Overview of DNA Sequencing Strategies

Jay A. Shendure,[1] Gregory J. Porreca,[2] and George M. Church[2]

[1]University of Washington, Seattle, Washington
[2]Harvard Medical School, Boston, Massachusetts

ABSTRACT

Efficient and cost-effective DNA sequencing technologies have been, and may continue to be, critical to the progress of molecular biology. This overview of DNA sequencing strategies provides a high-level review of six distinct approaches to DNA sequencing: (a) dideoxy sequencing; (b) cyclic array sequencing; (c) sequencing-by-hybridization; (d) microelectrophoresis; (e) mass spectrometry; and (f) nanopore sequencing. The primary focus is on 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 of this particular moment. *Curr. Protoc. Mol. Biol.* 81:7.1.1-7.1.11. © 2008 by John Wiley & Sons, Inc.

Keywords: DNA • sequencing • Sanger • dideoxy • polony

INTRODUCTION

In the mid-1960s, the first attempts at DNA sequencing followed the precedent set for protein (Ryle et al., 1955) and RNA (Holley et al., 1965): sequencing by detailed analysis of degradation products. However, the length and consequent complexity of the DNA polymer proved to be significantly problematic (Sanger, 1988). A key moment came in February, 1977, when groups led by Fred Sanger and Walter Gilbert independently published descriptions of methodologies for DNA sequencing, both of which relied on gel electrophoresis to separate DNA fragments with single-base-pair resolution (Maxam and Gilbert, 1977; Sanger et al., 1977). In the years that followed, the rapid dissemination of these technologies and their progression to robust protocols enabled a wide range of critical advances throughout the fields of genetics and molecular biology. The development of commercially available automated sequencing platforms in the mid-1980s represented a second key breakthrough that secured the dominance of the Sanger protocol (also known as "dideoxy sequencing") over the Maxam-Gilbert protocol (also known as "chemical sequencing") as the method of choice for the next several decades (Hunkapiller et al., 1991).

In addition to automation, a supporting cast of related technologies was developed to further reduce costs and improve sequencing throughput. These included a broad range of methods for efficient library construction and template preparation, dideoxynucleotides (ddNTPs) bearing fluorescent moieties (Prober et al., 1987), and thermostable polymerases engineered to accept them (Tabor and Richardson, 1995), as well as the implementation of efficient DNA sequence production workflows in core facilities and high-throughput sequencing centers. It is notable that much of this innovation was motivated by the Human Genome Project (HGP), which achieved completion of a draft of the canonical human genome sequence in 2001 (International Human Genome Sequencing Consortium, 2001). Consequent to the technological innovation that enabled the HGP, the per-base cost of dideoxy sequencing has followed an exponential decline (Collins et al., 2003; Shendure et al., 2004). Importantly, the read lengths and accuracy of sequencing traces have steadily improved as well. As community-wide capacity for high-throughput DNA sequence production has been maintained in the wake of the HGP, the number of sequenced nucleotides deposited in GenBank has continued its exponential rise. As of October 2007, genome sequences for 997 bacterial species and 164 eukaryotic species are available in at least draft assembly form.

In recent years, there has been a collective sense in the technology development field that optimization of dideoxy sequencing protocols may be approaching exhaustion, and that the trend of declining sequencing costs is unlikely to continue much further without a radical change in the underlying technology. This has sparked significant academic

Current Protocols in Molecular Biology 7.1.1-7.1.11, January 2008
Published online January 2008 in Wiley Interscience (www.interscience.wiley.com).
DOI: 10.1002/0471142727.mb0701s81

and commercial investment in alternative technological paradigms (Shendure et al., 2004). Several of these alternatives have quickly progressed to substantial proof-of-concept, demonstrating costs competitive with conventional dideoxy sequencing for certain applications (Margulies et al., 2005; Shendure et al., 2005). Some of these platforms have recently become, or are anticipated to become, widely available in an "open-source" format or as commercial products. Although dideoxy sequencing still accounts for the vast majority of DNA sequencing production, this is unlikely to be the case several years from now.

This unit provides a high-level overview of six distinct approaches to DNA sequencing. These are: (1) dideoxy sequencing, (2) cyclic array sequencing, (3) sequencing by hybridization, (4) microelectrophoresis, (5) mass spectrometry, and (6) nanopore sequencing. 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 represents a snapshot of this particular moment.

It is worthwhile to note that the research goals that motivate DNA sequencing may be undergoing a substantial shift as well, concurrent with the introduction of new technologies. Given that reference genome sequences for *H. sapiens* as well as all major model organisms are nearly complete, demand will likely shift away from de novo genome sequencing towards other areas of application, such as *resequencing* (identifying genetic variation in the genome of an individual for whose species a reference genome is already available) and *tag counting* (i.e., serial analysis of gene expression or chromatin occupancy by the sequencing of short but identifying DNA tags). The initial generation of new technologies will deliver sequence that is substantially shorter and less accurate than state-of-the-art Sanger sequencing. However, although the utility of such sequence may be limited for de novo sequencing, it will likely be compatible, and often preferable, for other areas of application.

DNA SEQUENCING STRATEGIES

Dideoxy Sequencing

Dideoxy sequencing, also known as Sanger sequencing, proceeds by primer-initiated, polymerase-driven synthesis of DNA strands complementary to the template whose sequence is to be determined (Fig. 7.1.1). Numerous identical copies of the sequencing template undergo the primer extension reaction within a single microliter-scale volume. Generating sufficient quantities of template for a sequencing reaction is typically achieved by either (1) miniprep of a plasmid vector into which the fragment of interest has been cloned, or (2) polymerase chain reaction (PCR) followed by a cleanup step. In the sequencing reaction itself, both the natural deoxynucleotides (dNTPs) and the chain-terminating dideoxynucleotides (ddNTPs) are present at a specific ratio that determines their relative probability of incorporation during the primer extension. Incorporation of a ddNTP instead of a dNTP results in termination of a given strand. Therefore, for any given template molecule, strand elongation will begin at the 3′ end of the primer and terminate upon incorporation of a ddNTP. In older protocols for dideoxy sequencing, four separate primer extension reactions are carried out, each containing only one of the four possible ddNTP species (ddATP, ddGTP, ddCTP, or ddTTP), along with template, polymerase, dNTPs, and a radioactively labeled primer. The result is a collection of many terminated strands of many different lengths within each reaction. As each reaction contains only one ddNTP species, fragments with only a subset of possible lengths will be generated, corresponding to the positions of that nucleotide in the template sequence. The four reactions are then 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 one to directly interpret the primary sequence of the template under analysis.

Current implementations of dideoxy sequencing differ in several key ways from the protocol described above. Only a single primer extension reaction is performed that includes all four ddNTPs. The four species of ddNTP are labeled with fluorescent dyes that have the same excitation wavelength but different emission spectra, allowing for identification by fluorescent energy resonance transfer (FRET). To minimize the required amount of template DNA, a "cycle sequencing" reaction is performed, in which multiple cycles of denaturation, primer annealing, and primer extension are performed to linearly increase the number of terminated strands. This requires

Figure 7.1.1 Schematic of the basic principle involved in dideoxy sequencing. The sequencing template consists of an unknown region whose sequence is to be determined, flanked by known sequence to which a sequencing primer can be hybridized. Cycle sequencing (multiple cycles of primer annealing, primer extension, and denaturation) are performed with polymerase, dNTPs, and fluorescently labeled ddNTPs (where a different label is present on each species of ddNTP). Products of the cycle sequencing reaction are run into a capillary containing a denaturing polymer. This yields size-based separation with single-base-pair resolution, with the shortest fragments running the fastest. Observation of the emission spectra in four channels (corresponding to the fluorescent labels for the four ddNTP species) over time, as fragments emerge from capillary electrophoresis, can be used to infer the primary sequence of the unknown template.

the use of engineered polymerases, such as ThermoSequenase, that are thermostable and that efficiently incorporate modified ddNTPs (Tabor and Richardson, 1995). The products of the cycle sequencing reaction are analyzed in an automated sequencing instrument via electrophoresis in a long capillary filled with a denaturing polymer that yields size separation with single-base-pair resolution. As fragments of each discrete length pass through a transparent component near the end of the capillary, a single wavelength of light excites the fluorophores linked to the ddNTPs. Labeled fragments fluoresce at one of four distinct wavelengths, revealing the identity of their terminal base via FRET. Simultaneous measurement of the emission spectra at these four wavelengths produces a four-color sequencing trace. Computer algorithms (“base callers”) interpret the peak heights in these traces to produce a DNA sequence. Importantly, sophisticated algorithms exist that also define the accuracy with which individual base-calls are made (Ewing and Green, 1998; Ewing et al., 1998). Although the per-base accuracy can vary substantially within a single sequencing read, the accuracy of the best base calls can be as high as 99.999%.

Nearly all dideoxy sequencing performed today makes use of automated capillary electrophoresis, which typically analyzes 96 to 384 sequencing reactions simultaneously via an array of capillaries. Major vendors include Applied Biosystems (e.g., the ABI 3730) and

GE Healthcare (e.g., the MegaBACE instrument series). There is a tradeoff between long read lengths and the overall throughput of an instrument. Depending on which parameter is being optimized, conventional instruments are capable of reads just over 1000 base pairs in length, or production throughputs of over 2.5 megabases per day. Because of variation in the levels of optimization and instrument uptime, the cost of dideoxy sequencing varies widely throughout the research community. The in-house costs of high-throughput sequencing centers may be as low as 50 cents per kilobase, while core facilities and commercial entities may charge anywhere from $1 to $20 per sequencing read.

Cyclic Array Sequencing

All of the recently released, or soon-to-be-released, non-Sanger commercial sequencing platforms, including systems from 454/Roche, Solexa/Illumina, Agencourt/Applied Biosystems, and Helicos BioSystems, fall under the rubric of a single paradigm, termed cyclic array sequencing (Fig. 7.1.2). Cyclic array platforms achieve low costs by simultaneously decoding a two-dimensional array bearing millions (potentially billions) of distinct sequencing features. The sequencing features are "clonal," in that each resolvable unit contains only one species of DNA (as a single molecule or in multiple 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 (distinct from the unknown sequence of other DNA features on the array) flanked by universal adaptor sequences. A key point in this approach is that the features are not necessarily separated into individual wells. Rather, because they 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 a single base position for all features in parallel. The enzymatic process is coupled to either the production of light or the incorporation of a fluorescent group. At the conclusion of each cycle, data are acquired by CCD-based imaging of the array. 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 feature 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 remarkably in the specifics of implementation. The primary areas of difference (summarized for several platforms in Table 7.1.1) are (1) the method used to generate the DNA sequencing features, and (2) the biochemistry used

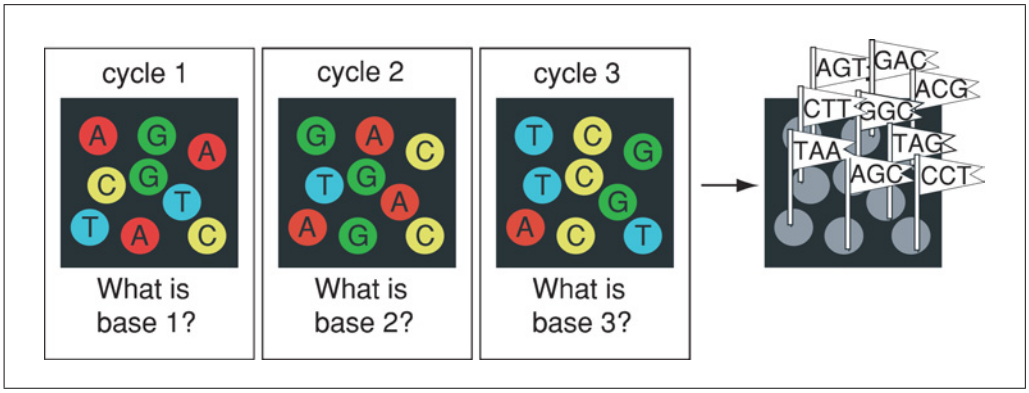

Figure 7.1.2 The concept of cyclic array sequencing platforms involves an array of DNA features to be sequenced, immobilized to constant locations on a solid substrate. At each cycle, the identity of a single base position is interrogated at each feature. Data are collected at each cycle by imaging of 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*.

Table 7.1.1 Cyclic Array Sequencing Platforms[a]

Platform	Polony amplification	Cycle sequencing
Harvard/Danaher Agencourt/Applied Biosystems	Emulsion PCR (1-μm beads)	Ligase
Solexa/Illumina	Bridge PCR	Polymerase (single base extension with reversible terminators)
454 Corp/Roche	Emulsion PCR (28-μm beads)	Polymerase (pyrosequencing)
Helicos	None (single molecule)	Polymerase (single base extension)

[a]Open-source and commercial cyclic array sequencing platforms that have been recently released, or are soon to be released. The primary areas of difference are the methods used to generate DNA sequencing features, and the biochemistry used for cyclic sequencing itself.

to perform cyclic sequencing. The following discussion briefly describes three distinct approaches that are currently available as commercial or open-source platforms, with commentary on some of the advantages and disadvantages of each.

"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). Both the polony sequencing system recently developed at Harvard (Shendure et al., 2005; *UNIT 7.8*) and the 454 system (Roche; Margulies et al., 2005) generate DNA sequencing features by performing an emulsion PCR amplification of a complex sequencing library, 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 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 may 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 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 454 system uses a different oil phase that yields much larger aqueous compartments, thus supporting emulsion PCR with capture to much larger beads (28 μm in diameter).

In the 454 system, amplified beads are arrayed on a prefabricated fiber optic bundle etched to contain over one million picoliter-scale wells. Sequencing is performed by the pyrosequencing method. 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. Incorporation events at each array feature are detected by real-time, luciferase-based monitoring of pyrophosphate release. Several hundred thousand 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 100 base pairs. As it was the first cyclic array platform to achieve commercialization, 454 sequencing has contributed to 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 dideoxy- 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). There are several disadvantages to consider. (1) There is a 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. As there is a systematic component to these errors, the extent to which they resolve with multiple reads covering the same positions is unclear. (2) The cost of sequencing with the 454 system is lower than that of conventional dideoxy sequencing, but not dramatically so. (3) Pyrosequencing is required for real-time observation of sequencing features during the enzymatic step of each cycle. This may limit the extent to which the system can evolve beyond its current feature density in future models of the instrument. (4) Performing emulsion PCR and bead recovery is cumbersome relative to bridge PCR or single-molecule methods (see below).

In the most recent implementation of the Harvard polony sequencing system (Shendure et al., 2005; *UNIT 7.8*), DNA sequencing features (1-μm beads generated by emulsion PCR) are dispersed to 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 population of fluorescently labeled nonamers is introduced. The population is composed such that the identity of a specific base position within each nonamer correlates with the identity of the fluorophore. When a position near the site of ligation is interrogated, each bead will predominately incorporate nonamers bearing a single type of fluorophore, 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 using libraries with mate-paired sequencing tags and sequencing into each tag in both the 5′ and 3′ directions, one can obtain at least 26 base pairs of sequence information per bead feature. 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 for commercial development to Agencourt/Applied Biosystems, which notably has developed more sophisticated sequencing-by-ligation chemistry that is capable of contiguous 25 to 35 base-pair reads per mate-paired tag. Applied Biosystems expects to release its instrument (also known as the SOLiD sequencing system) in 2008. 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 Solexa sequencing platform (which includes merged technology from Solexa, Lynx, and Manteia SA) generates polony sequencing features by a different method known as "bridge PCR" (Fedurco et al., 2006). 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 is also 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 50 bases for over 40 million features have been demonstrated on the Solexa system. Key advantages of this platform are: (1) at least a gigabase of sequence can be generated in a single sequencing run, (2) the simplicity of bridge PCR relative to emulsion PCR for feature generation, and (3) per-base costs are estimated to be at least two orders of magnitude lower than conventional dideoxy sequencing. Key disadvantages may include: (1) significantly shorter read lengths than dideoxy sequencing or the 454 system, and (2) the system was only recently released, so performance for key parameters such as raw and consensus accuracy are still under evaluation.

Several groups are working on cyclic array platforms for direct sequencing of single molecules, i.e., without any amplification step. These include the Helicos system, 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 Nanofluidics, Inc., based on technology developed in Watt Webb's lab (Levene et al., 2003), which proposes to use an array of zero-mode waveguides for real-time, single-molecule observation of polymerase-driven incorporation of fluorescently labeled nucleotides to a primed template.

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). This same concept is used for many genotyping array platforms, except that 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, each feature 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 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, yielding extensive discovery of 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, as is possible with other sequencing methods. A component of the problem lies in the difficulty that SBH approaches have with heterozygosity in diploid genomes. 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 true by Sanger-based confirmation, at a cost of ~$7,500 per strain (Herring et al., 2006).

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. The 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, on 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). A key advantage of this approach is the

retention of the dideoxy biochemistry, which has proven robustness for $>10^{11}$ bases of sequencing. Until alternative methods achieve significantly longer read lengths than they can today, 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.

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 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.

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 a synthetic solid-state device (Fologea et al., 2005). As individual nucleotides are expected to obstruct 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 $\sim$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). Additionally, a wide range of proposed approaches to nanopore sequencing were recently funded by the NIH and are under development. 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. Provided these challenges can be met, 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 non-Sanger 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?

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.

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.

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 less expensive per base, 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 read lengths of 25 bp. 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.

Mate-paired reads. The capacity to produce mate-paired reads (pairs of reads that are known to be 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.

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 and the creation of bioinformatics tools for interpreting massive amounts of short-read sequencing data are critical challenges that must be addressed if investigators are to make the most of these exciting new technologies.

LITERATURE CITED

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.

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.

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.

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.

Collins, F.S., Morgan, M., and Patrinos, A. 2003. The Human Genome Project: Lessons from large-scale biology. *Science* 300:286-290.

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

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.

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 ultrahigh-throughput genetic analysis. *Anal. Chem.* 74:5076-5083.

Ewing, B. and Green, P. 1998. Base-calling of automated sequencer traces using phred. II. Error probabilities. *Genome Res.* 8:186-194.

Ewing, B., Hillier, L., Wendl, M.C., and Green, P. 1998. Base-calling of automated sequencer traces using phred. I. Accuracy assessment. *Genome Res.* 8:175-185.

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.

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.

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.

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.

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.

Holley, R.W., Apgar, J., Everett, G.A., Madison J.T., Marquisee, M., Merrill, S.H, Penswick, J.R., and Zamir, A. 1965. Structure of a ribonucleic acid. *Science* 147:1462-1465.

Hunkapiller, T., Kaiser, R.J., Koop, B.F., and Hood, L. 1991. Large-scale and automated DNA sequence determination. *Science* 254:59-67.

International Human Genome Sequencing Consortium (2001). Initial sequencing and analysis of the human genome. *Nature* 409:860-921.

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.

Levene, M.J., Korlach, J., Turner S.W., Foquet, M., Craighead, H.G., and Webb, W.W. 2003. Zero-mode 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.

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.

Maxam, A.M. and Gilbert, W. 1977. A new method for sequencing DNA. *Proc. Natl. Acad. Sci. U.S.A.* 74:560-564.

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-Krzymanska-Olejnik, and Church, G.M. 2003. Fluorescent in situ sequencing on polymerase colonies. *Anal. Biochem.* 320:55-65.

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.

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.

Prober, J.M., Trainor, G.L., Dam, R.J., Hobbs, F.W., Robertson, C.W., Zagursky, R.J., Cocuzza, A.J., Jensen, M.A., and Baumeister, K. 1987. A system for rapid DNA sequencing with fluorescent chain-terminating dideoxynucleotides. *Science* 238:336-341.

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.

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.

Ryle, A.P., Sanger, F., Smith, L.F. and Kitai, R. 1955. The disulphide bonds of insulin. *Biochem. J.* 60:541-556.

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

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. 1977.

Nucleotide sequence of bacteriophage phi X174 DNA. *Nature* 265:687-695.

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.

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

Tabor, S. and Richardson, C.C. 1995. A single residue in DNA polymerases of the *Escherichia coli* DNA polymerase I family is critical for distinguishing between deoxy- and dideoxyribonucleotides. *Proc. Natl. Acad. Sci. U.S.A.* 92:6339-6343.

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

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.

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.