We are in a time of great change in genetics that may dramatically impact human biology and medicine. The completion of the human genome project,1,2 the development of low-cost, high-throughput parallel sequencing technology, and large-scale studies of genetic variation3 have provided a rich set of techniques and data for the study of genetic disease risk, treatment response, population diversity, and human evolution. Newly developed sequencing instruments now generate hundreds of millions to billions of short sequences per run, allowing for rapid complete sequencing of human genomes. These technological advances have facilitated a precipitous drop (Figure 1) in the cost per base pair of DNA sequenced. To capitalize on the potential of these technologies for research and clinical applications, translational scientists and clinicians must become familiar with a continuously evolving field. In this review, we will provide a historical perspective on human genome sequencing, summarize current and future sequencing technologies, highlight issues related to data management and interpretation, and finally consider research and clinical applications of high-throughput sequencing, with specific emphasis on cardiovascular disease.

**Historical Perspective**

Genome sequencing has become synonymous with high-throughput sequencing, but it is instructive to revisit historical milestones. Although James Watson and Francis Crick published the first description of the crystallographic double-helix DNA structure in 1953,4 it was not until 2 decades later, with the nearly simultaneous development of Maxam-Gilbert and Sanger sequencing,5,6 that DNA sequencing became widely available to the research community. The Sanger method, which is based on DNA chain termination with a small concentration of radio- or fluorescently labeled dideoxy nucleotide triphosphate (dNTP) molecules followed by size separation by gel electrophoresis, became the research and commercial standard because of technical ease and reliability of results. This was the standard sequencing technology for >3 decades and remains the method of choice for sequencing short segments of DNA and confirming genotypes from other technologies. Sanger sequencing, in conjunction with several methods for identifying the approximate genetic locations (loci) harboring variations in DNA associated with disease, was the method used to define the basis of many Mendelian, or single-gene disorders.

More recently, a modified Sanger approach was the main sequencing engine for the first draft human genome sequence, which was produced by sequencing 500- to 600-bp segments of DNA in parallel (shotgun sequencing) and assembly of these sequence fragments into contiguous stretches of DNA (contigs) based on sequence overlap.3,2,7 Two sequences were released nearly simultaneously; the first was a product of the decade-long publicly funded Human Genome Project,4 and the second was released by the Celera Corporation, led by Craig Venter and colleagues.7 The accuracy and read lengths generated by this technology were advantageous to sequencing projects in which no template or reference sequence was available, but the sequence time (years) and cost (estimated at between 300 million and 3 billion dollars) of these early efforts precluded the use of this technology for large-scale human genome sequencing. However, the completion of a human genome reference sequence allowed for the development of a next generation of sequencing instruments that substantially reduced DNA sequencing time and cost.

**Next-Generation Sequencing Technologies**

The development of a draft human genome sequence, which has subsequently been revised to constitute a reference human genome sequence, facilitated the development of next-generation sequencing (NGS). NGS is a broad term that refers to a set of methods for (1) genomic template preparation, or the methodology for processing genomic DNA for downstream sequencing; (2) near-simultaneous, or massively parallel, generation of millions to billions of short sequence reads; (3) alignment of sequence reads to a reference sequence; (4) sequence assembly from aligned sequence reads and genetic variant discovery (Figure 2). Most investigators use the output from this final step, a list of genotypes for positions with at least 1 allele that differs from a reference sequence (variants) in all downstream analysis. Thus, whole genome sequence data generally refers not to ∼3 billion diploid genotypes that cover the known chromosomal posi-

From the Center for Inherited Cardiovascular Disease, Division of Cardiovascular Medicine, Stanford University (F.E.D., S.P., M.T.W., E.A.A.); and Department of Bioengineering, Stanford University, Stanford, CA (S.R.Q.).

Correspondence to Euan A. Ashley, Stanford University School of Medicine, Falk CVRB, 300 Pasteur Dr, Stanford, CA 94305. E-mail euan@stanford.edu

(Circulation. 2012;125:931-944.)

© 2012 American Heart Association, Inc.

_Circulation is available at http://circ.ahajournals.org_

DOI: 10.1161/CIRCULATIONAHA.110.972828
tions, but the 3 to 4 million genotypes in each genome that differ from the reference sequence. Several NGS technologies exist that differ primarily in methods for clonal amplification of short fragments of DNA and sequencing the resulting short DNA fragments. Each has specific advantages in terms of read length, accuracy, and throughput (Table 1). All currently forego the time-consuming bacterial cloning step that was used for library preparation in the Human Genome Project. For full details of the technical aspects of each sequencing technology, we refer the reader to recent technological reviews.8,9 We will briefly review each technology here with a focus on advantages, disadvantages, and specific sequencing applications for each platform. One issue that deserves specific mention is that of read length. Shorter sequence reads (100 bp or shorter) are well suited to the biochemical reactions used by most of the sequencing technologies. However, the generation of short reads complicates sequence assembly, particularly in repetitive regions of the genome. The generation of longer sequence reads (1000 bp or longer) simplifies this task. Furthermore, the use of longer sequence reads spanning several variants aids in resolution of haplotype phase, which is the assignment of each allele in a heterozygous genotype to 1 chromosome of each homologous pair, eg, the assignment of an A allele in a A/G genotype to a paternally derived segment of chromosome 13.

Of the NGS platforms that are currently commercially available, the 454 (454 Life Sciences/Roche) instrument was developed first. This platform is based on pyrosequencing, which detects light emitted by secondary reactions initiated by the release of pyrophosphate during nucleotide incorporation.10 Advantages include long reads and facile mate-pair sequencing, a method that sequences both ends of a previously circularized DNA molecule. Pairing reads that span tens of kilobases of genomic template sequence further facilitates haplotype phasing and the identification of structural genetic variation such as deletions and insertions of large segments of DNA. Disadvantages include systematic errors in reading frame (“frameshift errors”) in certain circumstances and lower throughput and higher sequencing costs than other commercial technologies.

SOLiD (Applied Biosystems by Life Technologies) sequencing utilizes sequencing-by-ligation in which the sequence of a DNA template is read by competitive ligation of 2-base probes to the nascent DNA strand.11 Advantages include throughput (~20–30 Gbp per run), and base-level error information encoded in the 2-base sequences, both of which make the platform suitable for human whole genome and exome variant discovery. The main disadvantage is the necessity to work with unconventional data formats for sequence reads and the reference genome.

The Illumina/Solexa (Illumina, Inc) platform is widely used for a variety of applications, including human whole genome and exome variant discovery and transcriptome sequencing (“RNAseq”), by virtue of easily prepared paired-end sequencing libraries, high throughput, and ease of analysis of its short-read information. After genomic DNA isolation, fragmentation, and several enzymatic modification steps, sequencing libraries are amplified from single DNA strands on glass surfaces. The resultant templates are sequenced by the use of an approach in which fluorescently labeled end-blocked nucleotides, which do not allow further DNA polymerization, are incorporated by DNA polymerase, the base-specific fluorescent color is detected via fluorescence imaging, the end block and fluorescent tag is enzymat-
 życiakleaved, and the process is repeated following image storage, yielding image-encoded nucleotide sequences. Drawbacks include comparatively short sequence reads (≤100 bp) and practical limits to insert sizes for paired-end sequencing.

Complete Genomics, Inc provides a sequencing service, in contrast to other companies that have primarily focused on providing sequencing instruments, that is, targeted solely toward human whole genomes. The instrument uses sequencing-by-ligation of hundreds of DNA nano-balls, or chained-replicates of 70-bp sequences of sheared genomic DNA modified by adaptor inserts. Theoretical throughput exceeds that of any of the NGS technologies described thus far.

### Third-Generation Sequencing Technologies

A third generation of sequencing instruments has been developed that is defined by the lack of DNA or RNA amplification in template library preparation (single-molecule sequencing, Figure 2). By foregoing this step, these technologies require less genomic DNA, avoid polymerase chain reaction–introduced error and amplification bias, and may be superior for high-throughput sequencing applications, such as transcriptome sequencing (RNAseq), that depend on accurate quantification of relative DNA or RNA fragment abundance.

The first of these single-molecule sequencing technologies is the Helicos Heliscope (Helicos BioSciences). The specific
Helicos chemistry is based on single-molecule sequencing by cyclic reversible terminator nucleotide incorporation. A single dye molecule is used to label the dNTPs and fluorescence microscopy is used to image the dye in sequencing reactions performed on single-molecule templates on solid support. The order in which each fluorescently labeled dNTP is added to the sequencing reactor determines the base sequence at that position. Notably, the instrument is also suitable for direct RNA sequencing without conversion to cDNA, thus avoiding error and copy number bias associated with reverse transcription.

Pacific Biosciences have recently developed a method for imaging individual DNA polymerase molecules as they synthesize a nascent DNA molecule covalently attached to solid support. Advantages include read information that is theoretically as long as 1 kb or longer and real-time sequencing kinetics that reflect nucleotide methylation state and DNA secondary structure.

Life Technology’s Ion Torrent device is targeted toward individual laboratories interested in a small-footprint, medium-throughput sequencing platform. This sequencing engine is based on detection of hydrogen ions released from nucleotides incorporated into the growing DNA strand. This signal is detected in a solid-state semiconductor akin to a miniaturized pH meter, and the technology is theoretically suitable to single-molecule sequencing. Throughput is currently low (<1 Gbp per run), but the release of higher-density chips has made sequencing of transcriptomes and exomes feasible.

Nanopore sequencing technologies detect base-specific changes in ionic flux as DNA traverses small pores in solid surfaces that are placed in an electric field. Advantages to this method include theoretically unparalleled sequencing speed and minimal template preparation. At this point, however, detection speed and accuracy remain significant technological hurdles, because the transit speed of nucleic acids through nanopores in even minimal electric fields is several orders or magnitude higher than the highest detection frequency. Several enzymatic methods have been developed.

Table 1. Sequencing Platform Comparison

| Platform                        | Amplification                  | Sequencing                           | Detection                                                      | Read Length   | Output per Run | Run Time |
|---------------------------------|--------------------------------|--------------------------------------|                                                               |              |               |          |
| **Second-generation sequencing platforms** |                                |                                      |                                                               |              |               |          |
| 454                             | Emulsion PCR on beads          | Unlabeled nucleotide incorporation   | Detection of light emitted by release of PPi                 | Variable (400 bp for single end sequencing) | 400–600 Mbp  | 10 h     |
| SOLID                           | Emulsion PCR on beads          | Ligation of 2-base encoded fluorescent oligonucleotides | Fluorescence emission from labeled oligonucleotides          | 75–35 bp     | 20–30 Gbp     | 7 d      |
| Illumina                        | Array-based enzymatic amplification | Fluorescently labeled end-blocked nucleotide incorporation | Fluorescence emission from nucleotides | 2×100 bp     | 100–200 Gbp   | 8 d      |
| Complete                        | Rolling-circle replication of short segments of DNA into nanoballs | Ligation of fluorescently labeled oligonucleotide probes | Fluorescence emission from oligonucleotide probes          | 2×35 bp      | 20–60 Gbp    | 12 d     |
| **Third-generation sequencing platforms** |                                |                                      |                                                               |              |               |          |
| Helicos                         | NA                             | Single dye-labeled nucleotides are added sequentially and incorporated by polymerases by use of single DNA molecular templates | Microscopy of fluorescently labeled nucleotides | 2×25–55 bp   | 21–35 Gbp     | 8 d      |
| Pacific Biosciences             | NA                             | Incorporation of fluorescently labeled nucleotides by polymerases on solid support | Zero-mode waveguide imaging of fluorescent nucleotide incorporation by individual polymerases | 2×1000 bp    | 75–100 Mbp (projected); 5–10 Mbp (actual usable sequence) | 30 min   |
| Oxford nanopore                 | NA                             | Processive endo- or exonuclease activity feeds individual bases or whole DNA strands through protein or solid-state nanopores | Current disruption across nanopore corresponds to nucleotide structure | Variable     | Variable      | Variable |
| Ion Torrent                     | Variable                       | DNA polymerase incorporation of unlabeled nucleotides added sequentially to solid-state microwells | Solid-state detection of hydrogen ions released by nucleotide incorporation | 200 bp       | 10 Mbp to 1 Gbp | 2 h      |

bp indicates base pair; Gbp, one billion base pairs; Mbp, one million base pairs; PPi, pyrophosphate; and PCR, polymerase chain reaction.
to slow transit time and facilitate detection of changes in ionic flux.20,21

**Processing High-Throughput Sequence Data**

Data generation from high-throughput sequencing is becoming less expensive and time consuming. Generating sequence data, however, is only the first step in extracting usable information from high-throughput sequencing. For output from most currently available sequencing platforms, several tasks must be performed before downstream analysis: (1) short-read mapping, or alignment of each sequence read to a reference genome to identify the genomic sequence represented by the short read; (2) base calling at every genomic position covered by aligned short reads; and (3) identification of sequence variation from the reference genome. The percentage of base positions that are read by properly aligned short reads is described by coverage. The number of times that a single base position is read by short-read sequences is termed “depth of coverage” and most investigators currently consider 30-fold ($30 	imes$) average depth of coverage as a benchmark for high-quality genome sequence data. Before discussing these data management issues, it is worth highlighting some of the limitations of the current approach that utilizes a haploid reference sequence, ie, a sequence that has only 1 base for every genomic position.

**The Human Reference Genome and Its Limitations**

The human reference genome currently used for short-read alignment and variant calling (National Center for Biotechnology Information [NCBI] reference genome22) is derived from a collection of DNA samples from a small number of anonymous donors. It is currently the only finished-grade human genome in that it was assembled de novo from long sequence reads and covers ≈99% of known chromosomal positions with high fidelity. However, it represents a very small sampling of human genetic variation. Analysis in our laboratory using the 1000 genomes population variation data demonstrated that at ≈1.6 million genomic positions, the NCBI reference sequence differed from the major, or most frequent, allele in each of the 3 HapMap populations, including ≈800,000 positions at which all 3 population groups have major alleles that differ from the NCBI reference allele.23 In addition, the reference sequence contains thousands of common and rare disease risk alleles, including >20 rare disease susceptibility alleles such as the Factor V Leiden allele associated with hereditary thrombophilia.23,24 Various approaches to addressing these issues have been suggested, including the use of a “major allele” reference sequence. We have recently used this approach to identify the putative genetic basis for familial thrombophilia in a family quartet by use of whole genome sequencing.25 Notably, the multigenic risk for this trait we identified included the Factor V allele conferring activated protein C resistance, which would not have been identified in homozygous state by using the NCBI reference genome for variant identification.

**Aligning Sequence Reads to the Human Reference Genome**

There are several programs for mapping short reads to a reference genome; for an in-depth comparison of alignment programs, we direct the reader to a recent work by Li and Homer.25 Historically, mapping alignment with quality was the most widely used alignment algorithm,26 but this algorithm has been supplanted by other open-source solutions that are superior for longer (>35 bp) sequence reads. Although several alignment algorithms can be run on high-memory multiple core desktops and even laptops, parallel computing architecture, which utilizes multiple processors to perform alignment tasks simultaneously, reduces the time required for alignment severalfold. Unfortunately, few individual laboratories currently are able to provide this computing power. One solution is on-demand distributed or parallel computing architecture, ie, cloud computing. This approach is economical in the sense that elastic parallel computing environments allow users to select and utilize only processing and storage capacity necessary for current tasks.

**Identifying Single-Nucleotide Variants and Small Insertions/Deletions**

Following alignment to the reference genome, sequence reads are compared at every genomic position, producing a base call for each chromosomal position. For in-depth discussion of genotype calling from next-generation sequence data, including the use of linkage disequilibrium for genotype determination and probabilistic genotypes for low- and intermediate-coverage sequencing, such as that used in the 1000 genomes project, we direct the reader to a recent work by Nielsen et al.27 A variety of different algorithms incorporate base quality, which specifies the confidence of each base call within the individual short reads, mapping quality, or confidence of accurate mapping of each short read to the specified genomic locus, and the number of bases contributing to each of the possible 16 genotypes at a position, into a probabilistic score for genotypes at every chromosomal location. The most likely genotype is compared with the reference sequence, and, typically, only positions containing at least 1 base differing from the reference sequence are retained for downstream analysis. This fact has several important implications. First, the reference base is crucial to the identification of genetic variation: if the haploid reference base harbors the same allele predisposing to disease as the subject being sequenced, it will not appear in the variant list, potentially leading to underestimation of the burden of certain disease-associated alleles. Second, comparison between individuals, eg, in cosegregation and linkage studies, can be complicated by the degree of overlap between genetic variant sets such that the assumption of homozygous reference allele calls can bias exploratory studies for causative variants. Several variant calling solutions, notably, SAMtools28 and the Genome Analysis Toolkit29 have base calling algorithms that facilitate cohort-wide variant identification, which addresses this problem. Third, the reference sequence represents a small sampling of human genetic variation, and, as large-scale sequencing efforts are undertaken, ethnicity-specific major allele differences may impact alignment of
short reads against the current reference genome and subsequent variant identification.

**Identifying Large Structural Variants**

Large structural rearrangements >1 kb, termed structural variants (SVs), encompass large deletions, duplications, insertions, inversions, and transposons. Largely ignored in many early sequencing efforts, emerging evidence suggests that these SVs are strongly associated with several Mendelian and complex diseases, including familial dilated cardiomyopathy, autism spectrum disorders, idiopathic mental retardation, schizophrenia, and Crohn’s disease.30–35 In some cases, these large genetic variants underlie >15% of disease diagnoses.36 Several methods have been developed for identification of SVs, but 3 main methods have generally been accepted and are used for identification of specific types of SVs. A complementary, hybridization-based method for identifying SVs, comparative genomic hybridization, will not be discussed further here. Notably, however, because of high false-positive rates for SV detection with the use of high-throughput sequencing, this and other polymerase chain reaction–based methods are often used to confirm candidate SVs.

The first method for identification of SVs is mate-pair sequencing,37 which is based on sequencing 2 ends of a DNA molecule following circularization, providing paired short-read sequence information separated by hundreds to thousands of base pairs. A related technique, paired-end sequencing, is used routinely in most commercial sequencing technologies to provide paired short-sequence reads from each end of an amplified linear DNA molecule. Comparison of median insert size and orientation from paired-end reads to homologous chromosomal segments in the reference genome is used to identify structural rearrangements.38 Although sensitive for inversions and other copy neutral SVs, or SVs that do not change the copy number of the affected chromosomal region, and somewhat well suited to identifying start and end points of SVs (breakpoints), detection scope is limited by the size of the insert, in that only structural rearrangements spanned by the insert can be detected.

A second method for identification of structural rearrangements is based on regional variation in read depth, which in turn depends on copy number of the genomic region interrogated. Several methods have been developed for identification of significant differences in read depth in genomic regions relative to median read depth.39–43 This method for identification of SVs is ideally suited for identification of large insertions and deletions, but has limited capability to resolve breakpoints, and cannot distinguish copy neutral SVs from normal sequence.

The third method for identification of large SVs is split-read mapping, which is based on mapping elements with inserts in the reference genome or the sample genome to contiguous short-read sequences by using 1 end of the read as an anchor and the other end to search for possible breakpoints, yielding single-nucleotide level breakpoint resolution and novel sequence discovery in some cases.44 Finally, candidate SVs are often compared with known SVs identified by using population-scale sequencing or genotyping to provide probabilities of false discovery and improved breakpoint resolution.45

**Variant Quality Control and Genotype Validation**

Validation of sequence data has become a particularly difficult problem in interpretation of genetic variants discovered via high-throughput sequencing. Per genotype error rates for commercially available high-throughput sequencing technologies achieving an average depth of coverage of >30 times are currently between 1 in every 1000 to 1 in every 100 000 bases. By comparison, per genotype error rates for Sanger sequencing, the current standard for clinical applications, is between 1 in 100 000 and 1 in 1 000 000 bp. Filtering variants via a combination of quality score metrics for individual short reads and final genotypes can minimize errors. Roach et al46 and our group have demonstrated that leveraging family genotype information can also be useful for error identification, in that pedigree-based allele inheritance analysis can be used to identify not only inconsistencies with Mendel’s laws of inheritance, but also regions in which short reads have been incorrectly mapped or genotyped. We have recently demonstrated a >90% reduction in the error rate by sequestering variants identified in these regions.23

Despite these and other advances in error reduction, however, high-throughput sequencing platforms do not yet provide the level of confidence about individual variants that would be required for routine incorporation into clinical care. To date, clinically important variants have mostly been resequenced by using Sanger-based chemistry or confirmed with oligonucleotide genotyping arrays. Both approaches are time and resource intensive. Alternative capture-based approaches, in which either a standard commercial or custom oligonucleotide set is used to select genomic regions of interest for high-coverage, high-throughput resequencing, are also costly and time consuming. Validation of small SVs such as insertions and deletions is even more difficult, often requiring cloning of single chromosomal segments before resequencing. Until the accuracy of high-throughput sequencing improves such that primary data do not require orthogonal confirmation, data validation will continue to be a major barrier to widespread incorporation of high-throughput sequence data into clinical applications.

**Haplotype Phasing With Use of High-Throughput Sequence Data**

Resolution of haplotype phase is important to understanding shared disease-associate chromosomal segments containing variants that tend to be inherited en bloc, compound heterozygous (2 or more risk alleles in 1 gene) and oligogenic (2 or more risk alleles in multiple genes) genotype-phenotype associations, regulatory effects of genetic variation, and differential parent of origin effects in disease association studies.47 Furthermore, large databases of phased-sequence data will be important resources for genome-wide association studies that utilize imputation, or estimation of genotypes not assayed by other technologies such as chip-based genotyping. This practice has become commonplace as investigators combine datasets to improve power to detect disease associ-
ations of small magnitude, and will be important for investigating rare variant effects. Short-read, high-throughput sequence data alone do not provide information about haplotype phase. However, several statistical algorithms based on pedigree information, common population haplotypes, and paired short reads have been developed that are applicable to high-throughput sequence data.\(^{47-50}\) Moreover, several investigators have developed experimental methods for haplotype phasing based on sorting individual metaphase chromosomes and subsequent sequencing.\(^{51}\) or from a combination of long-insert cloning and next-generation sequencing.\(^{52}\) Further development of these methods will be critical to the use of this tool for investigating disease biology.

### High-Throughput Sequencing and Mendelian Disease Genetics

The utility of high-throughput sequencing for investigation of disease genetics is great. The application of NGS for the identification of cardiovascular disease–associated loci has resulted in several notable successes, including the identification of \(BAG3\) mutations as a cause of dilated cardiomyopathy, mutations in \(SMAD3\) associated with familial aortic aneurysms, and \(AARS2\) and \(ACAD9\) in familial mitochondrial cardiomyopathy.\(^{35,53-55}\) (Table 2). These studies have provided intriguing hypotheses for follow-up work characterizing novel pathways in human cardiovascular disease. The genetic basis for several noncardiovascular diseases has similarly been explored by using exome and whole genome sequencing (Table 3). It is noteworthy that 2 studies have demonstrated the promise of NGS in aiding clinical diagnosis and management. Choi et al.\(^{61}\) used exome sequencing to identify a mutation in \(SLC26A3\) in a patient with the suspected renal salt-wasting Bartter syndrome; this finding allowed them to make the unanticipated diagnosis of congenital chloride diarrhea and modify clinical care accordingly. Worthey et al.\(^{66}\) used exome sequencing to identify a missense mutation in the gene \(XIAP\) in a patient with intractable Crohn’s-like inflammatory bowel disease, establishing a diagnosis of X-linked inhibitor of apoptosis deficiency. Subsequent allogeneic stem cell transplant resulted in dramatic improvement in the patient’s gastrointestinal disease.

<table>
<thead>
<tr>
<th>Disease</th>
<th>Inheritance Model</th>
<th>Sequenced Subjects</th>
<th>Putative Loci Identified</th>
<th>Validation</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Complex I deficiency with hypertrophic cardiomyopathy</td>
<td>Autosomal recessive</td>
<td>One proband</td>
<td>(ACAD9)</td>
<td>Wild-type cDNA complementation in fibroblasts; compound heterozygous or homozygous mutations found in (ACAD9) in 120 index cases with complex I deficiency</td>
<td>54</td>
</tr>
<tr>
<td>Dilated cardiomyopathy</td>
<td>Autosomal dominant</td>
<td>Four affected family members</td>
<td>(BAG3)</td>
<td>Seven structural and single-nucleotide variants found in (BAG3) in 311 unrelated probands; knockdown of (bag3) in zebra fish recapitulated the phenotype</td>
<td>35</td>
</tr>
<tr>
<td>Familial thoracic aortic aneurysm</td>
<td>Autosomal dominant</td>
<td>Two distantly-related affected individuals in one family</td>
<td>(SMAD3)</td>
<td>Cosegregation in family; sequencing of 181 additional probands identified 3 additional (SMAD3) mutations in 4 families</td>
<td>53</td>
</tr>
<tr>
<td>Infantile mitochondrial cardiomyopathy</td>
<td>Autosomal recessive</td>
<td>One proband</td>
<td>(AARS2)</td>
<td>Cosegregation of mutation in a separate family; metabolomic analysis of postmortem heart and skeletal muscle of proband demonstrating increased alanine levels</td>
<td>55</td>
</tr>
<tr>
<td>Kabuki syndrome</td>
<td>Autosomal dominant</td>
<td>Ten unrelated cases</td>
<td>(MLL2)</td>
<td>Sanger sequencing confirmation in 26 of 43 additional cases</td>
<td>56</td>
</tr>
<tr>
<td>Moyamoya disease</td>
<td>Complex</td>
<td>Index case from each of 8 families</td>
<td>(RNF213)</td>
<td>Genome wide linkage analysis; combination of linkage analysis and sequencing of (RNF213) in 42 index cases; case-control study in 958 subjects of East Asian ancestry</td>
<td>57</td>
</tr>
<tr>
<td>Sick sinus syndrome</td>
<td>Complex</td>
<td>Seven individuals with SSS and the rs28730774[T] variant associated with SSS; 80 individuals without SSS</td>
<td>(MYH6)</td>
<td>Genotyping in 469 SSS cases and 1185 controls</td>
<td>58</td>
</tr>
</tbody>
</table>

SSS indicates sick sinus syndrome.
<table>
<thead>
<tr>
<th>Disease</th>
<th>Inheritance Model</th>
<th>Sequenced Subjects</th>
<th>Putative Loci Identified</th>
<th>Validation</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Exome sequencing projects</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Carnevale, Malpuech, Michel, and oculoskeletal-abdominal syndromes</td>
<td>Autosomal recessive</td>
<td>One proband</td>
<td>MASP1</td>
<td>Cosegregation in 2 additional families</td>
<td>59</td>
</tr>
<tr>
<td>Charcot-Marie-Tooth neuropathy</td>
<td>Autosomal recessive</td>
<td>Two affected family members</td>
<td>GJB1</td>
<td>Co-segregation in the family</td>
<td>60</td>
</tr>
<tr>
<td>Congenital chloride losing diarrhea*</td>
<td>Autosomal recessive</td>
<td>One proband with suspected Bartter syndrome</td>
<td>SLC26A3</td>
<td>Sanger sequencing identification of homozygous variants in SLC26A3 in 5 of 39 unrelated patients with suspected Bartter syndrome; clinical follow-up demonstrating evidence of chloride losing diarrhea</td>
<td>61</td>
</tr>
<tr>
<td>FADD deficiency</td>
<td>Autosomal recessive</td>
<td>One proband with biological features of ALPS</td>
<td>FADD</td>
<td>Cosegregation in the family; in vitro assay of FADD protein levels and apoptotic activity</td>
<td>62</td>
</tr>
<tr>
<td>Familial amyotrophic lateral sclerosis</td>
<td>Autosomal dominant</td>
<td>Two affected individuals</td>
<td>VCP</td>
<td>Sanger sequencing of a cohort of 210 familial ALS cases</td>
<td>63</td>
</tr>
<tr>
<td>Familial combined hypolipidemia</td>
<td>Autosomal recessive</td>
<td>Two affected siblings</td>
<td>ANGPTL3</td>
<td>Cosegregation in family and frameshift mutations at same locus associated with phenotype in previous work</td>
<td>64</td>
</tr>
<tr>
<td>Fowler syndrome</td>
<td>Autosomal recessive</td>
<td>Two unrelated cases</td>
<td>FLVCR2</td>
<td>Previous reports of cosegregation in 1 unrelated family</td>
<td>65</td>
</tr>
<tr>
<td>Intractable inflammatory bowel disease; X-linked inhibitor of apoptosis deficiency*</td>
<td>X-linked recessive</td>
<td>One proband</td>
<td>XIAP</td>
<td>Functional assay of PBMCs recapitulating XIAP deficiency and impaired immune reactivity. Bone marrow transplant of affected child improved disease</td>
<td>66</td>
</tr>
<tr>
<td>Joubert syndrome 2</td>
<td>Autosomal recessive</td>
<td>Proband and mother</td>
<td>TMEM216</td>
<td>Parallel linkage mapping and candidate resequencing in 13 cases from 8 kindreds</td>
<td>67</td>
</tr>
<tr>
<td>Mental retardation</td>
<td>Autosomal recessive</td>
<td>Two unaffected parents of 5 affected siblings</td>
<td>TECR</td>
<td>Cosegregation in family</td>
<td>68</td>
</tr>
<tr>
<td>Mental retardation</td>
<td>Sporadic</td>
<td>Ten unrelated cases and their unaffected parents</td>
<td>Several</td>
<td>Previous functional evidence suggests a role for the gene in mental retardation</td>
<td>69</td>
</tr>
<tr>
<td>Miller syndrome</td>
<td>Autosomal recessive</td>
<td>Four affected individuals in 3 kindreds</td>
<td>DHODH</td>
<td>Cosegregation by Sanger sequencing in 3 separate families</td>
<td>70</td>
</tr>
<tr>
<td>Neonatal diabetes mellitus</td>
<td>Autosomal dominant</td>
<td>One proband</td>
<td>ABCC8</td>
<td>Absence of mutation in healthy controls</td>
<td>71</td>
</tr>
<tr>
<td>Nonsyndromic hearing loss (DFNB82)</td>
<td>Autosomal recessive</td>
<td>Single proband</td>
<td>GPSM2</td>
<td>Cosegregation in family, not found in 192 controls and 192 unrelated cases</td>
<td>72</td>
</tr>
<tr>
<td>Primary lymphedema</td>
<td>Autosomal dominant</td>
<td>One proband</td>
<td>GJC2</td>
<td>Cosegregation by Sanger sequencing in 4 additional families</td>
<td>73</td>
</tr>
<tr>
<td>Schinzel-Giedion syndrome</td>
<td>Sporadic</td>
<td>Four unrelated cases</td>
<td>SETBP1</td>
<td>Sanger sequencing of 9 unrelated cases and 188 controls</td>
<td>74</td>
</tr>
<tr>
<td>Seckel syndrome</td>
<td>Autosomal recessive</td>
<td>Single proband</td>
<td>CEP152</td>
<td>Parallel linkage analysis and candidate region sequencing in a separate family; morphological analysis of CEP152-deficient mitotic cells</td>
<td>75</td>
</tr>
<tr>
<td>Spinocerebellar ataxia</td>
<td>Autosomal dominant</td>
<td>Four affected family members</td>
<td>TGM6</td>
<td>Cosegregation of mutations in the same gene in a second family</td>
<td>76</td>
</tr>
</tbody>
</table>

(Continued)
subsequently, these repositories now contain a small but definite subset of putative variants that are actually sequencing errors. Filtering of variants in cosegregation studies by their presence in these databases may thus lead to the misidentification of damaging mutations as benign polymorphisms. Further annotation of variants by identity-by-descent status, which seeks to identify common ancestral disease-associated haplotypes, represents an evolution of the cosegregation approach.80,81

### High-Throughput Sequencing and Complex Disease Genetics

More recently, there has been increasing interest in the use of high-throughput sequencing for association analysis with complex disease. Much of the focus has been on discovering the source of “missing heritability” of common complex diseases. Six years has elapsed since the first publication of a genome-wide association study of common genetic variants and common disease.82 Since then, hundreds of highly statistically significant, replicated associations with common disease have been found. However, most alleles identified via this technique confer modest risk, and the heritability of common disease explained by these alleles in isolation or aggregate is low.83 One of the hypotheses for this relative paucity of high-effect associations is that rare variants of large effect contribute in aggregate to common disease. By virtue of their rarity, these variants have not been included on current genotyping arrays, and, therefore, previous genome-wide association studies have been unable to assess their association with disease. Furthermore, several investigators have hypothesized that some of the modest associations between common variants and common disease are mediated via weak-linkage disequilibrium between common marker variants and rare, causative variants of large effect.84 Recently, Stefansson et al85 used a combination of large-scale chip-based genotyping and intermediate-depth (10X) whole genome sequencing of a smaller cohort of cases and controls to identify a rare variant in a novel locus strongly associated with sick sinus syndrome. Importantly, this is the first demonstration of the use of whole genome sequencing to identify an association between a rare variant and complex disease. Although it is not yet cost-effective to perform deep whole genome sequencing of large cohorts of individuals with common disease, as sequencing costs drop, genotype-phenotype association studies using whole genome sequencing may become feasible. Meanwhile, several efforts are currently underway to identify coding variants associated with complex phenotypes using exome sequencing.

An important advantage of whole genome over exome association studies is the ability to interrogate the noncoding genome. Despite systematic overrepresentation of protein coding regions on many genotyping arrays, 88% of significant genome-wide associations are located in intronic or intergenic regions.85 Thus, exome-targeted sequencing approaches are likely to miss the majority of significant genome-wide associations with common disease. For Mendelian disorders, the majority of underlying allelic variants identified thus far disrupt coding regions, and thus many early sequencing efforts have focused on the exome. However, it is likely that more comprehensive variant discovery will be required for discovery of many genotype-phenotype associations.

### Genome Sequencing and the Clinic

Applying the bulk of genetic predictive information to whole genome sequence data from individuals is one of the most difficult tasks in NGS data interpretation. We previously developed and applied a methodology for interpretation of genetic and environmental risk in a single subject by use of a combination of traditional clinical assessment, whole genome sequencing, and integration of genetic and environmental risk factors,86 and have recently done so for a family quartet.23 A similar approach has been applied to carrier testing for severe recessive childhood disease risk by use of NGS and for detection of fetal aneuploidy via sequencing of maternal blood samples.87,88 One of the main challenges to the widespread application of these analytic schemes is incomplete and inconsistent status of publicly available genome annotation databases. Several annotation sources exist for gene regions, including the consensus coding sequence database,89

---

**Table 3. Continued**

<table>
<thead>
<tr>
<th>Disease</th>
<th>Inheritance Model</th>
<th>Sequenced Subjects</th>
<th>Putative Loci Identified</th>
<th>Validation</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Charcot-Marie-Tooth neuropathy</td>
<td>Autosomal recessive</td>
<td>One proband</td>
<td>SH3TC2</td>
<td>Cosegregation with the phenotype by Sanger sequencing in family</td>
<td>77</td>
</tr>
<tr>
<td>Metachondromatosis</td>
<td>Autosomal dominant</td>
<td>One proband and partial linkage analysis in family</td>
<td>PTPN11</td>
<td>Cosegregation by Sanger sequencing in a second family</td>
<td>78</td>
</tr>
<tr>
<td>Miller syndrome</td>
<td>Autosomal recessive</td>
<td>Two affected offspring and both parents</td>
<td>DHODH, DNAH5, and KIAA0556</td>
<td>Previously published exome sequencing</td>
<td>46</td>
</tr>
<tr>
<td>Sitosterolemia*</td>
<td>Autosomal recessive</td>
<td>One proband</td>
<td>ABCG5</td>
<td>Presence of sitosterolemia in blood sample after weaning; previous association with sitosterolemia</td>
<td>79</td>
</tr>
</tbody>
</table>

FADD indicates Fas-associated death domain protein; ALPS, autoimmune lymphoproliferative syndrome; PBMC, peripheral blood mononuclear cell; and XIAP, X-linked inhibitor of apoptosis.

*Sequence information used to make clinical diagnosis.
RefSeq, the UCSC KnownGenes database, and the GENCODE and ENSEMBL databases. Each has advantages in terms of coverage and accuracy; however, the inconsistent use of these data in the literature is an issue for replicating research findings. Similarly, several variant databases exist for associations with Mendelian disorders, including the Human Gene Mutation Database, the Online Mendelian Inheritance in Man (http://www.ncbi.nlm.nih.gov/omim), and many disease-specific databases. None are well suited to variant-level annotation of whole genome sequence data, and many contain annotation errors and common polymorphisms, by some estimates comprising >25% of the entries. Furthermore, these databases are contaminated by descriptions of susceptibility loci of questionable impact, and mutation annotations are often based on differing builds of the reference genome or outdated gene and protein sequences. Several prediction algorithms exist for predicting variant pathogenicity that are based on different combinations of evolutionary conservation, structural prediction, and physical properties of amino acid substitutions. However, they are limited in specificity and sensitivity, and concordance between predictions from the various algorithms is low.100 Databases for common variant–common disease associations and pharmacogenomic associations are more complete, but there is a great need for comprehensive, easily searchable, and accurate variant-level association databases as whole genome sequence data becomes more widely available.

Other Applications of High-Throughput Sequencing

Although high-throughput sequencing has become synonymous with whole genome and exome sequencing, there are many other emerging applications for the technology. The first of these is whole transcriptome sequencing, which uses massively parallel sequencing to sequence RNA transcripts in various physiological conditions. This unique application allows for determination of allele-specific expression, information about alternative splicing, RNA-editing events, and, via read depth, accurate quantification of mRNA copy number, and, therefore, gene expression. In comparison with oligonucleotide expression arrays, RNAseq is able to quantify transcript abundance with a greater dynamic range and accuracy at extremes of transcript abundance, allowing for more accurate quantification of gene expression and rich functional genomics information. Matkovich et al recently used a unique combination of RNAseq and a new technology, RNA-induced silencing complexes sequencing, to characterize cardiac mRNA regulation by micro-RNAs, small noncoding RNAs that regulate diverse cellular functions by facilitating mRNA degradation or inhibiting translation. Technologies that do not require generation and amplification of a cDNA library, such as the Helicos platform, are particularly well suited to this application because they require no previous knowledge of the transcriptome and avoid biases in gene expression measurements and sequencing errors that are related to reverse transcription.

Second, subsets of exomes can be queried in a high-throughput manner in the next generation of candidate gene studies by use of custom oligonucleotide-based capture techniques coupled with high-throughput sequencing. Combined with a pooled case-control approach, these study designs may prove to be valuable to gene finding or comprehensive sequence interrogation (“fine mapping”) of genomic regions that have been linked with inherited disease by other technologies such as array-based genotyping or repetitive element mapping.

Third, there is increasing focus on the use of high-throughput sequencing in clinical diagnosis via the rapid identification of cell-free DNA. Specific to cardiovascular medicine is the recent demonstration of the use of cell-free sequencing of blood samples for the identification of an organ-specific transplant DNA signature correlating with acute cellular rejection in a pilot study of heart transplant recipients. With confirmation in larger cohorts, technologies such as these may be combined with other functional assays of the genome such as gene expression arrays to obviate the need for endomyocardial biopsy surveillance in select patients.

Last, although we have focused in much of this review on inherited genetic information, there is an entirely separate dimension of heritable information that researchers are just beginning to explore on a genome-wide scale. Epigenetic traits, or heritable traits that do not involve DNA sequence changes, are often due to chemical modifications of the DNA molecule such as cytosine methylation in CpG regions. To date, bisulfite sequencing, in which 5-methyl-cytosine bases are converted to uracil by bisulfite and subsequently sequenced, identifying CpG regions with high uracil content that correspond to methyl-cytosine bases, has been the standard technique. However, single-molecule sequencers yield polymerase kinetics information that correlates with methylation status and other structural information such as DNA polymerase footprint and RNA and DNA secondary structure.

Conclusions

The whole genome sequencing era is here. Challenges remain to widespread sequencing of individuals. However, advances in high-throughput sequencing technologies have made possible initial strides in understanding the fundamental genetic basis for inherited disease, and sequencing personal genomes may someday allow for individualization of health care to genetics. Ten years out from the completion of the human genome sequence, we are about to enter an era in which a vast amount of sequencing information will be available to medical researchers and, ultimately, healthcare professionals. It is incumbent on physicians and scientists as stewards of this technology to ensure that high-quality sequence data are incorporated appropriately into research and clinical endeavors.

Sources of Funding

Dr Dewey was supported by NIH/NHLBI training grant T32 HL094274-01A2 and the Stanford Dean’s Postdoctoral Research Fellowship. Dr Wheeler was supported by NIH National Research Service Award fellowship F32 HL097462. Dr Ashley was supported by NIH/NHLBI KO8 HL083914, NIH New Investigator DP2 Award OD004613, and a grant from the Breetwor Family Foundation.
Disclosures
Dr Quake is a founder, consultant, and equity holder in Helicos BioSciences and a founder, consultant, and shareholder of Fluidigm Corp. Dr Ashley is a founder and stockholder in Personalis. Dr Dewey is a stockholder and consultant for Personalis.

References
8. Dewey et al. DNA Sequencing and CV Disease 941


64. Dewey et al DNA Sequencing and CV Disease