Massively Parallel Sequencing: The Next Big Thing in Genetic Medicine

Tracy Tucker,[1,*] Marco Marra,[1,2] and Jan M. Friedman[1,3]

Massively parallel sequencing has reduced the cost and increased the throughput of genomic sequencing by more than three orders of magnitude, and it seems likely that costs will fall and throughput improve even more in the next few years. Clinical use of massively parallel sequencing will provide a way to identify the cause of many diseases of unknown etiology through simultaneous screening of thousands of loci for pathogenic mutations and by sequencing biological specimens for the genomic signatures of novel infectious agents. In addition to providing these entirely new diagnostic capabilities, massively parallel sequencing may also replace arrays and Sanger sequencing in clinical applications where they are currently being used.

Routine clinical use of massively parallel sequencing will require higher accuracy, better ways to select genomic subsets of interest, and improvements in the functionality, speed, and ease of use of data analysis software. In addition, substantial enhancements in laboratory computer infrastructure, data storage, and data transfer capacity will be needed to handle the extremely large data sets produced. Clinicians and laboratory personnel will require training to use the sequence data effectively, and appropriate methods will need to be developed to deal with the incidental discovery of pathogenic mutations and variants of uncertain clinical significance. Massively parallel sequencing has the potential to transform the practice of medical genetics and related fields, but the vast amount of personal genomic data produced will increase the responsibility of geneticists to ensure that the information obtained is used in a medically and socially responsible manner.

Introduction

DNA sequencing was first described by Maxim and Gilbert[1] and Sanger et al. in 1977.[2] Subsequent improvements to the Sanger method have increased the efficiency and accuracy more than three orders of magnitude. At each step, more sophisticated DNA sequencing instruments, programs, and bioinformatics have provided more automation and higher throughput. Several massively parallel sequencing methods have become available in the last couple of years, allowing larger-scale production of genomic sequence, and the number of human genomes sequenced with such instrumentation is now increasing rapidly.[3–6]

As the cost of massively parallel sequencing falls, it becomes feasible for smaller laboratories to adopt this technology, although doing so involves substantial initial costs. These include not just the massively parallel sequencing machines themselves, but also the associated costs of data storage and analysis. Massively parallel sequencing has had little impact on clinical diagnostics to date, but with the promise of the $1000 genome close at hand,[7] it seems only a matter of time before massively parallel sequencing becomes routinely available in clinical laboratories.

Massively parallel sequencing will allow simultaneous screening for mutations in hundreds of loci in genetically heterogeneous disorders, whole-genome screening for novel mutations, and sequence-based detection of novel pathogens that cause human disease. In addition, massively parallel sequencing will permit clinical application of our expanding knowledge of pharmacogenetics, cancer genetics, epigenetics, and complex traits. As with any new clinical test, analysis of clinical utility will have to be undertaken and clear standards and guidelines will need to be put in place before massively parallel sequencing can be routinely offered.

This review describes currently available massively parallel sequencing platforms and their potential impact on clinical testing in medical genetics, with consideration of technical issues that pertain to clinical laboratories and ethical issues that need to be addressed before massively parallel sequencing can be incorporated into routine clinical care.

Sanger Sequencing

Clinical DNA sequencing is currently performed by capillary-based, semiautomated Sanger sequencing. DNA is usually prepared by PCR amplification of a region of interest. The DNA is then sequenced by "cycle sequencing" that involves several rounds of template denaturation, primer annealing, and extension (Figure 1).[8] This approach can achieve read lengths of ~1 Kb and high accuracies at a cost of about $500 per megabase (Mb).[9]

Massively Parallel Sequencing Platforms

This section provides a brief overview of commercially available massively parallel sequencing platforms. For more detailed discussion of massively parallel platforms, readers are referred to recent reviews.[9,10]

The Illumina Genome Analyzer, which uses "sequencing by synthesis"[11,12] to produce single reads of 75+ basepairs (bp) (Figure 2), can currently generate about 17 gigabases

[1]Department of Medical Genetics, University of British Columbia, Vancouver, British Columbia V6H 3N1, Canada; [2]BC Cancer Agency Genome Sciences Centre, Vancouver, British Columbia V5Z 4S6, Canada; [3]Child & Family Research Institute, Vancouver, British Columbia V6H 3N1, Canada
*Correspondence: tbtucker@interchange.ubc.ca
DOI 10.1016/j.ajhg.2009.06.022.

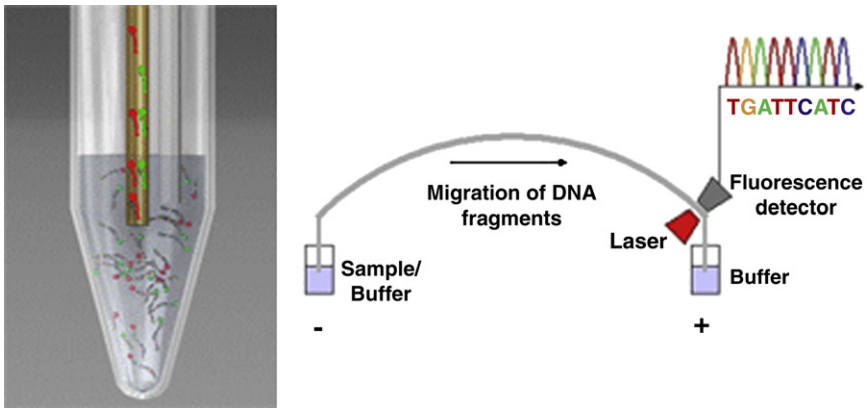

Figure 1. Sanger Sequencing Workflow
DNA fragments are enriched by PCR and sequenced with a combination of regular deoxynucleotides and terminating labeled dideoxynucleotides (ddNTPs), each with a base-specific color. Different fragment lengths are generated and size separated by capillary electrophoresis, and the location of each of the ddNTPs is identified by excitation with a laser. Reprinted with permission from Applied Biosystems.

(Gb) of sequence in 7 days at a cost of ~$6 per Mb (including consumables, labor, instrument costs, and disc storage).[10] The raw base accuracy is greater than 99.5% (Table 1).

The Applied Biosystems SOLiD Sequencer has read lengths of up to 50 bp and produces 10–15 Gb of sequence data in 3–7 days at a cost of ~$5.80 per Mb (including consumables, labor, instrument costs, and disc storage).[10] The raw base accuracy of the SOLiD System is 99.94% (Table 1).[13] This machine is unique in that it can process two slides at a time; one slide is receiving reagents while the other is being imaged. Each cycle of sequencing involves the hybridization of fluorescently labeled degenerate octomers to the DNA fragment sequence adjacent to the universal primer's 3′ end.[14] After several rounds of ligation, the extended primer is removed and the process is repeated with a universal primer that is offset by one base from the adaptor-fragment position (Figure 3). Offsetting the universal primer in five sets of cycles permits the entire fragment to be sequenced and provides an error-correction scheme because each base position is queried twice (once as a first base and again as the second base in the next or preceding set of cycles).[9]

The Roche GS-FLX 454 Genome Sequencer produces an average read length of 400 bp and generates ~400–600 Mb of sequence data per 10 hr run at a cost of $84.40 per Mb (including consumables, labor, instrument costs, and disc storage).[10] The raw base accuracy of the 454 Genome Sequencer is 99.5% (Table 1).[13] The sequencing process uses an enzymatic cascade to generate light from inorganic phosphate molecules released by the incorporation of nucleotides as the polymerase replicates the template DNA (Figure 4).[15]

The Helicos machine sequences single molecules of DNA without a prior amplification step.[16] Read lengths of 30–35 bp are obtained, and 20–28 Gb of sequence are generated in 8 days with a raw base accuracy greater than 99% (Table 1). The price for this equipment is not currently available. A highly sensitive fluorescence detection system is used to interrogate each nucleotide directly as it is synthesized (Figure 5).

Other Technologies

There are a number of other technologies that are currently under development, but these instruments are not yet commercially available. One such approach by Pacific Biosciences uses Single Molecule Real Time (SMRT) DNA sequencing. This method identifies nucleotide incorporation by DNA polymerase into a single DNA strand. Sequencing is performed on a chip containing thousands of tiny holes tens of nanometers in diameter that function as "zero-mode waveguides," defining the position at which light released by replication of a single tethered DNA molecule is detected. Nucleotides, each labeled through its phosphate chain with a different colored fluorophore, and Φ29 DNA polymerase, a highly accurate and efficient enzyme,[17] are added, and fluorescent light characteristic of each nucleotide is emitted as the DNA polymerase copies the tethered single-stranded sequence. The DNA polymerase cleaves the fluorophore as each new base is incorporated, returning the signal to baseline and permitting the addition of another nucleotide.[18]

Another approach being developed by Oxford Nanopore Technologies is nanopore sequencing. When voltage is applied across a nanopore, an electrical current is created. As a DNA fragment is electrophoretically pulled through the nanopore, each base creates a unique change in the magnitude of the electrical current.[19] Other unique massively parallel sequencing technologies are under development by other companies, including Visigen Biotechnology and Intelligent Biosystems. It is not clear which massively parallel sequencing technologies will gain greatest favor for clinical use, but it seems certain that further reductions in the cost of sequencing and the advantages conferred by these new technologies will assure that massively parallel sequencing becomes an essential clinical tool within the next decade.

Advantages of Massively Parallel Sequencing Technology

The advantages and limitations of massively parallel sequencing described below are presented in general terms, but the technological differences among the systems may make one particular massively parallel sequencing platform more or less well suited for a specific application.

Sanger sequencing has been used for many different applications, and improvements in chemistry, automation, and miniaturization over the years have permitted

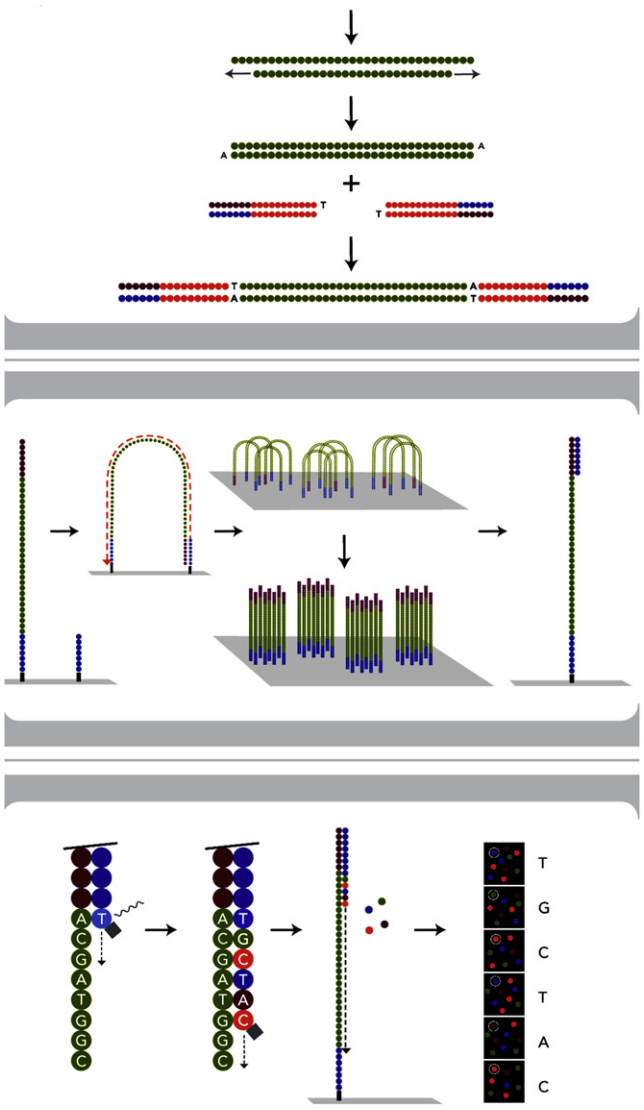

Figure 2. Illumina Genome Analyzer Workflow
Sequencing libraries are generated by fragmenting genomic DNA, denaturation, and adaptor ligation. Fragments are added to the flow cell chamber coated with oligonucleotides complementary to the adaptors. Hybridization forms a "bridge," and amplification is primed from the 3′ end and continues until it reaches the 5′ end. After several rounds of amplification, discrete clusters of fragments, all with the same sequence, are formed. The clusters are denatured, and sequencing primers, polymerase, and fluorescently labeled nucleotides, each with their 3′OH chemically inactivated, are added. After each base is incorporated, the surface is imaged, the 3′OH-inactivating residue and label are removed, and the process repeated. Reprinted with permission from Illumina, Inc.

it to be used for both small-scale (kilobase) and larger-scale (megabase) projects. Despite these advances, it seems unlikely that substantial further increases in throughput or decreases in cost will be possible with Sanger sequencing because of its dependence on lengthy procedures. The ability of massively parallel sequencing to overcome these limitations has allowed projects requiring many gigabases of sequence to be performed much more quickly and less expensively than with Sanger sequencing. For example, massively parallel sequencing has permitted uncovering a vast amount of germline and somatic variation in normal individuals.[4,12]

The increase in throughput and reduction in cost achieved by massively parallel sequencing are a result of three factors: (1) many thousands or millions of sequencing reactions are performed in parallel rather than just 1 to 96 at a time, as in conventional sequencing machines and (2) cloning or template amplification of the DNA fragments that are being sequenced is either unnecessary (in single-molecule sequencing) or fully automated within the same instrument that performs massively parallel sequencing.

Another advantage of massively parallel sequencing is the ability to detect minor alleles accurately. Each DNA fragment within the sequenced library is amplified and sequenced (or in the single-molecule technologies, just sequenced) independently of every other fragment, so if a sample is mosaic, as is the case for most tumors, rare somatic mutations can in principle be detected if depth of sequence coverage is sufficient. In addition, the "digital" nature of massively parallel sequencing means that the number of times any particular DNA segment is sequenced is proportional to the relative abundance of that segment compared to all of the other segments in the original sample. Thus, when a sample is sequenced to sufficiently high depth, the copy number of any particular segment can be inferred from the frequency with which that segment is found among the molecules sequenced. With conventional sequencing, rare mosaic variants may be lost and heterozygous deletions cannot be detected because sequencing is performed on the pool of templates, rather than on single molecules.

Limitations of Massively Parallel Sequencing Technology

All of the massively parallel sequencing platforms (except 454) produce read lengths of 50–100 bp, which are a fraction of those obtained with current-generation Sanger sequencing machines. Short read lengths make de novo sequence assembly more difficult and less complete, particularly for novel genomes or massively repetitive and rearranged DNA segments. Short read lengths also complicate interpretation in circumstances when it is necessary to determine the phase of variants (e.g., recessively inherited disorders). The implementation of paired-end or mate-paired reads in massively parallel sequencing, which are sequence reads from both ends of longer DNA molecules of known length, permits the analysis of genomic fragments up to 5–10 Kb in length, depending on the platform. Paired-end reads have been used to identify single-nucleotide mutations in *Caenorhabditis elegans*[20] and structural variants greater than 3 Kb in humans.[3]

Currently, the error rates of raw sequence data produced by all of the massively parallel sequencing platforms are higher than with Sanger sequencing, but the overall error rate is reduced because of the high degree of sequencing depth, typically 40-fold for a diploid genome, that is necessary to achieve complete coverage with massively parallel

Table 1. Comparing Massively Parallel Sequencing Technologies

	Sequencing Chemistry	Amplification Approach	Read Length	Run Time and Throughput	Raw Accuracy	Cost
Illumina	polymerase-based sequencing by synthesis	bridge PCR	75+ bp	17 Gb in 7 days	98.5%	$6/Mb
SOLiD	ligation-based	emulsion PCR	50 bp	10–15 Gb in 3-7 days	99.94%	$5.80/Mb
454	pyrosequencing	emulsion PCR	400 bp	400–600 Mb in 10 hr	99%	$84.40/Mb
Helicos	polymerase-based	none (single-molecule sequencing)	30–35 bp	21–28 Gb in 8 days	99%	not available

sequencing. Higher coverage is especially important when looking for mutations or sequence variants in repetitive or massively rearranged regions.[21,22] However, greater depth means more sequencing, thereby reducing the advantages of using massively parallel sequencing.

For clinical applications, there is a great need to increase the accuracy of raw massively parallel sequencing data. The introduction of a proof-reading polymerase in the sequencing process might increase the raw accuracy rate. The development of algorithms that take into account the data quality when making a base call[23–26] will no doubt be useful as well. These efforts will be enhanced by the development of standardized quality metrics for sequencing results, similar to those that have been implemented for microarray testing.[9] These include measures of (1) technical reproducibility, (2) distribution of estimated accuracies for raw base calls, (3) systematic error patterns in raw or consensus sequence data, and (4) bias and skewing of true ratios in tag counting applications.[9] It will be the responsibility of clinical laboratories that use massively parallel sequencing to include such quality metrics in their reports. This will not only permit standardization within the laboratory but also facilitate comparison of test results between different laboratories.

As with any new technology, the initial costs necessary to set up a massively parallel sequencing platform are high. Commercially available instruments cost $400,000–$1,350,000 each, and there are also costs associated with the software, training, and data transfer and storage required for the vast quantity of data generated by massively parallel sequencing platforms. In addition, massively parallel sequencing data interpretation requires much greater bioinformatic expertise than is available in most clinical laboratories. In recent years, a growing number of programs that vary in function and user-friendliness have been designed to align short read sequences to a reference and provide accurate base calling.[9] In order for clinical laboratories to adopt massively parallel sequencing, it will be necessary to develop data analysis and interpretation software that is geared toward clinical testing and that can be used by technicians and laboratory directors who do not have a background in bioinformatics.

Whole-genome sequencing is not yet practical in the clinical setting, and it is likely that massively parallel

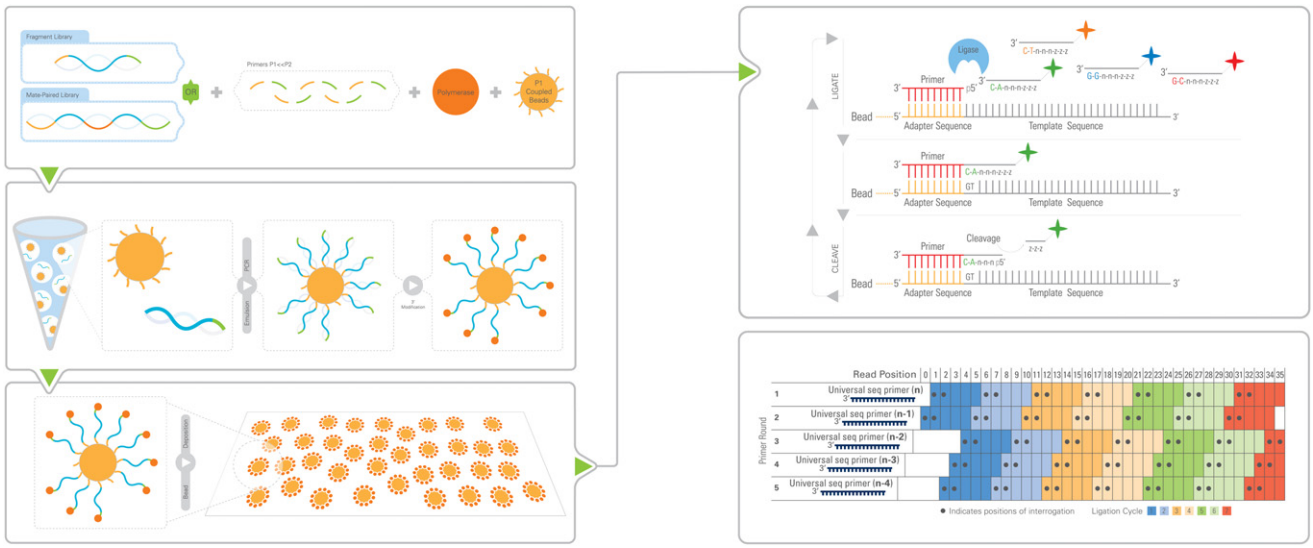

Figure 3. Applied Biosystems SOLiD Sequencer Workflow
DNA is fragmented and oligonucleotide adaptors are ligated to each end. The fragments are hybridized to complementary oligonucleotides attached to magnetic beads. The beads are contained within an oil emulsion where amplification is performed. When amplification is complete, the emulsion is broken, and the beads are attached to a glass surface and placed within the sequencer. A universal sequencing primer, complementary to the adaptor sequence, is added followed by subsequent ligation cycles with fluorescently labeled degenerate octomers. After each cycle, the glass surface is imaged and the octomer is cleaved between bases 5 and 6, removing the fluorescent tag, and a new octomer is added. After several rounds of sequencing, the extended universal primer is removed and a new universal primer is added that is offset by one base. Reprinted with permission from Applied Biosystems.

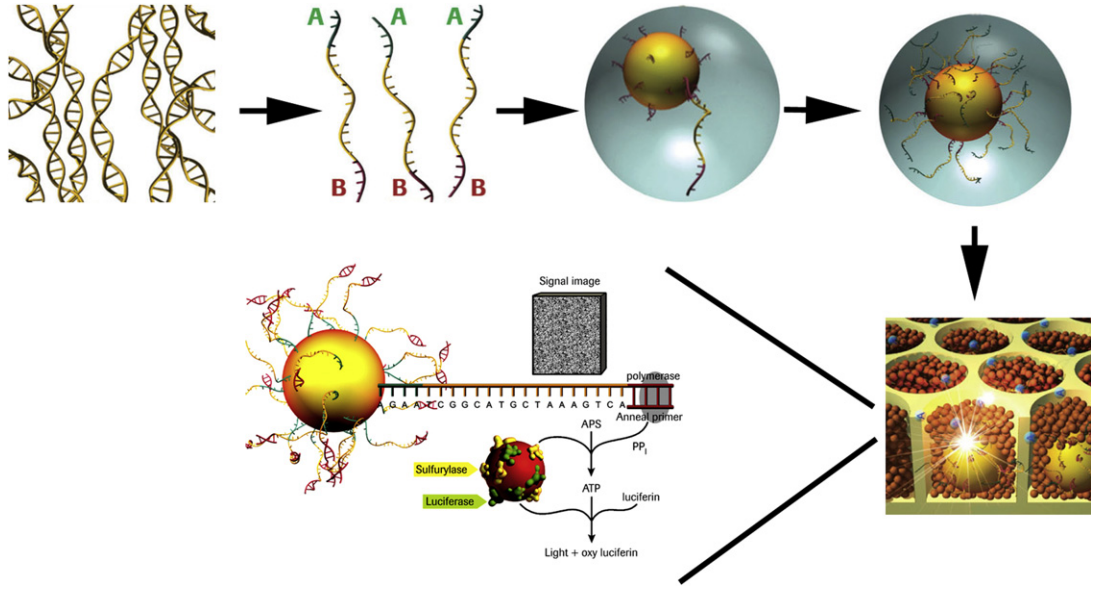

Figure 4. GS-FLX 454 Sequencer Workflow
DNA is fragmented and adaptors, one of which is biotinylated, are ligated to each end. Fragments are coupled to agarose beads by oligonucleotides complementary to the adaptor sequence and contained within an emulsion droplet for amplification. When amplification is completed, the beads are put into an individual well on a fiber optic slide and placed in the sequencer. Nucleotides and polymerase are sequentially added, and the sequence produced is monitored by the generation of light through an enzymatic reaction that is coupled to DNA synthesis. Modified with permission from 454 Sequencing, copyright 2009 Roche Diagnostics.

sequencing will initially be used to sequence selected genomic regions. Potential applications, which are discussed in more detail below, include testing many different loci for mutations simultaneously in a patient with a genetically heterogeneous disease or screening a large number of samples for mutations in a set of candidate genes. The latter is made possible by molecular "barcoding," which involves adding a short DNA sequence tag that is unique to a particular patient to every DNA fragment made from that patient's sample. Several patient samples can then be pooled and sequenced together, and the sequences obtained from each patient can be separated bioinformatically.[27]

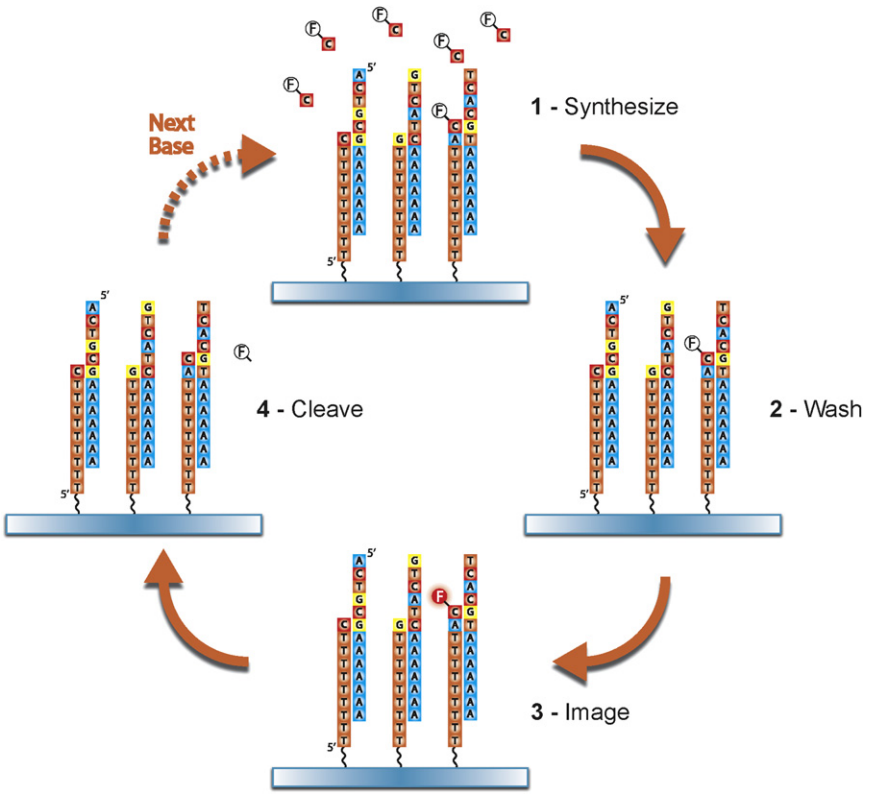

Figure 5. Helicos Heliscope Sequencer Workflow
Fragments are captured by poly-T oligomers tethered to an array. At each sequencing cycle, polymerase and single fluorescently labeled nucleotide are added and the array is imaged. The fluorescent tag is then removed and the cycle is repeated. Reprinted with permission from Helicos BioSciences Corporation.

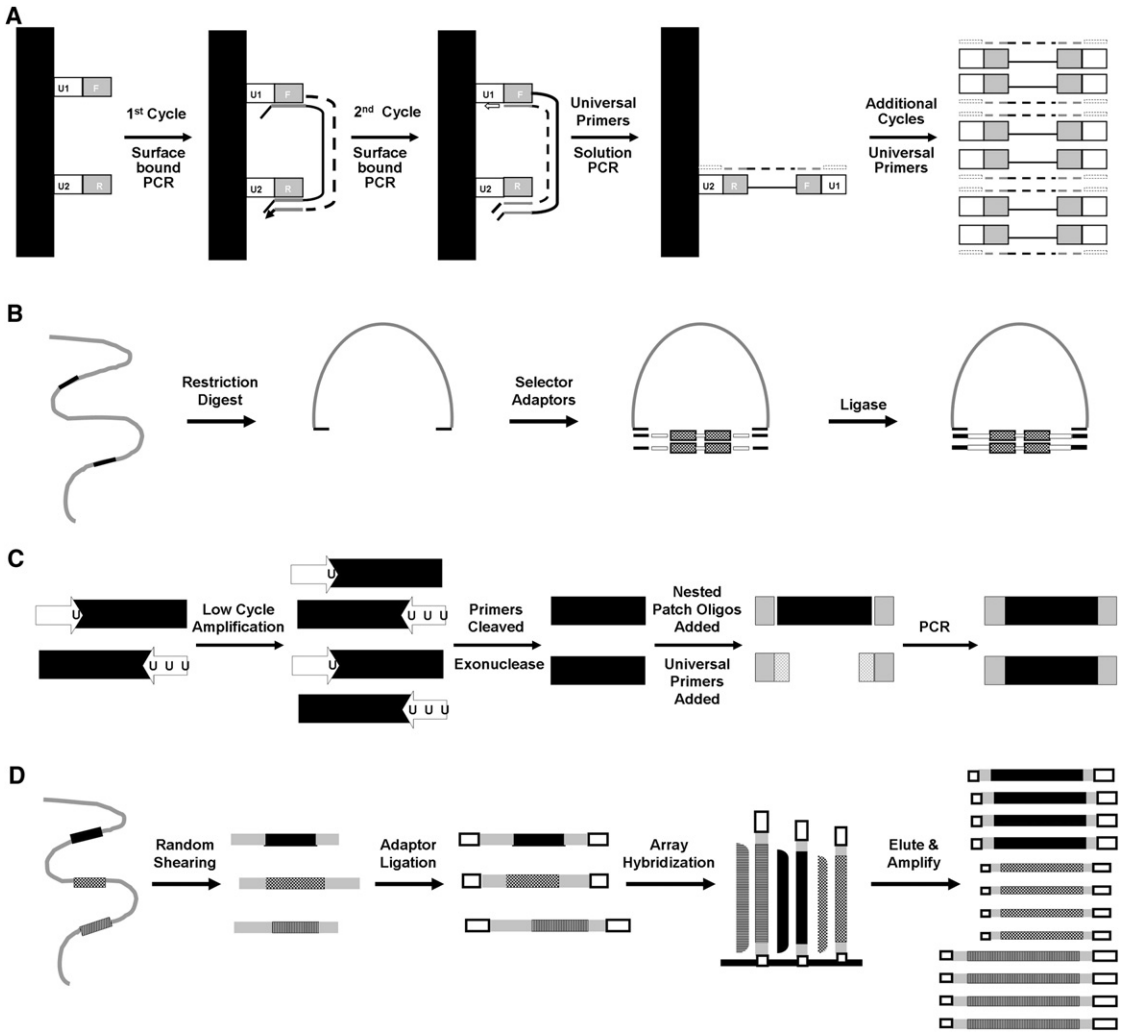

Figure 6. Genomic Enrichment Strategies

(A) Megaplex PCR. Surface-bound oligonucleotide primers (F & R) bind to DNA and amplify the sequence for the 1st and 2nd round of PCR. This reaction also incorporates a sequence that is complementary to a universal primer (U1 & U2), which is used for subsequent PCR cycles. Modified from ten Bosch and Grody.[13]

(B) Selector Probe Circularization. Genomic DNA (gray) is digested with restriction enzymes and circularized by hybridization of "selector probes" (black) with single stranded overhangs (white box) to the 3′ and 5′ ends of the digested DNA. DNA ligase fills in the gap, and universal primers (checkered box), complementary to the sequences within the selector probes, are used to amplify the circularized DNA. Modified from ten Bosch and Grody.[13]

(C) Nested-Patched PCR. Primer pairs containing uracil instead of thymine (wide white arrow) are constructed for all target regions. The primers amplify target regions for a low number of cycles. The primers are cleaved with uracil DNA glycosylase, nested patch oligonucleotides (gray and white checkered box) are annealed to target amplicons, universal primers (gray box) are ligated to the amplicons, and subsequent PCR cycles are primed with these universal primers. Modified from Varley and Mitra.[27]

(D) Microarray pull-down method. Genomic DNA is fragmented, and universal adaptor (white box) sequences are ligated to the ends of each fragment. The fragments of interest are captured by hybridization on the microarray (black line), which has been constructed with probes that are complementary to these sequences. The array is then denatured, and the fragments released are enriched by PCR with the universal adaptor sequence as primers. Modified from ten Bosch and Grody.[13]

Targeted sequencing requires substantial up-front preparation to select the DNA segments of interest. PCR with modifications to permit higher multiplexing is one useful way to do this. Examples of some available methods are shown in Figures 6A–6C. Such methods have been used to amplify hundreds of selected exons from a DNA sample.[27–29]

Alternatively, targeted regions can be obtained by direct hybridization to oligonucleotide arrays containing probes complementary to the regions of interest. The array is then denatured, and the fragments obtained can be amplified or directly sequenced, depending on the depth of hybridization (Figure 6D). This method has been successful on a large number of target regions with both the 454 FLX Sequencer[30] and the Illumina Genome Analyzer.[31–34] Several other novel methods have recently been developed to enrich segments of interest.[35–37] However, any enrichment strategy is unlikely to be effective for all genomic applications, highlighting the importance of further development of enrichment techniques.

Interpretation of Data

Six billion base pairs of DNA per patient are a lot of data to interpret. Computers and software can help—in fact, they are essential—but clinical interpretation of genome sequence data will always require a well-trained and experienced genetics professional. As a practical matter, only a subset of a person's genome can actually be examined for variants that cause or predispose to disease. This subset may either be selected before sequencing is done with a technique like those described in the previous section or, if whole-genome sequencing is performed, bioinformatically after all of the sequence data have been obtained. In either case, clinical sequencing requires decisions regarding what subset of the genome will be examined.

The subset examined for a particular patient may be different under different circumstances. If the goal is to obtain genotypes of a comprehensive set of single-nucleotide polymorphisms (SNPs), copy-number variants (CNVs), and other structural variants (SVs) for disease prediction, it would make most sense to focus on these polymorphic regions. If the goal is to identify a sequence mutation in an unidentified locus, sequencing all exons and adjacent promoter regions may be most informative. If the goal is to survey a more limited subset of loci for rare sequence mutations and copy number changes that are known to produce a genetically heterogeneous condition such as autism, a more targeted approach may be optimal. In other circumstances, such as identifying pathogenic mutations in patients with intellectual disability, it may be most productive to pursue a hybrid strategy that might include sequencing paired-end reads to assess copy number variation genome-wide as well as selective sequencing of all exons of loci known to cause recessive forms of intellectual disability.

Any selection process will be incomplete, and the possibility will always exist that a genetic variant outside the region examined in detail is actually pathogenic in the patient who is being tested. This would argue for the use of liberal inclusion criteria to select the genomic subset that will be analyzed in detail. On the other hand, the larger the fraction of the genome assessed, the greater the number of genetic variants that will have to be evaluated for pathogenicity in detail and the more likely that genetic variants of uncertain clinical significance will be encountered.

Variants encountered on sequencing may be of several kinds.[38] Some are known on the basis of extensive clinical experience to be pathogenic or, alternatively, not to be associated with disease. In many other instances, clinical experience with a particular variant is insufficient to provide an unequivocal interpretation with respect to pathogenicity, and other factors have to be considered. Such variants include:

- Those that are unreported but likely to be causative as determined by the type of mutation (e.g., mutations that create a stop codon or cause a frame shift);
- Those that are unreported and may or may not be causative of disease, including mutations that generate a cryptic splice site or are likely to affect transcription;
- Those that are unreported and are less likely to be causative of disease because they do not produce an amino acid change in the encoded protein.

Unfortunately, however, the pathogenicity of a sequence change sometimes cannot be predicted from its inferred effect on a gene's protein product. For example, a recent study that used Sanger sequencing to screen 718 coding exons on the X chromosome in 208 families with X-linked intellectual disability demonstrated that protein truncating variants, which usually result in loss of protein function and are a frequent cause of Mendelian diseases, occur in at least 1% of the X chromosome genes without any effect on normal intellectual ability.[39]

Disease- or locus-specific databases may be very helpful in determining whether a variant identified in a patient is causative of a particular disease. However, most of these databases are not designed to meet clinical standards, and they vary greatly in their completeness, rigor of interpretation, and currency. Critical information found in a database should be checked against the original source, and correlation of genomic findings with detailed phenotypic information on individual patients is essential.[7,40,41] Clinical sequencing will result in an explosion of data on variants, both pathogenic and benign, and clinical laboratories can improve their ability to interpret future data by carefully tracking their own patients in a local database and contributing both genotype and phenotype information to publicly available collaborative databases.

Genotyping of other family members is often very helpful in assessing pathogenicity, especially in the case of dominant diseases for which the presence of a mutation in an affected child and its absence in the normal parents suggests that the variant is causal. Functional testing of mutations is an effective means of determining pathogenicity of a given mutation but is beyond the scope of most clinical laboratories. Nevertheless, collaboration of clinicians, clinical laboratories, and research laboratories to perform functional studies is essential to the progress of clinical genetics.

Applications in Clinical Genetics Labs: Improvements in Current Diagnostic Capabilities

Sanger sequencing has been used clinically in conjunction with PCR for more than a decade to identify sequence mutations and other variants of selected Mendelian disease genes, but wider application of sequencing to clinical testing has been limited by cost and throughput. More recently, many clinical genetics laboratories have adopted array genomic hybridization as a means of detecting copy number changes that cause intellectual disability and other birth defects. Here we consider how massively parallel sequencing could be used to perform these and other forms of genetic testing that are currently available clinically more cost effectively and with higher throughput.

Mutation Detection in Mendelian Disease

Massively parallel sequencing could be used to sequence very large as well as smaller Mendelian disease genes fully, covering all exons as well as the associated 5′, 3′, and intronic sequences. This improved coverage could increase the sensitivity of mutation detection over current methods, which often employ a screening technique such as SSCP prior to sequencing and limit the segments sequenced to the exons that are most frequently affected by mutations or, at best, just to the exons and the immediately surrounding bases.[30,32,42] However, the capacity of current massively parallel sequencing platforms is too great for efficient sequencing of most single genes. For example, a single lane on an Illumina flow cell would provide more than 68× coverage of the entire genomic segment containing the *DMD* gene (MIM #300377) in a male. A solution to this problem would be to pool DNA samples from several family members or from unrelated patients that have been prepared by locus-specific PCR and barcoded so that the DNA from each individual can be distinguished.[27]

Recognizing Disease Predisposing and Protective Factors

Massively parallel sequencing has distinct advantages as a means of recognizing variants that may predispose to, or protect against, the development of common complex diseases. SNP arrays, even those with millions of features, provide genotypes of only a small fraction of the variants present in an individual. In contrast, massively parallel sequencing could provide complete information on all SNPs and other disease-associated variants that are present. This may be especially important in people whose origin is not from a population for which the tag-SNPs on most genotyping arrays provide optimal coverage. In addition, massively parallel sequencing could detect rare variants as well as the common ones for which genotyping chips are designed, and rare variants may be especially important in recognizing people with predispositions to developing certain complex diseases.[43,44]

Pharmacogenomics

Adverse drug reactions are one of the leading causes of death and illness.[45] Although many factors contribute, it is clear that genetic variation plays a key role in adverse reaction to drugs as well as to differences in the effectiveness of drug treatments. A good example of using pharmacogenetics clinically is testing for *CYP2C9* (MIM #601130) and *VKORC1* (MIM #607473) variants conjointly to determine dose requirements and hence susceptibility to adverse drug reactions related to warfarin.[46–49] Sequencing could permit the identification of these and all other pharmacogenetic variants (once we know them) in a single assay, thus permitting truly personalized drug treatment. This would be particularly valuable for many elderly patients and others with chronic diseases who must take many medications concurrently.

Improved Diagnosis and Treatment of Cancer

Somatic mutations of various kinds are present in almost all cancers. Microarray testing provides a useful means of surveying the entire genome for loss of heterozygosity and gene amplification,[50–52] and specialized molecular tests have been developed to assay for mutations and fusion genes produced by translocations that are associated with particular subtypes of cancer.[53,54] Massively parallel sequencing could provide information simultaneously on copy number changes, sequence mutations, and fusion genes anywhere in the genome that are associated with the development of malignancy.[55]

Cancer genomes are very heterogeneous because of the genetic instability and clonal evolution that characterize tumor development. Moreover, most tumors are composed of several different types of cells, some of which may not be part of the malignant clone. Consequently, mutations involved in tumor development or progression may be present in only a small fraction of the cells and may be missed with array studies and PCR-based assays of whole-tumor DNA. Massively parallel sequencing could accurately measure cancer-associated genetic alterations that occur in only a small fraction of the cells tested because the independent amplification and sequencing of millions of different DNA fragments from each specimen permits the accurate detection of rare sequences if the depth of coverage is sufficient.

Epigenetics

Somatic epigenetic changes are important in the development of some cancers[56] and constitutional epigenetic abnormalities cause congenital anomaly syndromes such as pseudohypoparathyroidism (MIM #612463),[57] Wiedemann-Beckwith syndrome (MIM #130650),[58] and Russell-Silver syndrome (MIM #180860).[59] Current clinical assays can demonstrate epigenetic alterations in individual genes, but massively parallel sequencing could be used to perform genome-wide tests of epigenetic changes that are known to cause particular disease states. For example, bisulfite sequencing was recently combined with massively parallel sequencing to examine methylation patterns throughout the genome in hematopoietic tumors,[60] and histone modifications have been identified genome-wide by combining chromatin immunoprecipitation with massively parallel sequencing.[61,62]

Identification of Structural Variants

Structural variants (SVs) include copy number variants (CNVs) as well as inversions and other chromosomal rearrangements that do not involve a change in copy number. Most CNVs occur as benign polymorphisms, but pathogenic CNVs have recently been found to be the most frequent recognizable cause of intellectual disability and some other birth defects.[63,64] Array genomic hybridization is now used clinically to detect such pathogenic CNVs, but this technology does not detect balanced SVs. Other CNVs are involved in modulation of complex traits[65] and disease susceptibility.[66,67] Massively parallel sequencing can reliably identify both balanced and unbalanced SVs[68] and provides much higher resolution than is possible with array genomic hybridization, permitting better genotype-phenotype correlation.

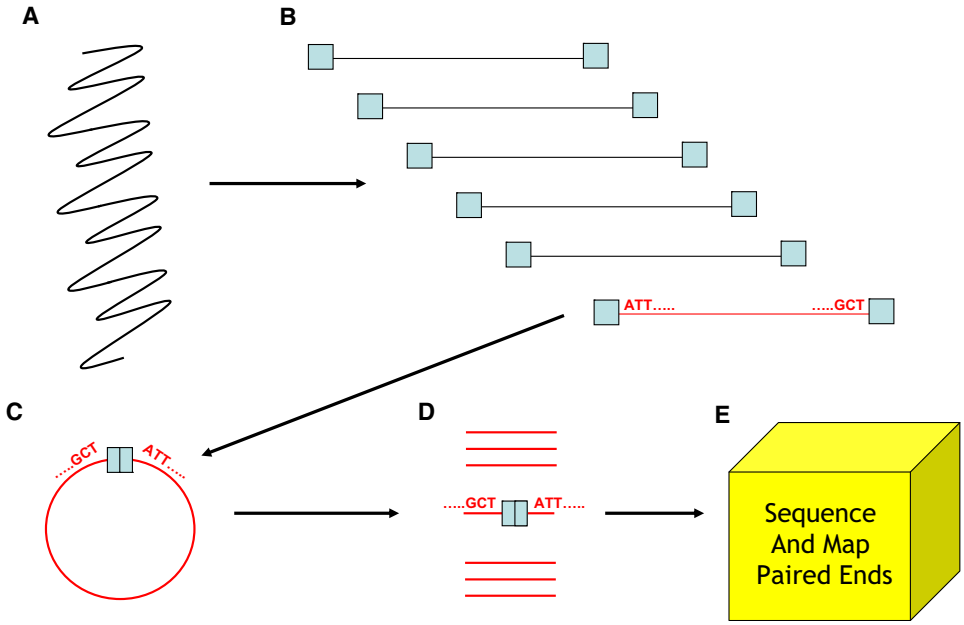

Figure 7. Paired-End Reads
DNA is isolated (A), fragmented into pieces of a standard size, e.g., about 3 kb, and ligated to adaptors (blue boxes) on both ends (B). Adaptors permit 3 kb pieces to be circularized (C). Circles are isolated, then broken into much smaller fragments (e.g., a few hundred base pairs) (D), and the fragments containing adaptors are isolated. In these fragments, the adaptor is flanked by the sequence that was at the opposite ends of the original 3 kb piece. The paired ends are sequenced and mapped back to the canonical human genome (E) so that structural variants can be identified (see text).

The short reads of most current massively parallel sequencing platforms limit their ability to map structural variants to the single-base-pair level because many short segments occur more than once in the genome and cannot be mapped uniquely. Mapping each segment to a unique genomic position is necessary to recognize balanced SVs and to count how many times each fragment or portion thereof appears in the original DNA sample. One solution to this problem is to use paired-end reads. DNA is fragmented in a manner that produces pieces of known size, and both ends of each fragment are ligated to adaptors, permitting circularization. The DNA circles containing the adaptors are then broken into much smaller fragments. A few of these smaller fragments include the adaptors flanked on either side by the DNA that lay on opposite ends of the original larger DNA fragment (Figure 7). These small fragments containing the adaptors and "paired-end tags" are then isolated, and the sequence at both ends is determined. Mapping these "paired-end reads" back to the canonical human genome sequence permits recognition of deletions as pairs of reads that map further apart than expected given the known length of the original fragment, duplications as pairs of reads that map closer to each other than expected given the length of the original fragment, inversions as pairs of reads that have a different orientation from the original fragment, and rearrangements as pairs that are not expected to lie together on the original fragment at all. Paired-end reads obtained by massively parallel sequencing have been used to map 853 deletions, 322 insertions, and 122 inversions identified in two individuals to an average breakpoint resolution of 644 bp in one recent study.[3]

Novel Diagnostic Capabilities by Massively Parallel Sequencing

The following discussion focuses on applications that are now being conducted in research laboratories but are not currently performed routinely, if at all, in clinical laboratories. Massively parallel sequencing offers the opportunity to implement these two applications as new clinical services. There are, of course, many other potential applications for massively parallel sequencing in clinical research, such as genome-wide association studies and linkage studies, but this discussion is limited to the applications that seem most likely to be used on a routine clinical basis in the next several years.

Simultaneous Screening for Mutations at Multiple Loci

Some conditions that are seen frequently by clinical geneticists may be caused by Mendelian mutations of many dozens or hundreds of different genetic loci. Examples include intellectual disability, deafness, familial cardiomyopathy, and retinitis pigmentosa. The current approach to identifying the causal mutation in a particular family involves recognition of a phenotypic subset that may be more or less specific, then mutation testing a series of genetic loci, individually or in small sets, based on the relative frequency of the mutations and the sensitivity of available assays. If there is no predominant mutation, as is the case for intellectual

disability, for example, this testing, no matter how extensive (and expensive), often fails to find the pathogenic mutation.

Massively parallel sequencing could provide the opportunity to test hundreds or even thousands of candidate loci simultaneously. This could be done by whole-genome sequencing and selective bioinformatic analysis or by sequencing candidate regions selected by array capture or one of the other methods of targeted amplification described above (Figure 6).

Metagenomics

Metagenomics is the brute force sequencing and bioinformatic analysis of DNA fragments obtained from an uncultured, unpurified microbial and or viral population. Humans live in symbiosis with billions of bacteria that inhabit both the outer and inner surfaces of our bodies (skin, respiratory tract, etc.).[69] These microorganisms are essential for our health, and alterations of their numbers or types can cause disease.[70,71] Massively parallel sequencing could provide the ability to recognize previously unidentified microorganisms that are associated with human disease by mass sequencing of an infected tissue or fluid, bioinformatically "subtracting out" all human sequences, recognizing the sequences of normal commensal organisms, and then analyzing what is left to identify the unknown pathogen. Longer paired-end reads could facilitate the de novo sequence assembly of the unknown pathogen by first mapping the smallest fragments onto larger fragments and then assembling the larger fragments into a whole genome sequence. The ability of massively parallel sequencing to characterize rare DNA fragments accurately and the ability to assemble de novo sequences via overlapping reads could permit the identification of a tiny amount of microbial DNA in the presence of a vast excess of human DNA. Several important advances have recently been achieved with this technology, including the identification of previously unrecognized microorganisms that are associated with a fatal febrile illness in organ transplant recipients,[72] infant diarrhea,[73] and a variety of other gastrointestinal diseases.[71]

Ethical Considerations

Massively parallel sequencing technology is rapidly advancing and is likely to enable these and other routine clinical applications in the near future. However, massively parallel sequencing, and especially whole-genome sequencing, raises a number of important ethical issues that need to be resolved prior to routine clinical implementation. None of these issues are unique to clinical massively parallel sequencing—all have been raised before in the context of genetic testing and other aspects of clinical genetics practice. Nevertheless, clinical use of this ultimate genetic technology raises these issues all at once and brings them into sharp focus. The huge amount of personal medical data produced by massively parallel sequencing, the fact that most of it will be irrelevant to any particular clinical problem but may be of importance to the patient in other ways or in the future, our inability to interpret much of the data, and the ability to link the information to an individual person despite the complete absence of any conventional personal identifying information all require careful consideration and the development of appropriate rules or guidelines prior to clinical implementation. The ethical issues raised by clinical use of massively parallel sequencing include the following:

- Consent
  - Does whole-genome sequencing require a different level or kind of consent than other genetic tests or medical assessments?
  - Should whole-genome sequencing be done when the same question can be answered by more limited (e.g., locus-specific) testing?
  - Should whole-genome sequencing be done in children or incompetent adults?
  - Is informed consent for whole-genome sequencing possible?
- Interpreting Sequence Data
  - Should patients be informed of results of uncertain clinical significance?
  - Should patients be informed of results that predict serious disease that cannot be prevented or treated?
  - Should patients be informed of results that do not have direct implications for them but do for other family members?
  - Should other family members be informed of findings that have direct implications for them that were found on analysis of a relative's genomic sequence?
- The Rest of the Data
  - Should patients be informed of incidental findings that unequivocally predict serious disease that can be prevented or ameliorated by early detection? What if the disease cannot be prevented or ameliorated?
  - Should patients be informed of incidental findings that indicate an increased (or reduced) risk for disease that can be prevented or ameliorated by early detection? What if the disease cannot be prevented or ameliorated?
  - Should physicians or clinical laboratories provide genomic information that has no medical importance but is of social or personal consequence to the patient (e.g., ancestry or paternity)?
  - Should physicians or clinical laboratories provide genomic information that has no medical importance but is of general interest to the patient (e.g., SNPs associated with athletic or musical ability)?
  - Is it appropriate to generate whole-genome data that may or may not be of clinical significance but analyze only a small portion of those data to answer a specific clinical question?
  - Do physicians or clinical laboratories have a duty to recontact patients if sequence data that were previously obtained are later found to have serious medical implications?

- Storing Sequence Data
  - Do physicians or clinical laboratories have a responsibility to retain a patient's genomic data for long periods of time (or throughout life) in case future reanalysis is necessary?
  - Where should individual sequence data be stored, and who should be responsible for the stored data?
  - Who should be able to obtain access to an individual's complete genomic sequence? The individual? Any treating physician? Insurance companies? Police (e.g., for criminal investigations)?
  - Under what circumstances should stored genomic data be used for purposes of identification (e.g., for identification of disaster victims or confirmation of citizenship)?

Conclusion

It is very likely that incremental improvements in currently available massively parallel sequencing technologies or the introduction of others that are currently in development will make sequencing an individual patient's entire genome at sufficient depth to identify almost all mutations and genetic variants practical for routine clinical applications in the near future. In order for massively parallel sequencing to be implemented clinically, the accuracy of sequencing needs to be increased and improvements in the methods available for selecting particular genomic subsets and for bioinformatic analysis of huge amounts of raw sequence data are necessary, but rapid progress is being made in these areas. The $10 million Archon X Prize for the first team to sequence 100 human genomes in 10 days may be won within the next year.

Clinical laboratory scientists, genetic counselors, clinical geneticists, and other physicians all must learn much more about this technology and its clinical application to use massively parallel sequencing safely and effectively. There is an urgent need for translational research regarding the clinical validity and clinical utility of massively parallel sequencing data,[74] as well as for professional education regarding the value, limitations, and appropriate clinical use of this powerful new technology.

Acknowledgments

This work was supported by a grant from the Canadian Institutes of Health Research to J.M.F.

Web Resources

The URLs for data presented herein are as follows:

Intelligent Bio-Systems, http://www.intelligentbiosystems.com
NCBI GeneTests, http://www.ncbi.nlm.nih.gov/sites/GeneTests/?db=GeneTests
Online Mendelian Inheritance in Man (OMIM), http://www.ncbi.nlm.nih.gov/Omim/
VisiGen Biotechnologies, Inc., http://visigenbio.com/
XPrize Foundation, http://www.xprize.org

References

1. Maxam, A.M., and Gilbert, W. (1977). A new method for sequencing DNA. Proc. Natl. Acad. Sci. USA *74*, 560–564.
2. Sanger, F., Nicklen, S., and Coulson, A.R. (1977). DNA sequencing with chain-terminating inhibitors. Proc. Natl. Acad. Sci. USA *74*, 5463–5467.
3. Korbel, J.O., Urban, A.E., Affourtit, J.P., Godwin, B., Grubert, F., Simons, J.F., Kim, P.M., Palejev, D., Carriero, N.J., Du, L., et al. (2007). Paired-end mapping reveals extensive structural variation in the human genome. Science *318*, 420–426.
4. Kidd, J.M., Cooper, G.M., Donahue, W.F., Hayden, H.S., Sampas, N., Graves, T., Hansen, N., Teague, B., Alkan, C., Antonacci, F., et al. (2008). Mapping and sequencing of structural variation from eight human genomes. Nature *453*, 56–64.
5. Wang, J., Wang, W., Li, R., Li, Y., Tian, G., Goodman, L., Fan, W., Zhang, J., Li, J., Zhang, J., et al. (2008). The diploid genome sequence of an Asian individual. Nature *456*, 60–65.
6. Levy, S., Sutton, G., Ng, P.C., Feuk, L., Halpern, A.L., Walenz, B.P., Axelrod, N., Huang, J., Kirkness, E.F., Denisov, G., et al. (2007). The diploid genome sequence of an individual human. PLoS Biol. *5*, e254.
7. Siva, N. (2008). 1000 Genomes project. Nat. Biotechnol. *26*, 256.
8. Swerdlow, H., Wu, S.L., Harke, H., and Dovichi, N.J. (1990). Capillary gel electrophoresis for DNA sequencing. Laser-induced fluorescence detection with the sheath flow cuvette. J. Chromatogr. *516*, 61–67.
9. Shendure, J., and Ji, H. (2008). Next-generation DNA sequencing. Nat. Biotechnol. *26*, 1135–1145.
10. Mardis, E.R. (2008). The impact of next-generation sequencing technology on genetics. Trends Genet. *24*, 133–141.
11. Fedurco, M., Romieu, A., Williams, S., Lawrence, I., and Turcatti, G. (2006). BTA, a novel reagent for DNA attachment on glass and efficient generation of solid-phase amplified DNA colonies. Nucleic Acids Res. *34*, e22.
12. Bentley, D.R., Balasubramanian, S., Swerdlow, H.P., Smith, G.P., Milton, J., Brown, C.G., Hall, K.P., Evers, D.J., Barnes, C.L., Bignell, H.R., et al. (2008). Accurate whole human genome sequencing using reversible terminator chemistry. Nature *456*, 53–59.
13. ten Bosch, J.R., and Grody, W.W. (2008). Keeping up with the next generation: Massively parallel sequencing in clinical diagnostics. J. Mol. Diagn. *10*, 484–492.
14. Shendure, J., Porreca, G.J., Reppas, N.B., Lin, X., McCutcheon, J.P., Rosenbaum, A.M., Wang, M.D., Zhang, K., Mitra, R.D., and Church, G.M. (2005). Accurate multiplex polony sequencing of an evolved bacterial genome. Science *309*, 1728–1732.
15. Ronaghi, M., Karamohamed, S., Pettersson, B., Uhlen, M., and Nyren, P. (1996). Real-time DNA sequencing using detection of pyrophosphate release. Anal. Biochem. *242*, 84–89.
16. Braslavsky, I., Hebert, B., Kartalov, E., and Quake, S.R. (2003). Sequence information can be obtained from single DNA molecules. Proc. Natl. Acad. Sci. USA *100*, 3960–3964.
17. Harris, T.D., Buzby, P.R., Babcock, H., Beer, E., Bowers, J., Braslavsky, I., Causey, M., Colonell, J., Dimeo, J., Efcavitch, J.W., et al. (2008). Single-molecule DNA sequencing of a viral genome. Science *320*, 106–109.
18. Eid, J., Fehr, A., Gray, J., Luong, K., Lyle, J., Otto, G., Peluso, P., Rank, D., Baybayan, P., Bettman, B., et al. (2009). Real-time

DNA sequencing from single polymerase molecules. Science *323*, 133–138.
19. Branton, D., Deamer, D.W., Marziali, A., Bayley, H., Benner, S.A., Butler, T., Di Ventra, M., Garaj, S., Hibbs, A., Huang, X., et al. (2008). The potential and challenges of nanopore sequencing. Nat. Biotechnol. *26*, 1146–1153.
20. Shen, Y., Sarin, S., Liu, Y., Hobert, O., and Pe'er, I. (2008). Comparing platforms for *C. elegans* mutant identification using high-throughput whole-genome sequencing. PLoS ONE *3*, e4012.
21. Srivatsan, A., Han, Y., Peng, J., Tehranchi, A.K., Gibbs, R., Wang, J.D., and Chen, R. (2008). High-precision, whole-genome sequencing of laboratory strains facilitates genetic studies. PLoS Genet *4*, e1000139.
22. Thomas, R.K., Nickerson, E., Simons, J.F., Janne, P.A., Tengs, T., Yuza, Y., Garraway, L.A., LaFramboise, T., Lee, J.C., Shah, K., et al. (2006). Sensitive mutation detection in heterogeneous cancer specimens by massively parallel picoliter reactor sequencing. Nat. Med. *12*, 852–855.
23. Brockman, W., Alvarez, P., Young, S., Garber, M., Giannoukos, G., Lee, W.L., Russ, C., Lander, E.S., Nusbaum, C., and Jaffe, D.B. (2008). Quality scores and SNP detection in sequencing-by-synthesis systems. Genome Res. *18*, 763–770.
24. Quinlan, A.R., Stewart, D.A., Stromberg, M.P., and Marth, G.T. (2008). Pyrobayes: An improved base caller for SNP discovery in pyrosequences. Nat. Methods *5*, 179–181.
25. Bryant, D.W., Jr., Wong, W.K., and Mockler, T.C. (2009). QSRA: A quality-value guided de novo short read assembler. BMC Bioinformatics *10*, 69.
26. Li, H., Ruan, J., and Durbin, R. (2008). Mapping short DNA sequencing reads and calling variants using mapping quality scores. Genome Res. *18*, 1851–1858.
27. Varley, K.E., and Mitra, R.D. (2008). Nested Patch PCR enables highly multiplexed mutation discovery in candidate genes. Genome Res. *18*, 1844–1850.
28. Dahl, F., Gullberg, M., Stenberg, J., Landegren, U., and Nilsson, M. (2005). Multiplex amplification enabled by selective circularization of large sets of genomic DNA fragments. Nucleic Acids Res. *33*, e71.
29. Krishnakumar, S., Zheng, J., Wilhelmy, J., Faham, M., Mindrinos, M., and Davis, R. (2008). A comprehensive assay for targeted multiplex amplification of human DNA sequences. Proc. Natl. Acad. Sci. USA *105*, 9296–9301.
30. Albert, T.J., Molla, M.N., Muzny, D.M., Nazareth, L., Wheeler, D., Song, X., Richmond, T.A., Middle, C.M., Rodesch, M.J., Packard, C.J., et al. (2007). Direct selection of human genomic loci by microarray hybridization. Nat. Methods *4*, 903–905.
31. Bau, S., Schracke, N., Kranzle, M., Wu, H., Stahler, P.F., Hoheisel, J.D., Beier, M., and Summerer, D. (2009). Targeted next-generation sequencing by specific capture of multiple genomic loci using low-volume microfluidic DNA arrays. Anal. Bioanal. Chem. *393*, 171–175.
32. Hodges, E., Xuan, Z., Balija, V., Kramer, M., Molla, M.N., Smith, S.W., Middle, C.M., Rodesch, M.J., Albert, T.J., Hannon, G.J., et al. (2007). Genome-wide in situ exon capture for selective resequencing. Nat. Genet. *39*, 1522–1527.
33. Gnirke, A., Melnikov, A., Maguire, J., Rogov, P., LeProust, E.M., Brockman, W., Fennell, T., Giannoukos, G., Fisher, S., Russ, C., et al. (2009). Solution hybrid selection with ultra-long oligonucleotides for massively parallel targeted sequencing. Nat. Biotechnol. *27*, 182–189.
34. Hodges, E., Rooks, M., Xuan, Z., Bhattacharjee, A., Benjamin Gordon, D., Brizuela, L., Richard McCombie, W., and Hannon, G.J. (2009). Hybrid selection of discrete genomic intervals on custom-designed microarrays for massively parallel sequencing. Nat. Protocols *4*, 960–974.
35. Herman, D.S., Hovingh, G.K., Iartchouk, O., Rehm, H.L., Kucherlapati, R., Seidman, J.G., and Seidman, C.E. (2009). Filter-based hybridization capture of subgenomes enables resequencing and copy-number detection. Nat. Methods *6*, 507–510.
36. Li, J.B., Gao, Y., Aach, J., Zhang, K., Kryukov, G., Xie, B., Ahlford, A., Yoon, J.K., Rosenbaum, A.M., Zaranek, A.W., et al. (2009). Multiplex padlock capture and sequencing reveal human hypermutable CpG variations. Genome Res., in press. Published online June 12, 2009. 10.1101/gr.092213.109.
37. White, R.A., 3rd, Blainey, P.C., Fan, H.C., and Quake, S.R. (2009). Digital PCR provides sensitive and absolute calibration for high throughput sequencing. BMC Genomics *10*, 116.
38. Richards, C.S., Bale, S., Bellissimo, D.B., Das, S., Grody, W.W., Hegde, M.R., Lyon, E., and Ward, B.E. (2008). ACMG recommendations for standards for interpretation and reporting of sequence variations: Revisions 2007. Genet. Med. *10*, 294–300.
39. Tarpey, P.S., Smith, R., Pleasance, E., Whibley, A., Edkins, S., Hardy, C., O'Meara, S., Latimer, C., Dicks, E., Menzies, A., et al. (2009). A systematic, large-scale resequencing screen of X-chromosome coding exons in mental retardation. Nat. Genet. *41*, 535–543.
40. Oetting, W.S. (2009). Clinical genetics & human genome variation: The 2008 human genome variation society scientific meeting. Hum. Mutat. *30*, 852–856.
41. Kuehn, B.M. (2008). 1000 Genomes Project promises closer look at variation in human genome. JAMA *300*, 2715.
42. Porreca, G.J., Zhang, K., Li, J.B., Xie, B., Austin, D., Vassallo, S.L., LeProust, E.M., Peck, B.J., Emig, C.J., Dahl, F., et al. (2007). Multiplex amplification of large sets of human exons. Nat. Methods *4*, 931–936.
43. Need, A.C., Ge, D., Weale, M.E., Maia, J., Feng, S., Heinzen, E.L., Shianna, K.V., Yoon, W., Kasperaviciute, D., Gennarelli, M., et al. (2009). A genome-wide investigation of SNPs and CNVs in schizophrenia. PLoS Genet *5*, e1000373.
44. Gratacos, M., Costas, J., de Cid, R., Bayes, M., Gonzalez, J.R., Baca-Garcia, E., de Diego, Y., Fernandez-Aranda, F., Fernandez-Piqueras, J., Guitart, M., et al. (2008). Identification of new putative susceptibility genes for several psychiatric disorders by association analysis of regulatory and non-synonymous SNPs of 306 genes involved in neurotransmission and neurodevelopment. Am. J. Med. Genet. B. Neuropsychiatr. Genet., in press. Published online December 11, 2008. 10.1002/ajmg.b.30902.
45. Lazarou, J., Pomeranz, B.H., and Corey, P.N. (1998). Incidence of adverse drug reactions in hospitalized patients: A meta-analysis of prospective studies. JAMA *279*, 1200–1205.
46. Anderson, J.L., Horne, B.D., Stevens, S.M., Grove, A.S., Barton, S., Nicholas, Z.P., Kahn, S.F., May, H.T., Samuelson, K.M., Muhlestein, J.B., et al. (2007). Randomized trial of genotype-guided versus standard warfarin dosing in patients initiating oral anticoagulation. Circulation *116*, 2563–2570.
47. Takahashi, H., Wilkinson, G.R., Nutescu, E.A., Morita, T., Ritchie, M.D., Scordo, M.G., Pengo, V., Barban, M., Padrini, R., Ieiri, I., et al. (2006). Different contributions of polymorphisms in VKORC1 and CYP2C9 to intra- and inter-population differences in maintenance dose of warfarin in Japanese,

Caucasians and African-Americans. Pharmacogenet. Genomics *16*, 101–110.
48. Vecsler, M., Loebstein, R., Almog, S., Kurnik, D., Goldman, B., Halkin, H., and Gak, E. (2006). Combined genetic profiles of components and regulators of the vitamin K-dependent gamma-carboxylation system affect individual sensitivity to warfarin. Thromb. Haemost. *95*, 205–211.
49. D'Andrea, G., D'Ambrosio, R.L., Di Perna, P., Chetta, M., Santacroce, R., Brancaccio, V., Grandone, E., and Margaglione, M. (2005). A polymorphism in the VKORC1 gene is associated with an interindividual variability in the dose-anticoagulant effect of warfarin. Blood *105*, 645–649.
50. Schwaenen, C., Viardot, A., Berger, H., Barth, T.F., Bentink, S., Dohner, H., Enz, M., Feller, A.C., Hansmann, M.L., Hummel, M., et al. (2009). Microarray-based genomic profiling reveals novel genomic aberrations in follicular lymphoma which associate with patient survival and gene expression status. Genes Chromosomes Cancer *48*, 39–54.
51. Sargent, R., Jones, D., Abruzzo, L.V., Yao, H., Bonderover, J., Cisneros, M., Wierda, W.G., Keating, M.J., and Luthra, R. (2009). Customized oligonucleotide array-based comparative genomic hybridization as a clinical assay for genomic profiling of chronic lymphocytic leukemia. J. Mol. Diagn. *11*, 25–34.
52. de Tayrac, M., Etcheverry, A., Aubry, M., Saikali, S., Hamlat, A., Quillien, V., Le Treut, A., Galibert, M.D., and Mosser, J. (2009). Integrative genome-wide analysis reveals a robust genomic glioblastoma signature associated with copy number driving changes in gene expression. Genes Chromosomes Cancer *48*, 55–68.
53. Rowley, J.D. (1973). Letter: A new consistent chromosomal abnormality in chronic myelogenous leukaemia identified by quinacrine fluorescence and Giemsa staining. Nature *243*, 290–293.
54. Zech, L., Haglund, U., Nilsson, K., and Klein, G. (1976). Characteristic chromosomal abnormalities in biopsies and lymphoid-cell lines from patients with Burkitt and non-Burkitt lymphomas. Int. J. Cancer *17*, 47–56.
55. Ley, T.J., Mardis, E.R., Ding, L., Fulton, B., McLellan, M.D., Chen, K., Dooling, D., Dunford-Shore, B.H., McGrath, S., Hickenbotham, M., et al. (2008). DNA sequencing of a cytogenetically normal acute myeloid leukaemia genome. Nature *456*, 66–72.
56. Lopez, J., Percharde, M., Coley, H.M., Webb, A., and Crook, T. (2009). The context and potential of epigenetics in oncology. Br. J. Cancer *100*, 571–577.
57. Liu, J., Nealon, J.G., and Weinstein, L.S. (2005). Distinct patterns of abnormal GNAS imprinting in familial and sporadic pseudohypoparathyroidism type IB. Hum. Mol. Genet. *14*, 95–102.
58. Reik, W., Brown, K.W., Schneid, H., Le Bouc, Y., Bickmore, W., and Maher, E.R. (1995). Imprinting mutations in the Beckwith-Wiedemann syndrome suggested by altered imprinting pattern in the IGF2–H19 domain. Hum. Mol. Genet. *4*, 2379–2385.
59. Yoshihashi, H., Maeyama, K., Kosaki, R., Ogata, T., Tsukahara, M., Goto, Y., Hata, J., Matsuo, N., Smith, R.J., and Kosaki, K. (2000). Imprinting of human GRB10 and its mutations in two patients with Russell-Silver syndrome. Am. J. Hum. Genet. *67*, 476–482.
60. Taylor, K.H., Kramer, R.S., Davis, J.W., Guo, J., Duff, D.J., Xu, D., Caldwell, C.W., and Shi, H. (2007). Ultradeep bisulfite sequencing analysis of DNA methylation patterns in multiple gene promoters by 454 sequencing. Cancer Res. *67*, 8511–8518.
61. Mikkelsen, T.S., Ku, M., Jaffe, D.B., Issac, B., Lieberman, E., Giannoukos, G., Alvarez, P., Brockman, W., Kim, T.K., Koche, R.P., et al. (2007). Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. Nature *448*, 553–560.
62. Barski, A., Cuddapah, S., Cui, K., Roh, T.Y., Schones, D.E., Wang, Z., Wei, G., Chepelev, I., and Zhao, K. (2007). High-resolution profiling of histone methylations in the human genome. Cell *129*, 823–837.
63. Shaffer, L.G., Bejjani, B.A., Torchia, B., Kirkpatrick, S., Coppinger, J., and Ballif, B.C. (2007). The identification of microdeletion syndromes and other chromosome abnormalities: cytogenetic methods of the past, new technologies for the future. Am. J. Med. Genet. C. Semin. Med. Genet. *145C*, 335–345.
64. Stankiewicz, P., and Beaudet, A.L. (2007). Use of array CGH in the evaluation of dysmorphology, malformations, developmental delay, and idiopathic mental retardation. Curr. Opin. Genet. Dev. *17*, 182–192.
65. de Smith, A.J., Tsalenko, A., Sampas, N., Scheffer, A., Yamada, N.A., Tsang, P., Ben-Dor, A., Yakhini, Z., Ellis, R.J., Bruhn, L., et al. (2007). Array CGH analysis of copy number variation identifies 1284 new genes variant in healthy white males: implications for association studies of complex diseases. Hum. Mol. Genet. *16*, 2783–2794.
66. de Smith, A.J., Walters, R.G., Froguel, P., and Blakemore, A.I. (2008). Human genes involved in copy number variation: Mechanisms of origin, functional effects and implications for disease. Cytogenet. Genome Res. *123*, 17–26.
67. Yang, T.L., Chen, X.D., Guo, Y., Lei, S.F., Wang, J.T., Zhou, Q., Pan, F., Chen, Y., Zhang, Z.X., Dong, S.S., et al. (2008). Genome-wide copy-number-variation study identified a susceptibility gene, UGT2B17, for osteoporosis. Am. J. Hum. Genet. *83*, 663–674.
68. Chen, W., Kalscheuer, V., Tzschach, A., Menzel, C., Ullmann, R., Schulz, M.H., Erdogan, F., Li, N., Kijas, Z., Arkesteijn, G., et al. (2008). Mapping translocation breakpoints by next-generation sequencing. Genome Res. *18*, 1143–1149.
69. Turnbaugh, P.J., Ley, R.E., Hamady, M., Fraser-Liggett, C.M., Knight, R., and Gordon, J.I. (2007). The human microbiome project. Nature *449*, 804–810.
70. Frank, D.N., and Pace, N.R. (2008). Gastrointestinal microbiology enters the metagenomics era. Curr. Opin. Gastroenterol. *24*, 4–10.
71. Peterson, D.A., Frank, D.N., Pace, N.R., and Gordon, J.I. (2008). Metagenomic approaches for defining the pathogenesis of inflammatory bowel diseases. Cell Host Microbe *3*, 417–427.
72. Palacios, G., Druce, J., Du, L., Tran, T., Birch, C., Briese, T., Conlan, S., Quan, P.L., Hui, J., Marshall, J., et al. (2008). A new arenavirus in a cluster of fatal transplant-associated diseases. N. Engl. J. Med. *358*, 991–998.
73. Holtz, L.R., Finkbeiner, S.R., Kirkwood, C.D., and Wang, D. (2008). Identification of a novel picornavirus related to cosaviruses in a child with acute diarrhea. Virol. J. *5*, 159.
74. Khoury, M.J., Gwinn, M., Yoon, P.W., Dowling, N., Moore, C.A., and Bradley, L. (2007). The continuum of translation research in genomic medicine: How can we accelerate the appropriate integration of human genome discoveries into health care and disease prevention? Genet. Med. *9*, 665–674.