Overview of Next-Generation Sequencing Technologies

Barton E. Slatko,[1] Andrew F. Gardner,[1] and Frederick M. Ausubel[2,3]

[1]New England Biolabs, Ipswich, Massachusetts
[2]Department of Molecular Biology, Massachusetts General Hospital, Boston, Massachusetts
[3]Corresponding author: *ausubel@molbio.mgh.harvard.edu*

High throughput DNA sequencing methodology (next generation sequencing; NGS) has rapidly evolved over the past 15 years and new methods are continually being commercialized. As the technology develops, so do increases in the number of corresponding applications for basic and applied science. The purpose of this review is to provide a compendium of NGS methodologies and associated applications. Each brief discussion is followed by web links to the manufacturer and/or web-based visualizations. Keyword searches, such as with Google, may also provide helpful internet links and information. © 2018 by John Wiley & Sons, Inc.

Keywords: next-generation sequencing • NGS • Sanger sequencing

How to cite this article:
Slatko, B. E., Gardner, A. F., & Ausubel, F. M. (2018). Overview of next-generation sequencing technologies. *Current Protocols in Molecular Biology*, *122*, e59. doi: 10.1002/cpmb.59

FOUNDING METHODOLOGY

The founding methods in DNA sequencing were the Sanger dideoxy synthesis (Sanger & Coulson, 1975; Sanger, Nicklen, & Coulson, 1977; *UNIT 7.4*) and Maxam-Gilbert chemical cleavage (Maxam & Gilbert, 1980; *UNIT 7.5*) methods. The Maxam-Gilbert method is based on chemical modification of DNA and subsequent cleavage of the DNA backbone at sites adjacent to the modified nucleotides. Sanger sequencing uses specific chain-terminating nucleotides (dideoxy nucleotides) that lack a 3′-OH group. Thus, no phosphodiester bond can be formed by DNA polymerase, resulting in termination of the growing DNA chain at that position. The ddNTPs are radioactively or fluorescently labeled for detection in "sequencing" gels or automated sequencing machines, respectively. Although the chemistry of the original Maxam-Gilbert method has been modified to help eliminate toxic reagents, the Sanger sequencing by synthesis (SBS) dideoxy method has become the sequencing standard.

The Sanger sequencing method was developed in 1977 and is described in detail in *UNIT 7.4* Although relatively slow by current NGS standards, improvements in the Sanger chain termination methodology, automation, and commercialization have enabled it to remain the most appropriate sequencing method for many current applications. Specifically, the replacement of ultrathin "slab gels" with multichannel capillary electrophoresis, the development of automated refillable reusable capillaries, and "electrokinetic" sample loading, have all contributed to the enhanced speed and ease of the Sanger process. The most significant innovations in Sanger sequencing have been: (1) the development of fluorescent (terminator) dyes, (2) the use of thermal-cycle sequencing to reduce the quantity of required input DNA and thermostable polymerases to efficiently and accurately incorporate the terminator dyes into the growing DNA strands, and (3) software developments to interpret and analyze the sequences. The leader in automated Sanger sequencing is Applied Biosystems (AB; now part of ThermoFisher). The current commercialized AB sequencers all utilize fluorescent dyes and capillary electrophoresis (CE). The machines vary in

Current Protocols in Molecular Biology e59, April 2018
Published online April 2018 in Wiley Online Library (wileyonlinelibrary.com).
doi: 10.1002/cpmb.59

capacity, 4 capillaries (SeqStudio Genetic Analyzer), 8 to 24 (3500 Series Genetic Analyzer), and 48 to 96 (3700 Series Genetic Analyzer), for DNA sequencing or fragment analysis protocols. All of these sequencers generate 600-1000 bases of accurate sequence. Although a variety of Sanger-sequencing-based sequencing machines have been introduced over the years, including instruments from Licor, Amersham, MilliGen, Perkin Elmer, and Dupont, all of them except the AB machines have been discontinued.

The Sanger sequencing technology remains very useful for applications where high throughput is not required. Many DNA sequencing core facilities and sequencing-for-profit companies provide Sanger sequencing services. The most common uses are for individual sequencing reactions using a specific DNA primer on a specific template, for example to verify plasmid constructs or PCR products. Now that molecular biology kits and reagents for DNA purification and relatively inexpensive high quality synthetic primers are available from many vendors, even relatively large Sanger sequencing projects can be completed in a reasonable time frame and cost.

In addition to sequencing DNA, another useful application of capillary electrophoresis on the AB machines has been the development of methods for measuring the activity of selected enzymes acting upon fluorescently labeled DNA substrates, by analysis, for example, of DNA fragment size (Greenough et al., 2016). Capillary electrophoresis can also be used to simultaneously analyze multiple substrates, products and/or reaction intermediates in a single reaction using different fluorescent labels (Greenough et al., 2016). For example, CE was used in high-throughput studies of DNA polymerase and DNA ligase kinetics and coupled enzyme pathways including Okazaki fragment processing and ribonucleotide excision repair (Greenough, Kelman, & Gardner, 2015; Schermerhorn & Gardner, 2015). AB CE is also useful in glycobiology for analyzing fluorescently labeled glycans (Callewaert, Geysens, Molemans, & Contreras, 2001; Laroy, Contreras, & Callewaert, 2006).

SECOND GENERATION SEQUENCING METHODS

The term "next generation" has implied a next step in the development of DNA sequencing technology and suggests there will be a "next-next" generation naming of new technologies in the future. We prefer to use the term second generation, third generation, etc., realizing that the automated AB sequencing machine technology described above is really the "second" generation after the original Sanger methods using radioactivity and slab gels. Given this naming convention, the need for higher throughput sequencing of large genomes at lower cost triggered the development of many second-generation or "nextgen" technologies using a variety of creative methodologies in addition to automated Sanger sequencing. As with the commercialization of automated Sanger sequencing, many of these technologies are no longer in use (for example, Solid™, Polinator™ Helicos™). These "second" generation sequencing technologies and associated methods are described below.

Second generation sequencing methods can be grouped into two major categories, sequencing by hybridization, and sequencing by synthesis (SBS). SBS methods are a further development of Sanger sequencing, without the dideoxy terminators, in combination with repeated cycles of synthesis, imaging, and methods to incorporate additional nucleotides in the growing chain. At first glance, these new methods may seem expensive, but the reactions are run in parallel often at nanoliter, picoliter, or zeptoliter volumes in small chambers, and thus the cost per base pair sequenced is nominal. Continual refinements and miniaturization are reducing costs even further.

A note about costs: costs for sequencing encompass many variables, some of which are often left out of commonly presented estimates of "cost per base". The reagent costs are usually dependent on the volume ordered and are often negotiated with the vendor. For example, core facilities and sequencing centers that order in larger quantities can obtain discounted pricing. Costs usually do not include labor and the bioinformatics pipeline at the end of the process. Nevertheless, goals such as the "$1,000 human genome" or reducing the "cost per base" are gold standards to be met by the sequencing technology and research community.

Sequencing by Hybridization

This method was originally developed in the 1980s, using arrayed DNA oligonucleotides of known sequence on filters that were hybridized to labeled fragments of the DNA to be sequenced. By repeatedly hybridizing and washing away the unwanted non-hybridized DNA, it was possible to determine

whether the hybridizing labeled fragments matched the sequence of the DNA probes on the filter. It was thus possible to build larger contiguous sequence information, based upon overlapping information from the probe hybridization spots. Sequencing by hybridization has been largely relegated to technologies that depend upon using specific probes to interrogate sequences, such as in diagnostic applications for identifying disease-related single-nucleotide polymorphisms (SNPs) in specific genes or identifying gross chromosome abnormalities (rearrangements, deletions, duplications, copy number variants or CNVs; Church, 2006; Drmanac et al., 2002; Hanna et al., 2000; Mirzabekov, 1994; Qin, Schneider, & Brenner, 2012).

Sequencing by Synthesis (SBS)

SBS methods have taken several distinct approaches (Fuller et al., 2009; Mardis, 2008; Metzker, 2010; Quail et al., 2012; Shendure & Ji, 2008). The "second generation" methods generally use a solid support containing micro channels or wells in which the sequencing reactions occur. In general, most of these new SBS methods do not use dideoxy terminators, although "reversible" terminators are used in some technologies, which allow the nucleotide incorporation reactions to proceed normally while imaging the incorporated nucleotides, and then removing synthesis blocking moieties on the labeled nucleotides to allow the incorporation of the next base in the sequence.

Current SBS methods differ from the approach of the original Sanger sequencing in that they rely on much shorter reads (currently up to about 300 to 500 bases). Further, they generally have an intrinsically higher error rate relative to Sanger sequencing, and rely on high sequence coverage ("massively parallel sequencing") of millions to billions of short DNA sequence reads (50 to 300 nucleotides) as a way to obtain an accurate sequence based upon the identification of a consensus (agreement) sequence. However, for some technologies, sequence context errors occur and cannot always be corrected by increasing the number of reads. For example, homopolymer sequences occur when DNA contains consecutive multiples of the same base, such as AAAAAAAAAA. In this case, sequencing platforms are limited in accurately determining the number of consecutive bases. Each platform generates its own unique set of potential sequence context errors and users need to be aware of these limitations. Using several platforms with different technologies is often used to circumvent these issues.

The shift from "longer read lengths" to "short read" technologies is now trending back toward developing technologies that generate longer primary read lengths, while maintaining the "massively parallel" nature of the technology. This is occurring in the "third", and especially the "fourth" generation technologies, described below. This is fueled in part by cost per reaction and in part by the desire to obtain as much primary sequence read information as possible to circumvent sequence context issues such repeated DNA elements.

Most SBS technologies utilize a method in which the individual DNA molecules to be sequenced are distributed to millions of separate wells or chambers, or tethered to specific locations on a solid substrate. The DNA molecules, amplified by PCR or by isothermal modified "rolling circle" amplification methods, are then subjected to DNA synthesis reactions in which labeled nucleotides, or chemical reactions based on the incorporation of a particular nucleotide, can be imaged or otherwise detected. Many creative technologies have been developed to enable the generation of millions of DNA sequence reads in a single sequence run. Sequence runs may last hours or several days depending on the throughput.

454 Pyrosequencing

Although discontinued, we include this as an example of the first of the "second" generation sequencing methods that came to market and utilized a novel approach, namely, the detection of pyrophosphate, a byproduct of nucleotide incorporation, to report whether a particular base was incorporated in a growing DNA chain ((Ronaghi, Karamohamed, Pettersson, Uhlen, & Nyren, 1996; see also *www.youtube.com/watch?v=WYBzbxIfuKs*). Individual DNA fragments, 400 to 700 base pairs (bp) long are ligated to adapters and amplified by PCR in an individual emulsion "bead" (emPCR) reaction. DNA sequences on the beads are complementary to sequences on the adaptors, allowing the DNA fragments to bind directly to the beads, ideally one fragment to each bead.

DNA synthesis followed by chemical detection of the DNA synthesis reactions then occurs in a picoliter-sized chamber where pyrophosphate release is measured. By consecutively flooding the chambers with sequencing reagents containing one of the 4 nucleotides, when the correct nucleotide is incorporated in the synthesized strand, pyrophosphate

release is measured utilizing a light-generating reaction. The intensity of light also provides information concerning homopolymer "runs" of nucleotides in the sequence, although difficulties are encountered with larger tracts of the same nucleotide. Pyrosequencing was developed in Sweden by Pyrosequencing AB, and subsequently acquired by Qiagen who licensed it to 454 Life Sciences, before it was ultimately acquired by Roche. 454 sequencers were discontinued in 2013, although reagents are still available from several suppliers. The "454 sequencing" technology was commonly used for genome sequencing and metagenome samples because of the long read lengths (up to 600 to 800 nt) that are typically achieved and relatively high throughput (25 million bases, at 99% or better accuracy in a 4 hr run), facilitating genome assembly.

Ion Torrent (*https://www.youtube.com/watch?v=WYBzbxIfuKs*)

Ion Torrent™ technology directly converts nucleotide sequence into digital information on a semiconductor chip (Rothberg et al., 2011). In a DNA synthesis reaction, when a correct nucleotide is incorporated across from its complementary base in a growing DNA chain, a hydrogen ion is released. This changes the pH of the solution which can be recorded as a voltage change by an ion sensor, much like a pH meter. If no nucleotide is incorporated, no voltage spike occurs. By sequentially flooding and washing out a "sequencing chamber" with sequencing regents which include only one of the 4 nucleotides at a time, voltage changes occur when the appropriate nucleotide is incorporated. When two adjacent nucleotides incorporate the same nucleotide, two hydrogens are released and the voltage doubles. Thus "runs" of a single nucleotide can also be determined. Large homopolymer strings of the same nucleotide are sometimes difficult to discern, however.

Ion Torrent sequencing reactions occur in millions of wells that cover a semiconductor chip containing millions of pixels that convert the chemical information into sequencing information. To begin the process, DNA is fragmented into 200 to 1500 base fragments which are ligated to adapters. The DNA fragments are attached to a bead by complementary sequences on the beads and adapters and are then amplified on the bead by emPCR. This process enables millions of beads to each have multiple copies of one DNA sequence. The beads are then flowed across the chip containing the wells such that only one bead can enter an individual well. When the sequencing reagents are then flowed across the wells, when the appropriate nucleotide is incorporated, a hydrogen ion is given off and the signal recorded. A major advantage of the system is that no camera, light source or scanner is needed; nucleotide incorporation is directly converted to voltage which is recorded directly, greatly speeding up the process.

The Ion Torrent system is sold by ThermoFisher and several versions of the platform are available, including the Ion Personal Genome Machine™ (PGM™) System, Ion Proton™ System, Ion S5 system and ION S5 XL system, each with different throughput characteristics.

An automated library and template preparation system is also available (Ion Chef™). A large number of applications are supported, including targeted and *de novo* DNA and RNA sequencing, transcriptome sequencing, microbial sequencing, copy number variation detection, small RNA and miRNA sequencing, and CHIP-seq (chromatin immunoprecipitation sequencing; Furey, 2012).

Illumina Technology

By far the major player in the second generation sequencing arena is Illumina, using technology first developed by Solexa and Lynx Therapeutics. Illumina sequencing is based on a technique known as "bridge amplification" wherein DNA molecules (about 500 bp) with appropriate adapters ligated on each end are used as substrates for repeated amplification synthesis reactions on a solid support (glass slide) that contains oligonucleotide sequences complementary to a ligated adapter. (See *www.youtube.com/watch?v=womKfikWlxM*). The oligonucleotides on the slide are spaced such that the DNA, following being subjected to repeated rounds of amplification, creates clonal "clusters" consisting of about 1000 copies of each oligonucleotide fragment. Each glass slide can support millions of parallel cluster reactions. During the synthesis reactions, proprietary modified nucleotides, corresponding to each of the four bases, each with a different fluorescent label, are incorporated and then detected. The nucleotides also act as terminators of synthesis for each reaction, which are unblocked after detection for the next round of synthesis. The reactions are repeated for 300 or more rounds. The use of fluorescent detection increases the speed of detection due to direct imaging, in contrast to camera-based imaging.

Illumina sequencing supports a variety of protocols including genomic sequencing,

exome and targeted sequencing, metagenomics, RNA sequencing, CHIP-seq, and methylome methods. Different Illumina sequencing machines provide varying levels of throughput, including the MiniSeq, MiSeq, NextSeq, NovaSeq and HiSeq models. The MiniSeq provides 7.5 Gb of information with 25 million reads/run in segments of 2 × 150bp reads. The MiSeq can perform 2 × 300bp reads, 25 million reads for an output of 15 Gb. The NextSeq can provide 120Gb with 400 million reads at 2 × 150bp read length. Details on each machine and its capabilities relative to particular sequencing projects can be found in *https://www.illumina.com/systems/sequencing-platforms.html*.

One issue that can arise with Illumina sequencing is a lack of synchrony in the synthesis reactions among the individual members of a cluster, interfering with the generation of an accurate consensus sequence and reducing the number of cycles that can be performed. Care also must be taken not to "overcluster" the support, which requires accurate quantitation of the amount of template DNA that is loaded onto the array. Because of the large amount of data generated, analysis of sequencing errors generated from the sequencing process can also be examined, aiding in the identification of "real" sequence variants versus protocol induced artifacts; see for example (Chen, Liu, Evans, & Ettwiller, 2016).

"THIRD" GENERATION (LARGE FRAGMENT SINGLE MOLECULE) SEQUENCING

In contrast to second generation sequencing methods, third generation sequencing methods aim to sequence long DNA (and RNA) molecules. The current commercialized technology leader in this area is Pacific Biosciences (PacBio) (*https://www.youtube.com/watch?v=v8p4ph2MAvI*), which has commercialized two sequencing systems, the original RSII model and more recently, the Sequel™ (see *https://www.pacb.com/products-and-services/pacbio-systems/*). PacBio sequencing, also referred to as Single Molecule Real Time (SMRT) sequencing, enables very long fragments to be sequenced, up to 30 to 50 kb, or longer. The SMRT method involves binding an engineered DNA polymerase, with bound DNA to be sequenced, to the bottom of a well (zero-mode waveguide, or ZMW, in a SMRT flow cell; see *https://www.pacb.com/smrt-science/smrt-sequencing/*). A ZMW is a small chamber that guides light energy into an area whose dimensions are small relative to the wavelength of the illuminating light. Because of the ZMW design and wavelength of light utilized, imaging occurs only at the bottom of the ZMW where the DNA polymerase, bound to the DNA, incorporates each base in a growing chain. The four nucleotides are labeled with different phospho-linked fluorophores for differential detection. When a nucleotide is incorporated into the growing chain, imaging occurs on the millisecond time scale as the correct fluorescently-labeled nucleotide is bound. After incorporation, the phosphate-linked fluorescent moiety is released, which "floats away" from the bottom of the ZMW and can no longer be detected. The next nucleotide can then be incorporated. Imaging is timed with the rate of nucleotide incorporation so that each base is identified as it is incorporated into the growing DNA chain. This simultaneously occurs in parallel in up to one million zeptoliter ZMWs, present on a single chip within the SMRT cells.

Template preparation is unique in the PacBio process as it involves production of a "SMRTbell", a circular double-stranded DNA molecule with a known adapter sequence complementary to the primers used to initiate the DNA synthesis on the template. This enables the polymerase to read through large templates numerous times by traversing the circular molecule in each ZMW, until the polymerase stops, to build up a consensus sequence (circular consensus sequence, CCS). As the adapters ligated to each side of the insert each have DNA synthesis priming sites, the sequencing polymerase can traverse the circular SMRTbell in the 5′->3′ direction on either DNA strand, providing complementary information from both strands of the dsSMRTbell.

An important advantage of the PacBio real time sequencing imaging and detection process is that the rate of each nucleotide addition during synthesis can be measured, termed the inter-pulse duration (IPD). Many (but not all) nucleotides with base modifications, such as some adenine and cytosine methylations, change the IPD and thus can be identified as a modified base (Fang et al., 2012; Flusberg et al., 2010; Murray et al., 2012; Rhoads & Au, 2015; Vilfan et al., 2013; Zhang, Sun, Menghe, & Zhang, 2015; see *https://s3.amazonaws.com/files.pacb.com/png/basemod_benefits_lg.png*). Many different modifications can be detected and catalogued for epigenetic studies. At this point, not all modifications can be identified, including

m^5C due to minor modification of the IPD. However, chemical modification of such nucleotides might enable their detection.

PacBio SMRT sequencing suffers from an inherently high error rate, but this is usually surmounted due to the depth of the number of read passes obtained in each ZMW for each SMRTbell template. Because errors are stochastic rather than systematic, sequencing the template multiple times and aligning the individual sequence reads results in high accuracy CCS reads. Further, similar sequencing of multiple templates provides additional consensus.

PacBio SMRT sequencing offers several advantages over previous methods. It enables rapid identification of methylation sites for epigenetic studies in addition to providing long reads for genome assemblies. For example, using PacBio technology, it is relatively straightforward to assemble a complete bacterial genome sequence using only a few SMRT cells and determine the methylation patterns within. Often, PacBio assemblies are combined with other methods, such as Illumina sequencing, for increased accuracy.

In terms of throughput, the PacBio RS II (2013) uses chips with 150,000 ZMWs, which when optimized, can generate as much as 350 megabases of sequence per SMRT cell. Optimization is carried out using a Poisson distribution so that only one polymerase bound DNA molecule should be in each well. The latest PacBio instrument, the Sequel, has 1 million ZMWs and can generate ~365,000 reads, with average reads of 10 to 15 kb (7.6 Gb of output). Continual upgrades of the chemistry and technology (such as "magnetic bead loading", and use of the SAGE Pippin to isolate large DNA fragments for PacBio sequencing (see ancillary methods below) are designed to provide more, longer, and more accurate reads.

TECHNOLOGIES ON THE "FOURTH" GENERATION CUSP

It is possible to pass long DNA molecules through small diameter "holes" and measure differing currents as each nucleotide passes by a linked detector (Benner et al., 2007; Branton et al., 2008; Cherf et al., 2012; Hornblower et al., 2007; Kasianowicz, Brandin, Branton, & Deamer, 1996; Liu, Wang, Deng, & Chen, 2016; see *https://www.youtube.com/watch?v=GUb1TZvMWsw*). In theory, more than a hundred kb of DNA could be threaded through the nanopore, and with many channels, tens to hundreds of Gb of sequence could be achieved at relatively low cost.

Two types of nanopore systems for DNA sequencing are being developed, biological membrane systems and solid-state sensor technology. Biological nanopore sequencing relies on the use of transmembrane proteins embedded in a lipid membrane to produce the pores. Two proteins that have been utilized to generate pores have been extensively studied: alpha hemolysin and *Mycobacterium smegmatis* porin A (MspA). The rate of DNA passage through the pores is regulated by the addition of motor proteins, such as a highly processive DNA polymerase (phi29) that ratchets DNA through upon nucleotide addition. Other accessory proteins, such as a DNA helicase, exonuclease I, or oligonucleotides to bind DNA strands, enable unwinding and "ratcheting" of the DNA nucleotides through the nanopore for detection. DNA can be moved through the pores at a constant rate for tens of thousands of nucleotides. Solid state sensor technology uses various metal or metal alloy substrates with nanometer sized pores that allow DNA or RNA to pass through.

Nanopore-based DNA sequencing was first proposed in the late 1990s and commercialization has recently been achieved by Oxford Nanopore Technologies (ONT) with a portable MinION (512 nanopore flowcell channels), benchtop GridION (5 mimIONs in a single module), and a high throughput PromethION (in development, 48 flow cells of 3000 nanopores each; Greninger et al., 2015; see *https://nanoporetech.com/applications/dna-nanopore-sequencing* and *https://nanoporetech.com/how-it-works*). These sequencers use protein nanopores in an electrically resistant polymer membrane through which characteristic current changes occur as each nucleotide passes thru the detector.

Long dsDNA molecules are first bound to a processive enzyme, such as phi29 polymerase. When the complex encounters a nanopore, one DNA strand enters the nanopore and the translocation rate through the pore is regulated by DNA polymerase synthesis and translocation. The processive enzyme enables the DNA to be continuously and processively "racheted" through it. As a nucleotide passes through the pore, it disrupts a current that has been applied to the nanopore. Each nucleotide provides a characteristic electronic signal that is recorded as a current disruption event. Recording is in real time and while 10 kb reads are now a reasonable output, in theory 100 kb of DNA can pass through each

nanopore and be detected. Once DNA has left a nanopore, the pore is available for use by a different DNA molecule.

Because of its small, handheld size, the MinION has potential for many applications where portability and or space requirements are at a premium. Currently the error rate is relatively high but as with other high throughput sequencing methods, this can be circumvented by the large number of molecules that can be sequenced. Nanopore technology has been used to sequence environmental and metagenomic samples, is currently on the space station, and has been used for bacterial strain identification. Viral genomes, environmental surveillance and haplotyping have been performed using the MinION. Nanopore technology is also able to identify base modifications, similar to PacBio technology, enabling epigenetic events to be readily identified. Nanopore sequencing also offers direct RNA sequencing, as well as PCR-free cDNA sequencing. Thus, nanopore sequencing has the potential to offer relatively low-cost DNA and RNA sequencing, environmental monitoring, and genotyping (Ammar, Paton, Torti, Shlien, & Bader, 2015; Cao et al., 2016; Loman, 2015; Schmidt et al., 2017).

ANCILLARY METHODS FOR HIGH THROUGHPUT SEQUENCING AND COMPLETING LARGE GENOME PROJECTS

Optical Mapping

Despite the recent advances in high throughput single molecule long-read sequencing methods, additional methods may be useful to complete or confirm the order of various DNA contigs (contiguous pieces) in a genome. The most popular of these methods is "optical mapping", using labeling methods distributed along long DNA molecules at nucleotide positions based upon sequence. These provide large scaffolds for "pinning" known sequences onto genome maps. By creating a "visual physical map" along large DNA molecules (similar in principle to a restriction enzyme map), one can correlate DNA sequence with physical location. The method was originally developed by David C. Schwartz in the 1990s (Cai et al., 1995; Jing et al., 1998; Meng, Benson, Chada, Huff, & Schwartz, 1995; Schwartz et al., 1993), but other methodologies have also been used to create the optical maps. The use of nanofludic methods, novel ways of visualizing and identifying specific sites in the DNA molecules, and bioinformatics-based image capturing/ analysis of fragments have led to improvements in the technology.

Optical mapping is useful for completing large genome projects (chromosome sized-contigs) and other applications, such as the identification of rearrangements in genomes. It can be used to assist in genome assembly, compare genomic structures, and correct genome assembly errors. It can also be used for comparing strain differences, for instance in medical microbiology applications. A major advantage of this method is that it avoids cloning or PCR artifacts and analyzes a single molecule at a time. Nevertheless, artifacts can be introduced, and thus multiple fragments are utilized to build a consensus map. The method has advantages over the (previous) industry standard of pulse field gel electrophoresis (PFGE) for creating large DNA fragment chromosome maps. Optical mapping is less time consuming and provides, in addition to fragment sizing information, "roadmap" locations that can be used for building consensus chromosome maps (Bogas et al., 2017; Lam et al., 2012; Mak et al., 2016; Mostovoy et al., 2016; Muller & Westerlund, 2017; Neely, Deen, & Hofkens, 2011).

Generally, optical mapping begins by shearing genomic DNA to a large size (100 mb to 1 gb). The key is to label these large DNA fragments at specific internal sites so that the labeled molecules can be imaged and lined up to provide contiguous overlapping maps. A number of approaches can be used to label specific genomic locations including digestion of the DNA with limited single-stranded nicking or a rare-cutting restriction enzyme or a set of enzymes which cleave at specific digestion sites. Labeling then can be performed on the free ends with specific dyes or by enzymatic "fill–in" reactions using single-stranded nucleases and polymerization dye incorporation. Enzymes which bind DNA at specific sites can also be used (such as DNA methyltransferases) where a label can be transferred to the DNA from the enzyme after binding. Labeling A+T or G+C rich DNA sites can also be used. The resulting fragments can be attached to the surface of glass slides or in elongated linear chambers for imaging using appropriate microscopy.

Once positioned on the solid substrates, the molecules are physically elongated. They can then be sized and the positions of the label(s), imaged, and recorded, creating an "optical map" of that molecule. By combining images from numerous DNA molecules, an

overlapping large "consensus optical map" can be made, which can then be used as a scaffold for the pinning of other physical markers, sequences, or contigs. A commercial leader in this field is Bionano Genomics (San Diego) (*https://bionanogenomics.com/products/*). Op Gen (*https://www.opgen.com/about-us/opgen-overview/*) will process samples as a commercial vendor.

Electron Microscopy DNA Sequencing

Electron microscopy DNA sequencing is another single-molecule sequencing technology, first considered in the 1960s and 70s. To be visualized, the DNA must be labeled with heavy atoms as the electron microscope cannot visualize individual nucleotides containing the standard carbon, hydrogen, nitrogen, oxygen, and phosphorus, isotopes in DNA. For these methods to work, the DNA must be denatured, labeled and stretched out on an electron microscope grid, using a "hypophase" method to keep the DNA denatured and linear. In theory, transmission electron microscopy DNA sequencing could provide extremely long read lengths, but the issue of electron beam damage has not been solved. A major company in this area is ZS Genetics (Wakefield MA) whose technology involves DNA nucleotides labeled with three heavy atom labels: bromine, iodine or trichloromethane. These appear as differential dark and light spots on the micrograph and the fourth DNA base is unlabeled. Sequencing by electron microscopy has not yet been commercialized.

ANCILLARY TECHNOLOGIES

DNA Shearing

A key step for all DNA sequencing methods is fragmenting DNA into a defined size. A number of shearing and large fragment DNA devices are available for isolation of DNA of various sizes for DNA library construction**.**

*Covaris, Inc. Shearing (***https://covarisinc.com***)*

Covaris, Inc. provides several acoustic shearing devices and consumables to enable fragmenting DNA and RNA to appropriate sizes for DNA and fragment library preparation (150 bp to 5 kb) and also provides protocols for chromatin and RNA shearing. For larger fragment isolation, they provide a G-TUBE™, designed to shear DNA into large fragments with a mean size ranging from 6 kbp to 20 kbp at room temperature (20°- 30°C). The g-TUBE works with a bench top centrifuge and the final size of the DNA fragments is controlled by the acceleration rate and speed of the centrifuge. g-TUBE uses centrifugal force (the "g" in G-TUBE) to push the sample through a precisely manufactured orifice. This produces shearing forces in the sample that fragment the DNA.

Diagenode, Inc. Megaruptor®

The Megaruptor was designed to provide a simple, automated, and reproducible device for the mechanical fragmentation of DNA from 2 kb to 75 kb**.** Shearing performance is independent of the source, concentration, temperature, or salt content of a DNA sample. The software allows two samples to be processed sequentially without additional user input and without cross-contamination. Several components for isolation of different sized DNA fragments are available. The Megaruptor base unit consists of an automated syringe pump with attached 9-port ceramic distribution valve and an integrated power supply. In order to control the device, a laptop with pre-loaded software is provided.

Enzymatic Fragmentation (NEBNext Ultra II FS; New England Biolabs Product Number E7805)

The NEBNext Ultra II FS enzyme in the NEBNext Ultra II FS DNA Library Prep Kit for Illumina contains the enzyme mix required to shear a broad range of input amounts of DNA into fragments for NextGen sequencing on the Illumina platform. The protocols use a minimum of 100 pg and create fragment sizes 100 bp to 1 kb. The reaction is generally 5 min, with a 30 min heat enzymatic heat kill step at 65°C.

*Sage Science Pippin (***www.sagescience.com/applications/dna-sequencing***)*

A novel technology for generating DNA fragments for sequencing has been developed by Sage Sciences. DNA from 100 bp to up to 2 mb can be size selected and isolated via a modified electrophoresis device that collects the selected DNA size in mini-chambers. The method can be used for any technique requiring size selected DNA, including library construction, removing low molecular weight content for long-read sequencing, and preparing ultra HMW DNA libraries for long-range genomic applications. Several versions of their machines are available, the Pippin Prep, BluePippin, or PippinHT, in which fragments of selected lengths are collected, eliminating all others.

FUTURE OUTLOOK

As DNA sequencing technology advances, the goal will be faster and more accurate sequencing (lower error rates, minimal artifacts) with lower amounts of input DNA and RNA at lower cost. Among these will be advances in sequencing from single cells and from circulating nucleic acids. Sequencing platforms that are smaller, require less power (battery operated), less reagents (zeptoliters or perhaps even just a few molecules of input reagents) and maintenance (perhaps disposable) will be utilized in medical, agricultural, ecological, and other settings (Shendure et al., 2017). Higher order multiplexing (barcoding) will enable more samples to be processed in a shorter time and at reduced cost. At the front end, advances in robotics, liquid handling, sample processing (nucleic acid preparation) will contribute to these advancements. Equally important will be advances in faster and more accurate bioinformatic data analysis as well as data transfer and storage.

ACKNOWLEDGEMENTS

F.M.A.'s laboratory is funded in part by NIH grants P30 DK040561 and P01 AI083214.

LITERATURE CITED

Ammar, R., Paton, T. A., Torti, D., Shlien, A., & Bader, G. D. (2015). Long read nanopore sequencing for detection of HLA and CYP2D6 variants and haplotypes. *F1000Res*, *4*, 17. doi: 10.12688/f1000research.6037.2.

Benner, S., Chen, R. J., Wilson, N. A., Abu-Shumays, R., Hurt, N., Lieberman, K. R., ... Akeson, M. (2007). Sequence-specific detection of individual DNA polymerase complexes in real time using a nanopore. *Nature Nanotechnology*, *2*(11), 718–724. doi: 10.1038/nnano.2007.344.

Bogas, D., Nyberg, L., Pacheco, R., Azevedo, N. F., Beech, J. P., Gomila, M., ... Westerlund, F. (2017). Applications of optical DNA mapping in microbiology. *Biotechniques*, *62*(6), 255–267. doi: 10.2144/000114555.

Branton, D., Deamer, D. W., Marziali, A., Bayley, H., Benner, S. A., Butler, T., ... Schloss, J. A. (2008). The potential and challenges of nanopore sequencing. *Nature Biotechnology*, *26*(10), 1146–1153. doi: 10.1038/nbt.1495.

Cai, W., Aburatani, H., Stanton, V. P., Jr., Housman, D. E., Wang, Y. K., & Schwartz, D. C. (1995). Ordered restriction endonuclease maps of yeast artificial chromosomes created by optical mapping on surfaces. *Proceedings of the National Academy of Sciences of the United States of America*, *92*(11), 5164–5168. doi: 10.1073/pnas.92.11.5164.

Callewaert, N., Geysens, S., Molemans, F., & Contreras, R. (2001). Ultrasensitive profiling and sequencing of N-linked oligosaccharides using standard DNA-sequencing equipment. *Glycobiology*, *11*(4), 275–281. doi: 10.1093/glycob/11.4.275.

Cao, M. D., Ganesamoorthy, D., Elliott, A. G., Zhang, H., Cooper, M. A., & Coin, L. J. (2016). Streaming algorithms for identification of pathogens and antibiotic resistance potential from real-time MinION(TM) sequencing. *Gigascience*, *5*(1), 32. doi: 10.1186/s13742-016-0137-2.

Chen, L., Liu, P., Evans, T. C., & Ettwiller, L. M. (2016). DNA damage is a major cause of sequencing errors, directly confounding variant identification. *bioRxiv*, doi: 10.1101/070334.

Cherf, G. M., Lieberman, K. R., Rashid, H., Lam, C. E., Karplus, K., & Akeson, M. (2012). Automated forward and reverse ratcheting of DNA in a nanopore at 5-A precision. *Nature Biotechnology*, *30*(4), 344–348. doi: 10.1038/nbt.2147.

Church, G. M. (2006). Genomes for all. *Scientific American*, *294*(1), 46–54. doi: 10.1038/scientificamerican0106-46.

Drmanac, R., Drmanac, S., Chui, G., Diaz, R., Hou, A., Jin, H., ... Little, D. (2002). Sequencing by hybridization (SBH): Advantages, achievements, and opportunities. *Advances in Biochemical Engineering/Biotechnology*, *77*, 75–101. doi: 10.1007/3-540-45713-5_5.

Fang, G., Munera, D., Friedman, D. I., Mandlik, A., Chao, M. C., Banerjee, O., ... Schadt, E. E. (2012). Genome-wide mapping of methylated adenine residues in pathogenic Escherichia coli using single-molecule real-time sequencing. *Nature Biotechnology*, *30*(12), 1232–1239. doi: 10.1038/nbt.2432.

Flusberg, B. A., Webster, D. R., Lee, J. H., Travers, K. J., Olivares, E. C., Clark, T. A., ... Turner, S. W. (2010). Direct detection of DNA methylation during single-molecule, real-time sequencing. *Nature Methods*, *7*(6), 461–465. doi: 10.1038/nmeth.1459.

Fuller, C. W., Middendorf, L. R., Benner, S. A., Church, G. M., Harris, T., Huang, X., ... Vezenov, D. V. (2009). The challenges of sequencing by synthesis. *Nature Biotechnology*, *27*(11), 1013–1023. doi: 10.1038/nbt.1585.

Furey, T. S. (2012). ChIP-seq and beyond: New and improved methodologies to detect and characterize protein-DNA interactions. *Nature Reviews Genetics*, *13*(12), 840–852. doi: 10.1038/nrg3306.

Greenough, L., Kelman, Z., & Gardner, A. F. (2015). The roles of family B and D DNA polymerases in *Thermococcus* species 9°N Okazaki fragment maturation. *Journal of Biological Chemistry*, *290*(20), 12514–12522. doi: 10.1074/jbc.M115.638130.

Greenough, L., Schermerhorn, K. M., Mazzola, L., Bybee, J., Rivizzigno, D., Cantin, E., ... Gardner, A. F. (2016). Adapting capillary gel electrophoresis as a sensitive, high-throughput method to accelerate characterization of

nucleic acid metabolic enzymes. *Nucleic Acids Research*, *44*(2), e15. doi: 10.1093/nar/gkv899.

Greninger, A. L., Naccache, S. N., Federman, S., Yu, G., Mbala, P., Bres, V., ... Chiu, C. Y. (2015). Rapid metagenomic identification of viral pathogens in clinical samples by real-time nanopore sequencing analysis. *Genome Medicine*, *7*, 99. doi: 10.1186/s13073-015-0220-9.

Hanna, G. J., Johnson, V. A., Kuritzkes, D. R., Richman, D. D., Martinez-Picado, J., Sutton, L., ... D'Aquila, R. T. (2000). Comparison of sequencing by hybridization and cycle sequencing for genotyping of human immunodeficiency virus type 1 reverse transcriptase. *Journal of Clinical Microbiology*, *38*(7), 2715–2721.

Hornblower, B., Coombs, A., Whitaker, R. D., Kolomeisky, A., Picone, S. J., Meller, A., & Akeson, M. (2007). Single-molecule analysis of DNA-protein complexes using nanopores. *Nature Methods*, *4*(4), 315–317. doi: 10.1038/nmeth1021.

Jing, J., Reed, J., Huang, J., Hu, X., Clarke, V., Edington, J., ... Schwartz, D. C. (1998). Automated high resolution optical mapping using arrayed, fluid-fixed DNA molecules. *Proceedings of the National Academy of Sciences of the United States of America*, *95*(14), 8046–8051. doi: 10.1073/pnas.95.14.8046.

Kasianowicz, J. J., Brandin, E., Branton, D., & Deamer, D. W. (1996). Characterization of individual polynucleotide molecules using a membrane channel. *Proceedings of the National Academy of Sciences of the United States of America*, *93*(24), 13770–13773. doi: 10.1073/pnas.93.24.13770.

Lam, E. T., Hastie, A., Lin, C., Ehrlich, D., Das, S. K., Austin, M. D., ... Kwok, P. Y. (2012). Genome mapping on nanochannel arrays for structural variation analysis and sequence assembly. *Nature Biotechnology*, *30*(8), 771–776. doi: 10.1038/nbt.2303.

Laroy, W., Contreras, R., & Callewaert, N. (2006). Glycome mapping on DNA sequencing equipment. *Nature Protocols*, *1*(1), 397–405. doi: 10.1038/nprot.2006.60.

Liu, Z., Wang, Y., Deng, T., & Chen, Q. (2016). Solid-State Nanopore-Based DNA Sequencing Technology. *Journal of Nanomaterials*, *2016*, 13. doi: 10.1155/2016/5284786.

Loman, N. (2015). How a small backpack for fast genomic sequencing is helping combat Ebola. *The Conversation*, available at *http://theconversation.com/how-a-small-backpack-for-fast-genomic-sequencing-is-helping-combat-ebola-41863*

Mak, A. C., Lai, Y. Y., Lam, E. T., Kwok, T. P., Leung, A. K., Poon, A., ... Kwok, P. Y. (2016). Genome-Wide Structural Variation Detection by Genome Mapping on Nanochannel Arrays. *Genetics*, *202*(1), 351–362. doi: 10.1534/genetics.115.183483.

Mardis, E. R. (2008). Next-generation DNA sequencing methods. *Annual Review of Genomics and Human Genetics*, *9*, 387–402. doi: 10.1146/annurev.genom.9.081307.164359.

Maxam, A. M., & Gilbert, W. (1980). Sequencing end-labeled DNA with base-specific chemical cleavages. *Methods in Enzymology*, *65*(1), 499–560. doi: 10.1016/S0076-6879(80)65059-9.

Meng, X., Benson, K., Chada, K., Huff, E. J., & Schwartz, D. C. (1995). Optical mapping of lambda bacteriophage clones using restriction endonucleases. *Nature Genetics*, *9*(4), 432–438. doi: 10.1038/ng0495-432.

Metzker, M. L. (2010). Sequencing technologies - the next generation. *Nature Reviews Genetics*, *11*(1), 31–46. doi: 10.1038/nrg2626.

Mirzabekov, A. D. (1994). DNA sequencing by hybridization–a megasequencing method and a diagnostic tool? *Trends in Biotechnology*, *12*(1), 27–32. doi: 10.1016/0167-7799(94)90008-6.

Mostovoy, Y., Levy-Sakin, M., Lam, J., Lam, E. T., Hastie, A. R., Marks, P., ... Kwok, P. Y. (2016). A hybrid approach for de novo human genome sequence assembly and phasing. *Nature Methods*, *13*(7), 587–590. doi: 10.1038/nmeth.3865.

Muller, V., & Westerlund, F. (2017). Optical DNA mapping in nanofluidic devices: Principles and applications. *Lab on A Chip*, *17*(4), 579–590. doi: 10.1039/c6lc01439a.

Murray, I. A., Clark, T. A., Morgan, R. D., Boitano, M., Anton, B. P., Luong, K., ... Roberts, R. J. (2012). The methylomes of six bacteria. *Nucleic Acids Research*, *40*(22), 11450–11462. doi: 10.1093/nar/gks891.

Neely, R. K., Deen, J., & Hofkens, J. (2011). Optical mapping of DNA: Single-molecule-based methods for mapping genomes. *Biopolymers*, *95*(5), 298–311. doi: 10.1002/bip.21579.

Qin, Y., Schneider, T. M., & Brenner, M. P. (2012). Sequencing by hybridization of long targets. *Plos One*, *7*(5), e35819. doi: 10.1371/journal.pone.0035819.

Quail, M. A., Smith, M., Coupland, P., Otto, T. D., Harris, S. R., Connor, T. R., ... Gu, Y. (2012). A tale of three next generation sequencing platforms: Comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers. *BMC Genomics [Electronic Resource]*, *13*, 341. doi: 10.1186/1471-2164-13-341.

Rhoads, A., & Au, K. F. (2015). PacBio Sequencing and Its Applications. *Genomics, Proteomics & Bioinformatics*, *13*(5), 278–289. doi: 10.1016/j.gpb.2015.08.002.

Ronaghi, M., Karamohamed, S., Pettersson, B., Uhlen, M., & Nyren, P. (1996). Real-time DNA sequencing using detection of pyrophosphate release. *Analytical Biochemistry*, *242*(1), 84–89. doi: 10.1006/abio.1996.0432.

Rothberg, J. M., Hinz, W., Rearick, T. M., Schultz, J., Mileski, W., Davey, M., ... Bustillo, J. (2011). An integrated semiconductor device enabling non-optical genome sequencing. *Nature*, *475*(7356), 348–352. doi: 10.1038/nature10242.

Sanger, F., & Coulson, A. R. (1975). A rapid method for determining sequences in DNA by primed

synthesis with DNA polymerase. *Journal of Molecular Biology*, *94*(3), 441–448.

Sanger, F., Nicklen, S., & Coulson, A. R. (1977). DNA sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of Sciences of the United States of America*, *74*(12), 5463–5467. doi: 10.1073/pnas.74.12.5463.

Schermerhorn, K. M., & Gardner, A. F. (2015). Pre-steady-state Kinetic Analysis of a Family D DNA Polymerase from *Thermococcus* sp. 9°N Reveals Mechanisms for Archaeal Genomic Replication and Maintenance. *Journal of Biological Chemistry*, *290*(36), 21800–21810. doi: 10.1074/jbc.M115.662841.

Schmidt, M., Vogel, A., Denton, A., Istace, B., Wormit, A., van de Geest, H., ... Usadel, B. (2017). Sequencing the gigabase plant genome of the wild tomato species *Solanum pennellii* using Oxford Nanopore single molecule sequencing. *Plant Cell*, *29*(10), 2336–2348.

Schwartz, D. C., Li, X., Hernandez, L. I., Ramnarain, S. P., Huff, E. J., & Wang, Y. K. (1993). Ordered restriction maps of Saccharomyces cerevisiae chromosomes constructed by optical mapping. *Science*, *262*(5130), 110–114.

Shendure, J., Balasubramanian, S., Church, G. M., Gilbert, W., Rogers, J., Schloss, J. A., & Waterston, R. H. (2017). DNA sequencing at 40: Past, present and future. *Nature*, *550*(7676), 345–353. doi: 10.1038/nature24286.

Shendure, J., & Ji, H. (2008). Next-generation DNA sequencing. *Nature Biotechnology*, *26*(10), 1135–1145. doi: 10.1038/nbt1486.

Vilfan, I. D., Tsai, Y. C., Clark, T. A., Wegener, J., Dai, Q., Yi, C., ... Korlach, J. (2013). Analysis of RNA base modification and structural rearrangement by single-molecule real-time detection of reverse transcription. *Journal of Nanobiotechnology*, *11*, 8. doi: 10.1186/1477-3155-11-8.

Zhang, W., Sun, Z., Menghe, B., & Zhang, H. (2015). Short communication: Single molecule, real-time sequencing technology revealed species- and strain-specific methylation patterns of 2 Lactobacillus strains. *Journal of Dairy Science*, *98*(5), 3020–3024. doi: 10.3168/jds.2014-9272.