Christopher P. Fredlake[1]
Daniel G. Hert[1]
Elaine R. Mardis[2]
Annelise E. Barron[1]

[1]Department of Chemical and Biological Engineering, Northwestern University, Evanston, IL, USA
[2]Genome Sequencing Center, Washington University School of Medicine, St. Louis, MO, USA

Received July 3, 2006
Revised July 10, 2006
Accepted July 11, 2006

Review

What is the future of electrophoresis in large-scale genomic sequencing?

Although a finished human genome reference sequence is now available, the ability to sequence large, complex genomes remains critically important for researchers in the biological sciences, and in particular, continued human genomic sequence determination will ultimately help to realize the promise of medical care tailored to an individual's unique genetic identity. Many new technologies are being developed to decrease the costs and to dramatically increase the data acquisition rate of such sequencing projects. These new sequencing approaches include Sanger reaction-based technologies that have electrophoresis as the final separation step as well as those that use completely novel, nonelectrophoretic methods to generate sequence data. In this review, we discuss the various advances in sequencing technologies and evaluate the current limitations of novel methods that currently preclude their complete acceptance in large-scale sequencing projects. Our primary goal is to analyze and predict the continuing role of electrophoresis in large-scale DNA sequencing, both in the near and longer term.

Keywords: DNA sequencing / Genome sequencing / Microchip electrophoresis
DOI 10.1002/elps.200600408

1 Introduction

With the completion of the draft sequence [1, 2] as well as subsequent finishing [3] of the human genome, biomedical researchers now have an ever-improving reference genome sequence with which to conduct research in diverse areas. Examples of these types of studies include exploring the genetic sources of disease susceptibility [4], developing medical and biotechnology applications [5, 6], and studying the history and mechanisms of human genetic evolution [7–9]. With a complete sequence of our genome now in hand, it may seem that the era of large-scale genomic sequencing projects is coming to an end. However, the National Institutes of Health (NIH)/National Human Genome Research Institute (NHGRI) continues to sponsor many research projects aimed at developing new technologies to enable large-scale sequencing of complex genomes. The goal of these projects is to reduce the total cost of sequencing either by two orders of magnitude (the $100 000 genome) or by four orders of magnitude (the $1000 genome) from the current cost of sequencing a human-sized genome (about $10 million) [9a]. The ultimate long-term goal of this investment is to create new technologies that will allow a full or partial genome sequence determination to become part of a standard physical examination and to facilitate personalized medical care tailored to an individual's unique genetic makeup.

It has generally been assumed that characterizing the unique aspects of an individual's genome at the single-nucleotide level will provide a sound genetic basis for personalized medical care. While there is no doubt that some of the 11 million hypothesized SNPs in the human genome [10] contribute greatly to phenotypic variations among individuals, other types of larger-scale genetic variations such as insertions, deletions, and large-scale genomic rearrangements are also likely to play a prominent role in determining our species' wide genetic diversity [11]. Characterization of this scale of genomic variability will, in many cases, require more than simple SNP

Correspondence: Professor Annelise E. Barron, Department of Chemical and Biological Engineering, Northwestern University, Evanston, IL 60208, USA
E-mail: a-barron@northwestern.edu
Fax: +1-847-491-3728

Abbreviations: BAC, bacterial artificial chromosome; **CAE**, capillary array electrophoresis; **IR**, infrared; **LPA**, linear polyacrylamide; **PMMA**, poly(methyl methacrylate); **WGS**, whole genome shotgun genotyping or even directed resequencing. In fact, whole human genome resequencing, along with an analysis of copy number polymorphism, loss of heterozygosity, methylation status, and chromosomal rearrangement, will be necessary to fully characterize any one person's genetic blueprint. A present-day example of such a project is The Cancer Genome Atlas project (TCGA) that aims to produce genome-wide data for multiple samples of each major type of human cancer (Recommendation for a Human Cancer Genome Project, Report of the Working Group on Biomedical Technology, February 2005, http://www.genome.gov/15015123). This and similar efforts will require significant reductions in sequencing costs, as the proposed 15 000 tumor samples are expected to be fully analyzed within a budget of $1.5 billion dollars (about $100 000 *per* sample).

The noncoding portions of the genome have long been ignored or often considered less "interesting" than protein-encoding regions of the genome, and for obvious reasons. Apart from encoding no known protein or RNA products, these regions are typically rich in repetitive sequences and, consequently, are not properly assembled by computer algorithms (especially for whole genome shotgun (WGS) strategies [12]). However, this so-called "junk DNA", which comprises half of the entire human genome [1], has conversely been shown to have importance in mediating DNA–protein interactions, in promoting or repressing transcription, and in forming the general architecture of the genome as it resides in the cell [13]. There are also clear examples in the literature, using genome-scale microarray technologies, that indicate these regions, once thought to be quiescent with respect to transcription, are quite active, likely encoding regulatory RNAs or unknown/unpredicted genes [14].

A special class of repetitive DNA sequence known as segmental duplications is currently receiving attention as these genomic regions can contribute significantly to genetic and, hence, phenotypic variation [11]. A recent study has shown that duplications of a region of the genome containing the *CCL3L1* gene affect an individual's susceptibility to acquiring HIV/AIDS [4]. In this study, the copy number of the duplicated region in relative proportion to that of the unaffected surrounding population determines an individual's susceptibility to the virus. Segmental duplication events also have contributed more to the human/chimpanzee genetic split than have single-base pair substitutions [15]. Hence, characterizing the segmental duplications clearly in an individual genome is very important.

Given the clearly of important, yet still poorly understood roles of repetitive content, it will be ideal if human genome resequencing projects can include both repetitive and protein-encoding regions in order to be maximally informative. Hence, novel whole genome sequencing technologies must be able to sequence repetitive regions of the genome with a minimum of finishing costs.

The development of automated and high-throughput processes, for both the sample preparation [16] and the capillary array electrophoresis (CAE) separation steps [17] of conventional sequencing pipelines, have reduced the cost of sequencing a human genome from the $2.7 billion required to obtain the first finished sequence [17a] to between $10 and 20 million (Recommendation for a Human Cancer Genome Project, Report of the Working Group on Biomedical Technology, February 2005, http://www.genome.gov/15015123) for unfinished raw data production and assembly. A number of technologies have been proposed and are being developed with the goal of attaining either the $100 000 or the $1000 genome over the next 5–10 years. This review will focus on and discuss developments in both traditional Sanger sequencing approaches that utilize electrophoresis as the final separation step as well as novel, nonelectrophoretic approaches. Additionally, we will evaluate the strengths of various technologies as well as the technical and economic challenges to supplanting current CAE approaches within large-scale genome sequencing centers, and anticipate the role of electrophoresis in future large-scale sequencing efforts.

2 The current paradigm for genomic DNA sequencing

2.1 Genomic sequencing strategies

The two common methods employed by genome sequencing centers at present are the WGS method [18] and the hierarchical shotgun, or physical map-based, method. The WGS strategy was employed by the commercially supported Celera human sequencing project [2], whereas the publicly funded Human Genome Project [7] used the hierarchical shotgun method to sequence the human genome. Here, we will briefly summarize the two approaches and discuss the challenges of current sequencing strategies that novel technologies are attempting to address.

The WGS strategy randomly shears the genomic DNA into smaller fragments, and then creates a genomic library of plasmid and fosmid subclones of various insert sizes, using combinations of cloning vectors and corresponding *Escherichia coli* strains to propagate the subclones. The insert sizes are relatively well controlled and can range in size from 1 to 50 kbp. Electroporation and plating of the genomic libraries are relatively manual processes that are then followed by automated steps. For example, colony picking is done in an automated fashion using complex robotics [16]. Once colonies are picked for sequencing, the subclones are isolated for use as the DNA template in the Sanger cycle sequencing reaction [19]. The Sanger reaction is carried out using fluorescently labeled dideoxynucleotide terminators (ddNTPs) to identify the terminating base in each fragment of the sequencing ladder from which the raw sequencing reads will be produced. Generally, the sequencing sample is concentrated by ethanol precipitation to remove excess fluorescent terminators prior to ladder separation and detection [20, 21]. The electrophoretic separation step is completed in high-throughput CAE instruments, and raw sequencing reads of 650 to 700 *phred* 20 (Q20) quality bases *per* capillary can be obtained in about 1–2 h [22, 23]. This time includes polymer matrix filling before sample injection and capillary clearing and rinsing following the sequencing run. On an average day, approximately 750 000–800 000 high-quality raw bases are produced *per* CAE instrument. Complex computer algorithms have been developed for the assembly of the raw sequencing reads into a draft genome assembly. For a draft sequence of the genome, usually 6 × coverage is acquired, implying that each nucleotide in the genome is sampled six times.

In the hierarchical shotgun strategy, as shown ideally in Fig. 1, the first step is to create a high-density physical map of the genome using fingerprint mapping strategies [24]. As in the WGS approach, the genomic DNA is subjected to partial digestion and inserted into bacterial artificial chromosomes (BACs) for transfection into *E. coli* cells. BAC physical maps are based on digesting each BAC with the same restriction enzyme and collecting data on the resulting fragment sizes for 10–15 × genome coverage. Computer algorithms are then used to assemble contigs of BACs to recapitulate regions of the genome, based on shared fragment sizes. On the basis of the assembled BAC contigs in the physical map, one can select a so-called “minimal tiling path” or MTP of BACs for sequencing. Each selected BAC is grown in culture, isolated and sheared in preparation for plasmid subclone library generation. Raw reads are then obtained in the same manner as the WGS with final genome assembly being somewhat easier since each BAC is assembled as a discrete entity, and subsequently, linked *in silico* to assemble the genome.

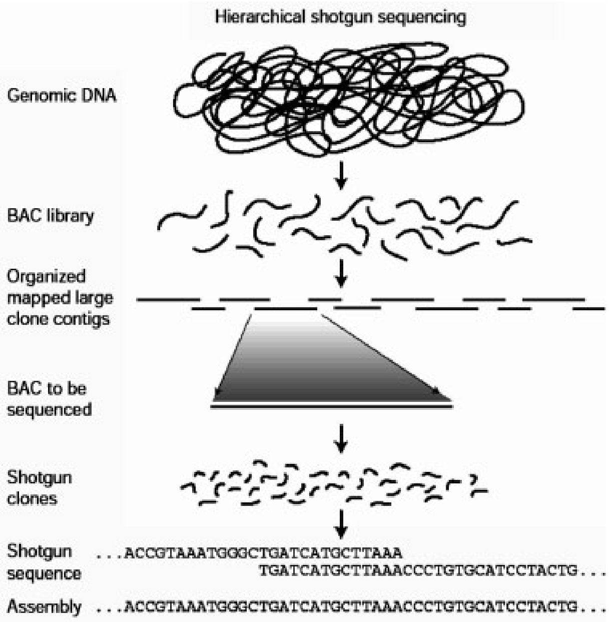
Hierarchical shotgun sequencing
Genomic DNA
BAC library
Organized mapped large clone contigs
BAC to be sequenced
Shotgun clones
Shotgun sequence
...ACCGTAAATGGGCTGATCATGCTTAAA
TGATCATGCTTAAACCCTGTGCATCCTACTG...
Assembly
...ACCGTAAATGGGCTGATCATGCTTAAACCCTGTGCATCCTACTG...

Figure 1. Idealized version of the hierarchical shotgun sequencing process used by the Human Genome Project. Reproduced with permission from [1].

2.2 Genome assembly

While assembly of the genome sequence is the last step of the sequencing process, it is generally one of the most difficult endeavors computationally. Most assembly algorithms use an “overlap-layout consensus” method where raw reads are compared to each other, and the algorithm searches for reads that overlap (Phrap Documentation: Algorithms, http://www.phrap.org) [25–27]. When reads containing some defined minimum level of overlap are found (usually only at the ends of the reads), the two fragments are condensed into a larger contig. Still other computational strategies for genome assembly are currently being developed [28].

The basic assembly strategy described above can be augmented by additional information as well. A common method in WGS sequencing strategies is to use paired-end reads or mate-pairs to provide additional linkage information on raw sequencing fragments to the algorithm. Paired-end sequencing encompasses obtaining raw reads from both ends of a cloned insert as part of the sequencing procedure. As the approximate size of the inserts are known, these mate-pairs provide the assembler with information about their relative orientation to each other as well as about the expected distance between the end reads. The assembly program can further use this information to help localize reads and contigs relative to one another [29]. Mapping information in the hierarchical shotgun strategy is an additional source of localization information that can greatly enhance confidence in the overall assembly fidelity. As more organisms are sequenced, it may be possible to use a comparative assembly program where reads from one organism are placed on a sequence of another organism with high expected sequence homology to help localize regions conserved during evolution [30]. A simplified version of this strategy compares reads from the organism being sequenced to the reference genome of the same species, as in human genome resequencing, to obtain information about a particular individual's genome. Clearly, large-scale insertions, deletions, and rearrangements, as well as high repetitive content, make this approach somewhat more computationally challenging [11].

2.3 Complications and challenges to current sequencing approaches

Problems arise in sequence assembly from two main sources. First, cloning biases occur, in which some regions of the genome are not replicated in *E. coli* cells [1, 24]. Thus, in spite of random genome fragmentation and high sequence coverage, certain parts of the genome are never represented as raw reads, resulting in gaps in the assembled sequence. A second problem results from the presence of highly repetitive sequence in complex genomes. Many animal and plant genomes are repetitive (*e.g.*, up to 50% in the human genome), and as mentioned above, the structure of these areas of the genome may be very important in understanding the biology of the organism. High repetitive sequence content results in incorrect assembly, since repeat regions that actually span large parts of the genome (longer than single read lengths) can result in multiple reads collapsing into a single contig even though they belong to different parts of the genome, thereby, creating gaps in the sequence [29].

Current sequencing strategies cannot avoid clone-biasing issues as both the WGS and hierarchical strategies rely on cloned inserts. However, some approaches can help to ameliorate the problems with repeat-rich sequences. The use of mate-pairs can help localize different repeat regions, but only when the repeat length is less than a typical insert length [29]. Mapping strategies can also provide a framework for localization and, thereby, can reduce the complexity introduced by repeat regions in the WGS strategy. Even so, most assemblies of repeat-rich genomes (both hierarchical and WGS) leave gaps that must be filled by directed sequencing approaches during finishing. The finishing procedures can be very costly as they rely mainly on iterative procedures such as primer walking, and identifying appropriate sample preparation conditions for difficult regions of the genome [3]. Thus, if possible, gaps in the assembled sequence need to be minimized prior to genome finishing.

3 Advances in electrophoresis-based sequencing

Large percentages of conventional sequencing costs derive from CAE instrument depreciation and sequencing sample preparation (E. R. Mardis, personal communication, 2006) [31]. New instruments that can dramatically eclipse the read capacity of traditional CAE instruments during their normal depreciation period (usually 3 years) and can condense the sample preparation process could greatly reduce the costs of genome sequencing. In this section, we describe current research aimed at the integration of conventional sample preparation and DNA sequencing fragment separation onto one device, and then we focus on technical advances that improve and optimize the separation process itself.

3.1 Microfabricated systems for genetic analysis

Many advantages could be gained in both time and costs by the commercial introduction of "lab-on-a-chip" DNA sequencing devices, including decreased analysis times that result from shorter separation channels, highly reduced sample and reagent volumes, and the integration of all sequencing steps from sample preparation and cleanup to separation and detection into a single analytical instrument. Many review articles have been published in recent years that focus on the various individual aspects of genetic analysis and DNA sequencing on microfluidic chips [32–34], and therefore we limit the discussion here to those systems that have succeeded in integrating two or more relevant functions into a single device. In principle, a totally integrated DNA sequencing platform could replace all robotics and other equipment-associated with sample preparation as well as use microfabricated devices in place of CAE instruments to provide raw sequencing reads with substantially increased throughput.

The first attempts to integrate sample preparation with biomolecule detection on a microfluidic chip focused on amplifying a DNA template by PCR followed by electrophoretic separation of the PCR products in a microchannel. Initially, this was done with a microscale PCR chamber interfaced to the separation chip [35], but subsequent devices included the PCR chamber as a part of the overall chip layout [36–49]. Although thermal cycling has been demonstrated for integrated devices either on the whole chip [38, 39, 42] or on certain portions of the chip containing the PCR mixture [41, 45, 49] by the use of external heating and cooling sources, the inclusion of heating and temperature measurement capabilities into the microfabricated structure provided a step forward for totally integrated systems [36, 37, 40, 43, 44, 47, 48]. However, external heating methods such as the Landers group's infrared (IR)-mediated on-chip PCR can be very fast, pushing PCR times to their lower limits [50]. Additionally, systems have been created that enable precise fluidic handling *via* pumps and valves that control the temporal and spatial placement of the sample as it passes through successive steps of preparation and analysis [37]. As an example of a highly integrated device for genetic analysis, the Burns laboratory has recently demonstrated an integrated, micromachined silicon system for PCR amplification, restriction enzyme digestion, and electrophoretic separation of DNA in a cross-linked polyacrylamide gel for detecting a specific strain of influenza virus [48]. This microfluidic chip includes both integrated heaters and reversible wax valves to isolate the PCR area and the restriction digestion area of the chip.

Integrated sample cleanup has been a focus for both the Landers group at Virginia and the Mathies group at U.C. Berkeley. The work of the Landers group has been aimed at on-chip integration of sample cleanup followed by PCR amplification and detection of the product for forensic analysis and pathogen detection. They have developed on-chip SPE protocols for isolating genomic DNA from cells lysed both off- and on-chip prior to DNA amplification [45, 49, 51]. In this approach, the DNA physically adsorbs to silica beads (packed in a silica sol–gel) while proteins and other impurities that can inhibit PCR are removed by flushing. The adsorbed DNA molecules can then be eluted from the SPE column by washing with a different buffer.

The Mathies group, on the other hand, has focused on achieving the necessary sample purification after DNA sequencing ladder preparation (*i.e.*, the Sanger cycle sequencing reaction) and prior to sample injection and electrophoretic separation on-chip [52, 53]. Here, methacrylate-modified oligonucleotides complementary to the 20-base stretch immediately 3' of the M13 universal priming site are covalently attached to a linear polyacrylamide (LPA), and a viscous solution of this LPA is placed within a chamber in the chip. The amplified DNA fragments hybridize to this "capture gel" at lower temperatures (50°C) while the Sanger reaction impurities are removed from the chamber by electrophoresis. The sample is then removed from the capture gel by raising the temperature above the melting point for the DNA strands (67°C) and electrophoresing again. Although each group performs a sample purification step, neither group has yet integrated sample purification both before and after DNA fragment amplification by PCR. Purification at each of these points is very important as both fragment generation by cycle sequencing and sample injection are sensitive to impurities present in the sample [20, 21].

The use of an oligo-modified LPA capture gel by the Mathies group is the only approach that has appeared in the literature attempting to integrate sample preparation with DNA separation for sequencing on one microfluidic device [52, 53], although earlier studies attempted to integrate sample preparation and sequencing in capillary-based devices [54, 55]. None of these technologies have appeared yet as commercial products for the research or high-throughput sequencing markets, although there is ongoing work in this direction. The major challenge to creating integrated devices for sequencing is that the read length is very sensitive to a high-quality sample injection, which requires excellent sample cleanup protocols. The Mathies group demonstrated that from an essentially unpurified sequencing sample, their chip system could call up to 560 bases of high-quality data (*phred* 20 or better) [52]. Further integration of their system has recently yielded a device that can perform the Sanger cycle sequencing reaction, sample purification using the capture gel, and sequencing ladder separation, providing a significant step forward for miniaturized sequencing systems [53]. This study also featured many of the advances in their microfabricated chips over the past few years including integrated heating and temperature control, layered, on-chip PDMS valves and pumps [56], and "hyperturns" in the separation channel to provide an extended separation distance [57]. Figure 2 shows the layout of this device. It is important to note that in addition to purification of the DNA sequencing ladder, the capture gel effectively concentrated the sample in this study, allowing for more sensitive LIF detection. The average read length in this study was 427 bases of high-quality sequence (≧ *phred* 20), with the longest read being 560 bases in approximately 35 min of on-chip electrophoresis. While this read length is lower than in the previous study, further optimization of the DNA capture and injection conditions should be able to increase this read length to greater than 600 bases.

3.2 Advances in on-chip sequencing separations

While the success of CAE instruments in the genome sequencing centers to date rested on both (i) creating reliable, easy-to-use, high-throughput automated systems and (ii) developing high-performance polymer matrices that separate DNA sequencing fragments with long read lengths, the literature on microchannel sequencing systems so far has mainly focused on the former. The leader in publishing reports of miniaturized DNA sequencing

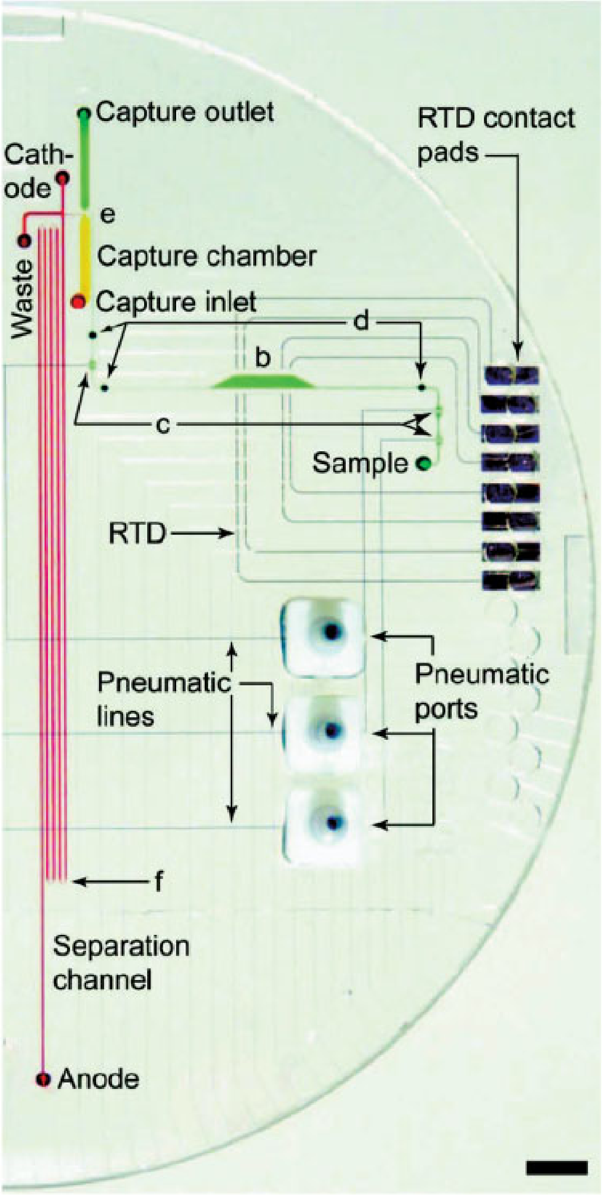

Figure 2. The Mathies group integrated sequencing chip. The individual parts of the device are labeled in the figure: (b) thermal cycling reactor; (c) microvalves; (d) *via* holes; (e) capture-inject region; (f) tapered turns for separation channel. Reproduced with permission from [53].

systems with many channels in parallel has been the Mathies laboratory. A paper by Woolley and Mathies [58] was the first to demonstrate four-color sequencing on chips, and Mathies' group later developed a high-performance single-channel chip system [59], reported to sequence 500 bases in 20 min, as well as a 96-channel system [60] that sequenced an average of 430 bases/channel in 24 min. Similarly, a group at Amersham Biosciences led by Stevan Jovanovich also reported parallel-channel sequencing in a 16-channel chip, achieving an average read length of 450 bases in 15 min [61]. As mentioned above, these systems continue to grow in sophistication as the Mathies laboratory has demonstrated highly integrated microfabricated sequencing systems achieving read lengths up to 560 bases in 35 min, although not in a highly parallel system [52, 53]. While the challenges of creating these systems and integrating certain processing steps have been successfully met in research settings, the need to further optimize the separations to match or surpass current CAE read lengths and to make a commercially available, highly reproducible system still exists. Although these systems have been limited in part by injection problems or other hardware issues, the only way to really extend read lengths to the 650–700 bases presently provided by commercial CAE instruments is to improve the polymer matrix and microchannel wall coating used for DNA separation.

The Ehrlich group has also published reports on microchannel sequencing systems [62–64] while putting a stronger emphasis on developing high-performance LPA matrices for use in improving separations. Salas-Solano *et al.* [63] demonstrated 580-base reads in 18 min and 630-base reads in 30 min, but interestingly reached the conclusion that device lengths were too short and that much longer channels, on the order of 40 cm, would be needed to facilitate the achievement of read lengths competitive with CAE. While later studies by this group have shown read lengths of up to 800 bases [64], the run times are on the same time-scale as CAE systems, such that the cost savings expected from increased throughput by using smaller chips may not be as great in these large glass plates, which are also more difficult and more expensive to manufacture.

DNA sequencing has also been demonstrated in plastic chips. The Soper group has reported one-color sequencing fragment separations on poly(methyl methacrylate) (PMMA) substrates using LIF detection and a near-IR detectable dye set to reduce the background fluorescence from the chip itself [65]. Although this was the first published report of DNA sequencing in PMMA channels, little optimization or the demonstration of four-color detection has followed as would be needed to achieve commercializable sequencing success. Four-color sequencing has been accomplished in plastic chips made from a polyolefin substrate by a group at ACLARA BioSciences [66]. Here, very short channels (4.5 cm) were used, and 320 bases were read in 13 min. A random linear copolymer of poly(*N,N*-dimethylacrylamide) and poly(*N,N*-diethylacrylamide) [67] was used to form a physically adsorbed polymer coating to reduce EOF and analyte–wall interactions. The authors stated that the key to their success in obtaining read lengths much longer than anything to-date in such short channels was the development of an optimized polymer matrix and wall coating formulations. They reached the interesting conclusion that the best matrix formulation in CAE systems is not necessarily the best one for plastic microchip systems, echoing a finding in a Salas-Solano *et al.* paper [63] for glass chips in which the optimal LPA concentration for glass microfluidic electrophoresis chips was found to differ from that which provided the best separations in fused silica capillaries.

While these reports have demonstrated remarkable progress in on-chip sequencing, substantial further development and optimization of polymer materials for both the sequencing matrix and the wall coating are needed to allow the robust performance of these devices. LPA has been the separation matrix used in all of the above studies, an obvious choice as this polymer has shown very long read lengths in capillaries [68]. However, no group has yet explored the use of other polymer matrix formulations as an alternative to achieving longer reads. Additionally, all studies with the exception of the ACLARA study used an LPA wall coating that was covalently linked to the channel surfaces [69]. These covalent coatings can be costly to create, do not last very long before their performance begins to degrade, and can cause channel clogging during the formation of the coating, rendering some channels on the chip unusable.

Most matrices used in the studies we have discussed were comprised of 3–4% LPA solutions where the polymers were either polymerized in the electrophoresis buffer with little attempt at purification and/or were neither characterized nor optimized for important properties such as polymer molar mass or polydispersity. These LPA matrices are also generally very viscous and can be difficult to load into microchannels where pressures are restricted to 200 psi, above which the thermally bonded glass chips will fail. Recent results from our laboratory show that, with the use of optimized sequencing matrices having a different chemical structure than LPA, DNA sequencing of up to 600 bases in ~6.5 min with 98.5% base-calling accuracy is possible on chips with separation distances of only 7.5 cm (Fredlake, C. P., Hert, D. G., Kan, C. W., Cheisl, T. N. *et al.*, submitted 2006). The results reduce the required sequencing time from previous on-chip studies by approximately 66%, and these faster separations are attributable to a unique separation mechanism of the DNA within the formulated matrices, as discussed in the paper. These results demonstrate that read lengths similar to those in the CAE instruments can be obtained on chips in a much shorter time, reducing the separation time by ~90%.

4 Alternative sequencing technologies

In an effort to reduce current sequencing costs, new technologies have been proposed as alternatives to the Sanger method [19] of DNA sequencing by electrophoresis. Many different approaches to sequencing DNA have been proposed (see [31] for detailed review), but here we will focus on two general approaches with the maximum potential to displace current CAE (and electrophoresis-based methods in general) from large-scale sequencing projects. The first approach takes several specific forms, and is characterized by simultaneously detecting sequencing data being produced in real time from large numbers of different DNA molecules. These highly, or massively, parallel sequencing methods involve sequential enzymatic or chemical steps that are repeated in cyclic fashion to produce raw sequencing reads. The second general approach is characterized by obtaining DNA sequence information from a polynucleotide as it passes through a nanometer-sized pore under an electric field. The goal is that each specific, measured response of the system (through a number of variables) during passage of the DNA strand through the nanopore will indicate the unique sequence of the polynucleotide.

4.1 Massively parallel sequencing systems

Massively parallel sequencing methods generally aim to detect the identity of a nucleotide base as it is incorporated into a synthesized DNA strand, and to do this for hundreds of thousands of individual DNA strands simultaneously. Such “sequencing by synthesis” (SBS) techniques are intrinsically very interesting since native DNA polymerases can synthesize DNA at rates of up to 1000 bases/s; however, current methods proceed much more slowly since fresh reagents must be provided to the reactions, *en masse*, at each cycle. Fluorescently labeled-nucleotide SBS methods can involve chemically cleavable [70, 71] or photocleavable [72–75] fluorescent groups covalently linked to the nucleotide to reveal its identity. Additionally, research groups have attempted to sequence DNA in real time at the single-molecule level using zero-mode waveguides [76] or molecular motors [77] to detect specific DNA polymerase activity with respect to labeled nucleotides. These latter, single-molecule approaches represent potentially very important long-term approaches to massively parallel sequencing, if the technical challenges to their practical and robust implementation in a commercial device can be met successfully.

While most of the initial research projects aimed at developing labeled-nucleotide SBS methods have struggled to reach sequencing read lengths greater than 12 bases [78, 79], with each monitored base addition requiring approximately 20 min, 454 Life Sciences recently introduced a new commercial instrument, the GS-20, that uses an enzymatic SBS technique known as pyrosequencing [80–82] to read up to 25 million bases in one 4-h run, with an average read length of 100 bases [83]. This was accomplished by creating a massively parallel SBS system where DNA fragments from sheared genomic DNA are attached to beads directly and amplified by emulsion PCR [84]. Beads with highly amplified, unique DNA templates are then placed into a fiber optic well, on a plate that offers up to 1.6 million individual wells. Figure 3 schematically shows the amplification process and the placement of beads into wells. The 454 GS-20 instrument is now being used and tested at several genome centers and other sequencing facilities worldwide, and at present seems best suited to sequencing bacterial and viral genomes, as well as other small DNA species such as ESTs, serial analysis of gene expression (SAGE) ditags, and microRNAs. However, for sequencing larger genomes, the significantly shorter read lengths obtained with this system relative to Sanger sequencing, the coverage model required to obtain a reasonably confident consensus sequence, and the need for paired-end reads make it overall less suitable than CAE systems at present.

Another highly parallel system for DNA sequencing based on emulsion PCR amplification of DNA attached to beads, but replacing polymerase sequencing with ligase-based sequencing, has also been demonstrated [85]. Here, DNA fragments resulting from genome fragmentation were amplified in segregated polymerase colonies, or "polonies", which have also proven useful in other DNA analysis techniques, such as SNP detection [86] and gene expression [87]. The polonies were formed by isolating emulsion PCR droplets in an acrylamide gel to essentially eliminate aggregation of the droplets by severely limiting diffusion [88].

The ligation sequencing method described, while still in the early stages of development with average raw read lengths of up to 26 bases and accuracy one to two orders of magnitude lower than Sanger sequencing, is able to probe the exact sequence of nucleotides at discrete locations. This is accomplished by attaching "universal" DNA sequences to genomic fragments, onto which universal forward and reverse primers can hybridize. After universal primer hybridization, degenerate, fluorescently labeled nonamers are hybridized and ligated to the universal forward or reverse primer at its 3'-end. As the nucleotide identity and position in each degenerate nonamer are known, ligation and subsequent fluorescence detection indicates the identity of the base. By massive parallelization, genome sequence information can be extracted. The authors state that this method would be useful in resequencing applications, though it is acknowledged that a human genome resequencing application would require extension of the average read

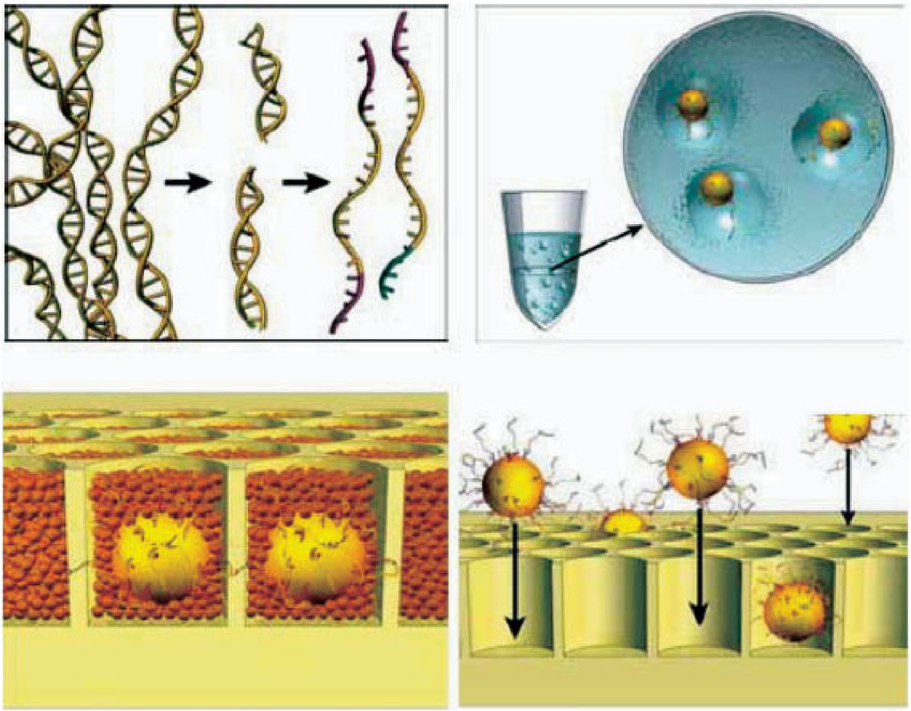

Figure 3. Scheme of 454 Life Science's emulsion-based amplification of sequencing templates and placement of beads into Picotitre plates. Reproduced with permission from [83].

length by at least eight toten bases in order to obtain high fidelity of placement onto the reference sequence. The development of this technology for commercialization is presently underway.

4.2 Nanopores

Another alternative technology that has generated much interest but has been slower and more difficult to develop is the use of nanopores to sequence polynucleotides as they move through narrow channels under an applied electric field [89]. Generally, nanopores are constructed by embedding the well-characterized [90] porin protein α-hemolysin in a lipid bilayer. A stable current can be measured across the pore as buffer ions pass through the channel. As a polynucleotide enters the pore, the measured current is significantly decreased due to blockage of the passage of the buffer ions. Measurable parameters, such as the degree of current reduction, the translocation time, and the translocation time distribution, depend on the composition of the polynucleotide. However, thus far, only the discrimination of large homopolymers, di-block copolymers, or alternating copolymers of nucleotides has been demonstrated by nanopore "sequencing" [91–93]. One of the great potentials that this technology holds is high throughput, since the polynucleotide can translocate through the pore very rapidly (~1–10 μs/base) [94, 95]. However, these translocation speeds also complicate the measurements needed to elucidate the sequence. Currently, research with nanopores is focusing on gaining a deeper general understanding of the underlying physics of DNA translocation.

A very interesting demonstration of sequencing up to three sequential nucleotides using an α-hemolysin pore has been accomplished recently by Howorka *et al.* [96]. In this study, the translocation time of a ssDNA fragment was increased when it interacted with a short oligonucleotide strand that was complexed to the interior of the α-hemolysin pore and whose sequence was complementary to the passing DNA strand. While this is an important result, the ability to sequence longer oligonucleotides by either increasing read lengths in individual pores or by creating many parallel pores is still unknown.

The use of protein pores likely will not be useful for single-nucleotide resolution sequencing of DNA, since the ability to engineer and optimize protein-based pore properties is not straightforward. Thus, recent research in this area has also focused on the fabrication of nanopores from polymers such as PDMS [97] and polyimide [98] and inorganic materials such as silicon nitride [99–102] and silicon oxide [103–106].

In silicon nitride pores, work has been done to slow down DNA translocation velocities by controlling the buffer viscosity, buffer concentration, electric field, and temperature [101]. Such an optimization eases the requirements of detection bandwidth by reducing the frequency from ~30 to ~3 bases/μs. However, operating silicon nitride nanopores in alkaline conditions can reduce the usable lifetime of the nanopores from days to hours by causing drifting baseline currents, increased noise, and permanent blockage of the nanopore entrance [107].

In the case of silicon oxide nanopores, significant deprotonation of the silanol groups at the nanopore surface occurs at pH 7–8, which induces the aggregation of buffer cations at the surface [103, 104, 106]. As an electric field is applied to induce translocation of DNA molecules, there is a concurrent EOF in the direction opposite to DNA migration, causing alterations in both the translocation time of DNA and the measured electrical current. The issue of achieving predictable single-base resolution then becomes even more difficult.

In general, much less specific work has been performed to distinguish individual nucleotides with synthetic nanopores relative to the work done on α-hemolysin. This can be attributed to the relatively recent emergence of this approach, and ongoing work will be focused on developing specialized methods needed to create well-defined and robust nanopore structures. More research and innovation in these areas will be necessary for nanopore technology to become a feasible alternative to current sequencing methods. Individual read lengths of thousands of bases or more may be feasible eventually, but unfortunately, the technical challenges to sequencing DNA beyond a few bases seem daunting. The major challenge for these systems is to obtain single-base resolution of the DNA as it moves through the pore. To improve the resolution, the systems need to decrease the translocation speed of the DNA and reduce the distance spanned by the nanopore to the size of a single base. DNA translocation speed must be reduced so that detection and data processing software can accurately resolve incoming signal to the single base level. Nanopore size must be carefully controlled such that the pore will only be occupied by a single nucleotide at any one time, and the measured signal results from only a single base and correlates clearly with the identity of that base. Alternatively, a synthetic nanopore could be created with embedded electrodes of width approximately that of the length of a single nucleotide of DNA, which could eliminate the effects of adjacent nucleotides on electrical current measurements [108].

5 Critical evaluation of the challenges for new sequencing technologies

Many of the technologies discussed in this review have the potential both to significantly improve the throughput and to reduce the cost of genomic sequencing. However, many technical and economic challenges remain that, at present, seem to either limit or prohibit these new technologies from ever replacing CAE for the sequencing of complex genomes. In this section, we will examine the benefits and potential pitfalls of these approaches as they attempt to mature into the next large-scale sequencing technology.

5.1 Electrophoretic sequencing on microfluidic devices

The main perceived advantages of these systems, as mentioned earlier, are that raw sequence reads can be obtained at much higher rates, and potentially at lower cost, and that sample preparation can be integrated with sequence acquisition. Thereby, integrated microfluidic sequencing systems could effectively replace both the robotic sample preparation systems and the CAE instruments presently in use, on a much smaller platform. As we discussed, while sequencing 650–700 bases on a CAE instrument requires 1–2 h, the same result could be generated in 7–10 min on a chip-based sequencer. As more reads could be produced *per* day on this instrument, more data would be generated over the lifetime of the instrument, and hence the capital equipment costs could be amortized over many more reads, making the *per-read* contribution to cost effectively lower than for CAE instruments. Furthermore, by reducing the additional instrumentation used in sample preparation and purification, the overall capital costs would be significantly lowered. By shrinking the entire process pipeline, reagents, and consumables consumption would also be greatly reduced, further driving down sequencing costs.

One of the main consumables in microfabricated systems will be the chips themselves, much like capillary arrays in current CAE instruments. As a comparison, the currently listed cost of a 50-cm-long capillary array of 96 capillaries for the ABI 3730xl is $3780. Assuming 300 runs *per* array and 24 runs *per* day, the yearly cost of new arrays alone is approximately $110 000 *per* instrument. A borosilicate glass wafer can be purchased for ~$100, and including fabrication costs and assuming some commercial price markup, a 96-channel chip can be expected to cost approximately $400. Assuming the chip will have the same lifetime as a capillary array and that the sequencing runs are 10 min in duration (a 6× increase in throughput), the yearly cost of chips for a single instrument is expected to total $73 000. Even though the yearly costs are not significantly reduced, many more sequencing reads are produced *per* instrument. Thus, on a *per-read* basis, the reduction in cost when replacing capillary arrays with glass chips is ~90%. Adding more channels on a chip (>96) will not greatly increase the chip substrate cost, whereas the cost of capillaries always scales linearly with the number of added capillaries. Furthermore, if these systems are adapted to use plastic chips fabricated from materials such as PMMA or cyclic olefins, the chips would then cost approximately $15 each (S. A. Soper, personal communication, 2006). The yearly cost for plastic chips would then total $2700, representing a greater than order-of-magnitude reduction over capillaries, presumably still with a 6× increase in throughput if these materials can perform comparably to borosilicate chips. On a *per-read* basis, PMMA plastic microfluidic chips could therefore offer a cost savings of 96% relative to glass chips, and 99.6% relative to currently used capillary arrays.

Although microfabricated systems potentially offer many cost and throughput advantages, technical issues have thus far prevented the new technology from displacing CAE. By combining many steps into a single genomic sequencing device, the total analysis time will be limited by the slowest step, since currently available microchip systems process samples serially. It may be advantageous for these systems to adopt a strategy wherein sample preparation steps operate independently from sequencing reaction preparation, sample injection and separation. An efficient system will be capable of taking advantage of increase in speed at every process step to maximize the sequencing throughput. This will require parallel operation at every step in the sequencing process; however, parallel processing has only been demonstrated for DNA separations thus far. In general, the slowest step will require a more parallel operation so that the other steps in the process can run continuously with little downtime. Specifically, microfabricated systems have been developed that have 96 lanes in parallel for sequencing or even 384 channels for genotyping applications. With these designs, these systems will most likely never match the throughput of massively parallel systems such as that demonstrated by 454 Life Sciences' picotitreplate (PTP) formats, which contain up to 1.6 million individual wells, although they may provide an important platform for medium-throughput sequencing and will be uniquely suited to the sequencing of large and complex genomes.

As microchip electrophoresis systems utilize the Sanger-sequencing method, established genome sequencing strategies will not require a global adaptation to take advantage of this new technology. These systems will likely obtain the high-accuracy raw read lengths that current CAE systems produce so that current genome assembly algorithms can utilize the data without modification and with the same output level contiguity. This technology, however, will also be subject to the same limitations as the CAE-based systems. Furthermore, any clone-based system will be subject to the same clone bias as current genome sequencing approaches. A fundamental change to the bacterial-based library construction procedures for genomic- or large clone-derived DNA prior to sequencing will be required to get around cloning bias. Techniques such as emulsion PCR [84] or other amplification techniques that can produce large numbers of templates in parallel directly from sheared genomic DNA apparently alleviate the problems with nonrandom representation of a genome (E. R. Mardis, personal communication, 2006).

Appropriate assembly of repetitive DNA regions will still be difficult without obtaining longer read lengths so that finishing costs may still be high. Apart from repetitive sequence content, which is problematic for assembly algorithms regardless of read type, since these systems have significantly higher throughputs than current systems, increasing the level of genome coverage may reduce finishing costs by reducing the number of gaps in the assembled sequence. Additionally, dedicated and customized microchip systems with specially designed materials to exclusively carry out finishing-related functions may also reduce finishing costs.

5.2 Massively parallel sequencing systems

Systems that sequence hundreds of thousands of DNA templates in parallel have increased sequencing throughput by nearly 100 times over current CAE instruments. Additionally, the use of emulsion PCR techniques or other noncloning based methods for template amplification avoids the cloning bias that exists when bacterial cells are used for genomic DNA amplification.

The production of tens of millions of bases in a single run represents the throughput obtainable for these massively parallel systems. In spite of this staggeringly increased throughput, these methods are characterized by much shorter individual read lengths, ranging from 12 to 100 bases (on average) for various approaches. One inherent limitation to read length for SBS methods is the loss of synchronicity for each sequenced copy of any given location (bead or cluster). Such a loss of synchronicity (or "phasing") results from the incomplete extension of one or more copies during the synthesis steps, or from incomplete removal of reagents by the intermediate washing step that then carries over to the next synthesis step. As increasing numbers of DNA copies move out of phase, the S/N for the correct incorporation is reduced, although the raw signal may be subject to improvement using computer algorithms for background subtraction (where background levels are typically measured after each wash step) [83]. A second limitation, loss of efficiency at each step, is analogous to the limited lengths of oligonucleotides than can be obtained by solid-phase synthesis. In general, the loss of efficiency can be calculated by $(0.005)^x$ where x represents the number of steps for each cycle and each step is 99.5% efficient. Using these assumptions, read lengths for a four-step process, for example, would be limited to about 230 bases assuming a 99% loss of template is the lower detection limit [31].

As read lengths in these systems will be limited to approximately 200 bases (or far less in many technologies), genome assembly becomes much more difficult for complex genomes, and one must instead utilize read (or read pair) mapping to a reference genome. Recent reports on the assembly of genomes using shorter reads have concluded that while smaller and repeat-poor genomes will present little problem for genome assembly, larger and more repeat-rich genomes will lead to many gaps and very high, perhaps prohibitive, genome finishing costs [109, 110]. Bacterial and viral genomes have been sequenced (both *de novo* and resequenced) by highly parallel methods [83, 85], but these genomes are both much smaller than a mammalian genome and contain very little repetitive sequence. When attempting to assemble human chromosome 1 using sets of reads with short or long average read lengths, only 80% of contigs contained more than 1000 nucleotides, and only 17% of contigs were longer than 10 000 nucleotides when read lengths were set to 50 bases. Read lengths of 500 bases, however, resulted in 98.4% of the contigs having lengths greater than 10 000 nucleotides [110].

As such, new technologies with read lengths of 25–50 bp will rely on read or read pair mapping to a reference genome and may not be suitable for *de novo* sequencing of large (Gbp) and complex genomes. Here, if a read or read pair finds a unique high-quality alignment in the reference genome, it is noted and set aside, whereas if high-quality alignments are found in multiple places in the genome, that read or read pair may represent a breakpoint in a reciprocal translocation, or an inversion, for example, and is therefore of interest. Read pairs, as such, have a distinct advantage over single read alignment and are more likely to be alignable by virtue of the longer effective read length and positional information. The challenge with short read lengths is the bioinformatics exercise of examining the data in an intelligent way, and then organizing the information that results from the analysis. In a complex genome, with high repetitive sequence content, it is likely that a large number of reads will not be placed uniquely due to the degenerate nature of repeats.

Our ability to estimate the cost of resequencing a complex genome with highly parallel methods is still somewhat hindered by the fact that an entire project, starting from isolating genomic DNA to complete read assembly, has not yet been attempted. The pyrosequencing instrument from 454 Life Sciences can be used as a basis for an economic estimate for these methods. The cost of this instrument is approximately $500 000, a somewhat higher figure than current CAE instruments (~$350 000). For the production of raw sequence data, the instrument can produce about 20 Mb of data for about $10 000 (this cost includes reagents, consumables, labor, and device depreciation) (E. R. Mardis, personal communication, 2006) resulting in a raw sequencing cost of $0.0005 *per* base. Thus, for 10 × coverage of the human genome, the cost of producing the raw data would be $15 million. However, since these instruments produce short read lengths without paired-end reads, 30 × coverage may ultimately be required for better genome assembly, leading to an increase in the raw sequencing costs to $45 million. Additional costs could be incurred for genome assembly and finishing; these costs will be higher than electrophoresis-based sequencing methods as the shorter read lengths will produce more gaps in the assembled sequence for repeat-rich genomes. Reduction of these costs remains a major challenge for this technology if broader applicability to the human or other complex genomes is to become a reality.

6 Conclusions and projections for the future

New sequencing technologies have been discussed in the context of their applicability to *de novo* genome sequencing and resequencing. Many challenges for these methods exist that must be met prior to their incorporation into the production pipelines of genome sequencing centers, but future scientific and engineering developments may reduce the current technical, practical, and economic limitations. Some of the technical challenges have been laid out for the various approaches; while devices based on highly parallel systems can produce unprecedented amounts of data quickly, the shortness of individual read lengths, and in some cases, the lack of paired-end reads, limits their application to the sequencing of small and less complex genomes or genome regions. Microfluidic sequencing devices utilizing electrophoretic DNA separations, on the other hand, may provide the longer read lengths produced by CAE instruments, but integration challenges as well as the current limitations in cloning bias and repeat region assembly remain and may or may not be addressable with future technological advances.

While this review has focused mainly on technical aspects of the advancements and challenges in developing new genomic sequencing strategies, an urgent need to reduce the overall cost of DNA sequencing is the driving force for new technology development. Current sequencing costs are targeted for a reduction by two orders of magnitude in the near future, while a four-orders-of-magnitude decrease is the ultimate goal of the NHGRI [7a]. Although microfabricated systems are inherently designed to reduce DNA sequencing costs through miniaturization, no microchip-based instrument currently exists in the commercial sector that can sequence even a simple genome. Even though a massively parallel system from 454 Life Sciences has been utilized to fully sequence small and noncomplex genomes, the total cost for a mammalian genome such as the human would be prohibitively higher than the current CAE-based systems. Even though both approaches will undoubtedly continue to drop in cost, realistic replacements for these proven and established technologies may still be many years away.

Prior to any technology achieving the requisite two- to four-orders-of-magnitude reduction in cost, new devices and sequencing strategies may continue to provide incremental cost reductions. It is likely that most of the novel platforms will be tested extensively during the upcoming few years, and sequencing centers will combine both improved electrophoresis-based instruments and massively parallel systems into their operation. In fact, the integrated use of different sequencing technologies may be the key to future reductions in the cost of genomic sequencing. While it is currently unclear which technology(ies) will ultimately be used in genome centers in the future, all technologies and research fields involved in the development of these technologies remain important and should continue to be supported toward the worthwhile goal of extremely low-cost DNA sequencing. Clearly, if large and complex genomes such as an individual human genome could be sequenced essentially "for free", our society would be completely transformed, yielding great benefits to humankind. While this utopian vision is rather far from being realized, the continuing evolution of DNA sequencing technology remains fascinating to follow and of tremendous importance.

This publication was made possible by grant no. 5 R01 HG001970-06 from the National Human Genome Research Institute. However, the views expressed in this paper are solely the responsibility of the authors and do

not necessarily represent the official views of the National Human Genome Research Institute, or the National Institutes of Health.

7 References

[1] Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C. *et al.*, *Nature* 2001, *409*, 860–921.

[2] Venter, J. C., Adams, M. D., Myers, E. W., Li, P. W. *et al.*, *Science* 2001, *291*, 1304–1351.

[3] Collins, F. S., Lander, E. S., Rogers, J., Waterston, R. H., *Nature* 2004, *431*, 931–945.

[4] Gonzalez, E., Kulkarni, H., Bolivar, H., Mangano, A. *et al.*, *Science* 2005, *307*, 1434–1440.

[5] Collins, F. S., McKusick, V. A., *JAMA* 2001, *285*, 540–544.

[6] Wolfsberg, T. G., Wetterstrand, K. A., Guyer, M. S., Collins, F. S., Baxevanis, A. D., *Nat. Genet.* 2002, *32*, 4–79.

[7] Mikkelsen, T. S., Hillier, L. W., Eichler, E. E., Zody, M. C. *et al.*, *Nature* 2005, *437*, 69–87.

[8] Newman, T. L., Tuzun, E., Morrison, V. A., Hayden, K. E. *et al.*, *Genome Res.* 2005, *15*, 1344–1356.

[9] Zhao, S. Y., Shetty, J., Hou, L. H., Delcher, A. *et al.*, *Genome Res.* 2004, *14*, 1851–1860.

[9a] NIH News Release, NHGRI Seeks Next Generation of Sequencing Technologies, 10 October 2004, http://genome.gov/12513210.

[10] Kruglyak, L., Nickerson, D. A., *Nat. Genet.* 2001, *27*, 234–236.

[11] Check, E., *Nature* 2005, *437*, 1084–1086.

[12] Venter, J. C., Adams, M. D., Sutton, G. G., Kerlavage, A. R. *et al.*, *Science* 1998, *280*, 1540–1542.

[13] Shapiro, J. A., von Sternberg, R., *Biol. Rev.* 2005, *80*, 227–250.

[14] Kapranov, P., Drenkow, J., Cheng, J., Long, J. *et al.*, *Genome Res.* 2005, *15*, 987–997.

[15] Cheng, Z., Ventura, M., She, X. W., Khaitovich, P. *et al.*, *Nature* 2005, *437*, 88–93.

[16] Meldrum, D., *Genome Res.* 2000, *10*, 1081–1092.

[17] Meldrum, D., *Genome Res.* 2000, *10*, 1288–1303.

[17a] NIH News Release, International Consortium Completes Human Genome Project, 14 April 2003, http://www.genome.gov/11006929.

[18] Fleischmann, R. D., Adams, M. D., White, O., Clayton, R. A. *et al.*, *Science* 1995, *269*, 496–512.

[19] Sanger, F., Nicklen, S., Coulson, A. R., *Proc. Natl. Acad. Sci. USA* 1977, *74*, 5463–5467.

[20] Ruiz-Martinez, M. C., Salas-Solano, O., Carrilho, E., Kotler, L., Karger, B. L., *Anal. Chem.* 1998, *70*, 1516–1527.

[21] Salas-Solano, O., Ruiz-Martinez, M. C., Carrilho, E., Kotler, L., Karger, B. L., *Anal. Chem.* 1998, *70*, 1528–1535.

[22] Ewing, B., Green, P., *Genome Res.* 1998, *8*, 186–194.

[23] Ewing, B., Hillier, L., Wendl, M. C., Green, P., *Genome Res.* 1998, *8*, 175–185.

[24] McPherson, J. D., Marra, M., Hillier, L., Waterston, R. H. *et al.*, *Nature* 2001, *409*, 934–941.

[25] Peltola, H., Soderlund, H., Ukkonen, E., *Nucleic Acids Res.* 1984, *12*, 307–321.

[26] Sutton, G. G., White, O., Admas, M. D., Kerlavage, A. R., *Genome Sci. Technol.* 1995, *1*, 9–19.

[27] Huang, X. Q., Madan, A., *Genome Res.* 1999, *9*, 868–877.

[28] Pevzner, P. A., Tang, H. X., Waterman, M. S., *Proc. Natl. Acad. Sci. USA* 2001, *98*, 9748–9753.

[29] Pop, M., Salzberg, S. L., Shumway, M., *Computer* 2002, *35*, 47–54.

[30] Pop, M., Phillippy, A., Delcher, A. L., Salzberg, S. L., *Brief. Bioinform.* 2004, *5*, 237–248.

[31] Chan, E. Y., *Mutat. Res.-Fundam. Mol. Mech. Mutagen.* 2005, *573*, 13–40.

[32] Paegel, B. M., Blazej, R. G., Mathies, R. A., *Curr. Opin. Biotechnol.* 2003, *14*, 42–50.

[33] Ugaz, V. M., Elms, R. D., Lo, R. C., Shaikh, F. A., Burns, M. A., *Philos. Transact. A Math. Phys. Eng. Sci.* 2004, *362*, 1105–1129.

[34] Kan, C. W., Fredlake, C. P., Doherty, E. A. S., Barron, A. E., *Electrophoresis* 2004, *25*, 3564–3588.

[35] Woolley, A. T., Hadley, D., Landre, P., deMello, A. J. *et al.*, *Anal. Chem.* 1996, *68*, 4081–4086.

[36] Burns, M. A., Mastrangelo, C. H., Sammarco, T. S., Man, F. P. *et al.*, *Proc. Natl. Acad. Sci. USA* 1996, *93*, 5556–5561.

[37] Burns, M. A., Johnson, B. N., Brahmasandra, S. N., Handique, K. *et al.*, *Science* 1998, *282*, 484–487.

[38] Waters, L. C., Jacobson, S. C., Kroutchinina, N., Khandurina, J. *et al.*, *Anal. Chem.* 1998, *70*, 158–162.

[39] Waters, L. C., Jacobson, S. C., Kroutchinina, N., Khandurina, J. *et al.*, *Anal. Chem.* 1998, *70*, 5172–5176.

[40] Lagally, E. T., Simpson, P. C., Mathies, R. A., *Sens. Actuators B* 2000, *63*, 138–146.

[41] Khandurina, J., McKnight, T. E., Jacobson, S. C., Waters, L. C. *et al.*, *Anal. Chem.* 2000, *72*, 2995–3000.

[42] Dunn, W. C., Jacobson, S. C., Waters, L. C., Kroutchinina, N. *et al.*, *Anal. Biochem.* 2000, *277*, 157–160.

[43] Lagally, E. T., Emrich, C. A., Mathies, R. A., *Lab Chip* 2001, *1*, 102–107.

[44] Lagally, E. T., Medintz, I., Mathies, R. A., *Anal. Chem.* 2001, *73*, 565–570.

[45] Ferrance, J. P., Wu, Q. R., Giordano, B., Hernandez, C. *et al.*, *Anal. Chim. Acta* 2003, *500*, 223–236.

[46] Koh, C. G., Tan, W., Zhao, M. Q., Ricco, A. J., Fan, Z. H., *Anal. Chem.* 2003, *75*, 4591–4598.

[47] Lagally, E. T., Scherer, J. R., Blazej, R. G., Toriello, N. M. *et al.*, *Anal. Chem.* 2004, *76*, 3162–3170.

[48] Pal, R., Yang, M., Lin, R., Johnson, B. N. *et al.*, *Lab Chip* 2005, *5*, 1024–1032.

[49] Legendre, L. A., Bienvenue, J. M., Roper, M. G., Ferrance, J. P., Landers, J. P., *Anal. Chem.* 2006, *78*, 1444–1451.

[50] Giordano, B. C., Ferrance, J., Swedberg, S., Huhmer, A. F. R., Landers, J. P., *Anal. Biochem.* 2001, *291*, 124–132.

[51] Wolfe, K. A., Breadmore, M. C., Ferrance, J. P., Power, M. E. *et al.*, *Electrophoresis* 2002, *23*, 727–733.

[52] Paegel, B. M., Yeung, S. H. I., Mathies, R. A., *Anal. Chem.* 2002, *74*, 5092–5098.

[53] Blazej, R. G., Kumaresan, P., Mathies, R. A., *Proc. Natl. Acad. Sci. USA* 2006, *103*, 7240–7245.

[54] Tan, H. D., Yeung, E. S., *Anal. Chem.* 1997, *69*, 664–674.

[55] Tan, H. D., Yeung, E. S., *Anal. Chem.* 1998, *70*, 4044–4053.

[56] Grover, W. H., Skelley, A. M., Liu, C. N., Lagally, E. T., Mathies, R. A., *Sens. Actuators B* 2003, *89*, 315–323.

[57] Paegel, B. M., Hutt, L. D., Simpson, P. C., Mathies, R. A., *Anal. Chem.* 2000, *72*, 3030–3037.

[58] Woolley, A. T., Mathies, R. A., *Anal. Chem.* 1995, *67*, 3676–3680.

[59] Liu, S. R., Shi, Y. N., Ja, W. W., Mathies, R. A., *Anal. Chem.* 1999, *71*, 566–573.

[60] Paegel, B. M., Emrich, C. A., Weyemayer, G. J., Scherer, J. R., Mathies, R. A., *Proc. Natl. Acad. Sci. USA* 2002, *99*, 574–579.

[61] Liu, S. R., Ren, H. J., Gao, Q. F., Roach, D. J. *et al.*, *Proc. Natl. Acad. Sci. USA* 2000, *97*, 5369–5374.

[62] Schmalzing, D., Adourian, A., Koutny, L., Ziaugra, L. *et al.*, *Anal. Chem.* 1998, *70*, 2303–2310.

[63] Salas-Solano, O., Schmalzing, D., Koutny, L., Buonocore, S. *et al.*, *Anal. Chem.* 2000, *72*, 3129–3137.

[64] Koutny, L., Schmalzing, D., Salas-Solano, O., El-Difrawy, S. *et al.*, *Anal. Chem.* 2000, *72*, 3388–3391.

[65] Llopis, S. D., Stryjewski, W., Soper, S. A., *Electrophoresis* 2004, *25*, 3810–3819.

[66] Shi, Y. N., Anderson, R. C., *Electrophoresis* 2003, *24*, 3371–3377.

[67] Sassi, A. P., Barron, A., AlonsoAmigo, M. G., Hion, D. Y. *et al.*, *Electrophoresis* 1996, *17*, 1460–1469.

[68] Zhou, H. H., Miller, A. W., Sosic, Z., Buchholz, B. *et al.*, *Anal. Chem.* 2000, *72*, 1045–1052.

[69] Hjertén, S., *J. Chromatogr.* 1985, *347*, 191–198.

[70] Bi, L. R., Kim, D. H., Ju, J. Y., *J. Am. Chem. Soc.* 2006, *128*, 2542–2543.

[71] Aksyonov, S. A., Bittner, M., Bloom, L. B., Reha-Krantz, L. J. *et al.*, *Anal. Biochem.* 2006, *348*, 127–138.

[72] Li, Z. M., Bai, X. P., Ruparel, H., Kim, S. *et al.*, *Proc. Natl. Acad. Sci. USA* 2003, *100*, 414–419.

[73] Bai, X. P., Li, Z. M., Jockusch, S., Turro, N. J., Ju, J. Y., *Proc. Natl. Acad. Sci. USA* 2003, *100*, 409–413.

[74] Seo, T. S., Bai, X. P., Ruparel, H., Li, Z. M. *et al.*, *Proc. Natl. Acad. Sci. USA* 2004, *101*, 5488–5493.

[75] Ruparel, H., Bi, L. R., Li, Z. M., Bai, X. P. *et al.*, *Proc. Natl. Acad. Sci. USA* 2005, *102*, 5932–5937.

[76] Levene, M. J., Korlach, J., Turner, S. W., Foquet, M. *et al.*, *Science* 2003, *299*, 682–686.

[77] Chan, E. Y., *US Patent 6210896*, 1999.

[78] Seo, T. S., Bai, X. P., Kim, D. H., Meng, Q. L. *et al.*, *Proc. Natl. Acad. Sci. USA* 2005, *102*, 5926–5931.

[79] Kartalov, E. P., Quake, S. R., *Nucleic Acids Res.* 2004, *32*, 2873–2879.

[80] Hyman, E. D., *Anal. Biochem.* 1988, *174*, 423–436.

[81] Gharizadeh, B., Nordstrom, T., Ahmadian, A., Ronaghi, M., Nyren, P., *Anal. Biochem.* 2002, *301*, 82–90.

[82] Ronaghi, M., Uhlen, M., Nyren, P., *Science* 1998, *281*, 363–365.

[83] Margulies, M., Egholm, M., Altman, W. E., Attiya, S. *et al.*, *Nature* 2005, *437*, 376–380.

[84] Dressman, D., Yan, H., Traverso, G., Kinzler, K. W., Vogelstein, B., *Proc. Natl. Acad. Sci. USA* 2003, *100*, 8817–8822.

[85] Shendure, J., Porreca, G. J., Reppas, N. B., Lin, X. X. *et al.*, *Science* 2005, *309*, 1728–1732.

[86] Butz, J. A., Yan, H., Mikkilineni, V., Edwards, J. S., *BMC Genet.* 2004, *5*, 1–5.

[87] Mikkilineni, V., Mitra, R. D., Merritt, J., DiTonno, J. R. *et al.*, *Biotechnol. Bioeng.* 2004, *86*, 117–124.

[88] Mitra, R. D., Church, G. M., *Nucleic Acids Res.* 1999, *27*, e34.

[89] Kasianowicz, J. J., Brandin, E., Branton, D., Deamer, D. W., *Proc. Natl. Acad. Sci. USA* 1996, *93*, 13770–13773.

[90] Song, L. Z., Hobaugh, M. R., Shustak, C., Cheley, S. *et al.*, *Science* 1996, *274*, 1859–1866.

[91] Meller, A., Nivon, L., Brandin, E., Golovchenko, J., Branton, D., *Proc. Natl. Acad. Sci. USA* 2000, *97*, 1079–1084.

[92] Meller, A., Branton, D., *Electrophoresis* 2002, *23*, 2583–2591.

[93] Akeson, M., Branton, D., Kasianowicz, J. J., Brandin, E., Deamer, D. W., *Biophys. J.* 1999, *77*, 3227–3233.

[94] Deamer, D. W., Akeson, M., *Trends Biotechnol.* 2000, *18*, 147–151.

[95] Deamer, D. W., Branton, D., *Acc. Chem. Res.* 2002, *35*, 817–825.

[96] Howorka, S., Cheley, S., Bayley, H., *Nat. Biotechnol.* 2001, *19*, 636–639.

[97] Saleh, O. A., Sohn, L. L., *Nano Lett.* 2003, *3*, 37–38.

[98] Mara, A., Siwy, Z., Trautmann, C., Wan, J., Kamme, F., *Nano Lett.* 2004, *4*, 497–501.

[99] Li, J., Stein, D., McMullan, C., Branton, D. *et al.*, *Nature* 2001, *412*, 166–169.

[100] Li, J. L., Gershow, M., Stein, D., Brandin, E., Golovchenko, J. A., *Nat. Mater.* 2003, *2*, 611–615.

[101] Fologea, D., Uplinger, J., Thomas, B., McNabb, D. S., Li, J. L., *Nano Lett.* 2005, *5*, 1734–1737.

[102] Chen, P., Gu, J. J., Brandin, E., Kim, Y. R. *et al.*, *Nano Lett.* 2004, *4*, 2293–2298.

[103] Chang, H., Kosari, F., Andreadakis, G., Alam, M. A. *et al.*, *Nano Lett.* 2004, *4*, 1551–1556.

[104] Fan, R., Karnik, R., Yue, M., Li, D. Y. *et al.*, *Nano Lett.* 2005, *5*, 1633–1637.

[105] Gracheva, M. E., Xiong, A. L., Aksimentiev, A., Schulten, K. *et al.*, *Nanotechnology* 2006, *17*, 622–633.

[106] Smeets, R. M. M., Keyser, U. F., Krapf, D., Wu, M. Y. *et al.*, *Nano Lett.* 2006, *6*, 89–95.

[107] Fologea, D., Gershow, M., Ledden, B., McNabb, D. S. *et al.*, *Nano Lett.* 2005, *5*, 1905–1909.

[108] Lagerqvist, J., Zwolak, M., Di Ventra, M., *Nano Lett.* 2006, *6*, 779–782.

[109] Chaisson, M., Pevzner, P., Tang, H. X., *Bioinformatics* 2004, *20*, 2067–2074.

[110] Whiteford, N., Haslam, N., Weber, G., Prugel-Bennett, A. *et al.*, *Nucleic Acids Res.* 2005, *33*, e171–e171(1).