang W, Li R, Li Y, Tian G, et al. 2008. The diploid genome sequence of an Asian individual. *Nature* 456:60–65

51. Levy S, Sutton G, Ng PC, Feuk L, Halpern AL, et al. 2007. The diploid genome sequence of an individual human. *PLoS Biol.* 5:e254

52. Kitzman JO, Mackenzie AP, Adey A, Hiatt JB, Patwardhan RP, et al. 2011. Haplotype-resolved genome sequencing of a Gujarati Indian individual. *Nat. Biotechnol.* 29:59–63

53. Bailey JA, Yavor AM, Massa HF, Trask BJ, Eichler EE. 2001. Segmental duplications: organization and impact within the current human genome project assembly. *Genome Res.* 11:1005–17

54. Redon R, Ishikawa S, Fitch KR, Feuk L, Perry GH, et al. 2006. Global variation in copy number in the human genome. *Nature* 444:444–54

55. Kidd JM, Cooper GM, Donahue WF, Hayden HS, Sampas N, et al. 2008. Mapping and sequencing of structural variation from eight human genomes. *Nature* 453:56–64

56. Hanson EH, Imperatore G, Burke W. 2001. HFE gene and hereditary hemochromatosis: a HuGE review. Human Genome Epidemiology. *Am. J. Epidemiol.* 154:193–206

57. Fullerton SM, Wolf WA, Brothers KB, Clayton EW, Crawford DC, et al. 2012. Return of individual research results from genome-wide association studies: experience of the Electronic Medical Records and Genomics (eMERGE) Network. *Genet. Med.* 14:424–31

58. Feder JN, Gnirke A, Thomas W, Tsuchihashi Z, Ruddy DA, et al. 1996. A novel MHC class I–like gene is mutated in patients with hereditary haemochromatosis. *Nat. Genet.* 13:399–408

59. Fletcher LM, Dixon JL, Purdie DM, Powell LW, Crawford DH. 2002. Excess alcohol greatly increases the prevalence of cirrhosis in hereditary hemochromatosis. *Gastroenterology* 122:281–89

60. Roberts AG, Whatley SD, Morgan RR, Worwood M, Elder GH. 1997. Increased frequency of the haemochromatosis Cys282Tyr mutation in sporadic porphyria cutanea tarda. *Lancet* 349:321–23

61. Kwak EL, Bang YJ, Camidge DR, Shaw AT, Solomon B, et al. 2010. Anaplastic lymphoma kinase inhibition in non-small-cell lung cancer. *N. Engl. J. Med.* 363:1693–703

62. Zhou C, Wu YL, Chen G, Feng J, Liu XQ, et al. 2011. Erlotinib versus chemotherapy as first-line treatment for patients with advanced *EGFR* mutation–positive non-small-cell lung cancer (OPTIMAL, CTONG-0802): a multicentre, open-label, randomised, phase 3 study. *Lancet Oncol.* 12:735–42

63. Mok TS, Wu YL, Thongprasert S, Yang CH, Chu DT, et al. 2009. Gefitinib or carboplatin-paclitaxel in pulmonary adenocarcinoma. *N. Engl. J. Med.* 361:947–57

64. Maemondo M, Inoue A, Kobayashi K, Sugawara S, Oizumi S, et al. 2010. Gefitinib or chemotherapy for non-small-cell lung cancer with mutated *EGFR*. *N. Engl. J. Med.* 362:2380–88

65. Amberger J, Bocchini C, Hamosh A. 2011. A new face and new challenges for Online Mendelian Inheritance in Man (OMIM®). *Hum. Mutat.* 32:564–67

66. Ng SB, Turner EH, Robertson PD, Flygare SD, Bigham AW, et al. 2009. Targeted capture and massively parallel sequencing of 12 human exomes. *Nature* 461:272–76

67. Toydemir RM, Rutherford A, Whitby FG, Jorde LB, Carey JC, Bamshad MJ. 2006. Mutations in embryonic myosin heavy chain (*MYH3*) cause Freeman–Sheldon syndrome and Sheldon–Hall syndrome. *Nat. Genet.* 38:561–65

68. Ng SB, Buckingham KJ, Lee C, Bigham AW, Tabor HK, et al. 2010. Exome sequencing identifies the cause of a mendelian disorder. *Nat. Genet.* 42:30–35

69. Ng SB, Bigham AW, Buckingham KJ, Hannibal MC, McMillin MJ, et al. 2010. Exome sequencing identifies *MLL2* mutations as a cause of Kabuki syndrome. *Nat. Genet.* 42:790–93

70. Krawitz PM, Schweiger MR, Rodelsperger C, Marcelis C, Kolsch U, et al. 2010. Identity-by-descent filtering of exome sequence data identifies *PIGV* mutations in hyperphosphatasia mental retardation syndrome. *Nat. Genet.* 42:827–29

71. Becker J, Semler O, Gilissen C, Li Y, Bolz HJ, et al. 2011. Exome sequencing identifies truncating mutations in human *SERPINF1* in autosomal-recessive osteogenesis imperfecta. *Am. J. Hum. Genet.* 88:362–71

72. Byun M, Abhyankar A, Lelarge V, Plancoulaine S, Palanduz A, et al. 2010. Whole-exome sequencing–based discovery of *STIM1* deficiency in a child with fatal classic Kaposi sarcoma. *J. Exp. Med.* 207:2307–12

73. Gilissen C, Arts HH, Hoischen A, Spruijt L, Mans DA, et al. 2010. Exome sequencing identifies *WDR35* variants involved in Sensenbrenner syndrome. *Am. J. Hum. Genet.* 87:418–23

74. Pierce SB, Walsh T, Chisholm KM, Lee MK, Thornton AM, et al. 2010. Mutations in the DBP-deficiency protein HSD17B4 cause ovarian dysgenesis, hearing loss, and ataxia of Perrault syndrome. *Am. J. Hum. Genet.* 87:282–88

75. Hoischen A, van Bon BW, Gilissen C, Arts P, van Lier B, et al. 2010. De novo mutations of *SETBP1* cause Schinzel–Giedion syndrome. *Nat. Genet.* 42:483–85

76. Hoischen A, van Bon BW, Rodriguez-Santiago B, Gilissen C, Vissers LE, et al. 2011. De novo nonsense mutations in *ASXL1* cause Bohring–Opitz syndrome. *Nat. Genet.* 43:729–31

77. Albers CA, Cvejic A, Favier R, Bouwmans EE, Alessi MC, et al. 2011. Exome sequencing identifies *NBEAL2* as the causative gene for gray platelet syndrome. *Nat. Genet.* 43:735–37

78. Saitsu H, Osaka H, Sasaki M, Takanashi J, Hamada K, et al. 2011. Mutations in *POLR3A* and *POLR3B* encoding RNA polymerase III subunits cause an autosomal-recessive hypomyelinating leukoencephalopathy. *Am. J. Hum. Genet.* 89:644–51

79. Haack TB, Danhauser K, Haberberger B, Hoser J, Strecker V, et al. 2010. Exome sequencing identifies *ACAD9* mutations as a cause of complex I deficiency. *Nat. Genet.* 42:1131–34

80. Musunuru K, Pirruccello JP, Do R, Peloso GM, Guiducci C, et al. 2010. Exome sequencing, *ANGPTL3* mutations, and familial combined hypolipidemia. *N. Engl. J. Med.* 363:2220–27

81. Fujimoto K, Koishi R, Shimizugawa T, Ando Y. 2006. *Angptl3*-null mice show low plasma lipid concentrations by enhanced lipoprotein lipase activity. *Exp. Anim.* 55:27–34

82. Depienne C, Bouteiller D, Meneret A, Billot S, Groppa S, et al. 2012. *RAD51* haploinsufficiency causes congenital mirror movements in humans. *Am. J. Hum. Genet.* 90:301–7

83. Lines MA, Huang L, Schwartzentruber J, Douglas SL, Lynch DC, et al. 2012. Haploinsufficiency of a spliceosomal GTPase encoded by *EFTUD2* causes mandibulofacial dysostosis with microcephaly. *Am. J. Hum. Genet.* 90:369–77

84. Zhou C, Zang D, Jin Y, Wu H, Liu Z, et al. 2011. Mutation in ribosomal protein L21 underlies hereditary hypotrichosis simplex. *Hum. Mutat.* 32:710–14

85. Lupski JR, Reid JG, Gonzaga-Jauregui C, Rio Deiros D, Chen DC, et al. 2010. Whole-genome sequencing in a patient with Charcot–Marie–Tooth neuropathy. *N. Engl. J. Med.* 362:1181–91

86. Choi M, Scholl UI, Ji W, Liu T, Tikhonova IR, et al. 2009. Genetic diagnosis by whole exome capture and massively parallel DNA sequencing. *Proc. Natl. Acad. Sci. USA* 106:19096–101

87. Worthey EA, Mayer AN, Syverson GD, Helbling D, Bonacci BB, et al. 2011. Making a definitive diagnosis: successful clinical application of whole exome sequencing in a child with intractable inflammatory bowel disease. *Genet. Med.* 13:255–62

88. Chapman MA, Lawrence MS, Keats JJ, Cibulskis K, Sougnez C, et al. 2011. Initial genome sequencing and analysis of multiple myeloma. *Nature* 471:467–72

89. Lee W, Jiang Z, Liu J, Haverty PM, Guan Y, et al. 2010. The mutation spectrum revealed by paired genome sequences from a lung cancer patient. *Nature* 465:473–77

90. Pasqualucci L, Trifonov V, Fabbri G, Ma J, Rossi D, et al. 2011. Analysis of the coding genome of diffuse large B cell lymphoma. *Nat. Genet.* 43:830–37

91. Ley TJ, Mardis ER, Ding L, Fulton B, McLellan MD, et al. 2008. DNA sequencing of a cytogenetically normal acute myeloid leukaemia genome. *Nature* 456:66–72

92. Mardis ER, Ding L, Dooling DJ, Larson DE, McLellan MD, et al. 2009. Recurring mutations found by sequencing an acute myeloid leukemia genome. *N. Engl. J. Med.* 361:1058–66

93. Graubert TA, Mardis ER. 2011. Genomics of acute myeloid leukemia. *Cancer J.* 17:487–91

94. Zhang J, Ding L, Holmfeldt L, Wu G, Heatley SL, et al. 2012. The genetic basis of early T cell precursor acute lymphoblastic leukaemia. *Nature* 481:157–63

95. Tiacci E, Trifonov V, Schiavoni G, Holmes A, Kern W, et al. 2011. *BRAF* mutations in hairy-cell leukemia. *N. Engl. J. Med.* 364:2305–15

96. Badalian-Very G, Vergilio JA, Degar BA, MacConaill LE, Brandner B, et al. 2010. Recurrent *BRAF* mutations in Langerhans cell histiocytosis. *Blood* 116:1919–23

97. Stephens PJ, Greenman CD, Fu B, Yang F, Bignell GR, et al. 2011. Massive genomic rearrangement acquired in a single catastrophic event during cancer development. *Cell* 144:27–40

98. Magrangeas F, Avet-Loiseau H, Munshi NC, Minvielle S. 2011. Chromothripsis identifies a rare and aggressive entity among newly diagnosed multiple myeloma patients. *Blood* 118:675–78

99. Kloosterman WP, Guryev V, van Roosmalen M, Duran KJ, de Bruijn E, et al. 2011. Chromothripsis as a mechanism driving complex de novo structural rearrangements in the germline. *Hum. Mol. Genet.* 20:1916–24

100. Locasale JW, Grassian AR, Melman T, Lyssiotis CA, Mattaini KR, et al. 2011. Phosphoglycerate dehydrogenase diverts glycolytic flux and contributes to oncogenesis. *Nat. Genet.* 43:869–74

101. Parsons DW, Jones S, Zhang X, Lin JC, Leary RJ, et al. 2008. An integrated genomic analysis of human glioblastoma multiforme. *Science* 321:1807–12

102. Morin RD, Johnson NA, Severson TM, Mungall AJ, An J, et al. 2010. Somatic mutations altering EZH2 (Tyr641) in follicular and diffuse large B cell lymphomas of germinal-center origin. *Nat. Genet.* 42:181–85

103. Morin RD, Mendez-Lago M, Mungall AJ, Goya R, Mungall KL, et al. 2011. Frequent mutation of histone-modifying genes in non-Hodgkin lymphoma. *Nature* 476:298–303

104. Ernst T, Chase AJ, Score J, Hidalgo-Curtis CE, Bryant C, et al. 2010. Inactivating mutations of the histone methyltransferase gene *EZH2* in myeloid disorders. *Nat. Genet.* 42:722–26

105. Makishima H, Jankowska AM, Tiu RV, Szpurka H, Sugimoto Y, et al. 2010. Novel homo- and hemizygous mutations in *EZH2* in myeloid malignancies. *Leukemia* 24:1799–804

106. Nikoloski G, Langemeijer SM, Kuiper RP, Knops R, Massop M, et al. 2010. Somatic mutations of the histone methyltransferase gene *EZH2* in myelodysplastic syndromes. *Nat. Genet.* 42:665–67

107. Ding L, Ley TJ, Larson DE, Miller CA, Koboldt DC, et al. 2012. Clonal evolution in relapsed acute myeloid leukaemia revealed by whole-genome sequencing. *Nature* 481:506–10

108. Kassem A, Schopflin A, Diaz C, Weyers W, Stickeler E, et al. 2008. Frequent detection of Merkel cell polyomavirus in human Merkel cell carcinomas and identification of a unique deletion in the *VP1* gene. *Cancer Res.* 68:5009–13

109. Feng H, Shuda M, Chang Y, Moore PS. 2008. Clonal integration of a polyomavirus in human Merkel cell carcinoma. *Science* 319:1096–100

110. Paschka P, Schlenk RF, Gaidzik VI, Habdank M, Kronke J, et al. 2010. *IDH1* and *IDH2* mutations are frequent genetic alterations in acute myeloid leukemia and confer adverse prognosis in cytogenetically normal acute myeloid leukemia with *NPM1* mutation without *FLT3* internal tandem duplication. *J. Clin. Oncol.* 28:3636–43

111. Marcucci G, Maharry K, Wu YZ, Radmacher MD, Mrozek K, et al. 2010. *IDH1* and *IDH2* gene mutations identify novel molecular subsets within de novo cytogenetically normal acute myeloid leukemia: a Cancer and Leukemia Group B study. *J. Clin. Oncol.* 28:2348–55

112. Ley TJ, Ding L, Walter MJ, McLellan MD, Lamprecht T, et al. 2010. *DNMT3A* mutations in acute myeloid leukemia. *N. Engl. J. Med.* 363:2424–33

113. Patel JP, Gonen M, Figueroa ME, Fernandez H, Sun Z, et al. 2012. Prognostic relevance of integrated genetic profiling in acute myeloid leukemia. *N. Engl. J. Med.* 366:1079–89

114. Thol F, Kade S, Schlarmann C, Loffeld P, Morgan M, et al. 2012. Frequency and prognostic impact of mutations in *SRSF2*, *U2AF1*, and *ZRSR2* in patients with myelodysplastic syndromes. *Blood* 119:3578–84

115. Graubert TA, Shen D, Ding L, Okeyo-Owuor T, Lunn CL, et al. 2012. Recurrent mutations in the *U2AF1* splicing factor in myelodysplastic syndromes. *Nat. Genet.* 44:53–57

116. Yoshida K, Sanada M, Shiraishi Y, Nowak D, Nagata Y, et al. 2011. Frequent pathway mutations of splicing machinery in myelodysplasia. *Nature* 478:64–69

117. Gelsi-Boyer V, Trouplin V, Adelaide J, Bonansea J, Cervera N, et al. 2009. Mutations of polycomb-associated gene *ASXL1* in myelodysplastic syndromes and chronic myelomonocytic leukaemia. *Br. J. Haematol.* 145:788–800

118. Boultwood J, Perry J, Pellagatti A, Fernandez-Mercado M, Fernandez-Santamaria C, et al. 2010. Frequent mutation of the polycomb-associated gene *ASXL1* in the myelodysplastic syndromes and in acute myeloid leukemia. *Leukemia* 24:1062–65

119. Delhommeau F, Dupont S, Della Valle V, James C, Trannoy S, et al. 2009. Mutation in *TET2* in myeloid cancers. *N. Engl. J. Med.* 360:2289–301

120. Bejar R, Stevenson K, Abdel-Wahab O, Galili N, Nilsson B, et al. 2011. Clinical effect of point mutations in myelodysplastic syndromes. *N. Engl. J. Med.* 364:2496–506

121. Puente XS, Pinyol M, Quesada V, Conde L, Ordonez GR, et al. 2011. Whole-genome sequencing identifies recurrent mutations in chronic lymphocytic leukaemia. *Nature* 475:101–5

122. Wang L, Lawrence MS, Wan Y, Stojanov P, Sougnez C, et al. 2011. *SF3B1* and other novel cancer genes in chronic lymphocytic leukemia. *N. Engl. J. Med.* 365:2497–506

123. Logan AC, Gao H, Wang C, Sahaf B, Jones CD, et al. 2011. High-throughput VDJ sequencing for quantification of minimal residual disease in chronic lymphocytic leukemia and immune reconstitution assessment. *Proc. Natl. Acad. Sci. USA* 108:21194–99

124. Leary RJ, Kinde I, Diehl F, Schmidt K, Clouser C, et al. 2010. Development of personalized tumor biomarkers using massively parallel sequencing. *Sci. Transl. Med.* 2:20ra14

125. Ksiazek TG, Erdman D, Goldsmith CS, Zaki SR, Peret T, et al. 2003. A novel coronavirus associated with severe acute respiratory syndrome. *N. Engl. J. Med.* 348:1953–66

126. Yozwiak NL, Skewes-Cox P, Stenglein MD, Balmaseda A, Harris E, Derisi JL. 2012. Virus identification in unknown tropical febrile illness cases using deep sequencing. *PLoS Negl. Trop. Dis.* 6:e1485

127. Wootton SC, Kim DS, Kondoh Y, Chen E, Lee JS, et al. 2011. Viral infection in acute exacerbation of idiopathic pulmonary fibrosis. *Am. J. Respir. Crit. Care Med.* 183:1698–702

128. Yozwiak NL, Skewes-Cox P, Gordon A, Saborio S, Kuan G, et al. 2010. Human enterovirus 109: a novel interspecies recombinant enterovirus isolated from a case of acute pediatric respiratory illness in Nicaragua. *J. Virol.* 84:9047–58

129. Greninger AL, Holtz L, Kang G, Ganem D, Wang D, DeRisi JL. 2010. Serological evidence of human klassevirus infection. *Clin. Vaccine Immunol.* 17:1584–88

130. Greninger AL, Runckel C, Chiu CY, Haggerty T, Parsonnet J, et al. 2009. The complete genome of klassevirus—a novel picornavirus in pediatric stool. *Virol. J.* 6:82

131. Cheval J, Sauvage V, Frangeul L, Dacheux L, Guigon G, et al. 2011. Evaluation of high-throughput sequencing for identifying known and unknown viruses in biological samples. *J. Clin. Microbiol.* 49:3268–75

132. Nakamura S, Yang CS, Sakon N, Ueda M, Tougan T, et al. 2009. Direct metagenomic detection of viral pathogens in nasal and fecal specimens using an unbiased high-throughput sequencing approach. *PLoS ONE* 4:e4219

133. Chiu CY, Greninger AL, Chen EC, Haggerty TD, Parsonnet J, et al. 2010. Cultivation and serological characterization of a human Theiler's-like cardiovirus associated with diarrheal disease. *J. Virol.* 84:4407–14

134. Wright AM, Beres SB, Consamus EN, Long SW, Flores AR, et al. 2011. Rapidly progressive, fatal, inhalation anthrax–like infection in a human: case report, pathogen genome sequencing, pathology, and coordinated response. *Arch. Pathol. Lab. Med.* 135:1447–59

135. Kennedy AD, Otto M, Braughton KR, Whitney AR, Chen L, et al. 2008. Epidemic community-associated methicillin-resistant *Staphylococcus aureus*: recent clonal expansion and diversification. *Proc. Natl. Acad. Sci. USA* 105:1327–32

136. Chin CS, Sorenson J, Harris JB, Robins WP, Charles RC, et al. 2011. The origin of the Haitian cholera outbreak strain. *N. Engl. J. Med.* 364:33–42

137. Rasko DA, Webster DR, Sahl JW, Bashir A, Boisen N, et al. 2011. Origins of the *E. coli* strain causing an outbreak of hemolytic-uremic syndrome in Germany. *N. Engl. J. Med.* 365:709–17

138. Nelson MI, Tan Y, Ghedin E, Wentworth DE, St George K, et al. 2011. Phylogeography of the spring and fall waves of the H1N1/09 pandemic influenza virus in the United States. *J. Virol.* 85:828–34

139. Ghedin E, Laplante J, DePasse J, Wentworth DE, Santos RP, et al. 2011. Deep sequencing reveals mixed infection with 2009 pandemic influenza A (H1N1) virus strains and the emergence of oseltamivir resistance. *J. Infect. Dis.* 203:168–74

140. Lataillade M, Chiarella J, Yang R, Schnittman S, Wirtz V, et al. 2010. Prevalence and clinical significance of HIV drug resistance mutations by ultra-deep sequencing in antiretroviral-naïve subjects in the CASTLE study. *PLoS ONE* 5:e10952

141. Turnbaugh PJ, Hamady M, Yatsunenko T, Cantarel BL, Duncan A, et al. 2009. A core gut microbiome in obese and lean twins. *Nature* 457:480–4

142. Backhed F, Ley RE, Sonnenburg JL, Peterson DA, Gordon JI. 2005. Host-bacterial mutualism in the human intestine. *Science* 307:1915–20

143. Wu YC, Kipling D, Leong HS, Martin V, Ademokun AA, Dunn-Walters DK. 2010. High-throughput immunoglobulin repertoire analysis distinguishes between human IgM memory and switched memory B cell populations. *Blood* 116:1070–78

144. Venturi V, Quigley MF, Greenaway HY, Ng PC, Ende ZS, et al. 2011. A mechanism for TCR sharing between T cell subsets and individuals revealed by pyrosequencing. *J. Immunol.* 186:4285–94

145. Campbell PJ, Pleasance ED, Stephens PJ, Dicks E, Rance R, et al. 2008. Subclonal phylogenetic structures in cancer revealed by ultra-deep sequencing. *Proc. Natl. Acad. Sci. USA* 105:13081–86

146. Liao HX, Chen X, Munshaw S, Zhang R, Marshall DJ, et al. 2011. Initial antibodies binding to HIV-1 gp41 in acutely infected subjects are polyreactive and highly mutated. *J. Exp. Med.* 208:2237–49

147. Wu X, Zhou T, Zhu J, Zhang B, Georgiev I, et al. 2011. Focused evolution of HIV-1 neutralizing antibodies revealed by structures and deep sequencing. *Science* 333:1593–602

148. Snyder TM, Khush KK, Valantine HA, Quake SR. 2011. Universal noninvasive detection of solid organ transplant rejection. *Proc. Natl. Acad. Sci. USA* 108:6229–34

149. Fan HC, Quake SR. 2010. Sensitivity of noninvasive prenatal detection of fetal aneuploidy from maternal plasma using shotgun sequencing is limited only by counting statistics. *PLoS ONE* 5:e10439

150. Fan HC, Blumenfeld YJ, Chitkara U, Hudgins L, Quake SR. 2010. Analysis of the size distributions of fetal and maternal cell–free DNA by paired-end sequencing. *Clin. Chem.* 56:1279–86

151. Fan HC, Blumenfeld YJ, Chitkara U, Hudgins L, Quake SR. 2008. Noninvasive diagnosis of fetal aneuploidy by shotgun sequencing DNA from maternal blood. *Proc. Natl. Acad. Sci. USA* 105:16266–71

152. Lo YM, Tein MS, Lau TK, Haines CJ, Leung TN, et al. 1998. Quantitative analysis of fetal DNA in maternal plasma and serum: implications for noninvasive prenatal diagnosis. *Am. J. Hum. Genet.* 62:768–75

153. Chiu RW, Akolekar R, Zheng YW, Leung TY, Sun H, et al. 2011. Non-invasive prenatal assessment of trisomy 21 by multiplexed maternal plasma DNA sequencing: large scale validity study. *Br. Med. J.* 342:c7401

154. Kitzman JO, Snyder MW, Ventura M, Lewis AP, Qiu R, et al. 2012. Noninvasive whole-genome sequencing of a human fetus. *Sci. Transl. Med.* 4:137ra76

155. Fan HC, Gu W, Wang J, Blumenfeld YJ, El-Sayed YY, Quake SR. 2012. Non-invasive prenatal measurement of the fetal genome. *Nature* 487:320–24

156. Shen Y, Zhu YM, Fan X, Shi JY, Wang QR, et al. 2011. Gene mutation patterns and their prognostic impact in a cohort of 1,185 patients with acute myeloid leukemia. *Blood* 118:5593–603

Annual Review of Pathology: Mechanisms of Disease

Volume 8, 2013

Contents

Indexes

Errata

An online log of corrections to *Annual Review of Pathology: Mechanisms of Disease* articles may be found at http://pathol.annualreviews.org