Massively parallel sequencing: the new frontier of hematologic genomics

Jill M. Johnsen, Deborah A. Nickerson and Alex P. Reiner
Massively parallel sequencing: the new frontier of hematologic genomics

Jill M. Johnsen,1,2 Deborah A. Nickerson,3 and Alex P. Reiner4,5

1Department of Medicine, University of Washington, Seattle, WA; 2Research Institute, Puget Sound Blood Center, Seattle, WA; 3Department of Genome Sciences, University of Washington, Seattle, WA; 4Department of Epidemiology, University of Washington School of Public Health, Seattle, WA; and 5Public Health Sciences Division, Fred Hutchinson Cancer Research Center, Seattle, WA

Genomic technologies are becoming a routine part of human genetic analysis. The exponential growth in DNA sequencing capability has brought an unprecedented understanding of human genetic variation and the identification of thousands of variants that impact human health. In this review, we describe the different types of DNA variation and provide an overview of existing DNA sequencing technologies and their applications. As genomic technologies and knowledge continue to advance, they will become integral in clinical practice. To accomplish the goal of personalized genomic medicine for patients, close collaborations between researchers and clinicians will be essential to develop and curate deep databases of genetic variation and their associated phenotypes. (Blood. 2013;122(19):3268-3275)

Introduction

Modern DNA sequencing technologies have opened the door to the large-scale characterization of human genomes.1-3 Application of these new technologies to individuals and populations offers the unprecedented opportunity to identify and characterize functional human DNA variants amid the diverse spectrum of genomic variation. Appreciation of DNA as a complex and dynamic molecular anthology is essential for the study of inherited and acquired biological processes. In this article, we review the fundamentals of DNA variation as well as several common sequencing approaches, with emphasis on the application and trajectory of next-generation DNA sequencing technology.

Review of terminology and DNA sequence variation

DNA is a living molecule in that it is constantly changing. DNA accumulates variation as time progresses longitudinally over generations (germline variation) and within a single individual over many cell divisions (somatic variation). The vast majority of DNA variants cause no observable phenotype. However, a small fraction of variants are functional and can alter phenotypes.5,6

DNA sequence variation

DNA is a long double-stranded polymer composed of 4 nucleotides which form complementary base pairs (bp) with each other: adenine (A) with thymine (T), and guanine (G) with cytosine (C). Connected 5’ end to 3’ end (referring to the fifth and third carbons of the sugar), these 4 nucleotides are the building blocks of DNA.

DNA is organized into huge, linear, highly structured molecules which form the chromosomes. Chromatin, the physical organization of DNA and associated proteins, participates in regulating DNA function. Genes are the regions of DNA which encode for proteins. Protein coding regions are defined by the presence of exons, read 5’ to 3’, which are made up of codons, triplets of nucleotides which specify amino acids or signal translation stop. The stretches of nonprotein coding DNA between exons are introns. Splice sites mark the exon-intron boundaries and direct the excision of introns from the RNA message.

Control of gene expression

There are numerous functional noncoding DNA elements which participate in gene expression. Promoters, located immediately upstream of genes, are required for gene transcription. DNA regulatory elements which enhance or repress gene expression are often located near (or within introns of) structural genes, but can also lie at great distance. Some elements can control large genomic regions which contain many genes, such as the globin locus control region.4 Additionally, there are numerous DNA regions which transcribe noncoding functional RNAs, for example, transfer RNAs, ribosomal RNAs, and microRNAs. DNA nucleotides can also be reversibly chemically modified, such as by methylation, to affect elements which influence developmental or tissue-specific gene expression, such as occurs during imprinting or cell-lineage differentiation.5,6

© 2013 by The American Society of Hematology
DNA sequencing technologies

In the pregenomic era, various technologies were used to localize disease susceptibility genes (cytogenetics, fluorescence in situ hybridization) or to identify susceptibility alleles using DNA sequence variation in linkage analysis in family-based studies (microsatellite markers) or in candidate gene genotyping in unrelated individuals (restriction fragment length polymorphism [RFLP] analysis, allele-specific polymerase chain reaction [PCR]). The completion of the Human Genome Project and development of dense, genome-wide SNP marker genotyping arrays resulted in dramatic improvements in the design of genetic association studies for complex traits (Figure 2). These technologic advancements made it possible to efficiently screen the human genome for common polymorphisms associated with clinically relevant traits and ushered in the era of genome-wide association studies (GWAS). In the GWAS design, a large fraction of the commonly varying sites across the human genome are assessed either directly or indirectly (through linkage disequilibrium) for association with quantitative or qualitative phenotypes. While some of the genetic variants associated with complex hematologic traits are located within or near genes known to be involved in disease etiology or trait physiology, the genome-wide approach of GWAS led to discovery of previously unknown loci that provided new insights into disease biology. Similarly, comparative genomic hybridization (CGH) arrays based on the comparative cohybridization of fluorescently labeled sample and control DNAs have found CNVs to be common and sometimes associated with disease.

GWAS have exhaustively tested common, usually noncoding, DNA sequence variants and identified many new loci related to hematologic traits. However, rare DNA sequence variants, particularly those within protein-coding sequence, likely also contribute to interindividual variability in the population for hematologic traits or locus heterogeneity for monogenic hematologic syndromes. Recent advances in next-generation DNA sequencing technology allow comprehensive detection of rare DNA sequence variants.

**First-generation DNA sequencing**

DNA sequencing has always been at the core of human genetic analysis because sequencing is the only method that can provide the genotype at every position. In fact, capillary-based, fluorescence-based sequencing, known as Sanger sequencing, continues to be a mainstay technology to rapidly analyze any small region across a handful of samples. For fluorescence-based (Sanger) sequencing, the region of interest is first amplified from the genome by PCR. The amplified target is added to standard nucleotides (A, C, G, T) containing a mix of terminators, which are modified nucleotides each labeled with a different fluorophore. A DNA polymerase copies the target starting from an oligonucleotide primer, and as the DNA is synthesized (extended) the incorporation of fluorescently labeled terminators randomly stops synthesis such that a ladder of differently sized products is generated ending at each base in the target sequence. This cycle is repeated similar to PCR, generating many copies of the laddered products and enhancing the detection of each modified base that terminates a fragment. By subjecting the resulting ladder to single-base-resolution capillary electrophoresis, the fluorescence of the terminator in each fragment (from shortest to longest) is detected (Figure 2). The resulting sequence can exceed 600 bp. Fluorescence-based (Sanger) sequencing is still considered the gold standard, particularly in diagnostic situations. Therefore, this technology is still in common use, although higher-throughput technologies are rapidly being integrated into clinical laboratories.

**Next-generation DNA sequencing**

The development of next-generation sequencing (NGS) has changed the comprehensiveness of human genetic analysis and significantly reduced the costs associated with sequencing a genome. Today, for clinical evaluations, whole-exome sequencing (coding regions only) or whole-genome sequencing (coding and noncoding) of the individual are thought to be appropriate and practical testing modalities. The choice between exome or genome ultimately depends
on cost and the need for noncoding data for clinical assessment. Exome sequencing is currently cheaper than whole-genome sequencing, although that may change in the future. In addition to exome and whole-genome sequencing, specific target gene panels can be optimized for NGS.

**Sample preparation and library construction.** There are several basic steps that are common to all massively parallel sequencing approaches (Figure 3). The first step is the generation of an in vitro library from the sample (DNA, or RNA converted to complementary DNA [cDNA]). The quality of the library is critical in determining sequencing efficiency. Originally, to prepare the sample, DNA fragments were physically sheared. Now, a number of enzymatic approaches have been developed which greatly simplify the process and increase the uniformity of library production. The throughput of NGS has increased to the point where samples from different individuals can often be sequenced together by uniquely bar-coding individual samples during library construction and then pooling samples prior to amplification. After sequencing, the barcodes permit the sequences to be separated or deconvoluted.

**Amplification.** After a library is constructed, the molecules within the library are amplified to generate additional copies, ensuring robust detection. Amplification is one of the steps in the process that introduces biases as it decreases sequence coverage for some regions, that is, GC-rich regions such as some promoters and first exons, and introduces errors prior to sequencing. Errors that arise during the copying process are random, and their presence makes each individual sequence read less accurate. These errors are not usually miscalled as variants because the sequence is ultimately determined by a consensus of multiple unique reads, which are reads distinguished by their unique genomic positions, sequences, and/or lengths. However,
when low-level detection is desired, such as in cancer sequencing, errors introduced during amplification must be considered. Low-level tumor variants are usually called only when they are seen more than twice, giving greater confidence that the observed variant is truly present and not an artifact of the process. Thus, greater depth (usually 300 times or more) among the unique reads is desirable in explorations of heterogeneous samples, as occurs in malignancy. 26,31

Sequencing. There are a number of formats and chemistries used in NGS. 27,28 Many use fluorescent dyes in a manner similar to fluorescence-based (Sanger) sequencing. Sequence detection for NGS is performed in channels, chambers, nanowells, or on assembled nanoballs. What varies substantially among the platforms is the approach used to obtain the sequence. One of the most widely applied technologies (available from Illumina) uses reversible dye terminator sequencing. 29 In this system, the molecular library is captured in a channel and then amplified to generate a small cluster from each captured molecule. Next, DNA polymerase and all 4 dye terminators are flowed through the channel, resulting in fluorescent base extension for each cluster. Fluorescence is then read for the hundreds of millions of clusters found in the channel simultaneously. The dye terminators are then reversed by flowing reagents through the channel to clip off the fluorophore and repair the nucleotide, readying the base to be extended again. This whole process is known as a cycle, which is then repeated. Typically, 100 bp sequence reads are obtained from each end of the cluster, although read lengths from 50 to 200 bp are possible. An entire run from multiple channels can generate ~600 gigabases (Gb) of sequence in an 11-day period. With a new upgrade, 120 Gb can be generated in ~27 hours, or an entire whole genome every day at >30-fold coverage. This is truly massively parallel sequencing, and these approaches continue to improve and evolve.

Other sequence systems routinely generate whole-genome sequence data, such as those from Complete Genomics and Life Technologies. Complete Genomics does not sell an instrument, but provides a service in which a DNA sample is sent to the company and the sequence returned. In their approach, library construction leads to the production of a massive array of nanoballs, which are sequenced by a combination of hybridization and DNA ligation. 32 Life Technologies’ SOLiD platform also uses DNA ligation, rather than DNA polymerization, to sequence on a massive scale. 33

Other next- and third-generation sequencing platforms. Many other NGS platforms are available, and more are under development. Some have sufficient throughput to sequence the human exome, and all are capable of handling targeted gene panels or RNA sequencing applications. The Ion Torrent (Life Technologies) is unique in detecting the slight change in pH that takes place when each base is added 1 nucleotide position at a time 34 and is similar in concept to 454 (Roche) which detects a base addition by the generation of a pyrophosphate. 35 Another unique platform from Pacific Biosciences measures fluorescent base incorporation in single molecules in real time. 35 Although throughput is not as high as other NGS platforms, it can produce long reads up to 25 kilobases (kb) in length and directly detect DNA methylation. 36

Long read length can be advantageous when sequencing more complex regions of the human genome. Single-molecule sequencing holds the promise of long read length capabilities and is an active area of technology development. For example, in nanopore sequencing single DNA molecules are moved through a narrow pore and each base is detected in real time, 38,39 which offers the potential of being able to generate molecule long reads. While this and other promising platforms are possible in the future, existing NGS technologies are also working toward longer reads to improve assembly of individual
genome sequences and detection of difficult to call variants, such as larger indels and structural variants.

**Sequence assembly.** After sequence reads are generated by an NGS platform, they are typically aligned and assembled on a human reference sequence. The human reference sequence serves as a scaffold for read placement using a rapid indexing approach that finds the best match taking into account errors and variants in the reads. Although this form of assembly is not perfect, it is effective and fast at assembling the vast amounts of NGS data. In fact, the majority of the genome (~90%) can be reliably mapped with this approach.

However, not all variants in the sequenced genomes are represented in the reference sequence, and short NGS read length can pose problems in assembly. Particularly challenging are large indels and other structural variations which cannot be assembled simply by aligning reads to the reference scaffold. Emerging tools to identify and characterize these variants include the addition of other sequences into the assembly and/or use of alternative algorithms.\(^{40,41}\) Genes with high sequence similarity (homology) also pose dilemmas by generating highly similar short sequence reads originating from different genes. Obtaining longer paired-end NGS reads will solve many of these issues. The use of hybrid approaches to genome sequence and assembly also holds promise for improving sequence assembly by combining short- and long-read NGS technologies or assembling short reads into longer ones molecularly.\(^{42}\) Additionally, the selected use of de novo sequence assembly, which assembles sequences without a reference scaffold\(^{43,44}\) and/or the application of new approaches to resolve human sequence haplotypes,\(^{45,46}\) could also improve our ability to analyze these challenging regions.

**Variant calling.** Once assembled, sequence variants are called in the dataset. In the early days of NGS, variants were identified by counting the number of times they appeared in unique reads to a set threshold. For example, if a region had 30 overlapping reads where 15 were called C at 1 position and 15 were called T at that same position, a heterozygous C/T variant would be called because each haplotype should be equally represented if the data were obtained in an unbiased fashion. Of course, in reality, sequencing reads obtained by any method are not unbiased, and these biases must be considered in analysis of the data.

The accuracy of SNV calling is high for most NGS platforms, although there is still significant variation among platforms because the error profiles and decoding schemes are not the same. Indels (small and large) and CNVs are more problematic both in specificity and sensitivity. Statistical and machine learning approaches are being applied to find variants and accurately and reproducibly call genotypes,\(^{47}\) and many new algorithms can now handle different types of variants, that is, indels or CNVs.\(^{48,50}\) Target gene panels are already in use for cancer sequencing, where the deeper sequence coverage obtained for each base increases the sensitivity and specificity of identifying variants.\(^{26}\) New approaches are also being developed to identify somatic mutations present at lower levels in samples.\(^{51}\)

Thus, variant-calling capabilities are quite good with some variant types, such as SNVs, but face challenges for calling other, particularly larger, variants. The rapid pace of method development promises even higher sensitivity and specificity in the identification of complex variation in NGS data in the near future.

### Applications of NGS to human genetics

Many lessons are emerging from the large-scale application of NGS in human genetics. Sequencing of the human exome and other large, targeted gene panels reveals that the level of rare variation in the human genome is far greater than expected.\(^{52,53}\) NGS data have also confirmed that the rarest variants in the human population are the youngest ancestrally and predicted to be the most deleterious.\(^{54}\) These findings have stimulated the development of high-throughput approaches to sequence genes in thousands of samples using the same capture methods successfully applied for exome sequencing. Simplifying and increasing the cost effectiveness is a key focus area of technology development. Critical to this will be optimizing approaches for multiplex PCR to capture hundreds or thousands of genomic targets in a single reaction.\(^{55}\) Combined with extensive sample bar coding, hundreds of samples could be sequenced with ease across thousands of regions.\(^{56}\) Already, systems such as molecular inversion probes are offering new ways to advance NGS capabilities and facilitate sequencing of the thousands of individuals required to pursue rare variant discoveries in human genetics.\(^{57}\)

The scale of massively parallel sequencing opens new avenues for all forms of biological analysis, including analysis of sequence variants (shown in Table 1).\(^{58-63}\) Variant discovery and RNA sequencing are the principal applications today for NGS.\(^{64}\) Exome sequencing and genome sequencing have been successful in discovering causal variants in individuals with rare, highly penetrant monogenic disorders.\(^{65}\) The application of exome sequencing and exome-based genotyping arrays to more complex phenotypes in large population samples is under way through consortia such as the National Heart, Lung, and Blood Institute (NHLBI) Exome Sequencing Project (ESP). Early results suggest exome sequencing and newer statistical approaches to analyzing rare variants can be used to further characterize genetically heterogeneous traits in large population-based studies.\(^{66,68}\)

The types of digital profiling that can be tackled by NGS are nearly unlimited, including methylation,\(^{36,65}\) chromatin profiling,\(^{70}\) structural DNA interactions,\(^{71,72}\) as well as many others.\(^{73,74}\) Among the alternative applications, RNA sequencing is the most widely applied. In RNA-seq, read counts are used to measure gene expression in the sample.\(^{75}\) Importantly, RNA sequencing also assesses alternative splicing and isoform usage, which, along with high sensitivity, offers major advantage over microarray analysis. RNA-seq brings new challenges for optimizing sample preparation and analyzing the resulting data, which are active areas of development.\(^{66}\) NGS is emerging as the method of choice for analyzing biology and genomic regulatory elements on a large-scale,\(^{64,77,78}\) highlighted by the myriad of new insights from the ENCODE project.\(^{79}\)

### Table 1. Selected applications of NGS

<table>
<thead>
<tr>
<th>SeQUENCING applications</th>
<th>Examples</th>
<th>Selected references</th>
</tr>
</thead>
<tbody>
<tr>
<td>Genomic variation</td>
<td>Clinical genetics, cancer analysis</td>
<td>24, 25</td>
</tr>
<tr>
<td>Coding and targeted resequencing</td>
<td>Mendelian genetics, rare variant detection, cancer sequencing</td>
<td>22, 23, 26, 53</td>
</tr>
<tr>
<td>RNA profiling</td>
<td>RNA-Seq for expression, alternative splicing and cancer analysis</td>
<td>56, 65</td>
</tr>
<tr>
<td>DNA methylation</td>
<td>Epigenetic profiling</td>
<td>36, 69</td>
</tr>
<tr>
<td>Active regulatory regions</td>
<td>Dnase-seq, FAIRE-Seq</td>
<td>59, 72</td>
</tr>
<tr>
<td>DNA-DNA interactions</td>
<td>Hi-C</td>
<td>70, 71</td>
</tr>
<tr>
<td>Protein-DNA interactions</td>
<td>Chip-Seq, Chia-PET</td>
<td>60, 73</td>
</tr>
<tr>
<td>Mutagenesis profiling</td>
<td>Functional profiling of proteins or DNA elements, ie, (promoters, enhancers)</td>
<td>64, 78, 79</td>
</tr>
<tr>
<td>Immunoprofiling</td>
<td>HLA, T-cell receptor profiling</td>
<td>61-63</td>
</tr>
</tbody>
</table>
Promise and challenges in the genomic era

Benefits of genomic-scale DNA sequencing are being realized as NGS is widely applied in research and being piloted for some clinical applications, such as pharmacogenomics, inherited congenital syndromes, inherited cancer risk genes, and tumor profiling. There are several examples where genome sequencing has been used for diagnosis of rare, Mendelian disorders, and new opportunities are available through the Centers for Mendelian Genomics. Noninvasive prenatal diagnostic screening by genome sequencing is under way. Detailed functional characterization of the consequences of sequence variation in genetic regulatory elements or protein function at single-nucleotide resolution are now possible through massively parallel reporter assay analysis. Genomic technologies are incredibly powerful but still hindered by the physical limits of the amount of DNA which can be sequenced at a time and the computing power needed to analyze the resulting data. As read lengths become longer and DNA sequencing becomes cheaper, the ability to deeply characterize entire genomes is expected to continue its current path of exponential growth. However, as the cost of storing raw data for long periods of time is offset by the uncertain practical utility of maintaining it, there is movement to only maintain analyzed results and, if needed, simply resequence at a future date. The ability to generate comprehensive personal sequence data also increases the likelihood of capturing large numbers of incidental findings; this will require development of consensus on procedures for maintaining and returning incidental results that vary in penetrance, clinical relevance, and medical actionability.

Large-scale genomic sequencing of well-phenotyped population-based cohort or case-control studies may help to define the role of rare or lower frequency genetic variants, perhaps explaining some of the “missing heritability” of common, complex diseases and quantitative traits. However, extremely large sample sizes (in the hundreds of thousands) may be required for adequate statistical power. Consortia such as the NHLBI ESP have begun to evaluate the association of lower-frequency coding variants with hematologic traits, demonstrating the benefits of sequencing approaches to characterize genetically heterogeneous traits in large population-based studies. Another goal of ESP is to share these datasets with the scientific community, both through National Institutes of Health (NIH) repositories of genetic variants (dbGaP, dbSNP) and through the Exome Variant Server (http://evs.gs.washington.edu/EVS/), a web-based application that can be queried by gene or chromosomal location for a detailed summary of all identified sequence variants. Another outgrowth of large exome sequencing consortia has been the development of the Illumina Infinium Human Exome BeadChip, a lower-cost genotyping array that interrogates lower-frequency nonsynonymous, nonsense, and splice-site variants.

The prospect of personalized medicine truly seems to be within reach as DNA sequencing technologies continue to become cheaper, faster, and provide more information. However, there remain...
significant challenges to fulfilling the promise of personalized medicine via deciphering of individual human genomes. Complete reference human genomes and maps of human genetic variation are incomplete, in part due to limitations in detecting structurally complex variants, and in part due to the need for comprehensive DNA sequence data on many individuals from ethnically diverse populations in order to represent human genetic diversity.

In summary, significant expansion of our knowledge of the human genome in conjunction with rapid technological advancements in DNA sequencing technologies has led to the identification of genetic variants responsible for hundreds of diseases. This number will only continue to grow as NGS capabilities are now within reach of most research laboratories and are being developed in clinical settings. Maintenance of the current steep trajectory in the understanding of human genetic diversity and successful application of that knowledge for the benefit of human health requires active collaboration between genomic experts and medical scientists to generate the accessible, deeply characterized, well-annotated, and diverse genomic and biological reference data to realize the potential of the human genome.

References


Acknowledgments

This work was supported by the National Institutes of Health NHLBI Exome Sequencing Project and its ongoing studies, which produced and provided exome variant calls available through the Exome Variant Server: the Lung GO Sequencing Project (HL-102923), the WHI Sequencing Project (HL-102924), the Broad GO Sequencing Project (HL-102925), the Seattle GO Sequencing Project (HL-102926), and the Heart GO Sequencing Project (HL-103010).

Authorship

Contribution: D.A.N., J.M.J., and A.P.R. contributed to writing the manuscript.
Conflict-of-interest disclosure: The authors declare no competing financial interests.
Correspondence: Jill M. Johnsen, Research Institute, Puget Sound Blood Center, Seattle, WA 98104; e-mail: JillJ@psbc.org.

From bloodjournal.hematologylibrary.org by RAUL RIBEIRO on November 25, 2013. For personal use only.


