Overview of DNA Sequencing Strategies

Jay A. Shendure,[1] Gregory J. Porreca,[2] George M. Church,[3] Andrew F. Gardner,[4] Cynthia L. Hendrickson,[5] Jan Kieleczawa,[6] and Barton E. Slatko[4]

[1]Department of Genome Sciences, University of Washington, Seattle, Washington
[2]Good Start Genetics, Inc., Cambridge, Massachusetts
[3]Department of Genetics, Harvard Medical School, Boston, Massachusetts
[4]New England Biolabs, Ipswich, Massachusetts
[5]HudsonAlpha Institute for Biotechnology, Huntsville, Alabama
[6]Wyzer Biosciences, Cambridge, Massachusetts

ABSTRACT

Efficient and cost-effective DNA sequencing technologies are critical to the progress of molecular biology. This overview of DNA sequencing strategies provides a high-level review of seven distinct approaches to DNA sequencing: (a) dideoxy sequencing; (b) solid phase sequencing; (c) sequencing-by-hybridization; (d) mass spectrometry; (e) cyclic array sequencing; (f) microelectrophoresis; and (g) nanopore sequencing. Other platforms currently in development are also briefly described. The primary focus here is on Sanger dideoxy sequencing, which has been the dominant technology since 1977, and on cyclic array strategies, for which several competitive implementations have been developed since 2005. Because the field of DNA sequencing is changing rapidly, this unit represents a snapshot as of September, 2011. *Curr. Protoc. Mol. Biol.* 96:7.1.1-7.1.23. © 2011 by John Wiley & Sons, Inc.

Keywords: DNA sequencing • next-generation sequencing • polony • cyclic array sequencing • genomics • sequencing by ligation • nanopore sequencing

INTRODUCTION

This unit provides a high-level overview of seven distinct approaches to DNA sequencing. These are: (a) dideoxy sequencing; (b) solid phase sequencing; (c) sequencing-by-hybridization; (d) mass spectrometry; (e) cyclic array sequencing; (f) microelectrophoresis; and (g) nanopore sequencing, with cursory mention of other emerging platforms. Additionally, this unit presents key parameters that should be considered when choosing the DNA sequencing strategy most appropriate for a given application. It should be emphasized that the DNA sequencing field is changing rapidly, so the information in this unit reflects the status of the technology as of this writing (September, 2011). It is worthwhile to note that the research goals that motivate DNA sequencing are undergoing substantial shifts as well, concurrent with the introduction of new technologies. Given that reference genome sequences for H. *sapiens* as well as numerous major model organisms are either complete or in published "draft" form, demand is shifting away from de novo genome sequencing toward applications such as resequencing (identifying genetic variation in the genome of an individual for whose species a reference genome is already available for population, evolutionary, or medical comparisons) and "tag counting" (i.e., gene expression analysis, chromatin occupancy, ChIP-Seq, or methylation analysis achieved by the sequencing of short but identifying DNA tags). While the initial generation of new technologies is delivering terabytes of sequence that is substantially shorter and somewhat less accurate than state-of-the-art Sanger sequencing, the ease of generating substantial depth of coverage enables high accuracy. Even so, while the utility of such "less accurate" sequence may be limited for de novo sequencing, it will likely be compatible, perhaps preferable, for other applications.

DNA SEQUENCING STRATEGIES

Dideoxy Sequencing: The Sanger Method

Sanger dideoxy or enzymatic DNA sequencing (Sanger, 1988; Sanger et al., 1977a,b) utilizes a DNA-dependent DNA polymerase to synthesize a complementary copy of a single-stranded DNA template

Current Protocols in Molecular Biology 7.1.1-7.1.23, October 2011
Published online October 2011 in Wiley Online Library (wileyonlinelibrary.com).
DOI: 10.1002/0471142727.mb0701s96

(sequencing by synthesis, SBS). The DNA polymerase synthesizes a new chain beginning at the 3′ end of a primer DNA complementary to a single-stranded "template" DNA strand. During synthesis, the deoxynucleotide added to the growing chain is complementary to the nucleotide in the template DNA. The creation of a phosphodiester bridge between the 3′ hydroxyl group at the growing end of the primer and the 5′ phosphate group of the incoming deoxynucleotide elongates the DNA chain, with the corresponding dNTP releasing pyrophosphate (PPi) when a dNMP is incorporated.

The principle behind the original dideoxy chain termination DNA sequencing technology takes advantage of the fact that DNA polymerases will incorporate a chain-terminating 2′,3′-dideoxynucleotide monophosphate (ddNMP) at the appropriate complementary position, in lieu of a deoxynucleotide monophosphate. When incorporated at the 3′ end of the extending single-stranded DNA chain, synthesis is stopped because the next nucleotide to be added cannot attach to the growing chain due to the lack of the required 3′ hydroxyl group for dNMP phosphodiester bond formation. In order to generate a continuing series of synthesis products that reflects each potential chain termination position, four reactions are performed, each with template, polymerase, all four dNTPs (one radioactively labeled), and primer. Each reaction also contains one of the four ddNTPs at a specific ratio that determines their relative probability of incorporation. The result is a collection of many terminated strands of different lengths within each of the four reactions. As each reaction contains only one ddNTP species, a set of different-length fragments is generated in each reaction, terminated at all of the positions corresponding to one of the four nucleotides in the template sequence. In this way, each of the four elongation reactions contains a population of extended primer chains, all of which have a fixed 5′ end determined by the annealed primer and a variable 3′ end terminating at a specific dideoxynucleotide. The four reactions are then individually electrophoresed in four lanes of a denaturing polyacrylamide gel to yield size separation with single-nucleotide resolution. The pattern of bands (with each band consisting of terminated fragments of a single length) across the four lanes allows direct readout of the primary sequence of the template under analysis (Fig. 7.1.1).

Thermal cycle sequencing was an adjunct protocol for dideoxy sequencing, developed as a method of sequencing double-stranded DNA templates without the need for independent pre-denaturation steps (Chen and Seeburg, 1985; Haltiner et al., 1985; Zagursky et al., 1985; Hattori and Sakaki, 1986; Fig. 7.1.1). The method is also used to sequence from a small number of template DNA molecules, which are repetitively utilized to generate a sufficient quantity of reaction products for subsequent detection. Numerous identical copies of the sequencing template undergo the primer extension reactions within microliter-scale volumes. Various templates can be used for sequencing, including single-stranded templates (M13 derived, asymmetric PCR reactions, etc.) or double-stranded templates (plasmid preparations, lambda phage, BACs, cosmids, fosmids, etc.). In thermal cycle sequencing, a dideoxy sequencing reaction mixture (consisting of template, primer, dNTPs, ddNTPs, and a thermostable DNA polymerase) is subjected to repeated rounds of denaturation, annealing, and synthesis steps, similar to PCR but using only a single primer (see Chapter 15 for PCR). In this manner, linear amplification of the sequencing products occurs, allowing much less template DNA to be used than is usually required. In addition, thermal cycle sequencing eliminates the requirements for a separate annealing reaction preceding the sequencing reaction itself and for denaturing double-stranded DNA templates, and is compatible with automation processes. The method also helps eliminate many false secondary priming events during the annealing step. Thermal cycle sequencing offers advantages of helping eliminate sequencing "compressions" or "false stops" where isothermal polymerases have difficulties processing through certain sequence contexts, such as high-GC regions. Various protocols have been developed for thermal cycle sequencing reactions (Applied Biosystems, 1989; Carothers et al., 1989; Murray, 1989; Adams and Blakesley, 1991; Krishnan et al., 1991; Young and Blakesley, 1991; Sears et al., 1992; Craxton, 1993; Slatko, 1996). Two such protocols are included in *UNIT 7.4A* and a protocol for automated fluorescent dideoxy DNA sequencing is provided in *UNIT 7.2*. Several thermal cycle sequencing kits are commercially available.

Two historical protocols for dideoxy sequencing using a radiolabeled dNTP are provided in *UNIT 7.4A*. The original dideoxy method (Sanger, 1988; Sanger et al., 1977a,b)

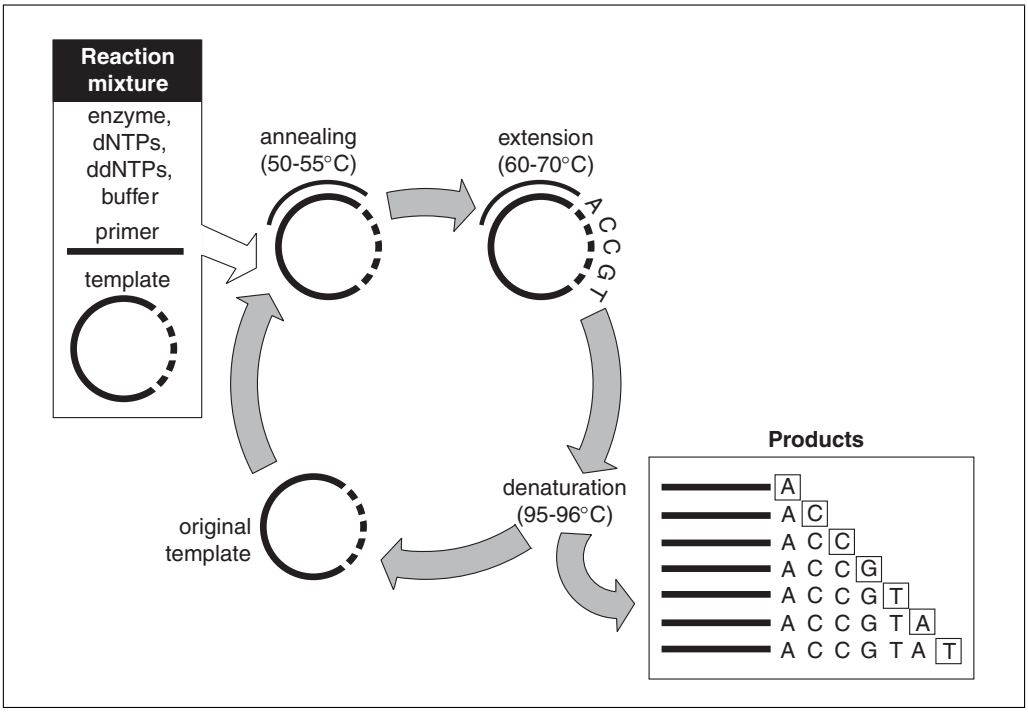

Figure 7.1.1 Cycle sequencing. This is a simple method in which successive rounds of denaturation, annealing and extension in a thermal cycler result in linear amplification of extension products. The products are then loaded onto a gel or injected into a capillary. Reprinted with permission from Applied Biosystems Automated DNA Sequencing Chemistry Guide, 2010 (*http://www3.appliedbiosystems.com/cms/groups/mcb_support/documents/generaldocuments/cms_041003.pdf*).

was developed for use with the large fragment of *E. coli* DNA polymerase I (Klenow fragment). In this method, as described above, the primer was extended and labeled until incorporation of a specific dideoxynucleotide caused termination. In a subsequent chase reaction, all four dNTPs were added at high concentrations so that all chains not terminated by a dideoxynucleotide would be elongated into high-molecular-weight DNA that remains unresolved at the top of the sequencing gel.

An alternative method was developed, termed "labeling/termination," for use with the highly processive modified T7 DNA polymerase (Sequenase; *UNIT 7.4A*; Tabor et al., 1987; Tabor and Richardson, 1987a,b, 1989a,b, 1990). In this case, labeling of extension products and termination by incorporation of a dideoxynucleotide occurred in two separate reactions (Fig. 7.1.2, left side). On average, the labeling-termination method was capable of yielding longer sequencing products than those obtained using the original Sanger protocol and was popular because of the advantage of obtaining the maximum amount of sequence information per template.

A variety of DNA polymerases have been commercially available for classical Sanger sequencing, and many companies supply reagent kits. Thermostable DNA polymerases are now the enzymes of choice for Sanger DNA sequencing because they can carry out a sequencing reaction at high temperatures, enabling "thermal cycling" to increase the yield of sequencing fragments, and aid in eliminating many "false terminations" due to destabilizing secondary structures of the DNA template, which can interfere with the elongation reaction.

The original radioactive Sanger dideoxy sequencing protocols, which used [α-^{32}P]dATP to label nascent DNA chains, were subsequently modified to use [α-^{35}S]dATP or [α-^{33}P]dATP because the lower-energy β emissions of ^{35}S and ^{33}P result in sharper autoradiographic bands compared to those generated by ^{32}P, allowing more (longer) sequence to be read, especially from the upper portion of the autoradiograph. ^{33}P has a maximum β-emission energy that is 50% stronger than ^{35}S, but 5-fold weaker than ^{32}P. Sequences generated using [α-^{33}P]dATP have short exposure times similar to ^{32}P, but band resolution comparable to that of ^{35}S (Zagursky et al., 1991). The lower-energy emissions of ^{35}S cause fewer breaks in the sugar-phosphate backbone of the

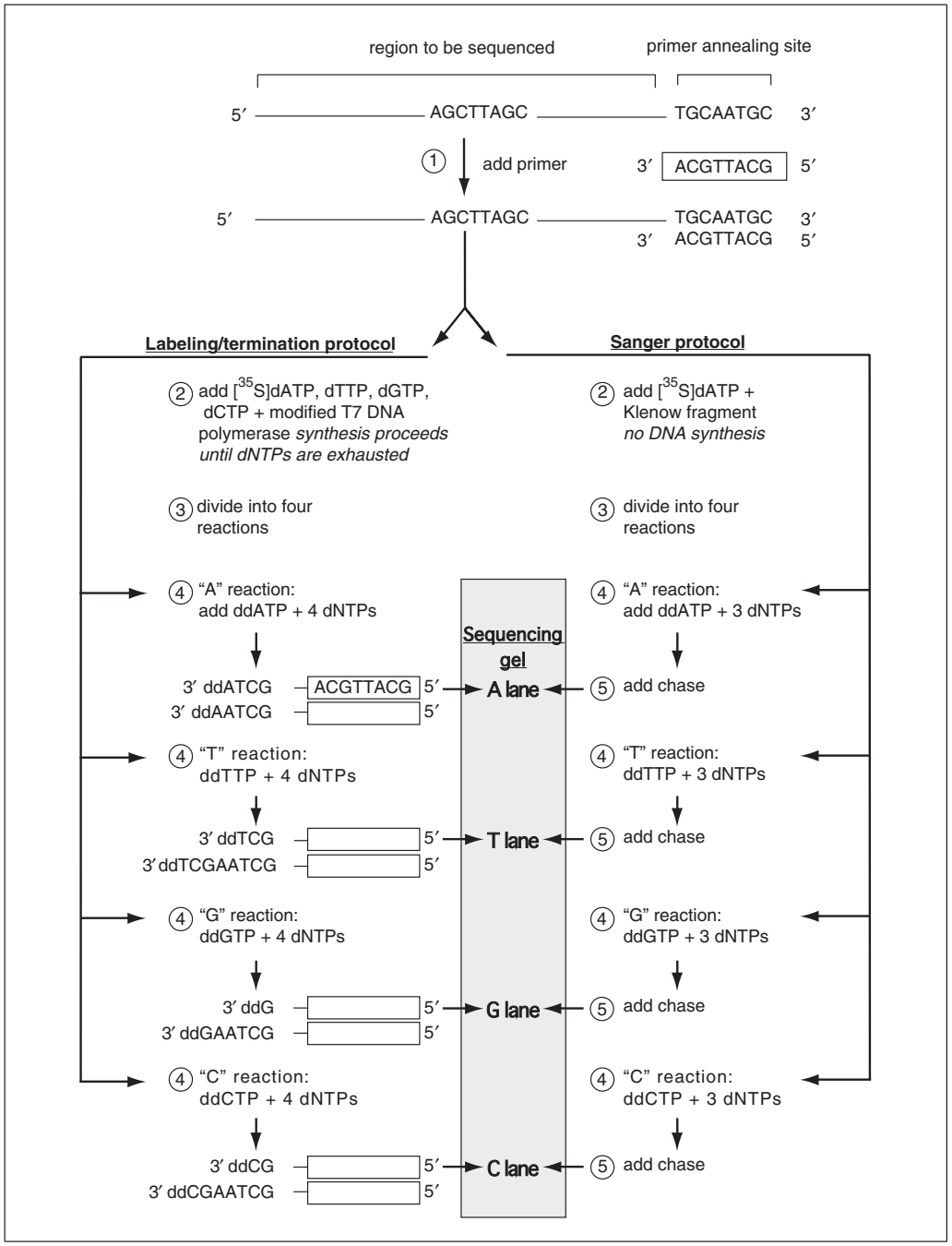

Figure 7.1.2 General strategy for dideoxy DNA sequencing methods. To sequence a fragment of DNA, a set of radiolabeled single-stranded oligonucleotides is generated in four separate reactions. In each of the four reactions, the oligonucleotides have one fixed end and one end that terminates sequentially at each A, T, G, or C, respectively. The products of each reaction are fractionated by electrophoresis on adjacent lanes of a high-resolution polyacrylamide gel. After autoradiography, the DNA sequence can be "read" directly from the gel image, either manually or with a digitizer. A single-stranded DNA fragment is annealed to an oligonucleotide primer for polymerization (step 1). In the main protocol (right side), a DNA-dependent DNA polymerase and radiolabeled dATP are added (step 2). The reaction is divided into four aliquots (step 3) and the other three dNTPs and either ddATP, ddTTP, ddGTP, or ddCTP are added (step 4). DNA synthesis occurs until terminated by the incorporation of a ddNTP. A "chase" of all four dNTPs (step 5) elongates chains not terminated by a ddNMP into higher-molecular-weight DNA. After the reactions are terminated, samples are loaded on adjacent lanes of a sequencing gel.

DNA meaning that ^{35}S-labeled reaction products can be stored at –20°C for several weeks without significant degradation; by contrast, ^{32}P products should be electrophoresed within a day. Another advantage of using ^{33}P and ^{35}S is that users are exposed to lower radiation doses than with ^{32}P. However, ^{32}P offered the advantage of short exposure times and was particularly useful in situations, such as verification of plasmid constructions, where maximizing resolution in the higher region of the sequencing gel autoradiograph was not a priority.

An alternative to labeling the nascent oligonucleotide with [α-^{35}S]dATP, [α-^{33}P]dATP, or [α-^{32}P]dATP is to use a 5′-end-labeled primer generated with T4 polynucleotide kinase and [γ-^{32}P]ATP or [γ-^{35}S]ATP (*UNIT 3.10*). Protocols for sequencing using end-labeled primers are provided in *UNIT 7.4A*.

In lieu of using radioactivity, a chemiluminescent detection method was developed that is comparable in sensitivity to traditional radiolabeling. A biotinylated primer is used in the dideoxy sequencing reactions, and, after electrophoresis of the biotinylated sequencing products on a sequencing gel, the products are transferred from the gel to a nylon membrane. After UV cross-linking the DNA to the membrane, the membrane is treated with streptavadin, biotinylated alkaline phosphatase, and a detection reagent, such as CDP* (Tropix/Applied Biosystems), which emits light upon dephosphorylation. After exposure to X-ray film, the resultant lumigram can be read in a manner similar to an autoradiogram when using radioactivity (Beck et al., 1989; Tizard et al., 1990; Creasey et al., 1991; Evans, 1991; Martin et al., 1991; *UNIT 7.4B*). Technology for end labeling DNA fragments with biotin allowed this detection method to also be used in chemical sequencing reactions (Tizard et al., 1990). Alternatively, the dideoxy sequencing reactions could be carried out with a standard, non-biotinylated primer. After electrophoresis, transfer, and cross-linking to a membrane, the sequencing products are hybridized with a biotinylated probe complementary to the primer before performing the detection method described above. This latter method can also be used with the products of chemical sequencing.

Hybridization of sequencing products with a probe complementary to the sequencing primer is the basis for "multiplex sequencing," which was developed to increase the throughput of DNA sequence information. In this method, sequencing products derived from a mixture of templates are subjected to electrophoresis on a sequencing gel, transferred to a membrane, and hybridized with a probe specific for one template. A set of sequencing vectors was commercially available for multiplex sequencing using the dideoxy method that had the identical priming region but contained unique sequences between the primer locus and the cloning site for the target DNA. These unique sequences were incorporated into the sequencing reaction products. After hybridization and detection of the sequencing ladder derived from a single template, the probe was removed at a high temperature and the membrane rehybridized with a different probe complementary to an independent set of sequencing products that were electrophoresed in the same lanes of the sequencing gel. Thus, the amount of sequence information available from one gel is multiplied by the number of times the membrane could be rehybridized (in practice, up to 20 times). Multiplex sequencing originally used radioactive probes and chemical sequencing technology (Church and Gilbert, 1984; Church and Kiefer-Higgins, 1988) but it was equally well adapted to chemiluminescent detection and/or dideoxy reactions.

A discussion of vectors used for dideoxy sequencing is provided in *UNIT 7.1*, and protocols for preparation of DNA templates derived from M13, plasmid, and bacteriophage λ vectors are provided in *UNIT 7.3*. The products of the polymerase chain reaction (PCR) and other amplification methods can also be sequenced by the dideoxy method, and several protocols for generating these templates are provided in *UNITS 15.2 & 15.5*.

The practical limit on the amount of information that could be obtained from a set of classical sequencing reactions was the resolution of the sequencing gel (refer to *UNIT 7.6* for protocols on setting up, running and processing sequencing gels). Generally, up to 700 nucleotides of sequence information could be reliably obtained in one set of radiolabeled or fluorescent sequencing reactions, although more information (up to 1000 nucleotides) could be obtained using first-generation automated DNA sequencers (see below). Thus, if the region of DNA to be sequenced was <500 to 600 nucleotides, a single cloning into the appropriate vector containing priming sites or a PCR product where the primer sequence is known, is all that was usually necessary to produce a molecule that could be sequenced in a single set of reactions.

For a larger region of DNA, it was generally necessary to break a large fragment into smaller ones that were then individually sequenced. This can be done in a random or an ordered fashion. A discussion of strategies for sequencing large regions of DNA using classical Sanger sequencing methods can be found in *UNIT 7.1*. Two protocols for subdividing large regions of DNA are provided in *UNIT 7.2*. These protocols are used to create a set of ordered, or nested, deletions for DNA sequencing using exonuclease III or Bal 31 nuclease. Analogous methods using transposons were also developed (Hoffman and Jendrisak, 1999; Strathmann et al., 1991).

First-Generation Automated Sanger DNA Sequencing

As technology moved to the first generation of automated sequencers, many time-consuming steps in the DNA sequencing process were eliminated (*UNIT 7.4A*). This included automation of most of the process including the gel electrophoresis steps, detection of the fluorescent DNA band patterns, and analysis of bands. The four sets of oligonucleotides generated by the sequencing reactions were loaded onto slab gels manually, and electrophoresis was then controlled automatically. Detection occurred in real time at a point near the bottom edge of the gel, and the bands of DNA, moving sequentially past a detector, were recorded. The replacement of radioactivity with labeled primers or ddNTPs and development of polymerases that could effectively incorporate these dyes were major steps, in the development of the automation. As such, dideoxy sequencing with four fluorescently labeled ddNTPs remains the general method of choice for these first-generation automated DNA sequencers. Nearly all first-generation automated DNA sequencing performed today makes use of automated capillary electrophoresis (*UNIT 2.8*), which typically analyzes 8 to 96 sequencing reactions simultaneously, in combination with the use of the latest generation of fluorescent dyes (resonant energy transfer, or ET dyes; Ju et al., 1995) that are efficiently excited at a common wavelength and exhibit strong and distinct fluorescent emissions (Ju et al., 1996; Lee et al., 1997; also see *UNIT 7.2* for a more detailed description). They are markedly superior to single dye–labeled primers and terminators for DNA sequencing and PCR fragment analysis. The strong emissions helped enable sequencing directly from large-insert clones (>30 kb), such as cosmid or BAC clones, which was very important for closing gaps in large genomic sequencing and mapping projects (Marra et al., 1996; Heiner et al., 1998).

Thus, implementations of "first-generation automated DNA sequencing" differ in several key ways from the methods described for the original Sanger sequencing. The "first generation" automated sequencers, pioneered by DuPont, Applied BioSystems (now part of Life Technologies), Pharmacia (now part of GE Healthcare), LiCor, Millipore, and other commercial companies, provided the advantages of eliminating radioactivity (or chemiluminescence), enabling "one-lane" sequencing, "one-tube" reactions, automated base calling, and elimination of slab-gel technology in favor of multi-capillary electrophoresis with automatic, electrokinetic-injection lane loading.

Although they revolutionized automated DNA sequencing technologies, fluorescent sequencing procedures have disadvantages, most notably false termination and background noise. In the dye primer method, all the extended DNA fragments—including false-terminated fragments—carry a fluorescent dye from the primer, and are thus all detected by the fluorescence sequencer. This causes background noise and results in inaccurate sequencing data. In the dye terminator method, the excess dye-labeled dideoxynucleotides need to removed prior to detection of the fluorescent dye molecules that have been incorporated into extended DNA templates (*UNIT 7.2*). Furthermore, if RNAs and nicked DNAs are present in the DNA templates, they may act as primers to generate false termination products or high background noise.

Using Sanger sequencing, DNA is sometimes encountered that cannot be accurately sequenced by the dideoxy method, especially G+C rich DNA. The composition or secondary structure of the template can sometimes cause premature termination by DNA polymerase. In addition, DNA molecules with significant secondary structure might not electrophorese properly on denaturing sequencing gels (Fig. 7.1.3). For these reasons, several techniques were developed to eliminate these problems ("false terminations" in the case of synthesis issues or gel "compressions" in electrophoresis). These "fixes" included more highly denaturing gels (formamide/urea), running gels at elevated temperatures, use of 7-deaza dGTP or 7-deaza dATP base analogs, and thermal cycle

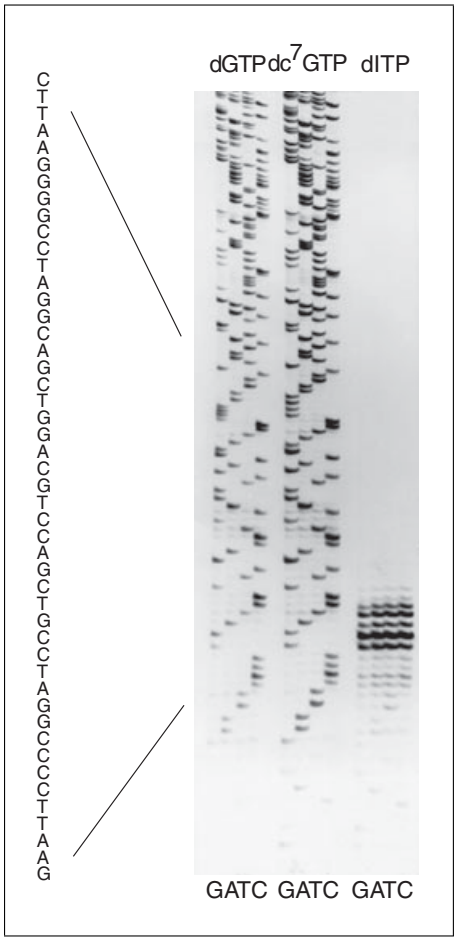

Figure 7.1.3 The left side of the figure shows the sequence of the polylinker region of M13mp7, which includes a 48-base inverted repeat (palindrome). The right side of the figure shows the products of sequencing reactions using M13mp7 DNA template, *Taq* DNA polymerase (*Taq* Version 2.0), the protocol and reagents from the TAQuence Version 2.0 DNA sequencing kit (U.S. Biochemical), and dGTP, 7-deaza-dGTP (dc^7GTP), or dITP. The secondary structure of the template DNA may inhibit elongation of the DNA polymerase under some conditions; synthesis in the presence of dITP is stopped completely within the palindromic region. No such stopping is seen on other templates lacking inverted repeat sequences. The palindromic region can also exhibit strong compression artifacts on sequencing gels; electrophoresis in these examples was in the presence of 40% formamide and 7 M urea. A mild compression near the 3′ end of the palindrome (GGGGAATTC) is still evident in lanes with samples from dGTP-containing reactions (compare to samples from 7-deaza-dGTP-containing reactions). Photo courtesy of Carl Fuller, U.S. Biochemical.

sequencing with thermostable polymerases, pyrophosphatases, and other additives. Refer to *UNIT 7.2* for further description of these techniques.

Solid-Phase Sequencing

An innovation that was applicable to both manual and automated DNA sequencing was the use of a solid-phase capture strategy to generate single-stranded DNA templates (Hultman et al., 1989, 1991; Jones et al., 1991; Kaneoka et al., 1991; Zimmerman et al., 1992). In this approach, one strand of a double-stranded DNA molecule is biotinylated (e.g., by amplification using PCR in which one of the two primers is biotinylated; see Chapter 15). The hemi-biotinylated DNA molecule is then bound to streptavidin-ferromagnetic beads. The strands are denatured by treating the beads with alkali and the biotinylated strands are separated from the nonbiotinylated strands using a magnet that traps the bead complex to which the biotinylated strands are bound. Sequencing reactions can be performed using either the biotinylated strand-bead complex or the nonbiotinylated strand preparation as the template. This allows both strands to be sequenced independently, if desired.

A sequencing chemistry using solid-phase-capturable dideoxynucleotides was developed that produces much cleaner sequencing data on both slab gel and capillary array sequencers, eliminating the disadvantages of the prevailing dye primer and dye terminator chemistries (Ju et al., 1997; Ju, 1999). The procedure involves coupling fluorescent ET primers that produce high fluorescent signals with solid phase–capturable terminators such as biotinylated dideoxynucleotides. After the sequencing reaction, the extended DNA fragments are captured on magnetic beads coated with streptavidin, while the other components in the sequencing reaction are washed away. Only the pure dideoxynucleotide-terminated extension products are released from the magnetic beads and loaded on the sequencing gel, producing very clean, high-quality data.

Sequencing by Hybridization

The principle of sequencing by hybridization (SBH) is that the differential hybridization of target DNA to an array of oligonucleotide

probes can be used to decode its primary DNA sequence. The most successful implementations of this approach rely on probe sequences based on the reference genome sequence of a given species, such that genomic DNA derived from individuals of that species can be hybridized to the array to reveal differences relative to the reference genome (i.e., resequencing, rather than de novo sequencing). The same concept is used for many genotyping array platforms, except that in this case SBH attempts to query all bases, rather than only bases at which common polymorphisms have been defined. In resequencing arrays developed by Affymetrix and Perlegen (no longer in operation and now part of Pfizer), each feature on the array consists of a 25-bp oligonucleotide of defined sequence. For each base pair to be resequenced, there are four features on the chip that differ only at their central (13^{th}) position (dA, dG, dC, or dT), while the flanking sequence is constant and is based on the reference genome. After hybridization of labeled target DNA to the chip, followed by imaging of the array, the relative intensities at each set of four features targeting a given position can be used to infer its identity. Perlegen developed and applied SBH arrays for resequencing the nonrepetitive portion of chromosome 21 in multiple individuals, thereby discovering many novel SNPs (Patil et al., 2001). A key limitation was a high false positive rate (3%), a significant problem as there is no possibility of redundant coverage to obtain higher consensus accuracies, which is possible with other sequencing methods.

Whereas it has been observed that SBH approaches have inherent difficulty with heterozygosity in large diploid genomes, they have also demonstrated a high level of accuracy in resequencing smaller haploid genomes. For example, a recent study demonstrated that Affymetrix arrays could be used for resequencing haploid *S. cerevisiae* strains with an impressively low false-positive rate (detection of 87% of ~30,000 SNPs with only eight false positives; Gresham et al., 2006). Nimblegen has developed a two-tiered SBH approach to genomic resequencing of microbial genomes: genome-wide discovery of approximate locations of mutations in the first array, followed by fine mapping in a second, custom array (Albert et al., 2005; Herring et al., 2006). In one recent study using Nimblegen resequencing arrays, 95 SNPs were predicted in the genomes of five evolved *E. coli* strains, 17 of which were confirmed by Sanger sequencing (Herring et al., 2006).

Mass Spectrometry

Over the past decade, mass spectrometry (MS) has established itself as the key data acquisition platform for the emerging field of proteomics. There are also applications for MS in genomics, including methods for genotyping, quantitative DNA analysis, gene expression analysis, analysis of indels (insertions/deletions) and DNA methylation, and DNA/RNA sequencing (Ragoussis et al., 2006). Sequencing with matrix-assisted laser desorption/ionization time-of-flight mass spectrometry (MALDI-TOF-MS) relies on the precise measurement of the masses of DNA fragments present within a mixture of nucleic acids (Edwards et al., 2005). With MS sequencing, fragmentation can also be achieved by primer extension with dideoxy termination; the primary difference is the use of MALDI-TOF MS rather than capillary electrophoresis to resolve fragment sizes. Alternatively, fragments are transcribed to RNA and subjected to base-specific cleavage prior to analysis. For de novo sequencing with MS, read lengths have generally been limited to <100 bp.

Applications of MS sequencing include deciphering sequences that appear as compression zones by gel electrophoresis, direct sequencing of RNA (including for identification of post-translational modifications of ribosomal RNA), the robust discovery of heterozygous frameshift and substitution mutations within PCR products in resequencing projects, and DNA methylation analysis (Ragoussis et al., 2006). MS sequencing will likely continue to have a role for specific problems not easily addressed with other methods, but is unlikely to displace conventional methods for most DNA sequencing applications. It is worth noting that at the beginning of the human genome project, MS was thought to carry great promise to deliver faster sequencing. However, with the advances of capillary electrophoresis and particularly with the emergence of "next generation," high-throughput methods, MS-based sequencing methods are no longer competitive.

NextGen Automated DNA Sequencing Strategies

Currently, there is a tradeoff between long read lengths and the overall throughput of a sequencing instrument. Depending on which parameter is being optimized, conventional dideoxy instruments are capable of reads just over 1000 base pairs in length, or production throughputs of over 1 megabase per day.

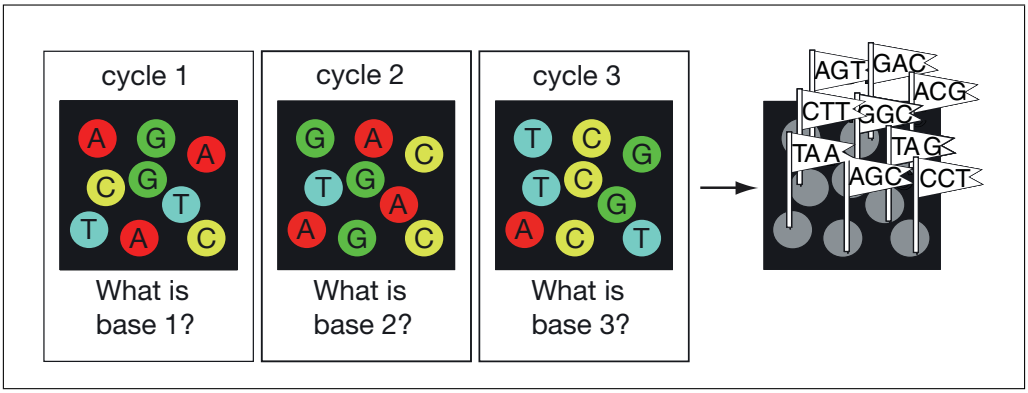

Figure 7.1.4 The concept of cyclic array sequencing platforms involves an array of DNA features to be sequenced, immobilized to constant locations on a solid substrate. At each cycle, the identity of a single base position is interrogated at each feature. Data are collected at each cycle by imaging of the array. At the conclusion of the experiment, imaging data for each feature collected over the full set of cycles can be used to infer contiguous stretches of sequence. The power of cyclic array methods to achieve low costs derives from the possibility of simultaneously sequencing millions to potentially billions of sequencing features in parallel. Also, microliter-scale reagent volumes can be used to manipulate all features in a single reaction, such that the effective reagent volume per sequencing feature is on the order of picoliters or femtoliters. For the color version of this figure go to *http://www.currentprotocols.com/protocol/mb0701*.

Because of variation in the levels of optimization and instrument up-time, the cost of dideoxy sequencing varies widely throughout the research community (see *UNIT 7.2* for protocols). The in-house costs of high-throughput sequencing centers may be as low as $0.50 per kilobase (kb), while core facilities and commercial entities may charge anywhere from $3.50 to $10 per sequencing read, partly depending upon whether reactions are performed in microcentrifuge tubes or 96-well plates.

Cyclic Array Sequencing

All of the recently developed non-Sanger commercial sequencing platforms, including systems from 454/Roche, Illumina, Applied Biosystems/Life Technologies, Dover, Helicos Biosciences, and Ion Torrent (now part of Life Technologies), fall under the rubric of a single paradigm termed cyclic array sequencing (Fig. 7.1.4; Shendure and Ji, 2008; Metzker, 2010). Cyclic array platforms achieve low costs by simultaneously decoding a two-dimensional array bearing millions (potentially billions) of distinct "features" (individual units that can be sequenced). The sequencing features are "clonal" in that each resolvable unit contains only one species of DNA (as a single molecule or in multiple identical copies) physically immobilized on the array. The features may be arranged in an ordered fashion or may be randomly dispersed. Each DNA feature generally includes an unknown sequence of interest flanked by universal adaptor sequences. A key point in this approach is that since the features are immobilized on a single surface, a single reagent volume is applied to simultaneously access and manipulate all features in parallel. The sequencing process is cyclic because in each cycle an enzymatic process is applied to interrogate the identity of the base at a particular position in each feature for all features in parallel. The enzymatic process is coupled to either the production of light or the incorporation of a fluorescent group, or as with the Ion Torrent system, measurement of H^+ ion release which is detected on a semiconductor chip acting as a very sensitive pH meter. At the conclusion of each cycle, data are acquired by CCD-based imaging of the array (except for the Ion Torrent system, as described above). Subsequent cycles are aimed at interrogating different base positions within the template. After multiple cycles of enzymatic manipulation, position-specific interrogation, and array imaging, a contiguous sequence for each DNA fragment can be derived from analysis of the full series of imaging data covering its position.

Although this basic paradigm serves to describe several different platforms for cyclic array sequencing, the platforms differ markedly in the specifics of implementation. The primary areas of difference are (1) the method used to generate the DNA sequencing features, (2) the biochemistry used

to perform cyclic sequencing, and (3) the detection method.

"Polymerase colony" or "polony" is a generic term to describe the polymerase-driven amplification of a complex library of sequencing templates, such that amplicons originating from any given template within the complex library remain locally clustered, analogous to a bacterial colony (Mitra and Church, 1999; Mitra et al., 2003; Edwards, 2008; *UNIT 7.8*). Both the polony sequencing system developed at Harvard Medical School (Shendure et al., 2004, 2005; *UNIT 7.8*), its derivative, the Life Technologies/ABI SOLiD system, the 454 system (Roche; Margulies et al., 2005), and Ion Torrent's PGM (Personal Genome Machine) generate DNA sequencing features by performing an emulsion PCR amplification of a complex library of sequencing templates, such that PCR products derived from individual templates are clonally captured onto the surface of micrometer-scale beads. Emulsion PCR is similar to standard PCR, but is performed in the context of a water-in-oil emulsion (Tawfik and Griffiths, 1998) such that the aqueous component is separated into millions of stable reaction chambers/droplets of uniform size. A complex library (consisting of unknown fragments to be sequenced, flanked by universal adaptors) can be simultaneously amplified with a single pair of primers, but PCR amplicons originating from a single molecule within the library remain compartmentalized to a single aqueous chamber. It has been demonstrated that the inclusion of 1-μm paramagnetic beads, where the beads bear one of the two PCR primers on their surface, enables the solid-phase capture of PCR amplicons generated within individual emulsion PCR compartments (Dressman et al., 2003). Individual templates are "clonally amplified" in that any single bead recovered from the emulsion PCR reaction should carry multiple copies of a single library species, while different beads (which were present in different compartments) carry multiple copies of other library species.

The Harvard Medical School platform uses emulsion PCR protocols adapted directly from work done by the group of Bert Vogelstein and Ken Kinzler (Dressman et al., 2003; Diehl et al., 2005, 2006; Li et al., 2006), using paramagnetic beads that are 1 μm in diameter. The SOLiD sequencer also uses 1-μm beads, but the 454 system uses a slightly different oil phase that yields much larger aqueous compartments, thus supporting emulsion PCR with capture on much larger beads (28 μm in diameter). In the 454 system, amplified beads are arrayed on a prefabricated fiber optic plate etched to contain over one million picoliter-scale wells. Sequencing is performed by the pyrosequencing method that was initially developed by Mostafa Ronaghi and colleagues (Valdar et al., 2006). Pyrosequencing involves polymerase extension of a primed template by sequential addition of a single nucleotide species at each cycle and the detection of a burst of light when pyrophosphate (PPi) is released and detected. Currently, about a million wells contain template-bearing beads that yield useful sequence. A key advantage of the 454 system, relative to other cyclic array platforms, is the clear demonstration of read lengths in excess of 400 base pairs, with a projected 1000-base read length in the near future. The 454 platform was the first commercialized massively parallel sequencing system, and it has contributed to a number of successful projects over a range of applications, including bacterial genome resequencing (Andries et al., 2005; Velicer et al., 2006), de novo bacterial genome sequencing (best done as a combination of Sanger dideoxy sequencing and pyrosequencing-based reads; Goldberg et al., 2006; Smith et al., 2007), small RNA discovery (Berezikov et al., 2006; Ruby et al., 2006), and metagenomic sampling (Edwards et al., 2006; Wiley et al., 2009).

There are several disadvantages to consider with the 454 system. (1) There is a relatively high error rate at homopolymeric sequences (consecutive runs of the same base) because it is difficult to interpret the precise number of consecutive incorporations in a single cycle. (2) The cost of sequencing with the 454 system is significantly lower than that of conventional dideoxy sequencing, but higher than that of platforms such as Illumina and Life Tech/SOLiD. The throughput of the instrument is also lower. These disadvantages are balanced by the major advantage of significantly longer read lengths.

Ion Torrent (*http://www.iontorrent.com*) as part of Life Technologies, has developed a DNA cycle sequencing method which identifies individual nucleotide incorporation during DNA synthesis by measuring a voltage change when hydrogen ions are released as the correct nucleotide is incorporated into a growing DNA chain. The method takes advantage of existing semiconductor technology, and is the first sequencing technology to use a voltage change instead of fluorescence during detection. The

Ion Torrent platform has several unique potential advantages over many other technologies because it uses unmodified nucleotides, does not image the DNA being sequenced (no photodamage), and eliminates storage (cost and space) issues related to the terabytes of image data produced with other methods. As a consequence, the sequencing runs can be completed much faster as compared to other cyclic array platforms. The Ion Torrent platform uses essentially the same protocol to generate templates for sequencing as the 454, Life Technologies/SOLiD, and the Harvard Medical School system, i.e., emulsion PCR. Like pyrosequencing, the method of sequencing involves sequential addition of non-terminating nucleotide species. As a consequence, some of the disadvantages of the 454 platform are also applicable to the Ion Torrent system, i.e., a high insertion-deletion rate, particularly at homopolymeric tracts. Currently, this system is capable of producing about 100,000 reads with an average read length of 120 bases at a cost of about $500. It is anticipated that the number of reads per run and the read length will increase significantly in the near future.

In the most recent implementation of the Harvard Medical School polony sequencing system (Shendure et al., 2005; *UNIT 7.8*), DNA sequencing features (1-μm beads generated by emulsion PCR) are dispersed on the surface of a glass coverslip as a disordered array, and immobilized either by a thin acrylamide gel or by direct covalent attachment to the surface. A unique aspect of this platform is that the enzyme conferring specificity during each sequencing cycle is a ligase, rather than a polymerase. At each cycle, a degenerate pool of fluorescently labeled oligonucleotides (e.g., nonamers or 9-mers) is introduced for sequencing by ligation. The oligonucleotides are structured such that the identity of a particular base position correlates with the identity of the fluorescent base attached to it. At the primer-template junction, sequencing features will primarily incorporate oligonucleotides bearing a single type of fluorophore (as oligonucleotide complementary to template is preferentially ligated), which reveals the identity of the base at that position. The method is successful for sequencing contiguous stretches of 6 to 7 bases from the site of ligation. By sequencing libraries that consist of mate-paired tags derived from restriction digestion with adaptors containing Type IIs restriction enzyme sites, and furthermore by sequencing each tag from both the 5′ and 3′ directions, one can obtain at least 26 base pairs of sequence information per bead feature (see Shendure et al., 2005, for further details).

Data are collected at each cycle by four-color imaging with a modified epifluoresence microscope. The system has demonstrated success in resequencing a bacterial genome with raw accuracies of up to 99.9%, and with very high consensus accuracies (<1 error per million bases), at a cost at least one order of magnitude below conventional sequencing (Shendure et al., 2005). An open-source version of the platform can be implemented with off-the-shelf instrumentation and reagents, with the instrument itself costing $150,000. The platform was licensed by Harvard Medical School for commercial development to Agencourt/Applied Biosystems, which notably has developed more sophisticated sequencing-by-ligation chemistry that generates up to 75 base-pair reads per mate-paired tag (now the Life Technologies SOLiD platform).

Advantages of these systems over other cyclic array platforms include: (1) very small feature sizes (1 μm), with a potential to fit more than one billion sequencing features on the surface of a standard microscope slide, and (2) high consensus accuracies, which are particularly important for resequencing applications. Notable disadvantages include: (1) significantly shorter read lengths than dideoxy sequencing or the 454 system, and (2) as with the 454 system, cumbersome emulsion PCR and bead recovery relative to bridge PCR or single-molecule sequencing (see below).

Illumina's sequencing platform (which includes merged technology from Solexa, Lynx, and Manteia SA) generates polony sequencing features by a different method known as "bridge amplification PCR" (Fedurco et al., 2006; *UNIT 25B.9*: Morrissy et al., 2010). In this approach, both forward and reverse PCR primers are immobilized to the two-dimensional surface of a glass slide. The primers are designed to target universal adaptors that flank a complex library of sequencing templates. PCR is performed by standard thermal cycling of the slide, with all reagents present in aqueous phase except for the primers, which are only present in surface-bound form. Because all primers are immobilized, copies remain local, and the result of amplification of each single template molecule is a tight cluster of ∼1000 copies. One species of primer is chemically released

from the slide, such that only one orientation of each amplified template remains.

Cyclic sequencing can be performed by a method distinct from those described above. A universal primer is hybridized to a position immediately adjacent to unknown sequence. At each cycle, polymerase extension is performed with modified dNTPs bearing unique fluorescent groups (identifying the dNTP species) and a reversibly terminating moiety in place of the 3′-hydroxyl position. Because of this terminating group, only a single base extension can occur at each cycle. The array is imaged in four colors to acquire data on all features for a single base position. After cleavage of the terminating moiety (leaving a 3′-hydroxyl group), the next cycle can begin. Read lengths of 35 to 150 bases for over 800 million features per slide have been demonstrated on the Illumina system.

Several groups have released cyclic array platforms for direct sequencing of single molecules, i.e., without any amplification step. These include the Helicos system (Heliscope) based on technology developed in Steven Quake's lab (Braslavsky et al., 2003), in which a library of single DNA molecules is dispersed to an array and sequenced by cyclic extensions with fluorescently labeled nucleotides. A different approach is being taken by Pacific Biosciences based on technology developed in Watt Webb's lab (Levene et al., 2003), which uses an array of zero-mode waveguides for real-time, single-molecule observation of polymerase-driven incorporation of fluorescently labeled nucleotides to a primed template. Zero-mode waveguides are structures that enable the illumination of extremely small (zeptoliter, or 10^{-21} liter, scale) observation volumes. Within such a small volume, fluorescent nucleotide incorporation events driven by DNA polymerase can be readily distinguished from background arising from other fluorescent nucleotides in solution. The current version of the PACBIO RS system delivers about 75,000 reads with an average read length of 1000 bases within 30 to 45 min. However, the reads have, on average, 8% to 10% error rate and, to increase the accuracy of sequencing, multiple passes are needed, which considerably decreases throughput. It is expected that with future releases of new chemistries and instrumentation, raw accuracy will be improved.

Microelectrophoresis

As mentioned above, conventional dideoxy sequencing is performed with microliter-scale reagent volumes, with most instruments running 96 or 384 reactions simultaneously in separate reaction vessels. A goal of microelectrophoretic methods is to make use of microfabrication techniques developed in the semiconductor industry to enable significant miniaturization of conventional dideoxy sequencing (Paegel et al., 2003), for example, by performing gel electrophoresis in nanoliter-scale channels cut into the surface of a silicon wafer (Emrich et al., 2002). A more ambitious goal, toward which much progress has been made, is the integration of a series of sequencing-related steps (e.g., PCR amplification, product purification, and sequencing) in a "lab-on-a-chip" format (Blazej et al., 2006; Forster et al., 2008). A key advantage of this approach is the retention of the dideoxy biochemistry, which has proven robustness for >1000 bases of sequencing.

While alternative methods are achieving significantly longer read lengths, relative to the first-generation automated DNA sequencing technology (*UNIT 7.2*), there will continue to be an important role for Sanger sequencing. Microelectrophoretic methods may prove critical to continuing the trend of reducing costs for this well-proven chemistry. There may also be a key role for "lab-on-a-chip" integrated sequencing devices for cost-effective, clinical "point-of-care" molecular diagnostics.

Nanopore Sequencing

A creative approach to single-molecule sequencing, first proposed in the 1980s, involves passing single-stranded DNA through a nanopore (Deamer and Akeson, 2000). The nanopore itself is a biological membrane protein, e.g., α-hemolysin (Kasianowicz et al., 1996) or MspA, an engineered *Mycobacterium smegmatis* porin A molecule (Derrington et al., 2010) or a synthetic solid-state device (Fologea et al., 2005). Chemical modifications to the nanopore by the covalent attachment of a cyclodextrin molecule to its inside surface acts as a binding site for individual DNA bases and allows accurate measurement of their passage through the nanopore binding site. Two methods are under development, an exonuclease method (Hornblower et al., 2007) and a polymerization method currently using a modified phi29 DNA polymerase (Lieberman et al., 2010). Because each of the four nucleotides obstructs the pore to varying degrees in a base-specific manner (or can potentially be encouraged to do so via base-specific modifications), the resulting fluctuations in electrical

conductance through the pore can, in principle, be measured and used to infer the primary DNA sequence. Published examples of nanopore-based characterization of single nucleic acid molecules include: (1) the measurement of duplex stem length, base-pair mismatches, and loop length within DNA hairpins (Vercoutere et al., 2001), (2) the classification of the terminal base pair of a DNA hairpin, with ~60% to 90% accuracy with a single observation, and >99% accuracy with 15 observations of the same species (Winters-Hilt et al., 2003), and (3) reasonably accurate (93% to 98%) discrimination of deoxynucleotide monophosphates from one another with an engineered protein nanopore sensor (Astier et al., 2006), including 5-methyl cytosine and 5 hydroxymethyl cytosine (Wanunu et al., 2010). It is probable that significant pore engineering and further technology development will be necessary to achieve accurate decoding of a complex mixture of DNA polymers with single-base-pair resolution and useful read lengths. Nanopore sequencing has great potential to enable extraordinarily rapid and cost-effective sequencing of populations of DNA molecules with comparatively simple sample preparation.

CHOOSING A SEQUENCING STRATEGY

Given the extent of flux in the sequencing technology field and the uncertainties surrounding the precise costs and performance parameters for several of the new Next Generation sequencing platforms, it is difficult to state with any certainty which system will be best for any given application. Some of the key parameters that should be considered in comparing technologies to one another include the following:

Cost per raw base. What is the all-inclusive cost (instrument amortization, reagents, labor, etc.) for producing each base pair of sequence?

Cost per consensus base. For example, one high-accuracy raw base call at a given position may be more valuable than several lower accuracy raw base calls.

Raw accuracy. What is the distribution of accuracies with which raw base calls are made? What is the dominant error modality? New technologies are clearly quite a bit behind conventional sequencing with respect to this parameter.

Consensus accuracy. If errors are systematic rather than random, then multiple reads covering a given position (raw base-calls) may not lead to higher consensus accuracies in a straightforward manner.

Read lengths. Although costs are considerably less expensive per base, most of the new platforms for DNA sequencing are currently at a significant disadvantage with respect to read length. Certain applications, such as de novo genome sequencing and assembly, may prove difficult with technologies limited to relatively short read lengths. On the other hand, short read lengths may be sufficient for resequencing (identifying variants in individuals for whom a canonical genome is already defined), as well as for tag counting.

Cost per read. For tag-counting applications such as serial analysis of gene expression (SAGE; *UNIT 25B.6*), read lengths add little information beyond a certain point. The key parameter here is the number of independently sequenced tags, each with sufficient information to identify the transcript from which it was derived. Nevertheless, cost may be an issue and some platforms are less expensive, per read, than others.

Mate-paired (paired-end) reads. The capacity to produce mate-paired reads (pairs of reads that are separated by a known distance distribution on the genome of origin) can be critical to certain applications, such as de novo genome assembly and the detection of structural rearrangements (see Chapter 7 introduction). These also effectively double the number of bases or read length per fragment, with short-read platforms.

Finally, it should be emphasized that the "pre-processing" protocols (e.g., library construction) and "post-processing" pipelines (data analysis) for new sequencing technology platforms are not nearly as mature as for conventional dideoxy sequencing. The development of robust, straightforward protocols for in vitro library construction for various applications (for example, target enrichment; Kothiyal et al., 2009; Igartua et al., 2010) and the creation of bioinformatics tools for interpreting massive amounts of short-read sequencing data are critical challenges that are being addressed to enable investigators to make the most of these exciting new technologies (McPherson, 2009).

PLATFORMS

Resequencing by Hybridization

Affymetrix: http://www.affymetrix.com; Roche/Nimblegen: http://www.nimblegen.com.

In these methods, a series of nucleic acid sequences (probes) are fixed onto solid supports

("chips"). These overlapping sequences represent the sum total of the genomic DNA to be interrogated. By hybridization of labeled fractionated DNA or RNA sequences to the chip, the hybridization patterns reveal the DNA sequence and any nucleotide changes by overlapping information. This method is especially useful for expression analysis and SNP (single nucleotide polymorphism) studies.

Roche/454 Platform

http://www.454.com

The Roche/454 sequencing method is based upon pyrophosphate detection (Nyrén and Lundin, 1985; Hyman, 1988) and further developed by Ronaghi et al. (1996) (see Du and Egholm, 2008; Wiley et al., 2009). In the library construction, two different adaptors are ligated to the ends of DNA fragments. An oil and water emulsion is created in which each water droplet contains one forward primer-coated bead and one adapter-ligated library fragment, as well as the reverse primer and PCR reagents. Thousands of copies of the library DNA fragment are amplified on the surface of the bead. After amplification (emPCR) and breaking the emulsion, the beads with attached DNA are loaded onto PicoTiter Plates (PTP; Roche) for sequencing. The plates allow only one bead per well due to size constraints, and thus each well contains a single clonally amplified set of DNA strands attached to a bead. After addition of sequencing enzymes, the fluidics subsystem of the sequencing instrument sequentially delivers individual nucleotides in a predetermined order across the entire plate with over 1 million wells, each containing one bead. Addition of one (or more) nucleotide(s) complementary to the template strand results in a chemiluminescent signal recorded by a CCD camera within the instrument. The combination of DNA polymerase, luciferase, and other enzymes generates flashes of light while synthesizing the second strand of the DNA on the beads.

The chemistry of the reaction, which gives off light upon nucleotide addition into a growing DNA chain, is based upon pyrophosphate release. When a nucleotide (dNTP) is encountered by the polymerase as the next base to be entered into the growing DNA chain during synthesis, it is incorporated as a dNMP, releasing PPi (pyrophosphate). In the presence of pyrophosphate and APS (adenosine 5′-phosphosulfate), the enzyme ATP sulfurylase converts the APS to ATP. The ATP catalyzes a reaction of luciferin to oxyluciferin in the presence of the luciferase enzyme, a byproduct of which is light. The light emission is recorded by the CCD camera at that fixed position on the PTP. Apyrase is added to the reaction to degrade any remaining ATP and dNTPs in the reaction before restarting the reaction with the next nucleotide. To reduce wash and addition steps, the sulfurylase and luciferase enzymes are linked to a bead deposited in each well along with the template on the emPCR bead.

Software then identifies the location of the beads and correlates the light flashes with each type of nucleotide that was incorporated into the synthesized DNA. The Titanium format of Roche/454 sequencing can generate sequences for about 1 million beads at lengths of about 400 bases. Because this platform produces one of the longest sequence reads of any of the commercially available HTS platforms, in addition to resequencing the genomes of individuals whose DNA is closely related to one whose sequence has already been determined, the Roche/454 platform is optimal for generating de novo whole-genome DNA sequences. It is also used for many other applications, as well, including ChIP-seq, RNA-seq, transcript analysis, and metagenomic projects, to name a few. Read lengths up to 500 nucleotides have been achieved, and 500 Mb of data can be obtained; 1-kb read lengths may be soon possible. Flowcells can be used for one sample, or the flowcell can be subdivided into 2, 4, 8, or 16 separate regions for multiple samples. Another method to increase the number of samples that can be run simultaneously is to use a barcoding protocol that is introduced in the library preparation step.

Illumina Platform

http://www.illumina.com/technology/sequencing_technology.ilmn

Illumina also builds its libraries by ligating adaptors to the ends of DNA fragments, but Illumina clones and amplifies these fragments by randomly capturing the fragments onto a forest of oligonucleotides covalently attached to channels in a flow cell apparatus, which has the dimensions of a microscope slide (Lakdawalla and VanSternhouse, 2008). The flow cell is divided into eight channels; therefore, eight separate samples may be sequenced on one flow cell or all the channels may be filled with the same sample to produce the maximum number of reads per flow cell for one genome. Each captured DNA fragment is amplified in place on the flow cell to yield thousands of copies of each fragment in a cluster on the surface of the flow cell by a process

called "bridge PCR." The result of cluster generation is about 1000 copies of single-stranded short fragments of DNA. The fragments are approximately 200 to 300 bases long and are randomly scattered over the surface of the flow cell. After the clusters have been amplified, a sequencing primer is annealed to the amplified DNA, and the flow cell is transferred to the sequencing machine where the sequencing occurs. The sequencer pumps DNA polymerase and a mixture of the four deoxynucleotides through the channels of the flow cell. Each of the four deoxynucleotides is labeled with a different fluorescent dye. One nucleotide is added to the 3′ end of each primer by the DNA polymerase, and DNA synthesis stops at this point because each deoxynucleotide has a chemical block on the 3′ hydroxyl group. The flow cell is washed, and the clusters on the flow cell are illuminated by lasers and filters while a CCD camera moves along all the channels of the flow cell taking pictures of the clusters. Typically, images are taken and software keeps track of the locations of the randomly placed clusters and the color of the nucleotide for each cluster. The chemical block and the fluorescent moiety are removed, and the next nucleotide is added to the growing DNA chain. Initially, 36 flows were the maximum in the process, so each cluster would yield 36 nucleotides of sequence. The output is commonly 50 million clusters per flow cell, yielding nearly 2 billion bases of sequence per flow cell. More recently, paired end reads of 100 bp or longer (i.e., two 100-bp reads per feature) have been achieved, and the newest HiSeq2000 platform can run two 8-channel slides capable of producing up to 500 Gb of high-quality data. The Illumina platform is very popular because of the simplicity of producing the DNA library and amplifying the clusters. Thus, it is widely used for SNP analysis, ChIP-seq, RNA-seq, and other methods that can exploit a previously sequenced genome (Quail et al., 2009; Morrissy et al., 2010).

ABI/SOLiD Platform

http://www.appliedbiosystems.com/absite/us/en/home/applications-technologies/solid-next-generation-sequencing.html

The third common commercially available DNA sequencing platform is the Life Technologies/ABI SOLiD system (Pandey et al., 2008). In this system the library preparation is very similar to the Roche/454 method in that the SOLiD system also uses beads and emulsion PCR for isolating single DNA fragments from the library and amplification on the beads. The main difference is the size of the beads. The beads in the SOLiD system are much smaller (~1 μm in diameter). Another difference between the bead sequencing methods of Roche and SOLiD is that Roche deposits their beads into the wells of a PicoTiterPlate, while SOLiD deposits their beads randomly onto the surface of a flow cell like the Illumina method. The SOLiD flow cells achieve a greater density of clone-beads than the Illumina or the Roche/454 systems.

A major difference in the sequencing method of the SOLiD system compared to the other platforms is that the SOLiD sequences by ligation whereas Illumina and 454 sequence by DNA polymerase synthesis of DNA. A primer is annealed to one of the adaptors on the bead clone, and a mixture of fluorescent short oligonucleotides is pumped into the flow cell. When the correct probe matches the sequence of the template strand, the fluorescent oligonucleotide is ligated onto the primer. After ligation, the fluorescence reports the first two nucleotides of the probe. Unincorporated oligos are washed away, and a CCD camera captures the different colors attached to the primer. Each fluorescence wavelength corresponds to a particular dinucleotide combination. After image capture, the fluorescent moiety is removed and new oligonucleotides are pumped into the flow cell for DNA ligation.

The most recent improvement to the system is the SOLiD 5500 series, which can yield 90 to 300 Gb of data from about 1 billion reads in a single run (about 7 days) of two full flow cells. The sequence reads average about 50 bases in length. The SOLiD platform is suitable for RNA-seq, ChIP-seq, SNP analysis, and genome resequencing.

Dover/Danaher Polonator

http://www.polonator.org/

The Polonator was developed to perform sequencing by ligation of polony libraries (Shendure et al., 2005; Edwards, 2008). To sequence a genome using this system, library construction occurs using paired-tag shotgun library construction techniques. Then, the library is amplified on beads using emulsion PCR (emPCR). The beads are then "enriched" for those containing amplified DNA. Finally, the beads are added to the flow cell and the sequence of the DNA on these beads is read using the Polonator instrument.

These general steps are shared with protocols from several "NextGen" platforms.

However, Dover, in collaboration with the Church Laboratory of Harvard Medical School, introduced the Polonator G.007 as an open platform with freely downloadable, open-source software and protocols, low-cost, off-the-shelf reagents, and dual, large-area flow cells. Users are able to choose standard protocols, reagents, and software, but are also able to modify protocols, as all aspects of the system enable the flexibility to support a wide range of alternative sequencing methods. The flow cells have been designed to permit sequential biochemistry cycles to be performed upon beads bound to the underside of the coverslip during its arraying protocol, which maximizes the imaging area (number of beads), minimizes reagent consumption, and provides good thermal conductivity to minimize the time required for thermal cycling. Users can elect to perform either a single- or a dual-flow cell sequencing run, depending on the throughput required.

The Polonator acquires 2,180 images per lane in each of four colors for every base that is read. The base caller outputs a separate read file for each lane, with each read file containing ~20 million reads.

Helicos Platform

http://www.helicosbio.com/

The fourth platform for next-generation DNA sequencing is the Heliscope from Helicos Biosciences, although it may not be commercially available in the near future. The Helicos system performs single-molecule DNA sequencing with DNA amplification (Pushkarev et al., 2009; Grabherr et al, 2011; Steinmann et al., 2011; Ozsolak and Milos, 2011a,b; Raz et al, 2011). Millions of single DNA fragments are sequenced independently in a massively parallel manner. Size-fractionated fragments of genomic DNA are denatured and poly(dA) extensions are added to the 3′ ends using terminal deoxynucleotidyl transferase (TDT). A fluorescently labeled ddNTP is also added to the terminus of the poly(dA) tail by the TDT enzyme. This completes the library preparation; there is no amplification and no ligation of adaptors. The single-stranded fragments are applied to a flow cell that has a forest (billions) of oligo(dT) molecules covalently bound to the glass surface. The flow cell captures the individual library fragments by hybridization of the poly(dA) tailed fragments to oligo(dT). The flow cell is assembled into the Heliscope sequencer where fluorescent nucleotides are pumped in one at a time, similar to the 454 system. The fluorescent nucleotides are incorporated into a growing DNA strand by DNA polymerase (SBS). After the image has been captured using a highly sensitive imaging system, the fluorescent moiety is removed, and the next fluorescent nucleotide is incorporated.

The output from the Heliscope is typically 600 million to 800 million reads per run at an average of 35 bases per read. This output is with two flow cells per run, and the flow cells can be divided into 50 independent channels. Typically, about 16 million reads are produced per channel in the flow cell. Library preparation is easiest on the Helicos system, but the image-capture time is up to 8 days. Helicos is suitable for complete genomes, ChIP-seq or SNP analysis, or any project where short reads can be mapped to a reference genome.

Ion Torrent

http://www.iontorrent.com (Life Technologies)

Ion Torrent has commercialized a technology that creates a direct connection between simple chemistry and semiconductor technology (Rothberg et al., 2011). Sequencing features generated by emulsion PCR to microbeads are deposited to an array of microwells. Sequencing by synthesis occurs by sequential addition of unmodified nucleotides, as with the 454 system. However, sequencing occurs by direct electrical detection rather than by the pyrosequencing method. Specifically, hydrogen ions are released as a result of the synthesis reaction, and these hydrogen ions (leading to a pH change) are translated into a voltage signal and subsequently into a base call. The Personal Genome Machine (PGM) sequencer sequentially floods the chip with one nucleotide after another. If the next nucleotide that floods the chip is not a match, no voltage change will be recorded and no base will be called. If there are two identical bases on the DNA strand, the voltage value will double, and the chip will record two identical bases. There are no cameras, light sources, or scanners involved. Instead, a digital voltage change is used for detection.

Oxford Nanopore

http://www.nanoporetech.com

Oxford Nanopore technology (not commercially available as of this writing) uses exonuclease I or a modified phi29 polymerase coupled to a plate in parallel spots, surrounded by a lipid bilayer incorporating MspA or alpha hemolysin (see Branton et al., 2008). These nanopores cover a series of microwells

with electrodes on both sides of the lipid bilayer. In the exonuclease method, when DNA is introduced, the exonuclease progressively cleaves the terminal nucleotide that enters the nanopore. With the polymerization method ("strand sequencing"), as the DNA passes through the pore, each base is recognized. Each nucleotide gives off a characteristic electric current signal distinguishable from the other three in DNA. Thus, the "electrical trace" of a DNA molecule can provide the DNA sequence of that individual molecule. Furthermore, cytosine-methylated bases can also be discriminated.

Pacific Biosciences

http://www.pacificbiosciences.com

Using Single Molecule Real Time (SMRT) DNA sequencing technology, the approach monitors processive DNA synthesis by a DNA polymerase in real time directly reading the DNA (Chin et al., 2011). DNA sequencing is performed in cells, each containing, in parallel, 80,000 zero-mode waveguides (ZMWs). A ZMW is a "hole," tens of nanometers in diameter, fabricated in a 100-nm thick metal film deposited on a silicon substrate. Each ZMW is a small chamber holding 20 zeptoliters (20×10^{-21} liters) where a single DNA polymerase molecule is attached to the bottom surface. Nucleotides, each type labeled with a different colored fluorophore, are then added and, as the polymerase incorporates the correct complementary nucleotide, it is held within the detection volume for tens of milliseconds, orders of magnitude longer than the amount of time it takes a nucleotide to diffuse in and out of the detection volume. During this time, the engaged fluorophore emits fluorescent light whose color corresponds to the base identity and can be detected at the bottom of the well. Then, as part of the natural incorporation cycle, the polymerase cleaves the bond holding the fluorophore in place and the dye diffuses out of the detection volume. Following incorporation, the signal immediately returns to baseline and the process is repeated on the template in the well, enabling reading of a DNA chain.

Intelligent Biosystems

http://www.intelligentbiosystems.com

Intelligent Biosystems (IBS) uses a DNA sequencing by synthesis (SBS) approach, based on technology from Jingyue Ju's lab at the Columbia University Genome Center (Ju et al., 2006). Millions of samples can be processed in parallel on a single chip. DNA is fragmented, amplified, attached to a DNA sequence primer, and then affixed as a high-density array of spots onto a glass chip. The array of fragments is subjected to DNA synthesis reagents including uniquely engineered nucleotides that contain a removable fluorescent dye and an end cap (DNA synthesis blocker). In the synthesis reaction, these nucleotides are added to the end of the growing strand of DNA in accordance with the base on the complementary strand. After the synthesis reaction, the array is scanned by a high-resolution electronic camera, and the fluorescent output of each of four dye colors at each array position is measured and recorded. Knowing which "color" corresponds to which nucleotide provides the knowledge of which base (A, C, G or T) was incorporated in the growing DNA fragment. After recording the "color," the array is exposed to reagents and washes that cleave off the fluorescent dye and end cap, such that no signal is left and the next labeled nucleotide synthesis reagents can be added and the correct nucleotide detected at that position on the chip. The cycles can then be repeated.

Life Technology/Starlight

http://www3.appliedbiosystems.com/cms/groups/global_marketing_group/documents/generaldocuments/cms_091831.pdf

The "Starlight" real-time, single-molecule technology uses an engineered DNA polymerase attached to a "quantum-dot" (or related fluorescent nanoparticle) as a sequencing engine. The quantum dot–DNA polymerase binds to a DNA strand tethered to a microscopy coverslip. A correct dye-labeled γ-phosphate-modified nucleotide binds to the DNA polymerase, FRET (fluorescence resonant energy transfer) between the donor nucleotide and the acceptor quantum dot–DNA polymerase is detected by an electron multiplying charged coupled device (emCCD) detection system. After nucleotide incorporation and detection, fluorescently labeled pyrophosphate (PPi) leaves the complex, generating unaltered DNA. The entire polymerase-nanoparticle then translocates to the next template nucleotide position and the entire process is repeated. DNA sequence is generated at a rate of $\sim$2 to 5 bases/second. The technology can also be used to generate multiple "ordered reads" from long DNAs (up to 1 Mb), as multiple sequencers operate on a single, long, nicked-DNA fragment in a flow cell. This technology also has the other advantages of not needing any cloning or amplification steps, and is based

upon the biologically "normal" DNA polymerization (synthesis) reaction (Hardin, 2008).

Complete Genomics

http://www.completegenomics.com

While the current business model for Complete Genomics is not to commercialize an instrument, but rather provide service-based sequencing of human genomes, it is perhaps worth noting their technology. They prepare libraries of DNA using multiple adaptors (four or more) and each template is clonally amplified resulting in DNA "nanoballs" (DNBs; Drmanac et al., 2010). The DNBs are then laid down on a solid microscope slide–sized surface, one DNB per spot. The approach then uses a "probe-anchor ligation" (PAL) technology wherein, first, an anchor probe is hybridized to the DNA, complementary to the adaptors. Then, a pool of probes, each with the discriminatory base labeled with one of four fluorescently labeled probes, is hybridized to the surface DNA. After ligation to the anchor probe and washing, only a probe with the correct complementary sequence remains bound, and the fluorescence is measured. The cycle is repeated by washing off the entire anchor-probe ligation complex, and the cycle repeated. The current cost for sequencing a human genome is about $20,000, but could drop as low as $5000/genome in the near future. The end user obtains data in the form of differences between genome in question and a reference human genome.

HANDLING NextGen SEQUENCE DATA

The cost of sequencing a nucleotide base is now less than the cost of storing a byte (Stein, 2010). In terms of base pairs per dollar, before the "NextGen sequencing revolution" started, the doubling time was about every 19 months; around 2005, this shifted on a more rapid trajectory to halve the cost about every 6 months. During both of these time periods (1990 to 2010), the cost in terms of megabytes per dollar for hard disc storage space has maintained a somewhat steady increase, doubling every 14 months. In addition, sequencing throughput has been increasing at a rate of five-fold per year, while computer performance has been improving according to Moore's law, doubling every 20 months or so (Moore, 1965). Moore's Law states that the number of transistors that can be placed on an integrated circuit board is increasing exponentially, with a doubling time of roughly 18 months. The trend has held up remarkably well for 35 years across multiple changes in semiconductor technology and manufacturing techniques. Similar laws for disk storage and network capacity have also been observed. Hard disk capacity doubles roughly annually (Kryder's Law; Walter, 2005) and the cost of sending a bit of information over optical networks halves every 9 months (Tehrani, 2000).

A number of programs are being developed for first-order data analysis (alignments, assembly, RNA-seq. ChIP-seq, SNP detection, methylation detection, etc. (Flicek and Birney, 2009; Medvedev et al., 2009; Pepke et al., 2009; Porter et al., 2009). One potential large-scale storage and accessible resource is "cloud computing" (Baker, 2010; Stein, 2010; Schatz et al., 2010). While there are issues of accessibility, interfacing, common formatting, and security, it will likely be necessary to have resources dedicated to, at the minimum, short-term storage of large amounts of data for analysis and dissemination. How data are collected, formatted, and stored, and how to interface different (or similar) applications are major considerations for "NextGen" technology.

ADDITIONAL TECHNOLOGIES ON THE HORIZON

A number of additional technologies for large-scale, parallel sequencing are under development. These include transmission electron microscopy (TEM) sequencing (*http://www.zsgenetics.com* and *http://www.halcyonmolecular.com*; Thomas and Glover, 2008), scanning tunnel microscopy (Tanaka and Kawa, 2009), and optical sequencing (Xiao and Kwok, 2008; Zhou et al., 2008). In TEM sequencing, long DNA pieces ARE arrayed on grids, having incorporated different nucleotides with different nuclear (proton) charges. Detection of the differential charge can be observed and cataloged. Optical sequencing is yet another technology on the horizon. Here, large pieces of DNA arrayed on glass slides can be treated with a limited "nicking enzyme" step, followed by an exonuclease step creating about 20- to 50-bp gaps in the DNA. In another version, once the nicking step has been completed, the DNA can be "stretched" due to its elasticity. The gaps that are created can then be filled in using a polymerase and labeled nucleotides, and imaged with fluorescence microscopy.

LITERATURE CITED

Adams, S. and Blakesley, R. 1991. Linear amplification sequencing. *Focus (BRL)* 13:56-57.

Albert, T.J., Dailidiene, D., Dailide, G., Norton, J.E., Kalia, A., Richmond, T.A., Molla, M., Singh, J., Green, R.D., and Berg, D.E. 2005. Mutation discovery in bacterial genomes: Metronidazole resistance in *Helicobacter pylori*. *Nat. Methods* 2:951-953.

Andries, K., Verhasselt, P., Guillemont, J., Gohlmann, H.W., Neefs, J.M., Winkler, H., Van Gestel, J., Timmerman, P., Zhu, M., Lee, E., Williams, P., de Chaffoy, D., Huitric, E., Hoffner, S., Cambau, E., Truffot-Pernot, C., Lounis, N., and Jarlier, V. 2005. A diarylquinoline drug active on the ATP synthase of *Mycobacterium tuberculosis*. *Science* 307:223-227.

Applied, Biosystems. 1989. Model 370 and 370A Automated Sequencer User Bulletin. Applied Biosystems, Foster City, Calif.

Astier, Y., Braha, O., and Bayley, H. 2006. 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.* 128:1705-1710.

Baker, M. 2010. Next-generation sequencing: Adjusting to data overload. *Nat. Methods* 7:495-499.

Beck, S., O'Keefe, T.O., Coull, J.M., and Koster, H. 1989. Chemiluminescent detection of DNA: Application for DNA sequencing and hybridization. *Nucleic Acids Res.* 17:5115-5123.

Berezikov, E., Thuemmler, F., van Laake, L.W., Kondova, I., Bontrop, R., Cuppen, E., and Plasterk, R.H. 2006. Diversity of microRNAs in human and chimpanzee brain. *Nat. Genet.* 38:1375-1377.

Blazej, R.G., Kumaresan, P., and Mathies, R.A. 2006. Microfabricated bioprocessor for integrated nanoliter-scale Sanger DNA sequencing. *Proc. Natl. Acad. Sci. U.S.A.* 103:7240-7245.

Branton, D., Deamer, D.W., Marziali, A., Bayley, H., Benner, S.A., Butler, T., Di Ventra, M., Garaj, S., Hibbs, A., Huang, X., Jovanovich, S.B., Krstic, P.S., Lindsay, S., Ling, X.S., Mastrangelo, C.H., Meller, A., Oliver, J.S., Pershin, Y.V., Ramsey, J.M., Riehn, R., Soni, G.V., Tabard-Cossa, V., Wanunu, M., Wiggin, M., and Schloss, J.A. 2008. The potential and challenges of nanopore sequencing. *Nat. Biotechnol.* 26:1146-1153.

Braslavsky, I., Hebert, B., Kartalov, E., and Quake, S.R. 2003. Sequence information can be obtained from single DNA molecules. *Proc. Natl. Acad. Sci. U.S.A.* 100:3960-3964.

Carothers, A.M., Urlab, G., Mucha, J., Grunburger, D., and Chasin, L.A. 1989. Point mutation analysis in a mammalian gene: Rapid preparation of total RNA, PCR amplification of cDNA and *Taq* sequencing by a novel method. *BioTechniques* 7:494-499.

Chen, E.Y. and Seeburg, P.H. 1985. Supercoil sequencing: A fast and simple method for sequencing plasmid DNA. *DNA (N.Y.)* 4:165-170.

Chin, C.S., Sorenson, J., Harris, J.B., Robins, W.P., Charles, R.C., Jean-Charles, R.R., Bullard, J., Webster, D.R., Kasarskis, A., Peluso, P., Paxinos, E.E., Yamaichi, Y., Calderwood, S.B., Mekalanos, J.J., Schadt, E.E., and Waldor, M.K. 2011. The origin of the Haitian cholera outbreak strain. *N. Engl. J. Med.* 364:33-42.

Church, G. and Gilbert, W. 1984. Genomic sequencing. *Proc. Natl. Acad. Sci. U.S.A.* 81:1991-1995.

Church, G. and Kiefer-Higgins, S. 1988. Multiplex DNA sequencing. *Science* 240:185-188.

Craxton, M. 1993. Cosmid sequencing. *Methods Mol. Biol.* 23:149-167.

Creasey, A., D'Angio, L.M., Dunne, T., Kissinger, C., O'Keefe, T., Perry-O'Keefe, H., Moran, L., Roskey, M., Shildkraut, I., Sears, L., and Slatko, B. 1991. Application of a novel chemiluminescent-based DNA detection method to single-vector and multiplex DNA sequencing. *BioTechniques* 11:102-109.

Deamer, D.W. and Akeson, M. 2000. Nanopores and nucleic acids: Prospects for ultrarapid sequencing. *Trends Biotechnol.* 18:147-151.

Derrington, I.M., Butler, T.Z., Collins, M.D., Manrao, E., Pavlenok, M., Niederweis, M., and Gundlach, J.H. 2010. Nanopore DNA sequencing with MspA. *Proc. Natl. Acad. Sci. U.S.A.* 107:16060-16065.

Diehl, F., Li, M., Dressman, D., He, Y., Shen, D., Szabo, S., Diaz, L.A. Jr., Goodman, S.N., David, K.A., Juhl, H., Kinzler, K.W., and Vogelstein, B. 2005. Detection and quantification of mutations in the plasma of patients with colorectal tumors. *Proc. Natl. Acad. Sci. U.S.A.* 102:16368-16373.

Diehl, F., Li, M., He, Y., Kinzler, K.W., Vogelstein, B., and Dressman, D. 2006. BEAMing: Single molecule PCR on microparticles in water-in-oil emulsions. *Nat. Methods* 3:551-559.

Dressman, D., Yan, H., Traverso, G., Kinzler, K.W., and Vogelstein, B. 2003. Transforming single DNA molecules into fluorescent magnetic particles for detection and enumeration of genetic variations. *Proc. Natl. Acad. Sci. U.S.A.* 100:8817-8822.

Drmanac, R., Sparks, A.B., Callow, M.J., Halpern, A.L., Burns, N.L., Kermani, B.G., Carnevali, P., Nazarenko, I., Nilsen, G.B., Yeung, G., Dahl, F., Fernandez, A., Staker, B., Pant, K.P., Baccash, J., Borcherding, A.P., Brownley, A., Cedeno, R., Chen, L., Chernikoff, D., Cheung, A., Chirita, R., Curson, B., Ebert, J.C., Hacker, C.R., Hartlage, R., Hauser, B., Huang, S., Jiang, Y., Karpinchyk, V., Koenig, M., Kong, C., Landers, T., Le, C., Liu, J., McBride, C.E., Morenzoni, M., Morey, R.E., Mutch, K., Perazich, H., Perry, K., Peters, B.A., Peterson, J., Pethiyagoda, C.L., Pothuraju, K., Richter, C., Rosenbaum, A.M., Roy, S., Shafto, J., Sharanhovich, U., Shannon, K.W., Sheppy, C.G., Sun, M., Thakuria, J.V., Tran, A., Vu, D., Zaranek, A.W., Wu, X., Drmanac, S., Oliphant, A.R., Banyai, W.C., Martin, B., Ballinger, D.G., Church, G.M., and Reid, C.A. 2010. Complete genomics: Human genome sequencing using unchained base reads on self-assembling DNA nanoarrays. *Science* 327:78-81.

Du, L. and Egholm, M. 2008. The next generation genome sequencing: 454/Roche GS FLX. *In* Next-generation Genome Sequencing and Analysis: Towards Personalized Medicine, (M. Janitz, ed). pp. 43-56. Wiley-Blackwell, Malden, Massachusetts.

Edwards, J. 2008. Polony sequencing: history, technology and applications. *In* Next-generation Genome Sequencing and Analysis: Towards Personalized Medicine, (M. Janitz, ed). pp. 57-76. Wiley-Blackwell, Malden, Massachusetts.

Edwards, J.R., Ruparel, H., and Ju, J. 2005. Mass spectrometry DNA sequencing. *Mutat. Res.* 573:3-12.

Edwards, R.A., Rodriguez-Brito, B., Wegley, L., Haynes, M., Breitbart, M., Peterson, D.M., Saar, M.O., Alexander, S., Alexander, E.C. Jr., and Rohwer, F. 2006. Using pyrosequencing to shed light on deep mine microbial ecology. *BMC Genomics* 7:57.

Emrich, C.A., Tian, H., Medintz, I.L., and Mathies, R.A. 2002. Microfabricated 384-lane capillary array electrophoresis bioanalyzer for ultra high-throughput genetic analysis. *Anal. Chem.* 74:5076-5083.

Evans, S. 1991. Millipore's system speeds up DNA sequencing and eliminates radioactivity. *Genet. Eng. News* 14:29-41.

Fedurco, M., Romieu, A., Williams, S., Lawrence, I., and Turcatti, G. 2006. BTA, a novel reagent for DNA attachment on glass and efficient generation of solid-phase amplified DNA colonies. *Nucleic Acids Res.* 34:e22.

Flicek, P. and Birney, E. 2009. Sense from sequence reads: Methods for alignment and assembly. *Nat. Methods* 6:S6-S12.

Fologea, D., Gershow, M., Ledden, B., McNabb, D.S., Golovchenko, J.A., and Li, J. 2005. Detecting single stranded DNA with a solid state nanopore. *Nano. Lett.* 5:1905-1909.

Forster, R., Fredlake, C. and Barron, A. 2008. Microchip-based Sanger sequencing of DNA. *In* Next-generation Genome Sequencing and Analysis: Towards Personalized Medicine, (M. Janitz, ed). pp. 153-163. Wiley-Blackwell, Malden, Massachusetts.

Goldberg, S.M., Johnson, J., Busam, D., Feldblyum, T., Ferriera, S., Friedman, R., Halpern, A., Khouri, H., Kravitz, S.A., Lauro, F.M., Li, K., Rogers, Y.H., Strausberg, R., Sutton, G., Tallon, L., Thomas, T., Venter, E., Frazier, M., and Venter, J.C. 2006. A Sanger/pyrosequencing hybrid approach for the generation of high-quality draft assemblies of marine microbial genomes. *Proc. Natl. Acad. Sci. U.S.A.* 103:11240-11245.

Grabherr, M.G., Haas, B.J., Yassour, M., Levin, J.Z., Thompson, D.A., Amit, I., Adiconis, X., Fan, L., Raychowdhury, R., Zeng, Q., Chen, Z., Mauceli, E., Hacohen, N., Gnirke, A., Rhind, N., di Palma, F., Birren, B.W., Nusbaum, C., Lindblad-Toh, K., Friedman, N., and Regev, A. 2011. Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat. Biotechnol.* 29:644-652.

Gresham, D., Ruderfer, D.M., Pratt, S.C., Schacherer, J., Dunham, M.J., Botstein, D., and Kruglyak, L. 2006. Genome-wide detection of polymorphisms at nucleotide resolution with a single DNA microarray. *Science* 311:1932-1936.

Haltiner, M., Kempe, T., and Tjian, R. 1985. A novel strategy for constructing clustered point mutations. *Nucleic Acids Res.* 13:1015-1025.

Hardin, S. 2008. Real-time DNA sequencing. *In* Next-generation Genome Sequencing and Analysis: Towards Personalized Medicine, (M. Janitz, ed). pp. 97-101. Wiley-Blackwell, Malden, Massachusetts.

Hattori, M. and Sakaki, Y. 1986. Dideoxy DNA sequencing method using denatured plasmid templates. *Anal. Biochem.* 152:232-238.

Heiner, C.R., Hunkapiller, K.L., Chen, S.M., Glass, J.I., and Chen, E.Y. 1998. Sequencing multimegabase-template DNA with BigDye terminator chemistry. *Genome Res.* 8:557-561.

Herring, C.D., Raghunathan, A., Honisch, C., Patel, T., Applebee, M.K., Joyce, A.R., Albert, T.J., Blattner, F.R., van den Boom, D., Cantor, C.R., and Palsson, B.Ø. 2006. Comparative genome sequencing of *Escherichia coli* allows observation of bacterial evolution on a laboratory timescale. *Nat. Genet.* 38:1406-1412.

Hoffman, L. and Jendrisak, J. 1999. Transposon-based strategies for efficient DNA sequencing and functional Genomics. *Epicentre Forum* 6:1-4.

Hornblower, B., Coombs, A., Whitaker, R.D., Kolomeisky, A., Picone, S.J., Meller, A., and Akeson, M. 2007. Single-molecule analysis of DNA-protein complexes using nanopores. *Nat. Methods* 4:315-317.

Hultman, T., Stahl, S., Hornes, E., and Uhlen, M. 1989. Direct solid phase sequencing of genomic and plasmid DNA using magnetic beads as solid support. *Nucleic Acids Res.* 17:4937-4946.

Hultman, T., Bergh, S., Moks, T., and Uhlen, M. 1991. Bidirectional solid phase sequencing of in vitro amplified plasmid DNA. *BioTechniques* 10:84-93.

Hyman, E.D. 1988. A new method of sequencing DNA. *Anal. Biochem.* 174:423-436.

Igartua, C., Turner, E.H., Ng, S.B., Hodges, E., Hannon, G.J., Bhattacharjee, A., Rieder, M.J., Nickerson, D.A., and Shendure, J. 2010. Targeted enrichment of specific regions in the human genome by array hybridization. *Curr. Protoc. Hum. Genet.* 66:18.3.1-18.3.14

Jones, D.S., Schofield, J.P., and Vaudin, M. 1991. Fluorescent and radioactive solid phase dideoxy sequencing of PCR products in microtitre plates. *J. DNA Seq. Map.* 1:279-283.

Ju, J. 1999. Nucleic acid sequencing with solid phase capturable terminators. U.S. Patent number. 5,876,936.

Ju, J., Ruan, C., Fuller, C.W., Glazer, A.N., and Mathies, R.A. 1995. Energy transfer fluorescent dye-labeled primers for DNA sequencing and

analysis. *Proc. Natl. Acad. Sci. U.S.A.* 92:4347-4351.

Ju, J., Glazer, A.N., and Mathies, R.A. 1996. Energy transfer primers: A new fluorescence labeling paradigm for DNA sequencing and analysis. *Nature Med.* 2:246-249.

Ju, J., Zaro, M., Doctorelo, M., Goralski, T., Konrad, K., Lachenmeier, E., and Cathcart, R. 1997. DNA sequencing with solid phase capturable terminators. *Microb. Comp. Genomics* 2:223.

Ju, J., Kim, D., Bi, L., Meng, Q., Bai, X, Li, Z., Li, X., marma, M., Shi, S., Wu, J., Edwards, J., Romu, A., and Turro, N. 2006. Four-color DNA sequencing by synthesis using cleavable fluorescent nucleotide reversible terminators. *Proc. Natl. Acad Sci. U.S.A.* 103:19635-19640.

Kaneoka, H., Lee, D.R., Hsu, K.-C., Sharp, G.C., and Hoffman, R.W. 1991. Solid phase DNA sequencing of allele specific polymerase chain reaction amplified HLA-DR genes. *BioTechniques* 10:30-40.

Kasianowicz, J.J., Brandin, E., Branton, D., and Deamer, D.W. 1996. Characterization of individual polynucleotide molecules using a membrane channel. *Proc. Natl. Acad. Sci. U.S.A.* 93:13770-13773.

Kothiyal, P., Cox, S., Ebert, J., Aronow, B., Greinwald, J., and Rehm, H. 2009. An overview of custom array sequencing. *Curr. Protoc. Hum. Genet.* 61:7.17.1-7.17.11.

Krishnan, B.R., Blakesley, R.W., and Berg, D.E. 1991. Linear amplification DNA sequencing directly from single phage plaques and bacterial colonies. *Nucleic Acids. Res.* 19:1153.

Lakdawalla, A. and VanSternhouse, H. 2008. Illumina genome analyzer II system. *In* Next-generation Genome Sequencing and Analysis: Towards Personalized Medicine, (M. Janitz, ed). pp. 13-28. Wiley-Blackwell, Malden, Massachusetts.

Lee, L.G., Spurgeon, S.L., Heiner, C.R., Benson, S.C., Rosenblum, B.B., Menchen, S.M., Graham, R.J., Constantinescu, A., Upadhya, K.G., and Cassel, J.M. 1997. New energy transfer dyes for DNA sequencing. *Nucleic Acids Res.* 25:2816-2822.

Levene, M.J., Korlach, J., Turner, S.W., Foquet, M., Craighead, H.G., and Webb, W.W. 2003. Zeromode waveguides for single-molecule analysis at high concentrations. *Science* 299:682-686.

Li, M., Diehl, F., Dressman, D., Vogelstein, B., and Kinzler, K.W. 2006. BEAMing up for detection and quantification of rare sequence variants. *Nat. Methods* 3:95-97.

Lieberman, K., Cherf, G., Doody, M., Olasagasti, F., Kolodji, Y., and Akeson, M. 2010. Processive replication of single DNA molecules in a nanopore catalyzed by phi29 DNA polymerase. *J. Am. Chem. Soc.* 132:17961-17972.

Margulies, M., Egholm, M., Altman, W.E., Attiya, S., Bader, J.S., Bemben, L.A., Berka, J., Braverman, M.S., Chen, Y.J., Chen, Z., Dewell, S.B., Du, L., Fierro, J.M., Gomes, X.V., Godwin, B.C., He, W., Helgesen, S., Ho, C.H., Irzyk, G.P., Jando, S.C., Alenquer, M.L., Jarvie, T.P., Jirage, K.B., Kim, J.B., Knight, J.R., Lanza, J.R., Leamon, J.H., Lefkowitz, S.M., Lei, M., Li, J., Lohman, K.L., Lu, H., Makhijani, V.B., McDade, K.E., McKenna, M.P., Myers, E.W., Nickerson, E., Nobile, J.R., Plant, R., Puc, B.P., Ronan, M.T., Roth, G.T., Sarkis, G.J., Simons, J.F., Simpson, J.W., Srinivasan, M., Tartaro, K.R., Tomasz, A., Vogt, K.A., Volkmer, G.A., Wang, S.H., Wang, Y., Weiner, M.P., Yu, P., Begley, R.F., and Rothberg, J.M. 2005. Genome sequencing in microfabricated high-density picolitre reactors. *Nature* 437:376-380.

Marra, M., Weinstock, L.A., and Mardis, E.R. 1996. End sequence determination from large insert clones using energy transfer fluorescent primers. *Genome Res.* 6:1118-1122.

Martin, C., Bresnick, L., Juo, R.-R., Voyta, J.C., and Bronstein, I. 1991. Improved chemiluminescence DNA sequencing. *BioTechniques* 11:110-113.

McPherson, J. 2009. Next-generation gap. *Nat Methods* 6:S2-S5.

Medvedev, P., Stanciu, M., and Brudno, M. 2009. Computational methods for discovering structural variation with next-generation sequencing. *Nat. Methods* 6:S13-S20.

Metzker, M.L. 2010. Sequencing technologies: The next generation. *Nat. Rev. Genet.* 11:31-46.

Mitra, R.D. and Church, G.M. 1999. In situ localized amplification and contact replication of many individual DNA molecules. *Nucleic Acids Res.* 27:e34.

Mitra, R.D., Shendure, J., Olejnik, J., Edyta, K.O., and Church, G.M. 2003. Fluorescent in situ sequencing on polymerase colonies. *Anal. Biochem.* 320:55-65.

Moore, G.E. 1965. Cramming more components onto integrated circuits. *Electronics* 38:4-7.

Morrissy, S., Zhao, Y., Delaney, A., Asano, J., Dhalla, N., Li, I., McDonald, H., Pandoh, P., Prabhu, A-L., Tam, A., and Hirst, M.M. 2010. Digital gene expression by Tag sequencing on the Illumina Genome Analyzer. *Curr. Protoc. Hum. Genet.* 65:11.11.1-11.11.36.

Murray, V. 1989. Improved double-stranded DNA sequencing using the linear polymerase chain reaction. *Nucleic Acids Res.* 17:88-89.

Nyrén, P. and Lundin, A. 1985. Enzymatic method for continuous monitoring of inorganic pyrophosphate synthesis. *Anal. Biochem.* 151:504-509.

Ozsolak, F. and Milos, P.M. 2011a. RNA sequencing: Advances, challenges and opportunities. *Nat. Rev. Genet.* 12:87-98.

Ozsolak, F. and Milos, P.M. 2011b. Transcriptome profiling using single-molecule direct RNA sequencing. *Methods Mol. Biol.* 733:51-61.

Paegel, B.M., Blazej, R.G., and Mathies, R.A. 2003. Microfluidic devices for DNA sequencing: Sample preparation and electrophoretic analysis. *Curr. Opin. Biotechnol.* 14:42-50.

Pandey, V., Nutter, R., and Prediger, E. 2008. Applied biosystems SOLiD™ system: Ligation-based sequencing. *In* Next-generation Genome Sequencing and Analysis: Towards Personalized Medicine, (M. Janitz, ed). pp. 20-42. Wiley-Blackwell Press, Malden, Massachusetts.

Patil, N., Berno, A.J., Hinds, D.A., Barrett, W.A., Doshi, J.M., Hacker, C.R., Kautzer, C.R., Lee, D.H., Marjoribanks, C., McDonough, D.P., Nguyen, B.T., Norris, M.C., Sheehan, J.B., Shen, N., Stern, D., Stokowski, R.P., Thomas, D.J., Trulson, M.O., Vyas, K.R., Frazer, K.A., Fodor, S.P., and Cox, D.R. 2001. Blocks of limited haplotype diversity revealed by high resolution scanning of human chromosome 21. *Science* 294:1719-1723.

Pepke, S., Wold, B., and Mortazavi, A. 2009. Computation for ChIP-seq and RNA-seq studies. *Nat. Methods*. 6:S22-S32.

Porter, S., Olson, N.E., and Smith, T. 2009. Analyzing gene expression data from microarray and next-generation DNA sequencing transcriptome profiling assays using GeneSifter Analysis Edition. *Curr. Protoc. Bioinformatics* 27:7.14.1-7.14.35.

Pushkarev, D., Neff, N.F., and Quake, S.R. 2009. Single-molecule sequencing of an individual human genome. *Nat. Biotechnol.* 27:847-850.

Quail, M., Swerdlow, H., and Turner, D. 2009. Improved protocols for the Illumina Genome Analyzer sequencing system. *Curr. Protoc. Hum. Genet*. 62:18.2.1-18.2.27.

Ragoussis, J., Elvidge, G.P., Kaur, K., and Colella, S. 2006. Matrix-assisted laser desorption/ionisation, time-of-flight mass spectrometry in genomics research. *PLoS Genet*. 2:e100.

Raz, T., Causey, M., Jones, D.R., Kieu, A., Letovsky, S., Lipson, D., Thayer, E., Thompson, J.F., Milos, P.M. 2011. RNA sequencing and quantitation using the Helicos Genetic Analysis System. *Methods Mol. Biol.* 733:37-49.

Ronaghi, M., Karamohamed, S., Pettersson, B., Uhlén, M., and Nyrén, P. 1996. Real-time DNA sequencing using detection of pyrophosphate release. *Anal. Biochem.* 242:84-89.

Rothberg, J.M., Hinz, W., Rearick, T.M., Schultz, J., Mileski, W., Davey, M., Leamon, J.H., Johnson, K., Milgrew, M.J., Edwards, M., Hoon, J., Simons, J.F., Marran, D., Myers, J.W., Davidson, J.F., Branting, A., Nobile, J.R., Puc, B.P., Light, D., Clark, T.A., Huber, M., Branciforte, J.T., Stoner, I.B., Cawley, S.E., Lyons, M., Fu, Y., Homer, N., Sedova, M., Miao, X., Reed, B., Sabina, J., Feierstein, E., Schorn, M., Alanjary, M., Dimalanta, E., Dressman, D., Kasinskas, R., Sokolsky, T., Fidanza, J.A., Namsaraev, E., McKernan, K.J., Williams, A., Roth, G.T., and Bustillo, J. 2011. An integrated semiconductor device enabling non-optical genome sequencing. *Nature* 47:348-352

Ruby, J.G., Jan, C., Player, C., Axtell, M.J., Lee, W., Nusbaum, C., Ge, H., and Bartel, D.P. 2006. Large-scale sequencing reveals 21U-RNAs and additional microRNAs and endogenous siRNAs in *C. elegans*. *Cell* 127:1193-1207.

Sanger, F. 1988. Sequences, sequences, and sequences. *Annu. Rev. Biochem*. 57:1-28.

Sanger, F., Nicklen, S., and Coulson, A.R. 1977a. DNA sequencing with chain-terminating inhibitors. *Proc. Natl. Acad. Sci. U.S.A.* 74:5463-5467.

Sanger, F., Air, G.M., Barrell, B.G., Brown, N.L., Coulson, A.R., Fiddes, C.A., Hutchison, C.A., Slocombe, P.M., and Smith, M. 1977b. Nucleotide sequence of bacteriophage phi X174 DNA. *Nature* 265:687-695.

Schatz, M.C., Langmead, B., and Salzberg, S.L. 2010. Cloud computing and the DNA data race. *Nat. Biotechnol*. 28:691-693.

Sears, L., Moran, L., Kissinger, C., Creasey, T., Perry-O'Keefe, H., Roskey, M., Sutherland, E., and Slatko, B. 1992. CircumVent™ thermal cycle sequencing and alternative manual and automated DNA sequencing protocols using the highly thermostable VentR (exo−) DNA polymerase. *BioTechniques* 13:626-633.

Shendure, J. and Ji, H. 2008. Next-generation DNA sequencing. *Nat. Biotechnol.* 10:1135-1145.

Shendure, J., Mitra, R.D., Varma, C., and Church, G.M. 2004. Advanced sequencing technologies: Methods and goals. *Nat. Rev. Genet*. 5:335-344.

Shendure, J., Porreca, G.J., Reppas, N.B., Lin, X., McCutcheon, J.P., Rosenbaum, A.M., Wang, M.D., Zhang, K., Mitra, R.D., and Church, G.M. 2005. Accurate multiplex polony sequencing of an evolved bacterial genome. *Science* 309:1728-1732.

Slatko, B. 1996. Thermal cycle dideoxy DNA sequencing. *Mol. Biotechnol*. 6:311-322.

Smith, M.G., Gianoulis, T.A., Pukatzki, S., Mekalanos, J.J., Ornston, L.N., Gerstein, M., and Snyder, M. 2007. New insights into *Acinetobacter baumannii* pathogenesis revealed by high density pyrosequencing and transposon mutagenesis. *Genes. Dev.* 21:601-614.

Stein, LD. 2010. The case for cloud computing in genome informatics. *Genome Biol*. 11:207.

Steinmann, K.E., Hart, C.E., Thompson, J.F., and Milos, P.M. 2011. Helicos single-molecule sequencing of bacterial genomes. *Methods Mol. Biol*. 733:3-24.

Strathmann, M., Hamilton, B., Mayeda, C., Simon, M., Meyerowitz, E., and Palazzolo, M. 1991. Transposon-facilitated DNA sequencing. *Proc. National Acad. Sci. U.S.A.* 88:1247-1250.

Tabor, S. and Richardson, C.C. 1987a. Selective oxidation of the exonuclease domain of bacteriophage T7 DNA polymerase. *J. Biol. Chem.* 262:15330-15333.

Tabor, S. and Richardson, C.C. 1987b. DNA sequence analysis with a modified bacteriophage T7 DNA polymerase. *Proc. Natl. Acad. Sci. U.S.A.* 84:4767-4771.

Tabor, S. and Richardson, C.C. 1989a. Selective inactivation of the exonuclease activity of bacteriophage T7 DNA polymerase by in vitro mutagenesis. *J. Biol. Chem.* 264:6447-6458.

Tabor, S. and Richardson, C.C. 1989b. Effect of manganese ions on the incorporation of dideoxynucleotides by bacteriophage T7 DNA polymerase and *Escherichia coli* DNA polymerase I. *Proc. Natl. Acad. Sci. U.S.A.* 86:4076-4080.

Tabor, S. and Richardson, C.C. 1990. DNA sequence analysis with a modified bacteriophage T7 DNA polymerase: Effect of pyrophosphorolysis and metal ions. *J. Biol. Chem.* 265:8322-8328.

Tabor, S., Huber, H., and Richardson, C.C. 1987. *Escherichia coli* thioredoxin confers processivity of the DNA polymerase activity of the gene 5 protein of bacteriophage T7. *J. Biol. Chem.* 262:16212-16223.

Tanaka, H. and Kawa, T. 2009. Partial sequencing of a single DNA molecule with a scanning tunnelling microscope. *Nat. Nanotechnol.* 4:518-522.

Tawfik, D.S. and Griffiths, A.D. 1998. Man-made cell-like compartments for molecular evolution. *Nat. Biotechnol.* 16:652-656.

Tehrani, R. 2000. As we may communicate. TMCNet, *http://www.tmcnet.com/articles/comsol/0100/0100pubout.htm*.

Thomas, W. and Glover, W. 2008. Direct sequencing by TEM of z-substituted DNA molecule. *In* Next-generation Genome Sequencing and Analysis: Towards Personalized Medicine (M. Janitz, ed.) pp. 103-116. Wiley-Blackwell, Malden, Massachusetts.

Tizard, R., Cate, R.L., Ramachandran, K.L., Wysek, M., Voyta, J.C., Murphy, O.J., and Bronstein, I. 1990. Imaging of DNA sequences with chemiluminescence. *Proc. Natl. Acad. Sci. U.S.A.* 87:4514-4518.

Valdar, W., Solberg, L.C., Gauguier, D., Burnett, S., Klenerman, P., Cookson, W.O., Taylor, M.S., Rawlins, J.N., Mott, R., and Flint, J. 2006. Genome-wide genetic association of complex traits in heterogeneous stock mice. *Nat. Genet.* 38:879-887.

Velicer, G.J., Raddatz, G., Keller, H., Deiss, S., Lanz, C., Dinkelacker, I., and Schuster, S.C. 2006. Comprehensive mutation identification in an evolved bacterial cooperator and its cheating ancestor. *Proc. Natl. Acad. Sci. U.S.A.* 103:8107-8112.

Vercoutere, W., Winters-Hilt, S., Olsen, H., Deamer, D., Haussler, D., and Akeson, M. 2001. Rapid discrimination among individual DNA hairpin molecules at single-nucleotide resolution using an ion channel. *Nat. Biotechnol.* 19:248-252.

Walter, C. 2005. Kryder's Law. *Sci. Am.* 293:32-33.

Wanunu, M., Cohen-Karni, D., Johnson, R., Fields, F., Benner, J., Peterman, N., Zheng, Y., Klein, M., and Drndic, M. 2010. Discrimination of methylcytosine from hydroxymethylcytosine in individual DNA Molecule *J. Am. Chem. Soc.* 133:486-492.

Wiley, G., Macmil, S., Qu, C., Wang, P., Xing, Y., White, D., Li, J., White, J.D., Domingo, A., and Roe, B.A. 2009. Methods for generating shotgun and mixed shotgun/paired-end libraries for the 454 DNA sequencer. *Curr. Protoc. Hum. Genet.* 61:18.1.1-18.1.21.

Winters-Hilt, S., Vercoutere, W., DeGuzman, V.S., Deamer, D., Akeson, M., and Haussler, D. 2003. Highly accurate classification of Watson-Crick basepairs on termini of single DNA molecules. *Biophys. J.* 84:967-976.

Xiao, M. and Kwok, P-Y. 2008. A single DNA molecule barcoding method with applications in DNA mapping and molecular haplotyping. *In:* Next-Generation Genome Sequencing and Analysis: Towards Personalized Medicine (M. Janitz, ed.) pp. 117-132. Wiley-Blackwell, Malden, Massachusetts.

Young, A. and Blakesley, R. 1991. Sequencing plasmids from single colonies with the ds-DNA cycle sequencing system. *Focus (BRL)* 13: 137.

Zagursky, R., Baumeister, K., Lomax, N., and Berman, M. 1985. Rapid and easy sequencing of large double-stranded DNA and supercoiled plasmid DNA. *Gene Anal. Tech.* 2:89-94.

Zagursky, R.J., Conway, P.S., and Kashdan, M.A. 1991. Use of ^{33}P for Sanger DNA sequencing. *BioTechniques* 11:36-38.

Zimmerman, J., Dietrich, T., Voss, H., Erfle, H., Schwager, C., Stegemann, J., Hewitt, N., and Ansorge, W. 1992. Fully automated Sanger sequencing protocol for double-stranded DNA. *Methods Mol. Cell Biol.* 3:39-42.