Daniel G. Hert[1]
Christopher P. Fredlake[1]
Annelise E. Barron[1,2]*

[1]Department of Chemical and Biological Engineering, Northwestern University, Evanston IL, USA
[2]Department of Bioengineering, Stanford University, Stanford, CA, USA

Received July 14, 2008
Revised August 29, 2008
Accepted August 29, 2008

Review

Advantages and limitations of next-generation sequencing technologies: A comparison of electrophoresis and non-electrophoresis methods

The reference human genome provides an adequate basis for biological researchers to study the relationship between genotype and the associated phenotypes, but a large push is underway to sequence many more genomes to determine the role of various specificities among different individuals that control these relationships and to enable the use of human genome data for personalized and preventative healthcare. The current electrophoretic methodology for sequencing an entire mammalian genome, which includes standard molecular biology techniques for genomic sample preparation and the separation of DNA fragments using capillary array electrophoresis, remains far too expensive ($5 million) to make genome sequencing ubiquitous. The National Human Genome Research Institute has put forth goals to reduce the cost of human genome sequencing to $100 000 in the short term and $1000 in the long term to spur the innovative development of technologies that will permit the routine sequencing of human genomes for use as a diagnostic tool for disease. Since the announcement of these goals, several companies have developed and released new, non-electrophoresis-based sequencing instruments that enable massive throughput in the gathering of genomic information. In this review, we discuss the advantages and limitations of these new, massively parallel sequencers and compare them with the currently developing next generation of electrophoresis-based genetic analysis platforms, specifically microchip electrophoresis devices, in the context of three distinct types of genetic analysis.

Keywords:
DNA sequencing / Genome sequencing / Massively parallel / Microchip electrophoresis
DOI 10.1002/elps.200800456

1 Introduction

The draft [1, 2] and finished [3] genomic DNA sequences resulting from the completion of the Human Genome Project were composed of genetic information from a collection of several individuals and have made a wealth of information accessible to biological researchers to use and expand upon in various fields, for projects such as recognizing disease susceptibility [4], exploring mechanisms of human evolution [5], and developing pharmaceuticals [6]. While having the reference sequence for ''the'' human genome may aid in understanding the correlation between DNA sequence and associated phenotypes, there would clearly be substantial scientific and medical benefits derived from obtaining the detailed, highly accurate sequences of many more individual human genomes, allowing researchers to fully understand these relationships and to develop truly personalized medicine by using the particular genome sequence of a patient as a diagnostic tool to help prevent and treat illness [7, 8].

Toward this end, the National Institutes of Health (NIH), specifically the NIH's National Human Genome Research Institute (NHGRI), has actively sponsored researchers with $99.5 million of external funding to develop sequencing technologies since the completion of the Human Genome Project [9–12]. These funds were disbursed to aid in developing the next generation of sequencing instruments, which aim at sequencing the full human genome at the much lower costs of $100 000 initially and then as little as $1000, as the new technologies mature.

Correspondence: Professor Annelise E. Barron, Department of Bioengineering, Stanford University, W300B James H. Clark Center, 318 Campus Drive, Stanford, CA 94305-5444, USA
E-mail: aebarron@stanford.edu
Fax: +1-650-723-9801

Abbreviations: CAE, capillary array electrophoresis; **HLA,** human leukocyte antigen; **NHGRI,** National Human Genome Research Institute

*Current address: Department of Bioengineering, Stanford University, Stanford, CA 94305, USA

© 2008 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim

These allocated research funds have provided grants to both academic groups and corporations for developing promising technologies. However, the majority of recent scientific breakthroughs toward developing these new technologies have been made by independent corporations.

One approach to reducing DNA sequencing costs has been to increase throughput by the use of massively parallel DNA sequencing systems that can determine the sequence of huge numbers of different DNA strands at one time [13, 14]. These systems allow millions of ''reads'' (contiguous regions of DNA sequence) to be gathered in a single experiment, in contrast to current capillary electrophoresis instruments, the throughput of which is limited by a 96-capillary array like the ones used to sequence the first human genome (currently~700 high-quality bases/capillary/hour). The first of these new, massively parallel sequencing systems arrived in the commercial marketplace only within the last few years and were designed and produced by 454 Life Sciences, Illumina, and Applied Biosystems [15–18]. Since then, a flurry of new genomic research projects have been undertaken that utilize the tremendous throughput advantages of these systems [19–29]. However, to date it seems that these new approaches are not yet much lower in cost than electrophoresis technologies, when everything is considered. This review will provide an overview of each of the new or ''next generation'' DNA sequencing instruments, with their different technological approaches, that are currently available commercially and will compare their cost and effectiveness with those of traditional Sanger-based sequencing by electrophoresis in the context of various, specific DNA sequencing applications. This field is currently developing and evolving with extreme rapidity, which means that certain quantitative information (*e.g.* cost estimates) that we provide in this review could be out of date rather quickly. However, it is not just cost that is worth discussing with regard to these new technologies and how they compare with the current best and best potential electrophoresis instruments. Our broader aim is to convey to the general reader, who may not be an expert in next-generation sequencing technologies, our understanding of the current state of sequencing technologies. What is written here is based primarily on peer-reviewed studies that have been published, which we feel is the best approach; we have cited only non-peer-reviewed sources when absolutely necessary because a lack of other sources and these instances are clearly mentioned.

2 Next-generation sequencing technologies

2.1 Pyrosequencing

The sequencer developed by 454 Life Sciences is based on a sequencing-by-synthesis technique called pyrosequencing [30–34]. This method does not utilize chemically bound fluorophores to detect incorporated bases of DNA. Rather, as a nucleotide is incorporated into the growing DNA strand, pyrophosphate is released, which is then enzymatically converted to ATP. When the ATP comes into contact with the enzyme luciferase, light is produced. dNTPs are added individually and sequentially to the growing DNA molecules, such that each flash of light signaling the incorporation of a nucleotide can be correlated with which specific nucleotide was incorporated. Drawbacks of this technology include a requirement for rigorous sample washing between nucleotide additions, the need for new enzyme addition with each nucleotide addition, and the fact that the signal intensity must be correlated with the number of bases incorporated, which proves problematic for the sequencing of homopolymeric regions that are greater than three bases in length.

The first instrument created using this technology, 454's GS20 instrument, was shown to be able to read up to 25 million bases of a bacterial genome in a single four-hour run [17]. The authors used emulsion PCR on 28-μm beads to amplify the genomic DNA. The beads were deposited with 30% efficiency on a glass Picotitre Plate [35] with approximately 1.6 million wells, a well center-to-center distance of 50 μm, and a well depth of 55 μm. The GS20 was able to obtain average sequencing read lengths of 110 bases with a raw read accuracy of 96%.

454 has since released a second, improved sequencer, the GS-FLX, which is able to obtain average read lengths of 250 bases and is able to perform mate-paired reads. According to the listed specifications for the GS-FLX system (http://www.454.com), the throughput specifications indicate that an average of 100 million DNA bases can be sequenced in a 7.5-h run. This instrument has been used to perform the resequencing of an entire human genome at 7.4 × coverage, reportedly for less than $1 million (although exactly how this calculation was made is not clear); and with a raw base accuracy of 99.5% [36]. However, based on current list prices for the GS-FLX system, the current cost for all the reagents, including the picotiter plate, library preparation kits, and emulsion PCR kits, to perform a single experiment is~$7000. Therefore, for each level of coverage of the 3 gigabases (Gb) human genome, an approximate cost of $210 000 is expected. For the human genome resequencing project completed at stated coverage, the expected cost should approach $1.54 million for a perfectly efficient process in which no analyses need to be repeated. However, deviations from the list price of reagents and other consumables, throughput higher than currently reported, or the specific manner of accounting various costs may account for this discrepancy.

2.2 Fluorescently labeled sequencing by synthesis

The massively parallel Genome Analyzer system developed by Illumina is also based on sequencing by synthesis [15]. Unlike 454 Life Sciences, fluorescently labeled, reversible nucleotide terminator chemistry is utilized on this platform,

© 2008 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim

similar to the sequencing-by-synthesis technologies described previously [37–39]. This system utilizes a dense array of small adapter molecules covalently bound to a glass surface. A small percentage of these adapters are also covalently bound to a DNA template. The tethered fragments are then replicated using several rounds of PCR to form a very dense cluster of the DNA templates tethered to adapters. The sequencing process begins with the addition of sequencing primers, fluorescently labeled reversible dNTP terminators, and DNA polymerase. After the corresponding complementary base has been incorporated into the first position, the polymerization is temporarily halted because of the 3′ terminating group attached to the incorporated dNTP. The unincorporated reagents are then washed away, a laser is used to excite the bound fluorescent labels, and the signal is recorded. The fluorescent label and the terminating group attached to the incorporated base are then removed, allowing for further extension of the DNA fragment. The DNA extension reaction is reinitiated by adding more labeled dNTPs and DNA polymerase to extend the growing fragment by one modified base. The reagents are washed away and the second position is interrogated by laser excitation of the fluorophore. The process is then repeated for 36–50 cycles, depending on the DNA library utilized.

According to product information available from Illumina (http://www.illumina.com), this system is capable of read lengths of up to 50 bases for fragment libraries and 36 bases for mate-paired libraries, with a raw base-calling accuracy of 98.5%. In addition, the reported throughput for this instrument is approximately 3 Gb *per* run and requires 5 days for each run. This throughput corresponds to approximately 7000 bases *per* second. Based on the listed reagent costs, each run of the instrument costs approximately $6300 (4800 bases for one cent).

2.3 Sequencing by hybridization and ligation

The massively parallel sequencing system developed by Applied Biosystems (ABI), called Sequencing by Oligonucleotide Ligation and Detection (SOLiD^{TM}), is based on a hybridization and ligation chemistry scheme [40]. This instrument uses a set of fluorescently labeled hybridization probes that each comprise eight bases, with the first three being degenerate bases and the final three being ''universal'' bases that are always the same. The fourth and fifth bases together are one of the 16 possible dinucleotide sequences and are responsible for the specificity of hybridization. All 16 probes are introduced to the system and hybridization occurs based on complementary base-pairing. A properly (fully) hybridized probe can be ligated to the primer by DNA ligase in the system. The incorporated probes are then identified by the detected fluorescence from excitation of the attached label. Cleavage then occurs after the fifth base of the probe, removing the three universal bases and the fluorescent label, and then the next hybridization process begins. The DNA bases are therefore always interrogated at positions five bases apart. After seven hybridization and ligation extensions, the probes are removed from the templates and a new round of sequencing begins with primers one base closer to the beads. This process is repeated until all bases of the template have been identified. This method leads to sequencing each position twice, enabling a reported 99.94% raw read accuracy, and a capping step that prevents any unligated strands in a round from undergoing hybridization and ligation in subsequent rounds should be able to eliminate dephasing.

No peer-reviewed publications are available that describe in detail the steps that are followed in DNA sequencing using the SOLiD^{TM} system. However, based on the information available from Applied Biosystems (http://solid.appliedbiosystems.com), this method allows for the sequencing of both fragment libraries (with read lengths up to 35 bases) and mate-paired libraries (with read lengths up to 25 bases). This system utilizes 1-μm beads that have DNA fragments covalently linked to the surface. The beads are then deposited on glass slides and covalently bound to the surface with a density of greater than 300 million beads *per* slide. After deposition of the beads, the sequential hybridization-based DNA sequencing process begins using the hybridization fragments or probes. For a mate-paired library, the average time required to run the instrument once is 10 days. At a throughput of 6 Gb *per* run, this translates to approximately 7000 bases *per* second.

In March 2008, Applied Biosystems issued a press release detailing a study performed by their researchers aimed at resequencing the genome of an anonymous Nigerian male, which they said cost $60 000 (http://press.appliedbiosystems.com). However, this study has not been published in a peer-reviewed journal; hence, the details are not publicly known. According to the press release, the sequencer performed seven runs to sequence 36 Gb (for 12× coverage). At the currently listed prices for reagents, the cost of a single run is approximately $7700 (7800 bases for one cent), which agrees reasonably well with the cost estimation given by Applied Biosystems for their resequencing effort.

2.4 Microchip electrophoresis-based sequencing

The current workhorse technology for genome sequencing centers has been and at the moment remains capillary array electrophoresis (CAE) instruments [41–44]. These systems utilize replaceable polymer networks to perform the separation of fluorescently labeled DNA sequencing fragments generated by the Sanger cycle sequencing reaction and provide nearly completely automated size-based separation and analysis of the ssDNA fragments. These CAE machines use bundled capillary arrays to perform the simultaneous electrophoretic separation of sequencing ladders in 16–96 capillaries in parallel. The separation

© 2008 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim

process, including polymer matrix formulation, temperature, and denaturant, has since been optimized such that it provides very long sequencing read lengths (>700 bases) in 1–2 h of electrophoresis [45, 46]. However, the cost reductions necessary to meet the goals put forth by the NHGRI are unlikely to come from advances in capillary electrophoresis technology; the preparation of one sample for each capillary is a very expensive step, proportionally. The introduction of an integrated microfluidic sequencing system, especially if a lower cost, higher throughput sample preparation method could be used, could reduce the required sample and reagent volumes, shorten analysis times due to shorter separation channels, and reduce the time needed for sample preparation. Such a system would increase throughput by streamlining the entire sequencing process.

Initial work toward developing integrated microfluidic devices focused on combining sample preparation steps with separation and detection by combining a microscale PCR chamber and separation microchannel on the same chip substrate [47–52], which has been advanced by the development of techniques to perform very rapid DNA amplification in microchannels using IR-mediated PCR [53, 54]. Other integrative work has focused on performing sample cleanup steps prior to electrophoresis and sample detection. The Mathies group has focused on developing protocols to purify DNA sequencing extension fragments between the amplification and separation steps [55–57]. In the most recent work, oligonucleotides complementary to the region of the sample directly 3′ of the primer site are covalently incorporated into a narrow region of the separation matrix near the cathodic end of the separation channel. The sample from the amplification step is then carried past this ''capture gel'' at low temperature, such that the extension products hybridize to the oligonucleotides while the buffer salts, unreacted primer, and free nucleotides that could prevent an efficient electrophoretic sample injection are carried to a waste well. The temperature of the system is then raised such that the captured fragments denature and electrophoresis is performed to separate the sequencing fragments. By demonstrating this ''DNA capture gel'' process, this group has shown that the detection of fluorescently labeled DNA sequencing fragments can be achieved down to attomolar concentrations, a unique and exciting contribution to the DNA sequencing field.

The Landers group at the University of Virginia has focused on integrating on-chip sample purification prior to DNA amplification and detection. The sample purification process that they have developed is a solid-phase extraction using silica beads, sol-gels, and silica/sol-gel hybrids [58–60] to remove genomic DNA from cell lysis products. The genetic material is then amplified, electrophoretically separated, and detected. In fact, the Landers group has shown that all of these processes can be achieved sequentially on a single microfluidic platform, making possible a system with ''sample-in, answer-out'' capability [61]. However, this group has not yet demonstrated the four-color sequencing of Sanger samples by utilizing this method.

Our group has recently reviewed these sample preparation methods and their potential for use in completely integrated microfluidic sequencing systems [62], as well as the potential for fine-tuning the chemical and physical properties of the polymer separation matrix [63–67] for more effective DNA sequencing separations. However, no group has yet developed a system that integrates all of the elements of the DNA sequencing process onto a single microfluidic platform to enable sequencing of entire complex genomes; this will be an exciting development when such a system is successfully created, especially for ''medical sequencing'', *i.e.* the rapid sequencing of specific human gene regions that have clinical significance.

3 Implications for genetic analysis applications

The NHGRI has set the target cost for sequencing the human genome at $100 000 in the short term and $1000 in the long term. The motivations for using the cost of a human genome as a benchmark are clear, since determining the entire genetic sequence of a human should also give the same information that can be gathered by other methods of genetic analysis, such as STR sizing and mutation detection *via* single-stranded conformation polymorphism, and with much more detail. In this section we discuss the advantages and disadvantages of using various new sequencing technologies for several different applications of genetic analysis.

3.1 *De novo* sequencing of the human genome

For *de novo* sequencing of the 3 Gb of the human genome (or, 6 Gb if one considers a diploid genome) with massively parallel sequencing technologies, the coverage requirements are expected to be much greater (>20 × coverage) than is needed for CAE due to the relatively short read lengths that these systems provide [68]. Thus, based on the calculated costs-*per*-base using the 454 GS-FLX instrument and estimating a necessary coverage of 30 ×, performing the raw reads required for *de novo* sequencing a 3 Gbp human genome would have a retail cost of approximately $6.1 million. Assuming similar coverage requirements are necessary using the Applied Biosystems and the Illumina sequencing instruments, performing the raw reads required for *de novo* sequencing of a human genome would cost approximately $80 K and $130 K, respectively.

The (admittedly very rough) prediction of these excitingly low values must be tempered by the fact that they are based on a number of assumptions. The coverage requirements were assumed to be the same for each technology, despite nearly an order of magnitude higher read lengths using instruments from 454 compared with those from ABI

© 2008 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim

and Illumina. Read length has been shown to be extremely important for the assembly of contigs and especially, for accurate final genome assembly when using these new technologies [68], which may indicate that much higher average coverage levels and finishing costs are necessary when using read lengths of 25–35 bases rather than read lengths of 250 bases, especially if the necessary contig overlap sequence approaches or surpasses the obtained read lengths. While mathematical distributions can be used to estimate the needed sequencing redundancy [69], the exact dependence of necessary coverage on read length is not known, and it may also be highly dependent on the type of genome assembly algorithm used and the accuracy of the reads, as well as how complex (or repetitive) various parts of the genome itself happen to be. The number of required sequencing runs, which largely determines the cost, is likely to be greatly affected by the determined minimum sequence coverage and the subsequent design of experiments, including using both fragment and mate-pair libraries, necessary to perform the genome sequencing and assembly. Even with extremely high raw coverage of each individual base, it may be extremely difficult and maybe impossible for very short reads to be correctly and fully assembled without mapping them to a reference genome [70, 71].

Sanger-based *de novo* sequencing of the human genome is being primarily performed by capillary sequencers run in parallel, since the introduction of a microfluidic electrophoresis platform for these types of analyses has not yet occurred. The current cost of sequencing a human genome is approximately $5 million, based on an NIH news release [12], which may or may not include other sequencing technologies in the calculation of this cost. The CAE method is the current standard due to its long read lengths (~600–900 bases), which enable relatively straightforward paths to genome assembly. However, the current methodology using capillary-based systems is too expensive for the routine collection of genome sequences. Electrophoresis-based Sanger sequencing has the potential to be streamlined by utilizing microfluidic platforms, which would lower reagent volumes, require smaller sample quantities [56], decrease analysis times [67], enable parallel sequencing separations [66], and integrate the individual steps of the sequencing process [57]. Since the sequencing steps are serialized, though, integrated microchip sequencers would be limited in throughput by the slowest step in the complete process. However, a microfluidic sequencing system should be able to obtain read lengths on the order of current CAE instruments in a fraction of the time, which should facilitate genome assembly and reduce the finishing costs relative to technologies with shorter read lengths.

3.2 Human leukocyte antigen (HLA) system compatibility testing

The HLA system is the region of the human genome that is responsible for encoding antigen-presenting proteins on cell surfaces, which are presented to T cells to prevent pathogens from infecting the body. The similarity of these sequences from one person to the next largely determines whether an organ transplant is rejected or succeeds based on a sufficient level of immunocompatibility. Thus, ascertaining (with ideally perfect accuracy) the DNA sequence of the HLA genes of a patient can help to determine *a priori* the viability of a prospective transplant [72]. The HLA system comprises an ~4 Mb region of human chromosome six that contains more than 200 genes [73, 74]. Of these 220 genes, 21 are highly polymorphic and contribute significantly to the outcome of organ transplants [73]. These polymorphic genes can be kilobases in length, with exons of 450 bases or less.

Technological requirements to make sequencing HLA a pervasive medical diagnostic tool include stringent accuracy of the sequencing of this region, speed, low costs, and virtually complete automation. A difference in one codon of the sequence in the HLA system could vastly alter the function of the resulting protein, which is critical in this clinical environment. To enable this tool to be widely used, the time from sample collection to receiving the results must be minimized and the sequencing capability must be near the clinic in which the information is used, preferably in the same hospital. For the instrument to be present in the hospital, sample automation is preferable to minimize excessive training requirements on hospital personnel.

Massively parallel sequencing systems could be effectively used for this application. While the instruments from 454, Applied Biosystems, and Illumina produce shorter read lengths, they have the ability to provide deep coverage of the sequences studied, which may allow for consensus sequencing of HLA for a patient, assuming feasible assembly of the read lengths obtained. Based on the throughput of each instrument and sequencing the entire 4 Mb HLA region, the expected coverage of one run from the GS-FLX is 25× (which is likely too low for such high accuracy requirements), from the SOLiD Sequencer is 750–1500×, and from the Genome Analyzer instrument is 750×. However, if the clinical need is to sequence a few polymorphic genes or exons of interest instead of the entire HLA region, the necessary assembled sequences change to the order of hundreds or thousands of bases instead of megabases. More coverage of these regions would then be possible, but a very small region of interest may lead to redundant oversampling and a more costly sequence. A possible solution in these instances would be to include several samples in the same run to reduce the cost *per* sample, which is theoretically possible due to the ability for each of the sequencing substrates to be physically subdivided (or in the case of the GS-FLX, using sample-identifying "barcodes" to eliminate the need for actual physical isolation of samples), but sample cross-contamination within the instrument or regulations imposed by a government agency, such as the FDA, may make this infeasible.

While all of the massively parallel instruments have the capacity to run "hands-free," each has intensive sample

© 2008 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim

preparation protocols, which would need a dedicated and well-trained staff to perform. The time required to run these instruments also inhibits their clinical acceptance. Since HLA compatibility is important for transplant cases, analysis of samples may be time-sensitive on the order of hours and the massively parallel systems require days to complete a run (sample preparation and instrument run time). The initial capital investment of the half a million dollars that each instrument costs may also prove infeasible for many clinical settings. Finally, if one run were performed using each of these instruments, the cost of the test can be approximated as simply the cost of the reagents. Therefore, the absolute minimum the test would cost using the GS-FLX is ~$7000, using the SOLiD Sequencer is ~$7500, and using the Genome Analyzer is ~$6300. Any application that would utilize these technologies would have such a minimum cost, which may be prohibitive in certain environments, specifically clinical settings. These costs could be lowered by running multiple samples at a time, but the challenge of limiting cross-contamination remains.

When comparing the cost of using electrophoresis-based sequencing to perform this test, scaling the current cost necessary to sequence the full human genome to the 4 Mb HLA region translates to a cost of ~$6700. The accuracy requirements for this application are higher than for the full genome, so the costs will most likely be higher than this value. However, for directed sequencing of individual genes or exons, the long read lengths of Sanger sequencing would require much lower coverage than the massively parallel sequencers and would enable more rapid assembly of contigs, lowering the overall costs. Furthermore, directed sequencing using a microfluidic electrophoresis instrument could be accomplished in a matter of hours, given the shorter times that will likely be required for sample isolation, PCR, and separation.

3.3 Human STR genotyping

Human identification by genetic methods was developed by observing characteristic sequence repeats in kilobase-sized ''minisatellites,'' repetitive regions that are present in each human genome, and comparing them to one another [75]. Since that time, methods have been developed to utilize shorter, but still highly polymorphic, microsatellites of the human genome for forensic DNA testing. These variable-length STR alleles, ranging in size from 50 to 500 bases, are most commonly amplified in a multiplex PCR and then the fragments are separated using capillary electrophoresis [76]. The samples are then genotyped by comparison with an allelic ladder standard included in the electrophoresis.

To use a massively parallel sequencing system for human identification, sequencing can be performed using the same, and possibly more, loci used for current STR analysis. The STR regions are repetitive by nature, which is one of the more difficult sample types for these systems to analyze properly due to short read lengths and possibly dephasing in the GS-FLX. Even with high degrees of coverage, assembly of these highly repetitive regions may prove very difficult. The accuracy requirements for these analyses are very rigorous due to their use in the legal system [77]. It is unclear whether multiple samples from multiple individuals could be run simultaneously on a single instrument without the possibility of cross-contamination. If multiple samples could not be run simultaneously, the minimum cost *per* run and time *per* run for the massively parallel sequencers would be prohibitively expensive, as in the HLA discussion, as would the capital cost of procuring one of these instruments for the resulting sample throughput.

Using electrophoresis-based methods is the current standard for performing genotyping assays. Several multiplexed loci kits are available for forensic genotyping, and most of these analyses are performed on capillary array instruments. Since these kits were designed for CAE, microsatellites were chosen such that many loci could be typed on a single run of the instrument with a multiplexed, multicolor system. As such, only a handful of runs may be necessary in order to obtain very accurate and reliable measurements, which would only necessitate one instrument. Since the cost of performing DNA separation by electrophoresis scales as the number of runs performed, the costs could be minimized by only running the absolute lowest number of experiments necessary for a certain error tolerance level. Furthermore, it has been previously shown that all 13 CODIS loci need not be interrogated to positively identify an individual [78], reducing the experimental time and cost requirements even further. Thus, electrophoresis holds a distinct cost and time advantage over massively parallel sequencing technologies for performing genotyping assays, owing to its highly scalable experimental cost and times.

4 Concluding remarks

Recent improvements in sequencing technology have yielded the advent of massively parallel DNA sequencing systems, which produce short read lengths (25–250 bases) at very high densities. The sequencing technologies from 454 Life Sciences, Applied Biosystems, and Illumina have been discussed, to the extent that detailed information is currently available, and compared to electrophoresis-based genetic analysis in the context of three separate case studies: *de novo* human genome sequencing, HLA system compatibility testing, and human STR genotyping. For large-scale sequencing applications, the newer, massively parallel systems hold distinct throughput and cost advantages. However, only one peer-reviewed article has been published detailing the work and results of performing a resequencing project, which has quite different specifications than a *de novo* genome sequencing project. The finishing costs of

© 2008 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim

obtaining a genome sequence *de novo* using a massively parallel sequencer are likely to be high, owing to difficulties inherent in assembling very short read lengths. However, it may prove most feasible and most cost-effective in the future to perform high-throughput sequencing on a large portion of the genome and also to perform directed sequencing using electrophoresis-based platforms on regions for which longer read lengths would ease assembly. This combined approach seems promising based on a recent study that describes the difficulty of successfully using massively parallel sequencing to identify copy number variations in the human genome, for repeats with sizes larger than 5 kbp, owing to the shorter read lengths obtained with these technologies and these instruments' inability to sequence and map within repeat-rich regions of the human genome [79]. Such considerations were also stated as the motivation for using CAE for the recent *de novo* sequencing and assembly of a different individual [80].

Directed genetic analyses such as HLA compatibility testing and forensic genotyping assays present separate challenges. Both are preferred to be inexpensive, easy to perform, and stringently accurate, with the potential for incorporation into acute clinical and perhaps military of disaster-site settings. Estimates of the costs for achieving consensus sequencing of the HLA complex are similar for both electrophoresis-based and non-electrophoresis-based platforms, but the desired high throughput would require several electrophoresis instruments to be used in parallel, increasing the overall capital cost, whereas a single massively parallel sequencer could be envisioned performing all of the sequencing for this 4 Mb region in a single run. For STR genotyping assays, very small, repetitive regions of the genome are interrogated. For these studies, as in the HLA scenario, one massively parallel run could be used to perform all of the analysis. If multiple samples are precluded from being genotyped on the same substrate, using one of these sequencers for only one sample may result in oversampling, leading to redundant coverage. The minimum cost for these experiments is fixed at the cost of reagents necessary for one sequencing run, and will likely be too expensive, especially for clinical use. Furthermore, massively parallel sequencers have an intrinsic limitation in sequencing highly repetitive regions, owing to the short read lengths produced by these systems. Electrophoresis-based platforms are capable of very long read lengths, which can easily and accurately read through repeat-rich sequences, and have an inherently scalable time and cost structure, based solely on the number of experiments run. Therefore, if only five runs are needed to provide accurate analysis, only five runs are performed with their associated costs.

We have shown that ''next-gen'', massively parallel DNA sequencers offer many potential benefits in performing genetic analyses, especially for large-scale projects. However, one of the major drawbacks that limits their use, especially in *de novo* sequencing, is the short sequencing read lengths they provide. A new technology, which combines the massive throughput of the next-gen sequencers with the long read lengths achieved by electrophoresis-based Sanger-sequencing, would enable rapid, high-quality production of *de novo* genome sequences. Two companies that are currently exploring single-molecule sequencing technologies with these goals in mind are Visigen and Pacific Biosciences [81]. Their technologies are similar to each other, with both Visigen [82–85] and Pacific Biosciences [86–90] using a single DNA polymerase molecule covalently bound to a glass surface. As a DNA molecule is synthesized, the identity of each incorporated base is recorded in real time. These methods aim to take advantage of the high synthesis rates of DNA polymerase and should be able to achieve read lengths limited only be the incorporation efficiency and fidelity of the DNA polymerase, which should be on the order of 1000 bases or probably more. These systems have great potential to impact the future of *de novo* human genome sequencing projects, but may have drawbacks for smaller scale genetic analysis applications that are similar to those of current massively parallel sequencers.

At the present time, human genome sequencing is still a great distance away from the lofty goal of a $1000 genome that was put forth by the NIH/NHGRI to pave the way for personalized medicine. The development of new technologies has certainly transformed how large-scale genetic research is done, but significant technological hurdles remain. Even if Visigen, Pacific Biosciences, or some other group succeeds in developing their technology to sequence millions of DNA sequences simultaneously with read lengths greater than 1000 bases, would the cost of a human genome finally be reduced to $1000? Even if only one run of the hypothetical instrument is necessary for sequencing the entire genome at required levels of coverage, a minimum cost for that single run exists. Such an instrument would be absolutely state of the art, and based on capitalized instrument cost as well, the cost for reagents and other consumables for the current massively parallel sequencers, the cost *per* run is likely to remain on the order of $10,000 or higher. While such a low price for a human genome would be amazing, it still does not reach the ultimate goal of $1000 *per* genome and likely will not until the technology fully matures and the cost of consumables is reduced significantly. Nevertheless, the spirit of the NHGRI goal of a $1000 genome is very much alive and is driving exciting innovations to make personalized genetics a reality that may affect us in our lifetime.

This work was made possible by grant # 2 R01 HG001970-07 from the NHGRI. However, the views expressed in this paper are solely the responsibility of the authors and do not necessarily represent the official views of the NHGRI or the NIH.

The authors have declared no financial or commercial conflict of interest.

© 2008 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim

5 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] Estivill, X., Armengol, L., *PLoS Genet.* 2007, *3*, 1787–1799.

[5] Uddin, M., Goodman, M., Erez, O., Romero, R. *et al.*, *Proc. Natl. Acad. Sci. USA* 2008, *105*, 3215–3220.

[6] Chin, K. V., Selvanayagam, Z. E., Vittal, R., Kita, T. *et al.*, *Drug Dev. Res.* 2004, *62*, 124–133.

[7] Ginsburg, G. S., Donahue, M. P., Newby, L. K., *J. Am. Coll. Cardiol.* 2005, *46*, 1615–1627.

[8] Garrison, L. P., Austin, M. J. F., *Health Aff.* 2006, *25*, 1281–1290.

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

[10] NIH News Release. NHGRI Expands Effort to Revolutionize Sequencing Technologies. August 2005 http://www.genome.gov/15015208

[11] NIH News Release. NHGRI Aims to Make DNA Sequencing Faster, More Cost Effective. October 2006 http://www.genome.gov/19518500

[12] NIH News Release. New Grants Bolster Efforts to Generate Faster and Cheaper Tools for DNA Sequencing. August 2007 http://www.genome.gov/25522229

[13] Rogers, Y. H., Venter, J. C., *Nature* 2005, *437*, 326–327.

[14] Hutchison, C. A., *Nucleic Acids Res.* 2007, *35*, 6227–6237.

[15] Bennett, S., *Pharmacogenomics* 2004, *5*, 433–438.

[16] Bennett, S. T., Barnes, C., Cox, A., Davies, L., Brown, C., *Pharmacogenomics* 2005, *6*, 373–382.

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

[18] Applied Biosystems Press Release. Applied Biosystems Surpasses Industry Milestone in Loweing the Cost of Sequencing Human Genom. March 12, 2008.

[19] Goldberg, S. M. D., Johnson, J., Busam, D., Feldblyum, T. *et al.*, *Proc. Natl. Acad. Sci. USA* 2006, *103*, 11240–11245.

[20] Oh, J. D., Kling-Backhed, H., Giannakis, M., Xu, J. *et al.*, *Proc. Natl. Acad. Sci. USA* 2006, *103*, 9999–10004.

[21] Smith, M. G., Gianoulis, T. A., Pukatzki, S., Mekalanos, J. J. *et al.*, *Genes Dev.* 2007, *21*, 601–614.

[22] Hiller, N. L., Janto, B., Hogg, J. S., Boissy, R. *et al.*, *J. Bacteriol.* 2007, *189*, 8186–8195.

[23] Mussmann, M., Hu, F. Z., Richter, M., de Beer, D. *et al.*, *PLoS Biol.* 2007, *5*, 1923–1937.

[24] Wicker, T., Schlagenhauf, E., Graner, A., Close, T. J. *et al.*, *BMC Genomics* 2006, *7*, 275.

[25] Green, R. E., Krause, J., Ptak, S. E., Briggs, A. W. *et al.*, *Nature* 2006, *444*, 330–336.

[26] Hillier, L. W., Marth, G. T., Quinlan, A. R., Dooling, D. *et al.*, *Nature Methods* 2008, *5*, 183–188.

[27] Van Tassell, C. P., Smith, T. P. L., Matukumalli, L. K., Taylor, J. F. *et al.*, *Nature Methods* 2008, *5*, 247–252.

[28] Baker, S., Holt, K., Whitehead, S., Goodhead, I. *et al.*, *Mol. Microbiol.* 2007, *66*, 1207–1218.

[29] Porreca, G. J., Zhang, K., Li, J. B., Xie, B. *et al.*, *Nature Methods* 2007, *4*, 931–936.

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

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

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

[33] Nyren, P., Pettersson, B., Uhlen, M., *Anal. Biochem.* 1993, *208*, 171–175.

[34] Ronaghi, M., Karamohamed, S., Pettersson, B., Uhlen, M., Nyren, P., *Anal. Biochem.* 1996, *242*, 84–89.

[35] Leamon, J. H., Lee, W. L., Tartaro, K. R., Lanza, J. R. *et al.*, *Electrophoresis* 2003, *24*, 3769–3777.

[36] Wheeler, D. A., Srinivasan, M., Egholm, M., Shen, Y. *et al.*, *Nature* 2008, *452*, 872–877.

[37] Ju, J. Y., Kim, D. H., Bi, L. R., Meng, Q. L. *et al.*, *Proc. Natl. Acad. Sci. USA* 2006, *103*, 19635–19640.

[38] Meng, Q. L., Kim, D. H., Bai, X. P., Bi, L. R. *et al.*, *J. Org. Chem.* 2006, *71*, 3248–3252.

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

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

[41] Kambara, H., Takahashi, S., *Nature* 1993, *361*, 565–566.

[42] Lu, X. D., Yeung, E. S., *Appl. Spectrosc.* 1995, *49*, 605–609.

[43] Anazawa, T., Takahashi, S., Kambara, H., *Anal. Chem.* 1996, *68*, 2699–2704.

[44] Crabtree, H. J., Bay, S. J., Lewis, D. F., Zhang, J. Z. *et al.*, *Electrophoresis* 2000, *21*, 1329–1335.

[45] Kotler, L., He, H., Miller, A. W., Karger, B. L., *Electrophoresis* 2002, *23*, 3062–3070.

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

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

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

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

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

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

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

[53] Roper, M. G., Easley, C. J., Legendre, L. A., Humphrey, J. A. C., Landers, J. P., *Anal. Chem.* 2007, *79*, 1294–1300.

© 2008 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim

[54] Easley, C. J., Humphrey, J. A. C., Landers, J. P., *J. Micromechanics Microengineering* 2007, *17*, 1758–1766.

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

[56] Blazej, R. G., Kumaresan, P., Cronier, S. A., Mathies, R. A., *Anal. Chem.* 2007, *79*, 4499–4506.

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

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

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

[60] Breadmore, M. C., Wolfe, K. A., Arcibal, I. G., Leung, W. K. *et al.*, *Anal. Chem.* 2003, *75*, 1880–1886.

[61] Easley, C. J., Karlinsey, J. M., Bienvenue, J. M., Legendre, L. A. *et al.*, *Proc. Natl. Acad. Sci. USA* 2006, *103*, 19272–19277.

[62] Fredlake, C. P., Hert, D. G., Mardis, E. R., Barron, A. E., *Electrophoresis* 2006, *27*, 3689–3702.

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

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

[67] Fredlake, C. P., Hert, D. G., Kan, C. W., Chiesl, T. N. *et al.*, *Proc. Natl. Acad. Sci. USA* 2008, *105*, 476–481.

[68] Sundquist, A., Ronaghi, M., Tang, H., Pevzner, P., Batzoglou, S., *PLoS ONE* 2007, *2*, e484.

[69] Lander, E. S., Waterman, M. S., *Genomics* 1988, *2*, 231–239.

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

[71] Whiteford, N., Haslam, N., Weber, G., Prugel-Bennett, A. *et al.*, *Nucleic Acids Res.* 2005, *33*, e171.

[72] Li, B. G., Hartono, C., Ding, R. C., Sharma, V. K. *et al.*, *N. Engl. J. Med.* 2001, *344*, 947–954.

[73] Robinson, J., Waller, M. J., Parham, P., Bodmer, J. G., Marsh, S. G. E., *Nucleic Acids Res.* 2001, *29*, 210–213.

[74] Robinson, J., Waller, M. J., Parham, P., de Groot, N. *et al.*, *Nucleic Acids Res.* 2003, *31*, 311–314.

[75] Jeffreys, A. J., Wilson, V., Thein, S. L., *Nature* 1985, *316*, 76–79.

[76] Butler, J. M., *Biotechniques* 2007, *43*, ii–v.

[77] Jobling, M. A., Gill, P., *Nat. Rev. Genet.* 2004, *5*, 739–751.

[78] Calafell, F., *Int. J. Legal Med.* 2000, *114*, 61–65.

[79] Kidd, J. M., Cooper, G. M., Donahue, W. F., Hayden, H. S. *et al.*, *Nature* 2008, *453*, 56–64.

[80] Levy, S., Sutton, G., Ng, P. C., Feuk, L. *et al.*, *PLoS Biol.* 2007, *5*, 2113–2144.

[81] Salisbury, M., 2008. http://www.genome-technology.com/issues/2_13/markers/145966-1.html

[82] Belosludtsev, Y., Battulga, N., Reddy, M., Kraltcheva, A. *et al.*, Visigen Biotechnologies Patent # US2008076189-A1.

[83] Hardin, S. H., Briggs, J. M., Tu, S., Gao, X. *et al.*, Visigen Biotechnologies Inc. Patent # WO200204680-A; EP1368460-A; WO200204680-A2; AU200182881-A; US2003064366-A1; EP1368460-A2; JP2004513619-W; CN1553953-A; US2005260614-A1; US2005266424-A1; US2007172864-A1; US2007172865-A1; US2007172861-A1; US2007172862-A1; US2007172866-A1; US2007172867-A1; US2007172868-A1; US2007172858-A1; US2007172863-A1; US2007184475-A1; AU2001282881-B2; EP1368460-B1; US2007275395-A1; DE60131194-E; US2007292867-A1; US7329492-B2; AU2007216737-A1.

[84] Hardin, S. H., Briggs, J. M., Xiaolian, G., Willson, R. *et al.*, Visigen Biotechnologies Inc. Patent # WO200244425-A; EP1354064-A; WO200244425-A2; AU200227156-A; US2003134807-A1; EP1354064-A2; AU2002227156-A8; US2007172819-A1; US2007172869-A1; US2007172860-A1; US2007172859-A1; US7211414-B2.

[85] Volkov, A., Colbert, C. M., Pan, I., Kraltcheva, A. *et al.*, Visigen Biotechnologies Inc. Patent # US2007250274-A1.

[86] Korlach, J., Turner, S., Pacific Biosciences Inc. Patent # US2007231804-A1; WO2007115205-A2.

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

[88] Rank, D., Korlach, J., Xu, Y., Turner, S. *et al.*, Pacific Biosciences Inc. Patent # US2008153100-A1.

[89] Turner, S., Korlach, J., Pacific Biosciences Inc. Patent # US2007206189-A1; US7313308-B2.

[90] Turner, S., Korlach, J., Dewinter, A., Pacific Biosciences Inc. Patent # US2007141598-A1.

© 2008 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim