Clinical Chemistry 55:4
641–658 (2009)

Reviews

Next-Generation Sequencing: From Basic Research to Diagnostics

Karl V. Voelkerding,[1,2*] Shale A. Dames,[1†] and Jacob D. Durtschi[1†]

BACKGROUND: For the past 30 years, the Sanger method has been the dominant approach and gold standard for DNA sequencing. The commercial launch of the first massively parallel pyrosequencing platform in 2005 ushered in the new era of high-throughput genomic analysis now referred to as next-generation sequencing (NGS).

CONTENT: This review describes fundamental principles of commercially available NGS platforms. Although the platforms differ in their engineering configurations and sequencing chemistries, they share a technical paradigm in that sequencing of spatially separated, clonally amplified DNA templates or single DNA molecules is performed in a flow cell in a massively parallel manner. Through iterative cycles of polymerase-mediated nucleotide extensions or, in one approach, through successive oligonucleotide ligations, sequence outputs in the range of hundreds of megabases to gigabases are now obtained routinely. Highlighted in this review are the impact of NGS on basic research, bioinformatics considerations, and translation of this technology into clinical diagnostics. Also presented is a view into future technologies, including real-time single-molecule DNA sequencing and nanopore-based sequencing.

SUMMARY: In the relatively short time frame since 2005, NGS has fundamentally altered genomics research and allowed investigators to conduct experiments that were previously not technically feasible or affordable. The various technologies that constitute this new paradigm continue to evolve, and further improvements in technology robustness and process streamlining will pave the path for translation into clinical diagnostics.

© 2009 American Association for Clinical Chemistry

In 1977, 2 landmark articles describing methods for DNA sequencing were published. Allan Maxam and Walter Gilbert reported an approach in which terminally labeled DNA fragments were subjected to base-specific chemical cleavage and the reaction products were separated by gel electrophoresis *(1)*. In an alternative approach, Frederick Sanger and colleagues described the use of chain-terminating dideoxynucleotide analogs that caused base-specific termination of primed DNA synthesis *(2)*. Refinement and commercialization of the latter method led to its broad dissemination throughout the research community and, ultimately, into clinical diagnostics. In an industrial, high-throughput configuration, Sanger technology was used in the sequencing of the first human genome, which was completed in 2003 through the Human Genome Project, a 13-year effort with an estimated cost of $2.7 billion. In 2008, by comparison, a human genome was sequenced over a 5-month period for approximately $1.5 million *(3)*. The latter accomplishment highlights the capabilities of the rapidly evolving field of "next-generation" sequencing (NGS)[3] technologies that have emerged during the past 5 years. Currently, 5 NGS platforms are commercially available, with additional platforms on the horizon. To add to this pace, the US National Human Genome Research Institute (NHGRI) announced funding in August 2008 for a series of projects as part of its Revolutionary Genome Sequencing Technologies program, which has as its goal the sequencing of a human genome for $1000 or less (http://www.genome.gov/27527585). This review describes NGS technologies, reviews their impact on basic research, and explores how they have the translational potential to substantially impact molecular diagnostics.

Fundamentals of NGS Platforms

NGS platforms share a common technological feature—massively parallel sequencing of clonally amplified or single DNA molecules that are spatially separated in a flow cell. This design is a paradigm shift from that of

[1] ARUP Institute for Experimental and Clinical Pathology, Salt Lake City, Utah; [2] Department of Pathology, University of Utah, Salt Lake City, Utah.
* Address correspondence to this author at: ARUP Laboratories, 500 Chipeta Way, Salt Lake City, Utah 84108. Fax (801) 584-5207; e-mail voelkek@aruplab.com.
† S.A. Dames and J.D. Durtschi contributed equally to the review.
Received October 7, 2008; accepted January 29, 2009.
Previously published online at DOI: 10.1373/clinchem.2008.112789

[3] Nonstandard abbreviations: NGS, next-generation sequencing; NHGRI, National Human Genome Research Institute; dNTP, deoxynucleoside triphosphate; Mb, million base pairs; Gb, billion base pairs; miRNA, microRNA.

Petitioner Illumina, Inc., EX1032, p. 641

Downloaded from https://academic.oup.com/clinchem/article/55/4/641/5629392 by Broad Institute user on 05 September 2024

Sanger sequencing, which is based on the electrophoretic separation of chain-termination products produced in individual sequencing reactions. In NGS, sequencing is performed by repeated cycles of polymerase-mediated nucleotide extensions or, in one format, by iterative cycles of oligonucleotide ligation. As a massively parallel process, NGS generates hundreds of megabases to gigabases of nucleotide-sequence output in a single instrument run, depending on the platform. These platforms are reviewed next.

ROCHE/454 LIFE SCIENCES

The 454 technology (http://www.454.com) is derived from the technological convergence of pyrosequencing and emulsion PCR. In 1993, Nyren et al. described a sequencing approach based on chemiluminescent detection of pyrophosphate released during polymerase-mediated deoxynucleoside triphosphate (dNTP) incorporation *(4)*. Refinement by Ronaghi et al. served as the foundation for the commercial development of pyrosequencing *(5, 6)*. On a separate front, Tawfik and Griffiths described single-molecule PCR in microcompartments consisting of water-in-oil emulsions *(7)*. In 2000, Jonathan Rothberg founded 454 Life Sciences, which developed the first commercially available NGS platform, the GS 20, launched in 2005. Combining single-molecule emulsion PCR with pyrosequencing, Margulies and colleagues at 454 Life Sciences performed shotgun sequencing of the entire 580 069 bp of the *Mycoplasma genitalia* genome at 96% coverage and 99.96% accuracy in a single GS 20 run *(8)*. In 2007, Roche Applied Science acquired 454 Life Sciences and introduced the second version of the 454 instrument, the GS FLX. Sharing the same core technology as the GS 20, the GS FLX flow cell is referred to as a "picotiter well" plate, which is made from a fused fiber-optic bundle. In its newest configuration, approximately 3.4×10^6 picoliter-scale sequencing-reaction wells are etched into the plate surface, and the well walls have a metal coating to improve signal-to-noise discrimination. For sequencing (Fig. 1), a library of template DNA is prepared by fragmentation via nebulization or sonication. Fragments several hundred base pairs in length are end-repaired and ligated to adapter oligonucleotides. The library is then diluted to single-molecule concentration, denatured, and hybridized to individual beads containing sequences complementary to adapter oligonucleotides. The beads are compartmentalized into water-in-oil microvesicles, where clonal expansion of single DNA molecules bound to the beads occurs during emulsion PCR. After amplification, the emulsion is disrupted, and the beads containing clonally amplified template DNA are enriched. The beads are again separated by limiting dilution, deposited into individual picotiter-plate wells, and combined with sequencing enzymes. Loaded into the GS FLX, the picotiter plate functions as a flow cell wherein iterative pyrosequencing is performed by successive flow addition of the 4 dNTPs. A nucleotide-incorporation event in a well containing clonally amplified template produces pyrophosphate release and picotiter-plate well–localized luminescence, which is transmitted through the fiber-optic plate and recorded on a charge-coupled device camera. With the flow of each dNTP reagent, wells are imaged, analyzed for their signal-to-noise ratio, filtered according to quality criteria, and subsequently algorithmically translated into a linear sequence output. With the newest chemistry, termed "Titanium," a single GS FLX run generates approximately 1×10^6 sequence reads, with read lengths of $\geq$400 bases yielding up to 500 million base pairs (Mb) of sequence. A recognized strength of the 454 technology is the longer read length, which facilitates de novo assembly of genomes *(9)*. An outstanding concern has been the accurate determination of homopolymers $>$3–4 bases in length. A 6-base homopolymer should theoretically yield twice the luminescence of a 3-base homopolymer. Operationally, this luminescence yield varies, and estimates of homopolymer length are less accurate with increasing length *(8, 10)*. 454 has reported that the metal coating of the walls of picotiter wells mentioned above improves the accuracy of homopolymer determination. Sequence coverage depth and accuracy for the 454 technology is discussed below in the NGS Data Analysis section.

ILLUMINA/SOLEXA

In 1997, British chemists Shankar Balasubramanian and David Klenerman conceptualized an approach for sequencing single DNA molecules attached to microspheres. They founded Solexa in 1998, and their goal during early development of sequencing single DNA molecules was not achieved, requiring a shift toward sequencing clonally amplified templates. By 2006, the Solexa Genome Analyzer, the first "short read" sequencing platform, was commercially launched. Acquired by Illumina (http://www.Illumina.com) in 2006, the Genome Analyzer uses a flow cell consisting of an optically transparent slide with 8 individual lanes on the surfaces of which are bound oligonucleotide anchors (Fig. 2). Template DNA is fragmented into lengths of several hundred base pairs and end-repaired to generate 5′-phosphorylated blunt ends. The polymerase activity of Klenow fragment is used to add a single A base to the 3′ end of the blunt phosphorylated DNA fragments. This addition prepares the DNA fragments for ligation to oligonucleotide adapters, which have an overhang of a single T base at their 3′ end to increase ligation efficiency. The adapter oligonucleotides are complementary to the flow-cell anchors. Un-

Downloaded from https://academic.oup.com/clinchem/article/55/4/641/5629392 by Broad Institute user on 05 September 2024

Petitioner Illumina, Inc., EX1032, p. 642

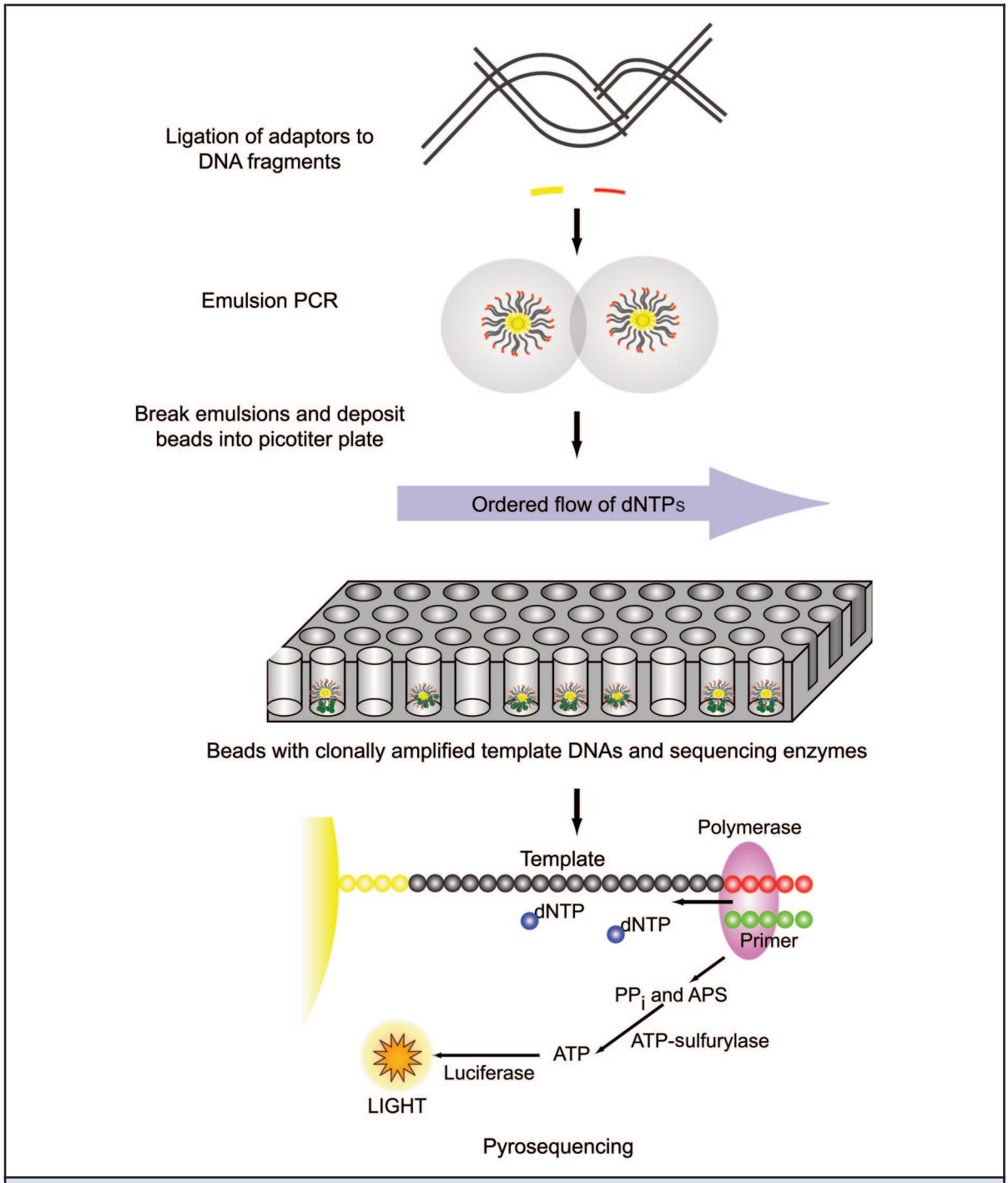

Fig. 1. Roche 454 GS FLX sequencing.

Template DNA is fragmented, end-repaired, ligated to adapters, and clonally amplified by emulsion PCR. After amplification, the beads are deposited into picotiter-plate wells with sequencing enzymes. The picotiter plate functions as a flow cell where iterative pyrosequencing is performed. A nucleotide-incorporation event results in pyrophosphate (PP_i) release and well-localized luminescence. APS, adenosine 5′-phosphosulfate.

Downloaded from https://academic.oup.com/clinchem/article/55/4/641/5629392 by Broad Institute user on 05 September 2024

Petitioner Illumina, Inc., EX1032, p. 643

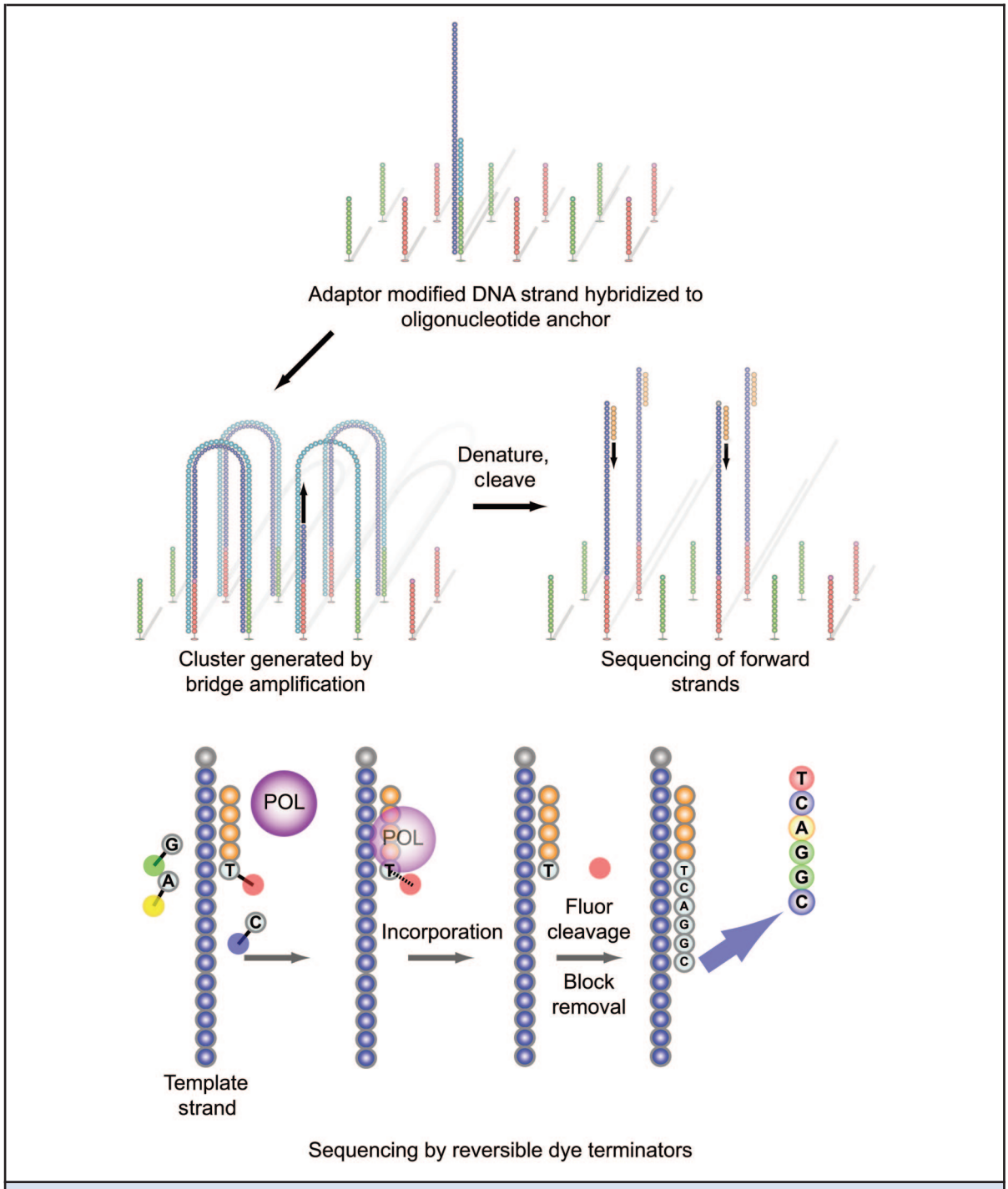

Fig. 2. Illumina Genome Analyzer sequencing.

Adapter-modified, single-stranded DNA is added to the flow cell and immobilized by hybridization. Bridge amplification generates clonally amplified clusters. Clusters are denatured and cleaved; sequencing is initiated with addition of primer, polymerase (POL) and 4 reversible dye terminators. Postincorporation fluorescence is recorded. The fluor and block are removed before the next synthesis cycle.

Downloaded from https://academic.oup.com/clinchem/article/55/4/641/5629392 by Broad Institute user on 05 September 2024

Petitioner Illumina, Inc., EX1032, p. 644

der limiting-dilution conditions, adapter-modified, single-stranded template DNA is added to the flow cell and immobilized by hybridization to the anchors. In contrast to emulsion PCR, DNA templates are amplified in the flow cell by "bridge" amplification, which relies on captured DNA strands "arching" over and hybridizing to an adjacent anchor oligonucleotide. Multiple amplification cycles convert the single-molecule DNA template to a clonally amplified arching "cluster," with each cluster containing approximately 1000 clonal molecules. Approximately 50×10^6 separate clusters can be generated per flow cell. For sequencing, the clusters are denatured, and a subsequent chemical cleavage reaction and wash leave only forward strands for single-end sequencing. Sequencing of the forward strands is initiated by hybridizing a primer complementary to the adapter sequences, which is followed by addition of polymerase and a mixture of 4 differently colored fluorescent reversible dye terminators. The terminators are incorporated according to sequence complementarity in each strand in a clonal cluster. After incorporation, excess reagents are washed away, the clusters are optically interrogated, and the fluorescence is recorded. With successive chemical steps, the reversible dye terminators are unblocked, the fluorescent labels are cleaved and washed away, and the next sequencing cycle is performed. This iterative, sequencing-by-synthesis process requires approximately 2.5 days to generate read lengths of 36 bases. With 50×10^6 clusters per flow cell, the overall sequence output is >1 billion base pairs (Gb) per analytical run *(11)*. The newest platform, the Genome Analyzer II, has optical modifications enabling analysis of higher cluster densities. Coupled with ongoing improvements in sequencing chemistry and projected read lengths of 50-plus bases, further increases in output should be realized. Illumina and other NGS technologies have devised strategies to sequence both ends of template molecules. Such "paired-end" sequencing provides positional information that facilitates alignment and assembly, especially for short reads *(12, 13)*. A technical concern of Illumina sequencing is that base-call accuracy decreases with increasing read length *(14)*. This phenomenon is primarily due to "dephasing noise." During a given sequencing cycle, nucleotides can be under- or overincorporated, or block removal can fail. With successive cycles, these aberrations accumulate to produce a heterogeneous population in a cluster of strands of varying lengths. This heterogeneity decreases signal purity and reduces precision in base calling, especially at the 3′ ends of reads. Modifications in sequencing chemistry and algorithms for data-image analysis and interpretation are being pursued to mitigate dephasing *(15)*. Investigators at the Wellcome Trust Sanger Institute, who have extensive experience with the Illumina platform, have published a series of technical improvements for library preparation, including methods for increasing the reproducibility of fragmentation by adaptive focused acoustic wave sonication, enhanced efficiency of adapter ligation by use of an alternate ligase, and reducing the G+C bias that has been observed in Illumina reads via a modified gel-extraction protocol *(16)*.

APPLIED BIOSYSTEMS/SOLiD

The SOLiD (Supported Oligonucleotide Ligation and Detection) System 2.0 platform, which is distributed by Applied Biosystems (http://www.solid.appliedbiosystems.com), is a short-read sequencing technology based on ligation. This approach was developed in the laboratory of George Church and reported in 2005 along with the resequencing of the *Escherichia coli* genome *(17)*. Applied Biosystems refined the technology and released the SOLiD instrumentation in 2007. Sample preparation shares similarities with the 454 technology in that DNA fragments are ligated to oligonucleotide adapters, attached to beads, and clonally amplified by emulsion PCR. Beads with clonally amplified template are immobilized onto a derivitized-glass flow-cell surface, and sequencing is begun by annealing a primer oligonucleotide complementary to the adapter at the adapter–template junction (Fig. 3). Instead of providing a 3′ hydroxyl group for polymerase-mediated extension, the primer is oriented to provide a 5′ phosphate group for ligation to interrogation probes during the first "ligation sequencing" step. Each interrogation probe is an octamer, which consists of (in the 3′-to-5′ direction) 2 probe-specific bases followed by 6 degenerate bases with one of 4 fluorescent labels linked to the 5′ end. The 2 probe-specific bases consist of one of 16 possible 2-base combinations (for example TT, GT, and so forth). In the first ligation-sequencing step, thermostable ligase and interrogation probes representing the 16 possible 2-base combinations are present. The probes compete for annealing to the template sequences immediately adjacent to the primer. After annealing, a ligation step is performed, followed by wash removal of unbound probe. Fluorescence signals are optically collected before cleavage of the ligated probes, and a wash is performed to remove the fluor and regenerate the 5′ phosphate group. In the subsequent sequencing steps, interrogation probes are ligated to the 5′ phosphate group of the preceding pentamer. Seven cycles of ligation, referred to as a "round," are performed to extend the first primer. The synthesized strand is then denatured, and a new sequencing primer offset by 1 base in the adapter sequence ($n - 1$) is annealed. Five rounds total are performed, each time with a new primer with a successive offset ($n - 2$, $n - 3$, and so on). By this approach, each template nucleo-

Downloaded from https://academic.oup.com/clinchem/article/55/4/641/5629392 by Broad Institute user on 05 September 2024

Petitioner Illumina, Inc., EX1032, p. 645

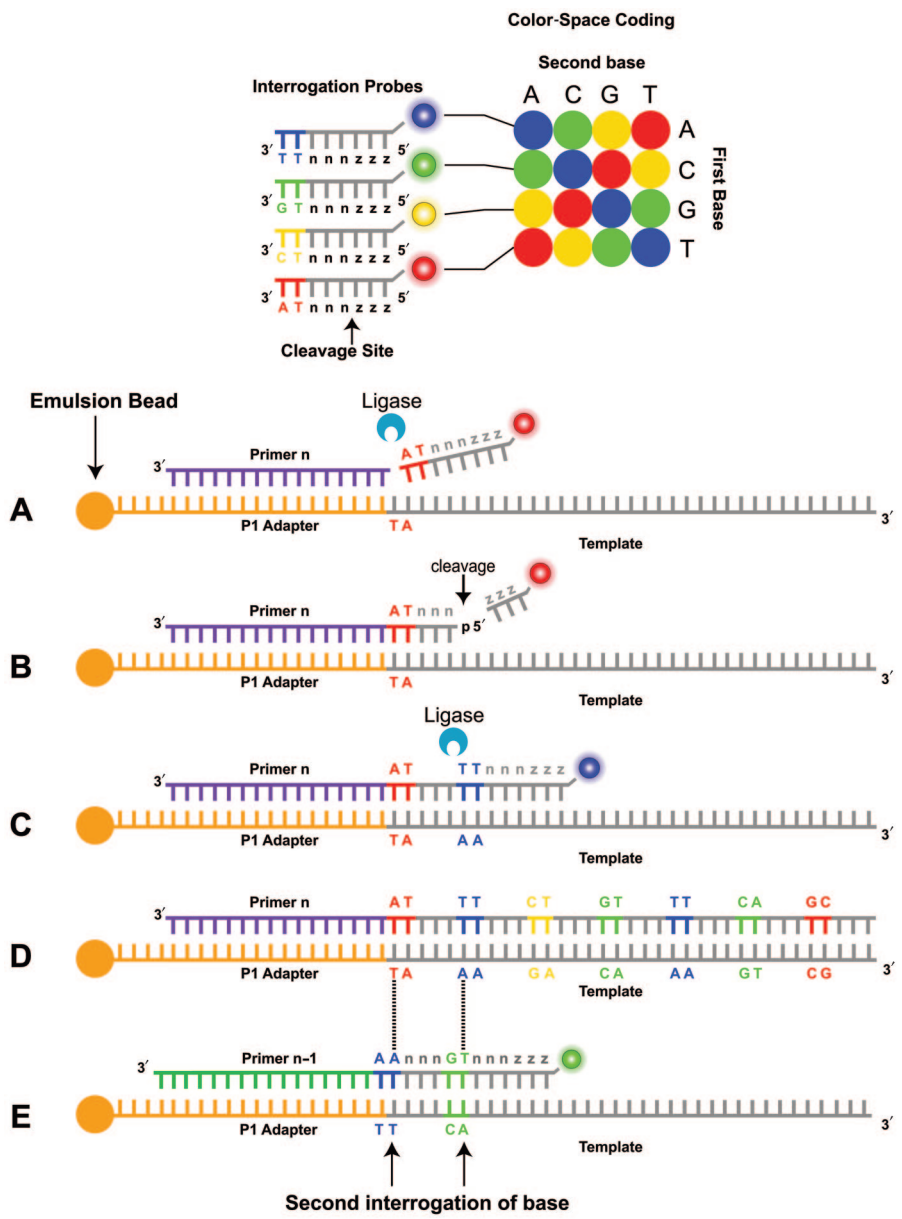

Fig. 3. Applied Biosystems SOLiD sequencing by ligation.

Top: SOLiD color-space coding. Each interrogation probe is an octamer, which consists of (3′-to-5′ direction) 2 probe-specific bases followed by 6 degenerate bases (nnnzzz) with one of 4 fluorescent labels linked to the 5′ end. The 2 probe-specific bases consist of one of 16 possible 2-base combinations. Bottom: (A), The P1 adapter and template with annealed primer (n) is interrogated by probes representing the 16 possible 2-base combinations. In this example, the 2 specific bases complementary to the template are AT. (B), After annealing and ligation of the probe, fluorescence is recorded before cleavage of the last 3 degenerate probe bases. The 5′ end of the cleaved probe is phosphorylated (not shown) before the second sequencing step. (C), Annealing and ligation of the next probe. (D), Complete extension of primer (n) through the first round consisting of 7 cycles of ligation. (E), The product extended from primer (n) is denatured from the adapter/template, and the second round of sequencing is performed with primer (n − 1). With the use of progressively offset primers, in this example (n − 1), adapter bases are sequenced, and this known sequence is used in conjunction with the color-space coding for determining the template sequence by deconvolution (see Fig. 1 in the online Data Supplement). In this technology, template bases are interrogated twice.

Downloaded from https://academic.oup.com/clinchem/article/55/4/641/5629392 by Broad Institute user on 05 September 2024

Petitioner Illumina, Inc., EX1032, p. 646

Table 1. Comparison of NGS platforms.

	Roche 454 GS FLX	Illumina Genome Analyzer	Applied Biosystems SOLiD	Sanger
Sequencing method	Pyrosequencing	Reversible dye terminators	Sequencing by ligation	Dye terminators
Read lengths	400 bases	36 bases	35 bases	800 bp
Sequencing run time	10 h	2.5 days	6 days	3 h
Total bases per run	500 Mb	1.5 Gb	4 Gb	800 bp

tide is sequenced twice. A 6-day instrument run generates sequence read lengths of 35 bases. Sequence is inferred by interpreting the ligation results for the 16 possible 2 base–combination interrogation probes. With the use of offset primers, several bases of the adapter are sequenced. This information provides a sequence reference starting point that is used in conjunction with the color space–coding scheme to algorithmically deconvolute the downstream template sequence (see Fig. 1 in the Data Supplement that accompanies the online version of this review at http://www.clinchem.org/content/vol55/issue4). Placing 2 flow-cell slides in the instrument per analytical run produces a combined output of 4 Gb of sequence or greater. Unextended strands are capped before the ligation to mitigate signal deterioration due to dephasing. Capping coupled with high-fidelity ligation chemistry and interrogation of each nucleotide base twice during independent ligation cycles yields a company-reported sequence consensus accuracy of 99.9% for a known target at a 15-fold sequence coverage over sequence reads of 25 nucleotides. On an independent track, the Church laboratory has collaborated with Danaher Motion and Dover Systems to develop and introduce an alternative sequencing-by-ligation platform, the Polonator G.007 (http://www.polonator.org). Table 1 summarizes GS FLX, Genome Analyzer, and SOLiD platform features.

HELICOS BIOSCIENCES AND SINGLE-MOLECULE SEQUENCING

The first single-molecule sequencing platform, the HeliScope, is now available from Helicos BioSciences (http://www.helicosbio.com) with a company-reported sequence output of 1 Gb/day. This technology stems from the work of Braslavsky et al., published in 2003 *(18)*. Having obviated clonal amplification of template, the method involves fragmenting sample DNA and polyadenylation at the 3′ end, with the final adenosine fluorescently labeled. Denatured polyadenylated strands are hybridized to poly(dT) oligonucleotides immobilized on a flow-cell surface at a capture density of up to 100×10^6 template strands per square centimeter. After the positional coordinates of the captured strands are recorded by a charge-coupled device camera, the label is cleaved and washed away before sequencing. For sequencing, polymerase and one of 4 Cy5-labeled dNTPs are added to the flow cell, which is imaged to determine incorporation into individual strands. After label cleavage and washing, the process is repeated with the next Cy5-labeled dNTP. Each sequencing cycle, which consists of the successive addition of polymerase and each of the 4 labeled dNTPs, is termed a "quad." The number of sequencing quads performed is approximately 25–30, with read lengths of up to 45–50 bases having been achieved. The Helicos platform was used to sequence the 6407-base genome of bacteriophage M13 *(19)*. This study demonstrated both the potential and important technical issues that may be relevant to all single-molecule sequencing methods that are based on sequencing by synthesis. First, sequencing accuracy was appreciably improved when template molecules were sequenced twice ("2-pass" sequencing). Second, the accuracy of sequencing homopolymers was compromised by the polymerase adding additional bases of the same identity in a homopolymeric stretch in a given dNTP addition. Helicos has since developed proprietary labeled dNTPs, termed "virtual terminators," which the company reports reduce polymerase processivity so that only single bases are added, improving the accuracy of homopolymer sequencing. Interestingly, the percentage of strands in which longer read lengths can be achieved (e.g., 50 nucleotides) is substantially lower than that obtained with shorter (e.g., 25 nucleotides) read-length sequencing, possibly reflecting secondary structures (e.g., hairpins) assumed by the template molecules.

The Impact of NGS on Basic Research

In the short 4 years since the first commercial platform became available, NGS has markedly accelerated multiple areas of genomics research, enabling experiments that previously were not technically feasible or affordable. We describe major applications of NGS and then review the analysis of NGS data.

Petitioner Illumina, Inc., EX1032, p. 647

Downloaded from https://academic.oup.com/clinchem/article/55/4/641/5629392 by Broad Institute user on 05 September 2024

GENOMIC ANALYSIS

The high-throughput capacity of NGS has been leveraged to sequence entire genomes, from microbes to humans *(3, 8, 9, 11, 20–24)*, including the recent sequencing of the genome of cytogenetically normal acute myeloid leukemia cells, which identified novel, tumor-specific gene mutations *(25)*. The longer read lengths of the 454 technology, compared with the Illumina and SOLiD short-read technologies, facilitate the assembly of genomes in the absence of a reference genome (i.e., de novo assembly). For resequencing, both long- and short-read technologies have been used successfully. In one comparative study, the 454, Illumina, and SOLiD technologies all accurately detected single-nucleotide variations when coverage depth was ≥15-fold per allele *(20)* (the critical issue of coverage depth is discussed further in the NGS Data Analysis section). The 454 read lengths provide nucleotide haplotype information over a range of several hundred base pairs and are predicted to be better suited for detecting larger insertions and deletions and for producing alignments in areas containing repetitive sequences. Further studies are needed to compare technology performance for detecting insertions and deletions. Each platform has an optional strategy for sequencing both ends of DNA libraries (paired-end sequencing). In addition to effectively doubling sequence output, knowing that reads are associated with each other on a given fragment augments alignment and assembly, especially for short reads. Paired-end sequencing has been used to map genomic structural variation, including deletions, insertions, and rearrangements *(12, 13, 26, 27)*. The ability to sequence complete human genomes at a substantially reduced cost with NGS has energized an international effort to sequence thousands of human genomes over the next decade (http://www.1000genomes.org), which will lead to the characterization and cataloging of human genetic variation at an unprecedented level.

TARGETED GENOMIC RESEQUENCING

Sequencing of genomic subregions and gene sets is being used to identify polymorphisms and mutations in genes implicated in cancer and in regions of the human genome that linkage and whole-genome association studies have implicated in disease *(28, 29)*. Especially in the latter setting, regions of interest can be hundreds of kb's to several Mb in size. To best use NGS for sequencing such candidate regions, several genomic-enrichment steps, both traditional and novel, are being incorporated into overall experimental designs. Overlapping long-range PCR amplicons (approximately 5–10 kb) can be used for up to several hundred kb's, but this approach is not practical for larger genomic regions. More recently, enrichment has been achieved by hybridizing fragmented, denatured human genomic DNA to oligonucleotide capture probes complementary to the region of interest and subsequently eluting the enriched DNA *(30–33)*. Capture probes can be immobilized on a solid surface (Roche NimbleGen, http://www.nimblegen.com; Agilent Technologies, http://www.agilent.com; and Febit, http://www.febit.com) or used in solution (Agilent). Current NimbleGen arrays contain 350 000 oligonucleotides of 60–90 bp in length that are typically spaced 5–20 nucleotides apart, with oligonucleotides complementary to repetitive regions being excluded. For enrichment, 5–20 μg of genomic DNA is fragmented and ligated to oligonucleotide linkers containing universal PCR priming sites. This material is denatured, hybridized to an array for 3 days, and eluted, with the enriched DNA amplified by the PCR before NGS library preparation. In reported studies, up to 5 Mb of sequence has been captured on the 350K array, with 60%–75% of sequencing reads mapping to targeted regions; other reads mapping to nontargeted regions reflect nonspecific capture. In development by NimbleGen is the use of an array of 2.1×10^6 features for capturing larger genomic regions. Agilent's solution-based technology uses oligonucleotides up to 170 bases in length, with each end containing sequences for universal PCR priming and with primer sites containing a restriction endonuclease–recognition sequence. The oligonucleotide library is amplified by the PCR, digested with restriction enzymes, and ligated to adapters containing the T7 polymerase promoter site. In vitro transcription is performed with biotinylated UTP to generate single-stranded biotinylated cRNA capture sequences. For capture, 3 μg of fragmented, denatured genomic DNA is hybridized with cRNA sequences for 24 h in solution. After hybridization, duplexes consisting of single-stranded DNA and cRNA are bound to streptavidin-coated magnetic beads; the cRNA is then enzymatically digested, leaving enriched single-stranded DNA that is subsequently processed for NGS. An alternative enrichment approach developed by RainDance Technologies (http://www.raindancetechnologies.com) uses a novel microfluidics technology in which individual pairs of PCR primers for the genomic regions of interest are segregated in water in emulsion droplets and then pooled to create a "primer library." Separately, emulsion droplets containing genomic DNA and PCR reagents are prepared. Two separate droplet streams are created, one with primer-library droplets and the other with droplets containing genomic DNA/PCR reagents. The 2 streams are merged and primer-library droplets and genomic DNA/PCR reagent droplets are paired in a 1:1 ratio. As paired droplets proceed through the microfluidic channel, they pass an electrical impulse that causes them to physically coalesce. The

Downloaded from https://academic.oup.com/clinchem/article/55/4/641/5629392 by Broad Institute user on 05 September 2024

Petitioner Illumina, Inc., EX1032, p. 648

coalesced droplets containing individual primer pairs and genomic DNA/PCR reagents are deposited in a 96-well plate and amplified by the PCR. After amplification, the emulsions are disrupted, and the amplicons are pooled and processed for NGS.

METAGENOMICS

NGS has had a tremendous impact on the study of microbial diversity in environmental and clinical samples. Operationally, genomic DNA is extracted from the sample of interest, converted to an NGS library and sequenced. The sequence output is aligned to known reference sequences for microorganisms that are predicted to be present in the sample. Closely related species can be discerned, and more distantly related species can be inferred. In addition, de novo assembly of the data set can yield information to support the presence of known and potentially new species. Qualitative genomic information is obtained, and analysis of the relative abundance of the sequence reads can be used to derive quantitative information on individual microbial species. To date, most NGS-based metagenomic analyses have used the 454 technology and its associated longer read lengths to facilitate alignment to microbial reference genomes and for de novo assembly of previously uncharacterized microbial genomes. Examples of metagenomic studies include the analysis of microbial populations in the ocean *(34, 35)* and soil *(36)*, the identification of a novel arenavirus in transplantation patients *(37)*, and the characterization of microflora present in the human oral cavity *(38)* and the guts of obese and lean twins *(39)*.

TRANSCRIPTOME SEQUENCING

NGS has provided a powerful new approach, termed "RNA-Seq," for mapping and quantifying transcripts in biological samples. Total, ribosomal RNA–depleted, or poly(A)$^+$ RNA is isolated and converted to cDNA. A typical protocol would involve the generation of first-strand cDNA via random hexamer–primed reverse transcription and subsequent generation of second-strand cDNA with RNase H and DNA polymerase. The cDNA is then fragmented and ligated to NGS adapters. For small RNAs such as microRNAs (miRNAs) and short interfering RNAs, preferential isolation via a small RNA–enrichment method, size selection on an electrophoresis gel, or a combination of these approaches is commonly used. RNA ligase is used to join adapter sequences to the RNA; this step is often followed by a PCR amplification step before NGS processing. After sequencing, reads are aligned to a reference genome, compared with known transcript sequences, or assembled de novo to construct a genome-scale transcription map. Although RNA-Seq is in its early stages as a technology, it has already shown some advantages over gene expression arrays *(40)*. First, arrays depend on tiling existing genomic sequences, whereas RNA-Seq is not constrained by this limitation, allowing characterization of transcription without prior knowledge of the genomic sites of transcription origin. RNA-Seq is capable of single-base resolution and, compared with arrays, demonstrates a greater ability to distinguish RNA isoforms, determine allelic expression, and reveal sequence variants. Expression levels are deduced from the total number of reads that map to the exons of a gene, normalized by the length of exons that can be uniquely mapped. Results obtained with this approach have shown close correlation with those of quantitative PCR and RNA-spiking experiments. The dynamic range of RNA-Seq for determining expression levels is 3–4 orders of magnitude, compared with 2 orders of magnitude for expression arrays. In this context, RNA-Seq has shown improved performance for the quantitative detection of both highly produced transcripts and transcripts produced at low levels. RNA-Seq is being used to confirm and revise gene annotation, including 5′ and 3′, and exon/intron boundaries; the latter is achieved by mapping reads to exon junctions defined by GT-AG splicing consensus sites. Both qualitative and quantitative information regarding splicing diversity can be deduced. RNA-Seq has been applied to a variety of organisms, including *Saccharomyces cerevisiae*, *Arabidopsis thaliana*, mice, and human cells *(40–51)*.

MAPPING OF DNA-BINDING PROTEINS AND CHROMATIN ANALYSIS

The delineation of regulatory proteins associated with genomes was substantially accelerated by the introduction of chromatin immunoprecipitation and microarray hybridization (ChIP-on-chip) technology *(52)*. In this approach, proteins in contact with genomic DNA are chemically cross-linked (typically with mild formaldehyde treatment) to their binding sites, and the DNA is fragmented by sonication or digestion with micrococcal nuclease. The proteins cross-linked with DNA are immunoprecipitated with antibodies specific for the proteins of interest. The DNA in the immunoprecipitate is purified and hybridized to an oligonucleotide array consisting of sequences from the genome, allowing identification of the protein-binding sites. This approach has been successfully used to identify binding sites for transcription factors and histone proteins. ChIP-on-chip technology is now being supplanted in a variety of experimental settings with ChIP-Seq, in which the DNA harvested from the immunoprecipitate is converted into a library for NGS. The obtained reads are mapped to the reference genome of interest to generate a genome-wide protein-binding map *(53–55)*. Studies to date that have exam-

Downloaded from https://academic.oup.com/clinchem/article/55/4/641/5629392 by Broad Institute user on 05 September 2024

Petitioner Illumina, Inc., EX1032, p. 649

ined the genomic-binding sites of the human NRSF (neuron restrictive silencer factor) and STAT1 (signal transducer and activator of transcription 1) proteins indicate the resolution of ChIP-Seq to be greater than for ChIP-on-chip, as evidenced by confirmation of previously identified binding sites and identification of novel binding sites *(56, 57)*. Analogous to RNA-Seq, ChIP-Seq has the important advantage of not requiring prior knowledge of genomic locations of protein binding. In addition to the study of transcription factors, NGS is being used to map genomic methylation. One approach involves traditional bisulfite conversion of DNA followed by NGS, which has been applied to the study of entire genomes or genomic subregions *(58, 59)*. Ongoing studies are attempting to develop a variant of ChIP-Seq in which genomic methylation is assayed by coupling immunoprecipitation with a monoclonal antibody directed against methylated cytosine and subsequent NGS *(60)*.

NGS Data Analysis

NGS experiments generate unprecedented volumes of data, which present challenges and opportunities for data management, storage, and, most importantly, analysis *(61)*. NGS data begin as large sets of tiled fluorescence or luminescence images of the flow-cell surface recorded after each iterative sequencing step (Fig. 4). This volume of data requires a resource-intensive data-pipeline system for data storage, management, and processing. Data volumes generated during single runs of the 454 GS FLX, Illumina, and SOLiD instruments are approximately 15 GB, 1 TB, and 15 TB, respectively. The main processing feature of the data pipeline is the computationally intensive conversion of image data into sequence reads, known as base calling. First, individual beads or clusters are identified and localized in an image series. Image parameters such as intensity, background, and noise are then used in a platform-dependant algorithm to generate read sequences and error probability–related quality scores for each base. Although many researchers use the base calls generated by the platform-specific data-pipeline software, alternative base-calling programs that use more advanced software and statistical techniques have been developed. Features of these alternative programs include the incorporation of ambiguous bases into reads, improved removal of poor-quality bases from read ends *(62)*, and the use of data sets for software training *(15)*. Incorporation of these features has been shown to reduce read error and improve alignment, especially as platforms are pushed to generate longer reads. These advantages, however, must be weighed against the substantial computer resources required by the large volumes of image data.

The quality values calculated during NGS base calling provide important information for alignment, assembly, and variant analysis. Although the calculation of quality varies between platforms, the calculations are all related to the historically relevant phred score, introduced in 1998 for Sanger sequence data *(63, 64)*. The phred score quality value, q, uses a mathematical scale to convert the estimated probability of an incorrect call, e, to a log scale:

$$q = -10 \cdot \log_{10}(e).$$

Miscall probabilities of 0.1 (10%), 0.01 (1%), and 0.001 (0.1%) yield phred scores of 10, 20, and 30, respectively. The NGS error rates estimated by quality values depend on several factors, including signal-to-noise levels, cross talk from nearby beads or clusters, and dephasing. Substantial effort has been made to understand and improve the accuracy of quality scores and the underlying error sources *(10, 14)*, including inaccuracies in homopolymer run lengths on the 454 platform and base-substitution error biases with the Illumina format. Study of these error traits has led to examples of software that require no additional base calling but that improve quality-score accuracies and thus improve sequencing accuracy *(65, 66)*. Quality values are an important tool for rejecting low-quality reads, trimming low-quality bases, improving alignment accuracy, and determining consensus-sequence and variant calls *(67)*.

Alignment and assembly are substantially more difficult for NGS data than for Sanger data because of the shorter reads lengths in the former. One limitation of short-read alignment and assembly is the inability to uniquely align large portions of a read set when the read length becomes too short. Similarly, the number of uniquely aligned reads is reduced when aligning to larger, more complex genomes or reference sequences because of their having a higher probability of repetitive sequences. A case in point is a modeling study that indicated that 97% of the *E. coli* genome can be uniquely aligned with 18-bp reads but that only 90% of the human genome can be uniquely aligned with 30-bp reads *(68, 69)*. Unique alignment or assembly is reduced not only by the presence of repeat sequences but also by shared homologies within closely related gene families and pseudogenes. Nonunique read alignment is handled in software by read distribution between multiple alignment positions or leaving alignment gaps. De novo assembly will reject these reads, leading to shorter and more numerous assembled contigs. These factors are relevant when choosing an appropriate sequencing platform with its associated read length, particularly for de novo assembly *(9)*.

Error rates for individual NGS reads are higher than for Sanger sequencing. The higher accuracy of

Downloaded from https://academic.oup.com/clinchem/article/55/4/641/5629392 by Broad Institute user on 05 September 2024

Petitioner Illumina, Inc., EX1032, p. 650

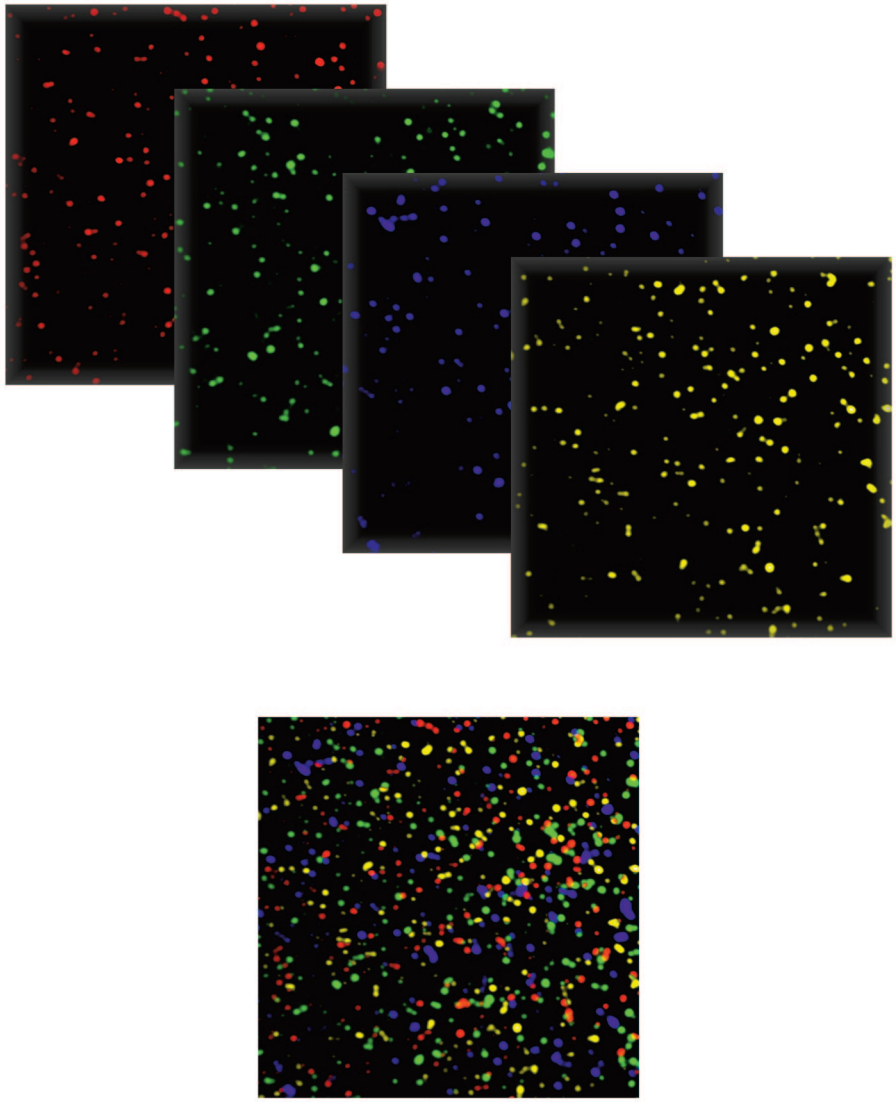

Fig. 4. Pseudocolor image from the Illumina flow cell.

Each fluorescence signal originates from a clonally amplified template cluster. Top panel illustrates 4 emission wavelengths of fluorescent labels depicted in red, green, blue, and yellow. Images are processed to identify individual clusters and to remove noise or interference. The lower panel is a composite image of the 4 fluorescence channels.

Sanger sequencing reflects not only the maturity of the chemistry but also the fact that a Sanger trace peak represents highly redundant, multiple terminated extension reactions. Accuracy in NGS is achieved by sequencing a given region multiple times, enabled by the massively parallel process, with each sequence contributing to "coverage" depth. Through this process, a "consensus" sequence is derived. To assemble, align, and analyze NGS data requires an adequate number of overlapping reads, or coverage. In practice, coverage across a sequenced region is variable, and factors other than the Poisson-like randomness of library preparation that may contribute to this variability include differential ligation of adapters to template sequences and differential amplification during clonal template generation *(11, 70)*. Beyond sequence errors, inadequate coverage can cause failure to detect actual nucleotide variation, leading to false-negative results for heterozygotes *(3, 11)*. Studies have shown that coverages of less than 20- to 30-fold begin to reduce the accuracy of single-nucleotide polymorphism calls in data on the 454 platform *(65)*. For the Illumina system, higher

Petitioner Illumina, Inc., EX1032, p. 651

Downloaded from https://academic.oup.com/clinchem/article/55/4/641/5629392 by Broad Institute user on 05 September 2024

coverage depths (50- to 60-fold) have been used in an effort to improve short-read alignment, assembly, and accuracy, although coverage in the 20- to 30-fold range may be sufficient for certain resequencing applications *(14)*. As noted above, one comparative study of a yeast genome showed that the 454, Illumina, and SOLiD technologies all accurately detected single-nucleotide variations when the coverage depth was ≥15-fold per allele *(20)*. Coverage gaps can occur when sequences are not aligned because of substantial variance from a reference. Alignment of repetitive sequences in repeat regions of a target sequence can also affect the apparent coverage. Reads that align equally well at multiple sites can be randomly distributed to the sites or in some cases discarded, depending on the alignment software. In de novo–assembly software, reads with ambiguous alignments are typically discarded, yielding multiple aligned read groups, or contigs, with no information regarding relative order.

A large variety of software programs for alignment and assembly have been developed and made available to the research community (see Table 1 in the online Data Supplement). Most use the Linux operating system, and a few are available for Windows. Many require a 64-bit operating system and can use ≥16 MB of RAM and multiple central-processing unit cores. The range of data volumes, hardware, software packages, and settings leads to processing times from a few minutes to multiple hours, emphasizing the need for sufficient computational power. Although a growing set of variations in alignment and assembly algorithms are available, there remains the trade-off between speed and accuracy in which many but not all possible alignments are evaluated, with a balance having to be struck between ideal alignment and computational efficiency.

NGS software features vary with the application and in general may include alignment, de novo assembly, alignment viewing, and variant-discovery programs. In addition some NGS statistical data-analysis tools are being developed (such as JMP Genomics; SAS Institute). Software packages available for alignment and assembly to a reference sequence include Zoom *(71)*, MAQ *(67)*, Mosaik *(72)*, SOAP *(73)*, and SHRiMP (http://compbio.cs.toronto.edu/shrimp/), which supports SOLiD color-space analysis. Software for de novo assembly includes Edina *(70)*, EULER-SR *(74)*, SHARCGS *(75)*, SSAKE *(69)*, Velvet *(76)*, and SOAPdenovo (http://soap.genomics.org.cn/). Recently released commercial software for alignment and de novo assembly includes packages from DNAStar (www.dnastar.com), SoftGenetics (www.softgenetics.com), and CLC bio (www.clcbio.com) that feature data viewers that allow the user to see read alignments, coverage depth, genome annotations, and variant analysis. Fig. 5 presents some examples of NGS data viewed in 2 different software systems.

RNA-Seq data analysis poses unique challenges and requires sequence alignment across spliced regions of transcripts as well as poly(A) tails. Current software has made strong inroads, however, with incorporation of motif recognition at splice junctions and identification of intron–exon borders through regions of low alignment coverage *(41)*. Deciphering multiple transcript isoforms involves mapping reads to known and putative splicing junctions and, in one approach, requires that each isoform be supported by multiple independent splice-junction reads with independent start sites *(51)*. ERANGE software has been used in the analysis of the mouse transcriptome *(43)*. ERANGE maps unique reads to their genomic site of origin and maps reads that match to several sites, or multireads, to a most likely site of origin. Reads that do not map to a known exon are grouped together by homology into candidate exons or parts of exons. The near proportional nature of NGS transcriptome data allows quantification of RNA production from the coverage of the assembled or aligned data. ERANGE uses normalized counts of unique reads, spliced reads, and multireads to quantify transcripts. Additional analytical considerations are needed for miRNA studies, including RNA secondary-structure analysis for hairpins, alignment to known miRNA databases, and searches within the NGS data set for complementary miRNA strands, as described in studies of developing rice grains *(77)* and chicken embryos *(78)*.

Research with ChIP-Seq has led to analysis methods and software that exploit the advantages over ChIP-in-chip, namely a larger, more information-rich data set. The single-base resolution of the data allows improved estimation of binding-site positions in the programs QuEST *(79)* and MACS *(80)*. Aligned data at the protein-binding regions typically have 2 characteristic offset peaks, each of which is populated by only forward or only reverse reads. These peaks are hallmarks of the immunoprecipitated short ChIP-Seq DNA fragments with a binding site near the center and are used by the software to estimate binding-site location near the mean peak position. Additional program features include advancements in statistical analysis to minimize miscalled binding sites, error probability estimation, and motif analysis (see Table 1 in the online Data Supplement).

A Clinical Future for NGS

From the impact that NGS has made at the basic-research level, we can anticipate its translation into molecular diagnostics. Key issues that will need to be addressed in this transition will include complexity of

Downloaded from https://academic.oup.com/clinchem/article/55/4/641/5629392 by Broad Institute user on 05 September 2024

Petitioner Illumina, Inc., EX1032, p. 652

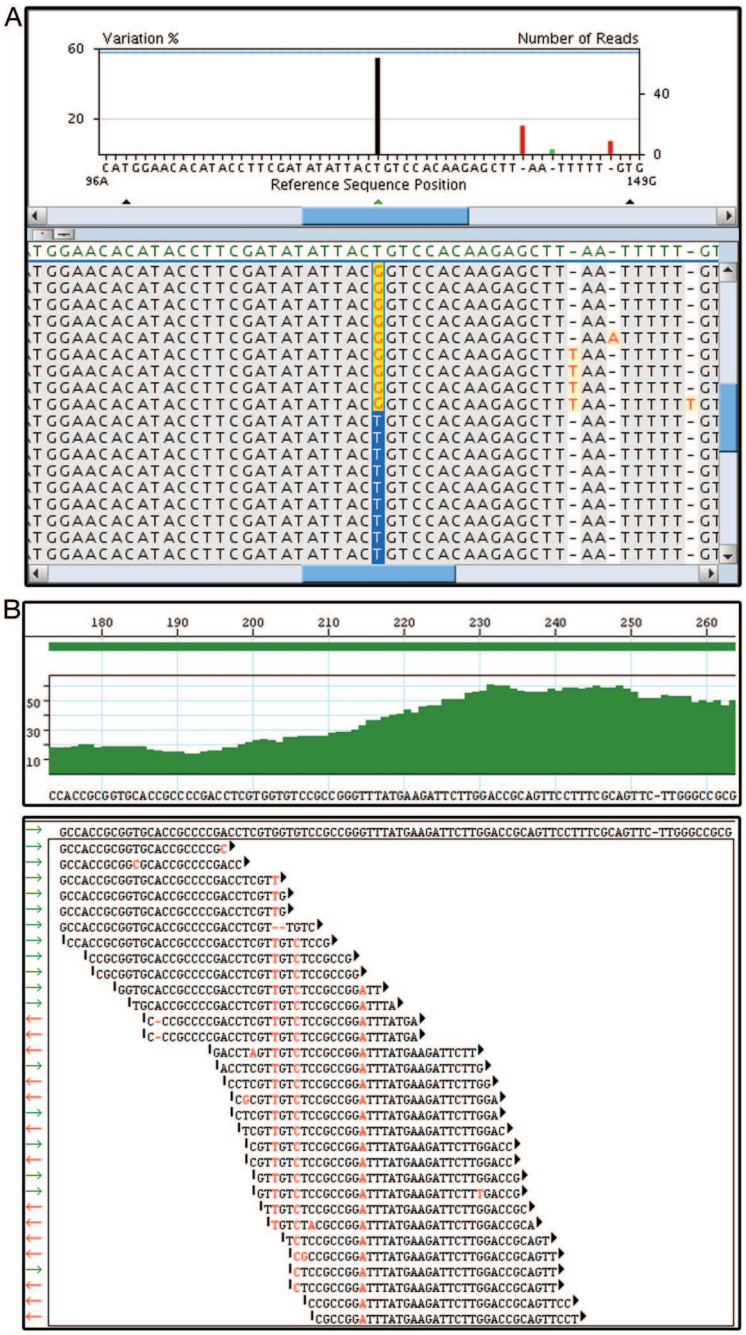

Fig. 5. Examples of NGS data viewed in 2 different software systems.

(A), Roche Amplicon Variant Analyzer software displaying GS FLX data from the *CFTR* gene [cystic fibrosis transmembrane conductance regulator (ATP-binding cassette sub-family C, member 7)]. Lower pane shows reference sequence (green) above 18 of 68 aligned reads. Column highlighted in yellow and blue shows a heterozygous single-nucleotide polymorphism (SNP). Single T/A insertions (red) may represent errors. Upper pane shows percent variation from reference (vertical bars) and coverage (pale blue line). (B), DNAStar SeqMan Pro software displaying Illumina data from *Mycobacterium massiliense*. Lower pane shows reference sequence above aligned reads. Green and red arrows show direction of sequencing; base calls at variance with reference are indicated (red). Three columns in agreement (red) indicate presumptive SNPs. Other bases in red may be errors. Upper pane shows read coverage, with relative alignment positions above the graph. See the Acknowledgments for disclosure information on the *CFTR*-gene analysis performed on residual, deidentified DNA.

Downloaded from https://academic.oup.com/clinchem/article/55/4/641/5629392 by Broad Institute user on 05 September 2024

Petitioner Illumina, Inc., EX1032, p. 653

technical procedures, robustness, accuracy, and cost. By all these measures, NGS platforms will benefit from continued process streamlining, automation, chemistry refinements, cost reductions, and improved data handling. The cost of NGS is currently substantial in terms of the investment in capital equipment (from approximately $600 000 for the Roche/454 Life Science, Illumina, and Applied Biosystems SOLiD platforms to $1.35 million for the HeliScope platform) and costs of sequencing reagents (from approximately $3500–$4500 for the Illumina, Applied Biosystems, and Roche/454 platforms to $18 000 for the HeliScope platform). Nonetheless, the cost per base is substantially lower than for Sanger sequencing, and combined with the tremendous output, it is straightforward to see why genome centers, core facilities, and commercial contract-sequencing enterprises have readily adopted this new technology. Work flow considerations include the fact that preparation of a sample library requires multiple molecular biology steps and 2–4 days to complete, depending on the platform. In addition to the required molecular biology expertise, data analysis requires expertise in bioinformatics facilitated by a knowledge of Linux operating systems. Leveraging the high-throughput capacity of NGS platforms can be facilitated by analyzing multiple samples with separate flow-cell lanes or compartments. In addition, unique identifier sequences or "bar codes" can be ligated to individual samples, which can subsequently be pooled and sequenced. After sequencing, sequences of individual samples are derived by data deconvolution *(81–83)*.

The transition of NGS into clinical diagnostics is in the early stages of development in large reference laboratories and is being leveraged for applications that require large amounts of sequence information, relative quantification, and high-sensitivity detection. Examples that meet these criteria include the aforementioned detection of mutations in tumor cells from biopsies or in the circulation. In the area of mitochondrial disorders, NGS can be used to sequence the entire 16.5-kb mitochondrial genome, determine mutation heteroplasmy percentage, and analyze nuclear genes whose protein products affect mitochondrial metabolism—all in a single analytical run. In the authors' laboratory, sequencing of mycobacterial genomes is ongoing as an approach to refine organism identification and support clinical epidemiologic investigations. HIV quasi-species detection and relative quantification have been demonstrated and can be used to monitor emerging drug resistance *(84)*. For human genetics, there is an increasing need to analyze multiple genes that, when mutated, lead to overlapping physical findings and clinical phenotypes. For example, 16 different genes are implicated in the pathogenesis of hypertrophic cardiomyopathy *(85, 86)*. For a comprehensive diagnostic evaluation in such settings, it will be necessary to sequence upwards of 100 000 to 200 000 bp. The coupling of NGS with the genomic-enrichment techniques described above offers a promising approach to this technical challenge.

Recently, investigative groups led by Y.M. Dennis Lo and Stephen Quake have applied NGS to the detection of fetal chromosomal aneuploidy *(87, 88)*. Prior work had demonstrated that cell-free fetal nucleic acids (DNA and RNA) are present in maternal blood during pregnancy, along with maternally derived cell-free nucleic acids. Several analytical approaches that use cell-free fetal nucleic acids have been developed to determine fetal aneuploidy, including the analysis of placental mRNA derived from the chromosomes of interest (e.g., chromosome 21) and the determination of relative chromosomal dosage via digital PCR analysis of a large number of target chromosomal loci compared with reference chromosomal loci *(89–91)*. Building on the concept of relative chromosome dosage, the Lo and Quake groups have independently shown the feasibility of converting cell-free DNA from maternal blood into an Illumina library, followed by sequencing and mapping the reads to the reference human genome. Counting the number of reads that map to each chromosome allows the relative dosage of each chromosome to be ascertained. If fetal aneuploidy is present, the number of sequence reads mapping to the affected chromosome would be expected to be statistically overrepresented in the data set. This expectation was confirmed in trisomy 21 pregnancies, with additional supporting evidence obtained for trisomy 18 and 13 pregnancies. These studies open a new avenue for assessing fetal aneuploidy and provide a foundation for NGS-based analysis of cell-free DNA in both nonpathologic and pathophysiological states.

Technologies on the Horizon

New single-molecule sequencing technologies in development may decrease sequencing time, reduce costs, and streamline sample preparation. Real-time sequencing by synthesis is being developed by VisiGen (http://www.visigenbio.com) and Pacific Biosciences (http://www.pacificbiosciences.com). VisiGen's approach uses DNA polymerase modified with a fluorescent donor molecule. Attached to a glass slide surface, the polymerase directs strand extension from primed DNA templates. Nucleotides are modified with fluorescent acceptor molecules, and light energy is used during incorporation to invoke fluorescence resonance energy transfer between polymerase and nucleotide fluorescent moieties, the latter being in the

Downloaded from https://academic.oup.com/clinchem/article/55/4/641/5629392 by Broad Institute user on 05 September 2024

Petitioner Illumina, Inc., EX1032, p. 654

γ-phosphate position and cleaved away during incorporation. The company envisions its platform will consist of a massively parallel array of tethered DNA polymerases that will generate 1×10^6 bp of sequence per second.

Pacific Biosciences performs single-molecule real-time sequencing and uses phospholinked fluorescently labeled dNTPs. DNA sequencing is performed in thousands of reaction wells 50–100 nm in diameter that are fabricated with a thin metal cladding film deposited on an optical waveguide consisting of a solid, transparent silicon dioxide substrate. Each reaction well is a nanophotonic chamber in which only the bottom third is visualized, producing a detection volume of approximately 20 zL (20×10^{-21} L). DNA polymerase/template complexes are immobilized to the well bottoms, and 4 differently labeled dNTPs are added. As the DNA polymerase incorporates complementary nucleotides, each base is held within the detection volume for tens of milliseconds, orders of magnitude longer than the amount of time it takes for a nucleotide to diffuse in and out of the detection volume. Laser excitation enables the incorporation events in individual wells to be captured through the optical waveguide, with the fluorescent color detected reflecting the identity of the dNTP incorporated. For sequencing, Pacific Biosciences uses a modified phi29 DNA polymerase that has enhanced kinetic properties for incorporating the system's phospholinked fluorescently labeled dNTPs. In addition, phi29 DNA polymerase is highly processive, with strand-displacement activity. By taking advantage of these properties, Pacific Biosciences has demonstrated sequencing reads exceeding 4000 bases when a circularized single-stranded DNA molecule is used as template. In this configuration, the phi29 DNA polymerase carries out multiple laps of DNA strand-displacement synthesis around the circular template. The mean DNA-synthesis rate was determined to be approximately 4 bases/s. The observed errors, including deletions, insertions, and mismatches can be addressed by developing a consensus sequence read derived from the multiple rounds of template sequencing. Further refinement of the chemistries and platform instrumentation are ongoing, with a 2010 target date for commercial launch *(92–94)*.

Farther out toward the horizon is sequencing based on monitoring the passage of DNA molecules through nanopores 2–5 nm or greater in diameter. Nanopores are being fabricated in inorganic membranes (solid-state nanopores), assembled from protein channels in lipid membranes, or configured in polymer-based nanofluidic channels. In some configurations, current is applied across nanopore membranes to drive the translocation of negatively charged DNA molecules through pores while monitoring changes in membrane electrical conductance measured in the picoampere range. NABsys (http://www.nabsys.com) is pursuing a combination of nanopores with sequencing by hybridization in which single-stranded DNA molecules are hybridized with a library of hexamers of known sequence. The hybridized DNA is interrogated through a nanopore, with the current changes being different in regions of hexamer hybridization. The patterns of hybridization are used to map annealing regions and determine sequence. Oxford Nanopore Technologies (http://www.nanoporetech.com) is developing nanopore-based sequencing that uses an α-hemolysin protein channel in reconstituted lipid bilayers. The nanopores are situated in individual array wells, and single DNA molecules are introduced into the wells and progressively digested by exonuclease. The released single-nucleotide bases enter the nanopore and alter the electrical current, creating a characteristic current change for each individual base *(95, 96)*. Although the technology is seemingly futuristic, considerable NHGRI funding is being directed toward a variety of nanopore technologies under development as part of the goal of achieving the $1000 genome. For further descriptions of nanopore technologies, the reader is referred to recent reviews *(97, 98)*.

Conclusions

The past few years have witnessed the emergence of NGS technologies that share a common basis, massively parallel sequencing of clonally amplified DNA molecules. In 2008, the first NGS platform based on single-molecule DNA sequencing was launched. On the horizon are real-time single-molecule DNA-sequencing technologies and approaches based on nanopores. NGS has had a substantial impact on basic genomics research in terms of scale and feasibility. Over the next several years, NGS is anticipated to transition into clinical-diagnostics use. Essential elements to make this transition successful will be the requirement of streamlining the processes, especially sample preparation, coupled with improvements in technology robustness and characterization of accuracy through validation studies. The large amounts of sequence-data output will pose a bioinformatics challenge for the clinical laboratory. In addition to data processing, the interpretation of sequencing results will require further characterization of the genomic variation present in the regions analyzed. Although considerable work lies ahead to implement NGS into clinical diagnostics, the potential applications are exciting and numerous.

Author Contributions: *All authors confirmed they have contributed to the intellectual content of this paper and have met the following 3 re-*

Downloaded from https://academic.oup.com/clinchem/article/55/4/641/5629392 by Broad Institute user on 05 September 2024

Petitioner Illumina, Inc., EX1032, p. 655

quirements: (a) significant contributions to the conception and design, acquisition of data, or analysis and interpretation of data; (b) drafting or revising the article for intellectual content; and (c) final approval of the published article.

Authors' Disclosures of Potential Conflicts of Interest: *No authors declared any potential conflicts of interest.*

Role of Sponsor: The funding organizations played a direct role in the preparation of the manuscript and in the final approval of the manuscript.

Acknowledgments: The analysis of the *CFTR* gene illustrated in Fig. 5 was performed on residual, deidentified DNA under the approval of University of Utah Institutional Review Board, human subjects protocol number 7275.

References

1. Maxam AM, Gilbert W. A new method for sequencing DNA. Proc Natl Acad Sci U S A 1977; 74:560–4.
2. Sanger F, Nicklen S, Coulson AR. DNA sequencing with chain-terminating inhibitors. Proc Natl Acad Sci U S A 1977;74:5463–7.
3. Wheeler DA, Srinivasan M, Egholm M, Shen Y, Chen L, McGuire A, et al. The complete genome of an individual by massively parallel DNA sequencing. Nature 2008;452:872–6.
4. Nyren P, Pettersson B, Uhlen M. Solid phase DNA minisequencing by an enzymatic luminometric inorganic pyrophosphate detection assay. Anal Biochem 1993;208:171–5.
5. Ronaghi M, Karamohamed S, Pettersson B, Uhlen M, Nyren P. Real-time DNA sequencing using detection of pyrophosphate release. Anal Biochem 1996;242:84–9.
6. Ronaghi M, Uhlen M, Nyren P. A sequencing method based on real-time pyrophosphate. Science 1998;281:363–5.
7. Tawfik DS, Griffiths AD. Man-made cell-like compartments for molecular evolution. Nat Biotechnol 1998;16:652–6.
8. Margulies M, Egholm M, Altman WE, Attiya S, Bader JS, Bemben LA, et al. Genome sequencing in microfabricated high-density picolitre reactors. Nature 2005;437:376–80.
9. Pearson BM, Gaskin DJ, Segers RP, Wells JM, Nuijten PJ, van Vliet AH. The complete genome sequence of *Campylobacter jejuni* strain 81116 (NCTC11828). J Bacteriol 2007;189:8402–3.
10. Huse SM, Huber JA, Morrison HG, Sogin ML, Welch DM. Accuracy and quality of massively parallel DNA pyrosequencing. Genome Biol 2007; 8:R143.
11. Bentley DR, Balasubramanian S, Swerdlow HP, Smith GP, Milton J, Brown CG, et al. Accurate whole human genome sequencing using reversible terminator chemistry. Nature 2008; 456:53–9.
12. Korbel JO, Urban AE, Affourtit JP, Godwin B, Grubert F, Simons JF, et al. Paired-end mapping reveals extensive structural variation in the human genome. Science 2007;318:420–6.
13. Campbell PJ, Stephens PJ, Pleasance ED, O'Meara S, Li H, Santarius T, et al. Identification of somatically acquired rearrangements in cancer using genome-wide massively parallel paired-end sequencing. Nat Genet 2008;40:722–9.
14. Dohm JC, Lottaz C, Borodina T, Himmelbauer H. Substantial biases in ultra-short read data sets from high-throughput DNA sequencing. Nucleic Acids Res 2008;36:e105.
15. Erlich Y, Mitra PP, delaBastide M, McCombie WR, Hannon GJ. Alta-Cyclic: a self-optimizing base caller for next-generation sequencing. Nat Methods 2008;5:679–82.
16. Quail MA, Kozarewa I, Smith F, Scally A, Stephens PJ, Durbin R, et al. A large genome center's improvements to the Illumina sequencing system. Nat Methods 2008;5:1005–10.
17. Shendure J, Porreca GJ, Reppas NB, Lin X, McCutcheon JP, Rosenbaum AM, et al. Accurate multiplex polony sequencing of an evolved bacterial genome. Science 2005;309:1728–32.
18. Braslavsky I, Hebert B, Kartalov E, Quake SR. Sequence information can be obtained from single DNA molecules. Proc Natl Acad Sci U S A 2003;100:3960–4.
19. Harris TD, Buzby PR, Babcock H, Beer E, Bowers J, Braslavsky I, et al. Single-molecule DNA sequencing of a viral genome. Science 2008;320: 106–9.
20. Smith DR, Quinlan AR, Peckham HE, Makowsky K, Tao W, Woolf B, et al. Rapid whole-genome mutational profiling using next-generation sequencing technologies. Genome Res 2008;18: 1638–42.
21. Quinn NL, Levenkova N, Chow W, Bouffard P, Boroevich KA, Knight JR, et al. Assessing the feasibility of GS FLX Pyrosequencing for sequencing the Atlantic salmon genome. BMC Genomics 2008;9:404.
22. Satkoski JA, Malhi R, Kanthaswamy S, Tito R, Malladi V, Smith D. Pyrosequencing as a method for SNP identification in the rhesus macaque (*Macaca mulatta*). BMC Genomics 2008;9:256.
23. Borneman AR, Forgan AH, Pretorius IS, Chambers PJ. Comparative genome analysis of a *Saccharomyces cerevisiae* wine strain. FEMS Yeast Res 2008;8:1185–95.
24. Wang J, Wang W, Li R, Li Y, Tian G, Goodman L, et al. The diploid genome sequence of an Asian individual. Nature 2008;456:60–5.
25. Ley TJ, Mardis ER, Ding L, Fulton B, McLellan MD, Chen K, et al. DNA sequencing of a cytogenetically normal acute myeloid leukaemia genome. Nature 2008;456:66–72.
26. Kim PM, Lam HY, Urban AE, Korbel JO, Affourtit J, Grubert F, et al. Analysis of copy number variants and segmental duplications in the human genome: evidence for a change in the process of formation in recent evolutionary history. Genome Res 2008;18:1865–74.
27. Chen J, Kim YC, Jung YC, Xuan Z, Dworkin G, Zhang Y, et al. Scanning the human genome at kilobase resolution. Genome Res 2008;18:751–62.
28. Yeager M, Xiao N, Hayes RB, Bouffard P, Desany B, Burdett L, et al. Comprehensive resequence analysis of a 136 kb region of human chromosome 8q24 associated with prostate and colon cancers. Hum Genet 2008;124:161–70.
29. Ding L, Getz G, Wheeler DA, Mardis ER, McLellan MD, Cibulskis K, et al. Somatic mutations affect key pathways in lung adenocarcinoma. Nature 2008;455:1069–75.
30. Albert TJ, Molla MN, Muzny DM, Nazareth L, Wheeler D, Song X, et al. Direct selection of human genomic loci by microarray hybridization. Nat Methods 2007;4:903–5.
31. Hodges E, Xuan Z, Balija V, Kramer M, Molla MN, Smith SW, et al. Genome-wide in situ exon capture for selective resequencing. Nat Genet 2007; 39:1522–7.
32. Okou DT, Steinberg KM, Middle C, Cutler DJ, Albert TJ, Zwick ME. Microarray-based genomic selection for high-throughput resequencing. Nat Methods 2007;4:907–9.
33. Porreca GJ, Zhang K, Li JB, Xie B, Austin D, Vassallo SL, et al. Multiplex amplification of large sets of human exons. Nat Methods 2007;4: 931–6.
34. Huber JA, Mark Welch DB, Morrison HG, Huse SM, Neal PR, Butterfield DA, Sogin ML. Microbial population structures in the deep marine biosphere. Science 2007;318:97–100.
35. Sogin ML, Morrison HG, Huber JA, Mark Welch D, Huse SM, Neal PR, et al. Microbial diversity in the deep sea and the underexplored "rare biosphere." Proc Natl Acad Sci U S A 2006;103: 12115–20.
36. Urich T, Lanzen A, Qi J, Huson DH, Schleper C, Schuster SC. Simultaneous assessment of soil microbial community structure and function through analysis of the meta-transcriptome. PLoS ONE 2008;3:e2527.
37. Palacios G, Druce J, Du L, Tran T, Birch C, Briese T, et al. A new arenavirus in a cluster of fatal transplant-associated diseases. N Engl J Med 2008;358:991–8.
38. Keijser BJ, Zaura E, Huse SM, van der Vossen JM, Schuren FH, Montijn RC, et al. Pyrosequencing analysis of the oral microflora of healthy adults. J Dent Res 2008;87:1016–20.
39. Turnbaugh PJ, Hamady M, Yatsunenko T, Cantarel BL, Duncan A, Ley RE, et al. A core gut microbiome in obese and lean twins. Nature 2008;457:480–4.
40. Wang Z, Gerstein M, Snyder M. RNA-Seq: a revolutionary tool for transcriptomics. Nat Rev Genet 2009;10:57–63.
41. Nagalakshmi U, Wang Z, Waern K, Shou C, Raha D, Gerstein M, Snyder M. The transcriptional landscape of the yeast genome defined by RNA sequencing. Science 2008;320: 1344–9.
42. Wilhelm BT, Marguerat S, Watt S, Schubert F, Wood V, Goodhead I, et al. Dynamic repertoire of a eukaryotic transcriptome surveyed at single-nucleotide resolution. Nature 2008;453: 1239–43.

Downloaded from https://academic.oup.com/clinchem/article/55/4/641/5629392 by Broad Institute user on 05 September 2024

Petitioner Illumina, Inc., EX1032, p. 656

43. Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B. Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nat Methods 2008; 5:621–8.
44. Lister R, O'Malley RC, Tonti-Filippini J, Gregory BD, Berry CC, Millar AH, Ecker JR. Highly integrated single-base resolution maps of the epigenome in *Arabidopsis*. Cell 2008;133:523–36.
45. Cloonan N, Forrest AR, Kolle G, Gardiner BB, Faulkner GJ, Brown MK, et al. Stem cell transcriptome profiling via massive-scale mRNA sequencing. Nat Methods 2008;5:613–9.
46. Marioni JC, Mason CE, Mane SM, Stephens M, Gilad Y. RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays. Genome Res 2008;18:1509–17.
47. Morin R, Bainbridge M, Fejes A, Hirst M, Krzywinski M, Pugh T, et al. Profiling the HeLa S3 transcriptome using randomly primed cDNA and massively parallel short-read sequencing. Biotechniques 2008;45:81–94.
48. Morin RD, O'Connor MD, Griffith M, Kuchenbauer F, Delaney A, Prabhu AL, et al. Application of massively parallel sequencing to microRNA profiling and discovery in human embryonic stem cells. Genome Res 2008;18:610–21.
49. Emrich SJ, Barbazuk WB, Li L, Schnable PS. Gene discovery and annotation using LCM-454 transcriptome sequencing. Genome Res 2007;17: 69–73.
50. Pan Q, Shai O, Lee LJ, Frey BJ, Blencowe BJ. Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. Nat Genet 2008;40:1413–5.
51. Wang ET, Sandberg R, Luo S, Khrebtukova I, Zhang L, Mayr C, et al. Alternative isoform regulation in human tissue transcriptomes. Nature 2008;456:470–6.
52. Ren B, Robert F, Wyrick JJ, Aparicio O, Jennings EG, Simon I, et al. Genome-wide location and function of DNA binding proteins. Science 2000; 290:2306–9.
53. Barski A, Cuddapah S, Cui K, Roh TY, Schones DE, Wang Z, et al. High-resolution profiling of histone methylations in the human genome. Cell 2007; 129:823–37.
54. Schones DE, Zhao K. Genome-wide approaches to studying chromatin modifications. Nat Rev Genet 2008;9:179–91.
55. Valouev A, Ichikawa J, Tonthat T, Stuart J, Ranade S, Peckham H, et al. A high-resolution, nucleosome position map of *C. elegans* reveals a lack of universal sequence-dictated positioning. Genome Res 2008;18:1051–63.
56. Johnson DS, Mortazavi A, Myers RM, Wold B. Genome-wide mapping of in vivo protein-DNA interactions. Science 2007;316:1497–502.
57. Robertson G, Hirst M, Bainbridge M, Bilenky M, Zhao Y, Zeng T, et al. Genome-wide profiles of STAT1 DNA association using chromatin immunoprecipitation and massively parallel sequencing. Nat Methods 2007;4:651–7.
58. Korshunova Y, Maloney RK, Lakey N, Citek RW, Bacher B, Budiman A, et al. Massively parallel bisulphite pyrosequencing reveals the molecular complexity of breast cancer-associated cytosine-methylation patterns obtained from tissue and serum DNA. Genome Res 2008;18: 19–29.
59. Cokus SJ, Feng S, Zhang X, Chen Z, Merriman B, Haudenschild CD, et al. Shotgun bisulphite sequencing of the *Arabidopsis* genome reveals DNA methylation patterning. Nature 2008;452: 215–9.
60. Marguerat S, Wilhelm BT, Bahler J. Next-generation sequencing: applications beyond genomes. Biochem Soc Trans 2008;36:1091–6.
61. Pop M, Salzberg SL. Bioinformatics challenges of new sequencing technology. Trends Genet 2008; 24:142–9.
62. Rougemont J, Amzallag A, Iseli C, Farinelli L, Xenarios I, Naef F. Probabilistic base calling of Solexa sequencing data. BMC Bioinformatics 2008;9:431.
63. Ewing B, Green P. Base-calling of automated sequencer traces using phred. II. Error probabilities. Genome Res 1998;8:186–94.
64. Ewing B, Hillier L, Wendl MC, Green P. Base-calling of automated sequencer traces using phred. I. Accuracy assessment. Genome Res 1998;8:175–85.
65. Brockman W, Alvarez P, Young S, Garber M, Giannoukos G, Lee WL, et al. Quality scores and SNP detection in sequencing-by-synthesis systems. Genome Res 2008;18:763–70.
66. Smith AD, Xuan Z, Zhang MQ. Using quality scores and longer reads improves accuracy of Solexa read mapping. BMC Bioinformatics 2008; 9:128.
67. Li H, Ruan J, Durbin R. Mapping short DNA sequencing reads and calling variants using mapping quality scores. Genome Res 2008;18: 1851–8.
68. Whiteford N, Haslam N, Weber G, Prugel-Bennett A, Essex JW, Roach PL, et al. An analysis of the feasibility of short read sequencing. Nucleic Acids Res 2005;33:e171.
69. Warren RL, Sutton GG, Jones SJ, Holt RA. Assembling millions of short DNA sequences using SSAKE. Bioinformatics 2007;23:500–1.
70. Hernandez D, Francois P, Farinelli L, Osteras M, Schrenzel J. De novo bacterial genome sequencing: millions of very short reads assembled on a desktop computer. Genome Res 2008;18:802–9.
71. Lin H, Zhang Z, Zhang MQ, Ma B, Li M. ZOOM! Zillions Of Oligos Mapped. Bioinformatics 2008; 24:2431–7.
72. Smith DR, Quinlan AR, Peckham HE, Makowsky K, Tao W, Woolf B, et al. Rapid whole-genome mutational profiling using next-generation sequencing technologies. Genome Res 2008;18: 1638–42.
73. Li R, Li Y, Kristiansen K, Wang J. SOAP: short oligonucleotide alignment program. Bioinformatics 2008;24:713–4.
74. Chaisson MJ, Pevzner PA. Short read fragment assembly of bacterial genomes. Genome Res 2008;18:324–30.
75. Dohm JC, Lottaz C, Borodina T, Himmelbauer H. SHARCGS, a fast and highly accurate short-read assembly algorithm for de novo genomic sequencing. Genome Res 2007;17:1697–706.
76. Zerbino DR, Birney E. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. Genome Res 2008;18:821–9.
77. Zhu QH, Spriggs A, Matthew L, Fan L, Kennedy G, Gubler F, Helliwell C. A diverse set of microRNAs and microRNA-like small RNAs in developing rice grains. Genome Res 2008;18:1456–65.
78. Glazov EA, Cottee PA, Barris WC, Moore RJ, Dalrymple BP, Tizard ML. A microRNA catalog of the developing chicken embryo identified by a deep sequencing approach. Genome Res 2008; 18:957–64.
79. Valouev A, Johnson DS, Sundquist A, Medina C, Anton E, Batzoglou S, et al. Genome-wide analysis of transcription factor binding sites based on ChIP-Seq data. Nat Methods 2008;5: 829–34.
80. Zhang Y, Liu T, Meyer CA, Eeckhoute J, Johnson DS, Bernstein BE, et al. Model-based analysis of ChIP-Seq (MACS). Genome Biol 2008;9: R137.
81. Binladen J, Gilbert MT, Bollback JP, Panitz F, Bendixen C, Nielsen R, Willerslev E. The use of coded PCR primers enables high-throughput sequencing of multiple homolog amplification products by 454 parallel sequencing. PLoS ONE 2007; 2:e197.
82. Meyer M, Stenzel U, Hofreiter M. Parallel tagged sequencing on the 454 platform. Nat Protoc 2008;3:267–78.
83. Meyer M, Stenzel U, Myles S, Prufer K, Hofreiter M. Targeted high-throughput sequencing of tagged nucleic acid samples. Nucleic Acids Res 2007;35:e97.
84. Wang C, Mitsuya Y, Gharizadeh B, Ronaghi M, Shafer RW. Characterization of mutation spectra with ultra-deep pyrosequencing: application to HIV-1 drug resistance. Genome Res 2007;17: 1195–201.
85. Fokstuen S, Lyle R, Munoz A, Gehrig C, Lerch R, Perrot A, et al. A DNA resequencing array for pathogenic mutation detection in hypertrophic cardiomyopathy. Hum Mutat 2008;29:879–85.
86. Morita H, Rehm HL, Menesses A, McDonough B, Roberts AE, Kucherlapati R, et al. Shared genetic causes of cardiac hypertrophy in children and adults. N Engl J Med 2008;358:1899–908.
87. Chiu RW, Chan KC, Gao Y, Lau VY, Zheng W, Leung TY, et al. Noninvasive prenatal diagnosis of fetal chromosomal aneuploidy by massively parallel genomic sequencing of DNA in maternal plasma. Proc Natl Acad Sci U S A 2008;105: 20458–63.
88. Fan HC, Blumenfeld YJ, Chitkara U, Hudgins L, Quake SR. Noninvasive diagnosis of fetal aneuploidy by shotgun sequencing DNA from maternal blood. Proc Natl Acad Sci U S A 2008;105: 16266–71.
89. Fan HC, Quake SR. Detection of aneuploidy with digital polymerase chain reaction. Anal Chem 2007;79:7576–9.
90. Lo YM, Lun FM, Chan KC, Tsui NB, Chong KC, Lau TK, et al. Digital PCR for the molecular detection of fetal chromosomal aneuploidy. Proc Natl Acad Sci U S A 2007;104:13116–21.
91. Dennis Lo YM, Chiu RW. Prenatal diagnosis: progress through plasma nucleic acids. Nat Rev Genet 2007;8:71–7.
92. Korlach J, Marks PJ, Cicero RL, Gray JJ, Murphy DL, Roitman DB, et al. Selective aluminum passivation for targeted immobilization of single DNA polymerase molecules in zero-mode waveguide nanostructures. Proc Natl Acad Sci U S A 2008;105:1176–81.

Downloaded from https://academic.oup.com/clinchem/article/55/4/641/5629392 by Broad Institute user on 05 September 2024

Petitioner Illumina, Inc., EX1032, p. 657

93. Levene MJ, Korlach J, Turner SW, Foquet M, Craighead HG, Webb WW. Zero-mode waveguides for single-molecule analysis at high concentrations. Science 2003;299:682–6.
94. Eid J, Fehr A, Gray J, Luong K, Lyle J, Otto G, et al. Real-time DNA sequencing from single polymerase molecules. Science 2008;323:133–8.
95. Astier Y, Braha O, Bayley H. Toward single molecule DNA sequencing: direct identification of ribonucleoside and deoxyribonucleoside 5′-monophosphates by using an engineered protein nanopore equipped with a molecular adapter. J Am Chem Soc 2006;128:1705–10.
96. Wu HC, Astier Y, Maglia G, Mikhailova E, Bayley H. Protein nanopores with covalently attached molecular adapters. J Am Chem Soc 2007;129:16142–8.
97. Gupta PK. Single-molecule DNA sequencing technologies for future genomics research. Trends Biotechnol 2008;26:602–11.
98. Rhee M, Burns MA. Nanopore sequencing technology: research trends and applications. Trends Biotechnol 2006;24:580–6.

Downloaded from https://academic.oup.com/clinchem/article/55/4/641/5629392 by Broad Institute user on 05 September 2024

Petitioner Illumina, Inc., EX1032, p. 658