High-throughput DNA sequencing – concepts and limitations

*Martin Kircher and Janet Kelso**

Recent advances in DNA sequencing have revolutionized the field of genomics, making it possible for even single research groups to generate large amounts of sequence data very rapidly and at a substantially lower cost. These high-throughput sequencing technologies make deep transcriptome sequencing and transcript quantification, whole genome sequencing and resequencing available to many more researchers and projects. However, while the cost and time have been greatly reduced, the error profiles and limitations of the new platforms differ significantly from those of previous sequencing technologies. The selection of an appropriate sequencing platform for particular types of experiments is an important consideration, and requires a detailed understanding of the technologies available; including sources of error, error rate, as well as the speed and cost of sequencing. We review the relevant concepts and compare the issues raised by the current high-throughput DNA sequencing technologies. We analyze how future developments may overcome these limitations and what challenges remain.

Keywords:

- ABI/Life Technologies SOLiD; Helicos HeliScope; Illumina Genome Analyzer; Roche/454 GS FLX Titanium; Sanger capillary sequencing

Introduction

In 1977 the first genome, that of the 5,386 nucleotide (nt), single-stranded bacteriophage φX174, was completely sequenced [1] using a technology invented just a few years earlier [2–5]. Since then the sequencing of whole genomes as well as of individual regions and genes has become a major focus of modern biology and completely transformed the field of genetics.

At the time of the sequencing of φX174, and for almost another decade, DNA sequencing was a barely automated and very tedious process which involved determining only a few hundred nucleotides at a time. In the late 1980s, semi-automated sequencers with higher throughput became available [6, 7], still only able to determine a few sequences at a time. A breakthrough in the early 1990s was the development of capillary array electrophoresis and appropriate detection systems [8–12]. As recently as 1996, these developments converged in the production of a commercial single capillary sequencer (ABI Prism 310). In 1998, the GE Healthcare MegaBACE 1000 and the ABI Prism 3700 DNA Analyzer became the first commercial 96 capillary sequencers, a development which was termed high-throughput sequencing.

Over the last decade, alternative sequencing strategies have become available [13–18] which force us to completely redefine ''high-throughput sequencing.'' These technologies outperform the older Sanger-sequencing technologies by a factor of 100–1,000 in daily throughput, and at the same time reduce the cost of sequencing one million nucleotides (1 Mb) to 4–0.1% of that associated with Sanger sequencing. To reflect these huge changes, several companies, researchers, and recent reviews [19–24] use the term ''next-generation sequencing'' instead of high-throughput sequencing, yet this term itself may soon be outdated considering the speed of ongoing developments.

DOI 10.1002/bies.200900181

Department of Evolutionary Genetics, Max Planck Institute for Evolutionary Anthropology, Leipzig, Germany

***Corresponding author:**
Janet Kelso
E-mail: kelso@eva.mpg.de

Abbreviations:
A/C/G/T, Deoxyadenosine, Deoxycytosine, Deoxyguanosine, Deoxythymidine; **ATP**, Adenosine triphosphate; **dATPαS**, Deoxy-adenosine-5'-(alpha-thio)-triphosphate; **CCD**, Charge-coupled Device, i.e. semi-conductor device used in digital cameras; **ChIP-Seq**, Chromatin Immuno-Precipitation sequencing; **CNV**, Copy Number Variation; **dNTPs/NTPs**, deoxy-nucleotides; **ddNTPs**, dideoxy-nucleotides (modified nucleotides missing a hydroxyl group at the third carbon atom of the sugar); **GA**, Short for Illumina Genome Analyzer; **InDel**, Insertion/Deletion; **kb/Mb/Gb**, kilo base (10^3 nt)/mega base (10^6 nt)/giga base (10^9 nt); **MeDIP-Seq**, Methyl- ation-Dependent Immuno-Precipitation sequen- cing; nt nucleotide(s); **PCR**, Polymerase Chain Reaction; **RNA-Seq**, Sequencing of mRNAs/transcripts; **SAGE**, Serial Analysis of Gene Expression; **SNP**, Single Nucleotide Polymorphism; **mRNA**, messenger RNA/transcripts.

 Bioessays 32: 524–536,© 2010 WILEY Periodicals, Inc.

Illumina Exhibit 1010, Page 1 of 13

Here we review the five sequencing technologies currently available on the market (capillary sequencing, pyrosequencing, reversible terminator chemistry, sequencing-by-ligation, and virtual terminator chemistry), discuss the intrinsic limitations of each, and provide an outlook on new technologies on the horizon. We explain how the vast increases in throughput are associated with both new and old types of problems in the resulting sequence data, and how these limit the potential applications and pose challenges for data analysis.

Sanger capillary sequencing

Current Sanger capillary sequencing systems, like the widely used Applied Biosystems 3xxx series or the GE Healthcare MegaBACE instrument, are still based on the same general scheme applied in 1977 for the φX174 genome [1, 3]. First, millions of copies of the sequence to be determined are purified or amplified, depending on the source of the sequence. Reverse strand synthesis is performed on these copies using a known priming sequence upstream of the sequence to be determined and a mixture of deoxy-nucleotides (dNTPs, the standard building blocks of DNA) and dideoxy-nucleotides (ddNTP, modified nucleotides missing a hydroxyl group at the third carbon atom of the sugar). The dNTP/ddNTP mixture causes random, non-reversible termination of the extension reaction, creating from the different copies molecules extended to different lengths. Following denaturation and clean up of free nucleotides, primers, and the enzyme, the resulting molecules are sorted by their molecular weight (corresponding to the point of termination) and the label attached to the terminating ddNTPs is read out sequentially in the order created by the sorting step. A schematic representation of this process is available in Fig. 1.

Sorting by molecular weight was originally performed using gel electrophoresis but is nowadays carried out by capillary electrophoresis [7, 25]. Originally, radioactive or optical labels were applied in four different terminator reactions (each sorted and read out separately), but today four different fluorophores, one per nucleotide (A, C, G, and T) are used in a single reaction [6]. Additionally, the advent of more sensitive detection systems and several rounds of primer extensions (equivalent to a linear amplification) permit smaller amounts of starting DNA to be used for modern sequencing reactions. Unfortunately, there is still little automation for creation of the high copy input DNA with known priming sites. Typically this is done by cloning, *i.e.*, introducing the target sequence into a known vector sequence using restriction and ligation procedures and using a bacterial strain to amplify the target sequence *in vivo* – thereby exploiting the low amplification error due to inherent proof-reading and repair mechanisms. However, this process is very tedious and is sometimes hampered by difficulties such as cloning specific sequences due to their base composition, length, and interactions with the bacterial host system. Although not yet widely used, integrated microfluidic devices have been developed which aim to automate the DNA extraction, *in vitro* amplification, and sequencing on the same chip [26–29].

Using current Sanger sequencing technology, it is technically possible for up to 384 sequences [29, 30] of between 600 and 1,000 nt in length [23, 31] to be sequenced in parallel. However, these 384-capillary systems are rare. The more standard 96-capillary instruments yield a maximum of approximately 6 Mb of DNA sequence per day, with costs for consumables amounting to about $500 per 1 Mb.

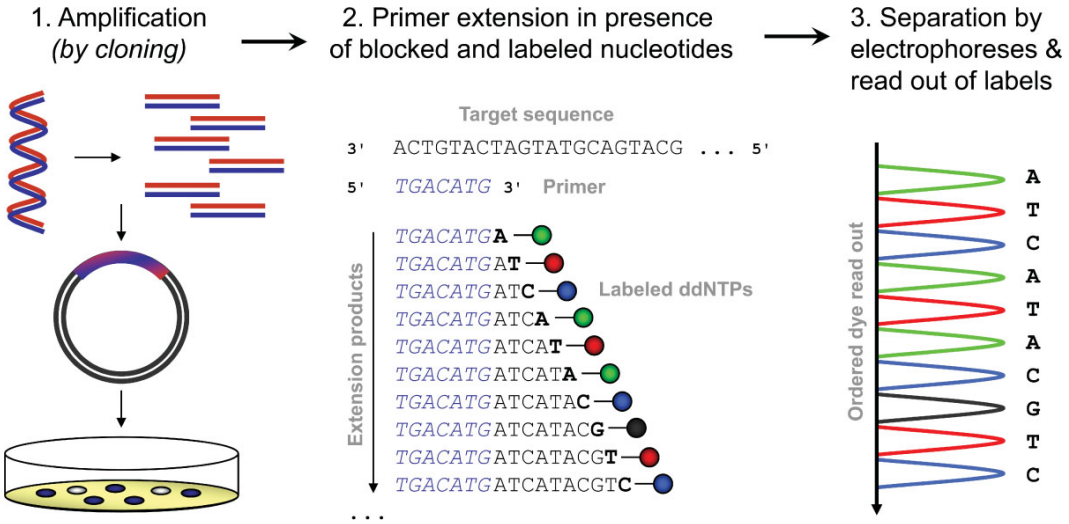

Figure 1. Schematic representation of the Sanger sequencing process. Input DNA is fragmented and cloned into bacterial vectors for *in vivo* amplification. Reverse strand synthesis is performed on the obtained copies starting from a known priming sequence and using a mixture of deoxy-nucleotides (dNTPs) and dideoxy-nucleotides (ddNTPs). The dNTP/ddNTP mixture randomly causes the extension to be non-reversibly terminated, creating differently extended molecules. Subsequently, after denaturation, clean up of free nucleotides, primers, and the enzyme, the resulting molecules are sorted using capillary electrophoresis by their molecular weight (corresponding to the point of termination) and the fluorescent label attached to the terminating ddNTPs is read out sequentially.

Illumina Exhibit 1010, Page 2 of 13

The sequencing error observed for Sanger sequencing is mainly due to errors in the amplification step (a low rate when done *in vivo*), natural variance, and contamination in the sample used, as well as polymerase slippage at low complexity sequences like simple repeats (short variable number tandem repeats) and homopolymers (stretches of the same nucleotide). Further, lower intensities and missing termination variants tend to lead to sequencing errors accumulating toward the end of long sequences. In combination with reduced separation by the electrophoresis, base miscalls [32] and deletions increase with read length. However, the average error rate (the average over all bases of a sequence) after sequence end trimming is typically very low, with an error every 10,000–100,000 nt [33].

Roche/454 GS FLX Titanium sequencer

The 454 sequencing platform was the first of the new high-throughput sequencing platforms on the market (released in October 2005). It is based on the pyrosequencing approach developed by Pål Nyrén and Mostafa Ronaghi at the Royal Institute of Technology, Stockholm in 1996 [34]. In contrast to the Sanger technology, pyrosequencing is based on iteratively complementing single strands and simultaneously reading out the signal emitted from the nucleotide being incorporated (also called sequencing by synthesis, sequencing during extension). Electrophoresis is therefore no longer required to generate an ordered read out of the nucleotides, as the read out is now done simultaneously with the sequence extension.

In the pyrosequencing process (Fig. 2), one nucleotide at a time is washed over several copies of the sequence to be determined, causing polymerases to incorporate the nucleotide if it is complementary to the template strand. The incorporation stops if the longest possible stretch of complementary nucleotides has been synthesized by the polymerase. In the process of incorporation, one pyrophosphate per nucleotide is released and converted to ATP by an ATP sulfurylase. The ATP drives the light reaction of luciferases present and the emitted light signal is measured. To prevent the dATP provided for sequencing reaction from being used directly in the light reaction, deoxy-adenosine-5′-(α-thio)-triphosphate (dATPαS), which is not a substrate of the luciferase, is used for the base incorporation reaction. Standard deoxyribose nucleotides are used for all other nucleotides. After capturing the light intensity, the remaining unincorporated nucleotides are washed away and the next nucleotide is provided.

In 2005, pyrosequencing technology was parallelized on a picotiter plate by 454 Life Sciences (later bought by Roche Diagnostics) to allow high-throughput sequencing [16]. The sequencing plate has about two million wells – each of them able to accommodate exactly one 28-μm diameter bead covered with

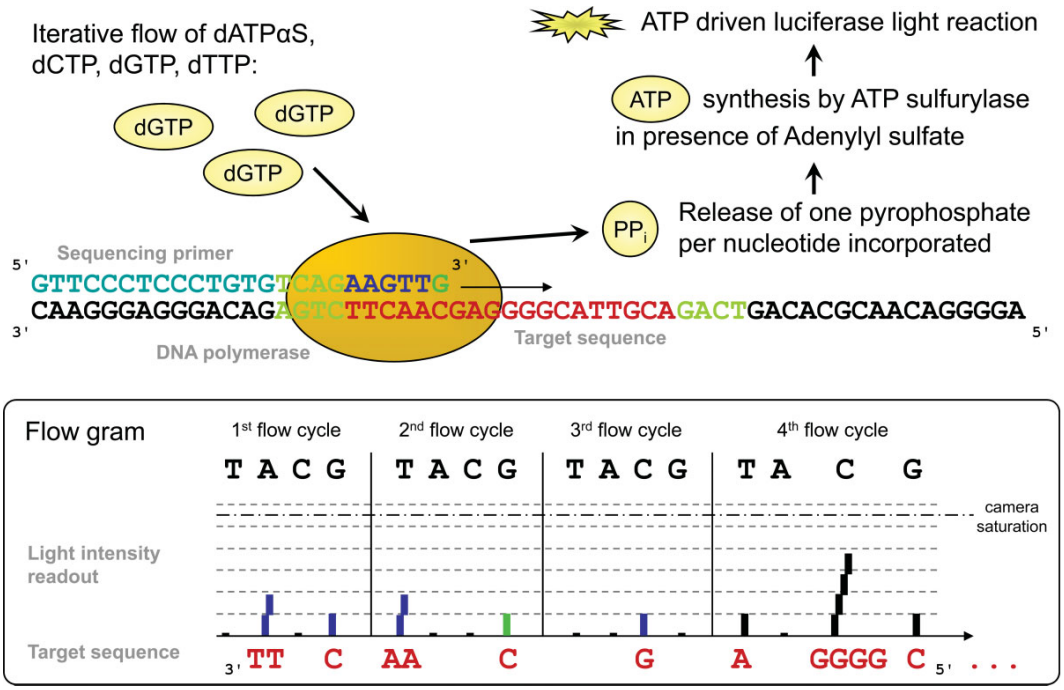

Figure 2. The pyrosequencing process. One of four nucleotides is washed sequentially over copies of the sequence to be determined, causing polymerases to incorporate complementary nucleotides. The incorporation stops if the longest possible stretch of the available nucleotide has been synthesized. In the process of incorporation, one pyrophosphate per nucleotide is released and converted to ATP by an ATP sulfurylase. The ATP drives the light reaction of luciferases present and a light signal proportional (within limits) to the number of nucleotide incorporations can be measured.

Illumina Exhibit 1010, Page 3 of 13

single-stranded copies of the sequence to be determined. The beads are incubated with a polymerase and single-strand binding proteins and, together with smaller beads carrying the ATP sulfurylases and luciferases, gravitationally deposited in the wells. Free nucleotides are then washed over the flow cell and the light emitted during the incorporation is captured for all wells in parallel using a high-resolution charge-coupled device (CCD) camera, exploiting the light-transporting features of the plate used.

One of the main prerequisites for applying this array-based pyrosequencing approach is covering individual beads with multiple copies of the same molecule. This is done by first creating sequencing libraries in which every individual molecule gets two different adapter sequences, one at the 5′ end and one at the 3′ end of the molecule. In the case of the 454/Roche sequencing library preparation [16], this is done by sequential ligation of two pre-synthesized oligos. One of the adapters added is complementary to oligonucleotides on the sequencing beads and thus allows molecules to be bound to the beads by hybridization. Low molecule-to-bead ratios and amplification from the hybridized double-stranded sequence on the beads (kept separate using emulsion PCR) makes it possible to grow beads with thousands of copies of a single starting molecule. Using the second adapter, beads covered with molecules can be separated from empty beads (using special capture beads with oligonucleotides complementary to the second adapter) and are then used in the sequencing reaction as described above.

The average substitution (excluding insertion/deletion, InDel) error rate is in the range of 10^{-3}–10^{-4} [16, 35], which is higher than the rates observed for Sanger sequencing, but is the lowest average substitution error rate of the new sequencing technologies discussed here. As mentioned earlier for Sanger sequencing, *in vitro* amplifications performed for the sequencing preparation cause a higher background error rate, *i.e.*, the error introduced into the sample before it enters the sequencer. In addition, in bead preparation (*i.e.*, emulsion PCR) a fraction of the beads end up carrying copies of multiple different sequences. These "mixed beads" will participate in a high number of incorporations per flow cycle, resulting in sequencing reads that do not reflect real molecules. Most of these reads are automatically filtered during the software post-processing of the data. The filtering of mixed beads may, however, cause a depletion of real sequences with a high fraction of incorporations per flow cycle.

A large fraction of the errors observed for this instrument are small InDels, mostly arising from inaccurate calling of homopolymer length, and single base-pair deletions or insertions caused by signal-to-noise thresholding issues [35]. Most of these problems can be resolved by higher coverage. For long (>10 nt) homopolymers, however, there is often a consistent length miscall that is not resolvable by coverage [35–37]. Strong light signals in one well of the picotiter plate may also result in insertions in sequences in neighboring wells. If the neighboring well is empty, this can generate so-called ghost wells, *i.e.*, wells for which a signal is recorded even though they contain no sequence template; hence, the intensities measured are completely caused by bleed-over signal from the neighboring wells. Computational post-processing may correct for these artifacts [38]. As for Sanger sequencing, the error rate increases with the position in the sequence. In the case of 454 sequencing, this is caused by a reduction in enzyme efficiency or loss of enzymes (resulting in a reduction of the signal intensities), some molecules no longer being elongated and by an increasing phasing effect. Phasing is observed when a population of DNA molecules amplified from the same starting molecule (ensemble) is sequenced, and describes the process whereby not all molecules in the ensemble are extended in every cycle. This causes the molecules in the ensemble to lose synchrony/phase, and results in an echo of the preceding cycles to be added to the signal as noise.

The current 454/Roche GS FLX Titanium platform makes it possible to sequence about 1.5 million such beads in a single experiment and to determine sequences of length between 300 and 500 nt. The length of the reads is determined by the number of flow cycles (the number of times all four nucleotides are washed over the plate) as well as by the base composition and the order of the bases in the sequence to be determined. Currently, 454/Roche limits this number to 200 flow cycles, resulting in an expected average read length of about 400 nt. This is largely due to limitations imposed by the efficiency of polymerases and luciferases, which drops over the sequencing run, resulting in decreased base qualities. Currently the platforms allows about 750 Mb of DNA sequence to be created per day with costs of about 20$/Mb.

Illumina Genome Analyzer II/IIx

The reversible terminator technology used by the Illumina Genome Analyzer (GA) employs a sequencing-by-synthesis concept that is similar to that used in Sanger sequencing, *i.e.* the incorporation reaction is stopped after each base, the label of the base incorporated is read out with fluorescent dyes, and the sequencing reaction is then continued with the incorporation of the next base [13, 39] (Fig. 3).

Like 454/Roche, the Illumina sequencing protocol requires that the sequences to be determined are converted into a special sequencing library, which allows them to be amplified and immobilized for sequencing [13, 40]. For this purpose two different adapters are added to the 5′ and 3′ ends of all molecules using ligation of so-called forked adapters.[1] The library is then amplified using longer primer sequences, which extend and further diversify the adapters to create the final sequence needed in subsequent steps.

This double-stranded library is melted using sodium hydroxide to obtain single-stranded DNAs, which are then pumped at a very low concentration through the channels of a flow cell. This flow cell has on its surface two populations of immobilized oligonucleotides complementary to the two different single-stranded adapter ends of the sequencing library. These oligonucleotides hybridize to the single-

[1] Hybrids of partially complementary oligonucleotides creating one double-stranded end with a T overhang, with a single-stranded and a different sequence at the other end.

Illumina Exhibit 1010, Page 4 of 13

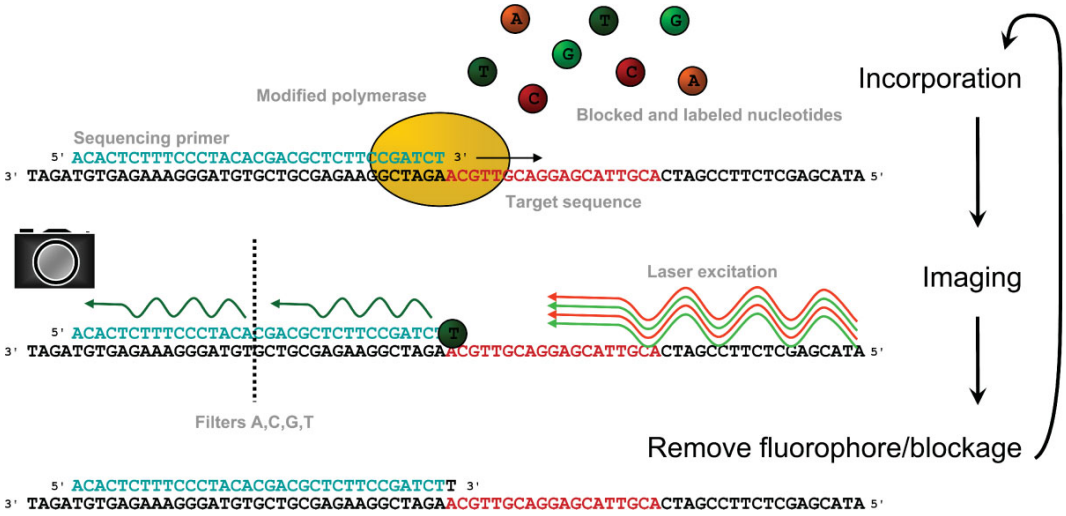

Figure 3. Reversible terminator chemistry applied by the Illumina GA. Sequencing primers are annealed to the adapters of the sequences to be determined. Polymerases are used to extend the sequencing primers by incorporation of fluorescently labeled and terminated nucleotides. The incorporation stops immediately after the first nucleotide due to the terminators. The polymerases and free nucleotides are washed away and the label of the bases incorporated for each sequence is read with four images taken through different filters (T nucleotide filter is indicated in the figure) and using two different lasers (red: A, C and green: G, T) to illuminate fluorophores. Subsequently, the fluorophores and terminators are removed and the sequencing continued with the incorporation of the next base.

stranded library molecules. By reverse strand synthesis starting from the hybridized (double-stranded) part, the new strand being created is covalently bound to the flow cell. If this new strand bends over and attaches to another oligonucleotide complementary to the second adapter sequence on the free end of the strand, it can be used to synthesize a second covalently bound reverse strand. This process of bending and reverse strand synthesis, called bridge amplification, is repeated several times and creates clusters of several 1,000 copies of the original sequence in very close proximity to each other on the flow cell [13, 40].

These randomly distributed clusters contain molecules that represent the forward as well as reverse strands of the original sequences. Before determining the sequence, one of the strands has to be removed to prevent it from hindering the extension reaction sterically or by complementary base pairing. Strands are selectively cleaved at base modifications of oligonucleotides on the flow cell. Following strand removal, each cluster on the flow cell consists of single stranded, identically oriented copies of the same sequence; which can be sequenced by hybridizing the sequencing primer onto the adapter sequences and starting the reversible terminator chemistry.

"Solexa sequencing", as it was introduced in early 2007, initially allowed for the simultaneous sequencing of several million very short sequences (at most 26 nt) in a single experiment. In recent years there have been several technical, chemical, and software updates. The product, which is now called the Illumina Genome Analyzer, has increased flow cell cluster densities (more than 200 million clusters per run), a wider range of the flow cell is imaged, and sequence reads of up to 100 nt can be generated. A technical update also enabled the sequencing of the reverse strand of each molecule. This is achieved by chemical melting and washing away the synthesized sequence, repeating a few bridge amplification cycles for reverse strand synthesis, and then selectively removing the starting strand (again using base modifications of the flow cell oligonucleotide populations), before annealing another sequencing primer for the second read. Using this "paired-end sequencing" approach, approximately twice the amount of data can be generated. The Illumina library and flow cell preparation includes several *in vitro* amplification steps, which cause a high background error rate and contribute to the average error rate of about 10^{-2}–10^{-3} [41, 42]. Further, the flow cell preparation creates a fraction of ordinary-looking clusters that are initiated from more than one individual sequence. These results in mixed signals and mostly low quality sequences for these clusters. Similar to the 454 ghost wells, the Illumina image analysis may identify chemistry crystals, dust, and lint particles as clusters and call sequences from these. In such cases the resulting sequences typically appear to be of low sequence complexity.

As is the case for the other platforms, the error rate increases with increasing position in the determined sequence. This is mainly due to phasing, which increases the background noise as sequencing progresses. While the ensemble sequencing process for pyrosequencing creates uni-directional phasing, reversible terminator sequencing creates bi-directional phasing [41, 43] as some incorporated nucleotides may also fail to be correctly terminated – allowing the extension of the sequence by another nucleotide in the same cycle.

Illumina Exhibit 1010, Page 5 of 13

With increasing cycle numbers, the intensities extracted from the clusters decline [41, 43, 44]. This is due to fewer molecules participating in the extension reaction as a result of non-reversible termination, or due to dimming effects of the sequencing fluorophores. In early versions of the chemistry, one of the fluorophores could become stuck to the clusters creating another source of increased background noise [41]. The simultaneous identification of four different nucleotides is also an issue. The GA uses four fluorescent dyes to distinguish the four nucleotides A, C, G, and T. Of these, two pairs (A/C and G/T) excited using the same laser, are similar in their emission spectra and show only limited separation using optical filters. Therefore, the highest substitution errors observed are between A/C and G/T [41, 42].

Even though the Illumina GA reads show a higher average error rate, a wider average error range, and are considerably shorter than 454/Roche reads, the GA instrument determines more than 5,000 Mb/day with a price of about 0.50$/Mb. This is more than six times higher daily throughput and for a considerably lower price per megabase.

Applied Biosystems SOLiD

The prototype of what was further developed and later sold by Life Technologies/Applied Biosystems (ABI) as the SOLiD sequencing platform, was developed by Harvard Medical School and the Howard Hughes Medical Institute and published in 2005 [17]. With its commercial release in late 2007, SOLiD was only the third new high-throughput system entering a highly competitive market with all three vendors selling their instruments for around half a million dollars. The Church lab at Harvard Medical School continued the development of the system and now offers a cheaper (<$200,000) open source version of the system (called Polonator) in collaboration with Dover System. In the third quarter of 2008, a biotechnology company from Mountain View, California, named Complete Genomics started offering a human genome sequencing service. Their technology is also based on the Church lab sequencing-by-ligation concept, but combines it with a new strategy of sequencing library construction and sequence immobilization using rolling circle amplification [45]. Here, we focus on the commercial SOLiD system as this is the most widespread application of this concept.

The principle behind sequencing-by-ligation is very different from the approaches discussed thus far. The sequence extension reaction is not carried out by polymerases but rather by ligases [17] (see Fig. 4 for a schematic representation of the SOLiD 2/3 platform). In the sequencing-by-ligation process, a sequencing primer is hybridized to single-stranded copies of the library molecules to be sequenced. A mixture of 8-mer probes carrying four distinct fluorescent labels compete for ligation to the sequencing primer. The fluorophore encoding, which is based on the two 3′-most nucleotides of the probe, is read. Three bases including the dye are cleaved from the 5′ end of the probe, leaving a free 5′ phosphate on the extended (by five nucleotides) primer, which is then available for further ligation. After multiple ligations (typically up to 10 cycles), the synthesized strands are melted and the ligation product is washed away before a new sequencing primer (shifted by one nucleotide) is annealed. Starting from the new sequencing primer the ligation reaction is repeated. The same process is followed for three other primers, facilitating the read out of the dinucleotide encoding for each start position in the sequence. Using specific fluorescent label encoding, the dye read outs (*i.e.* colors) can be converted to a sequence [46]. This conversion from color space to sequence requires a known first base, which is the last base of the used library adapter sequence. Given a reference sequence, this encoding system allows detection of machine errors and the application of an error correction to reduce the average error rate. In the absence of a reference sequence, however, color conversion fails with an error in the dye read out and causes the sequence downstream of the error to be incorrect.

For parallelization, the sequencing process uses beads covered with multiple copies of the sequence to be determined. These beads are created in a similar fashion to that described earlier for the 454/Roche platform. In contrast to the 454/Roche technology, the SOLiD system does not use a picotiter plate for fixation of the beads in the sequencing process; instead the 3′ ends of the sequences on the beads are modified in a way that allows them to be covalently bound onto a glass slide. As for the Illumina GA system, this creates a random dispersion of the beads in the sequencing chamber and allows for higher loading densities. However, random dispersion complicates the identification of bead positions from images, and results in the possibility that chemical crystals, dust, and lint particles can be misidentified as clusters. Further, dispersal of the beads results in a wide range of inter-bead distances, which then have different susceptibility to be influenced by signals from neighboring beads.

Types and causes of sequence errors are diverse: first, the *in vitro* amplification steps cause a higher background error rate. Secondly, beads carrying a mixture of sequences and beads in close proximity to one another create false reads and low quality bases. Further, signal decline, a small regular phasing effect, and incomplete dye removal result in increasing error as the ligation cycles progress [47]. Phasing, as described earlier, is a minor issue on this platform as sequences not extended in the last cycle are non-reversibly terminated using phosphatases. Since hybridization is a stochastic process, this causes a considerable reduction in the number of molecules participating in subsequent ligation reactions, and therefore substantial signal decline. On the other hand, given the efficiency of phosphatases the remaining phasing effect can be considered very low. However, incomplete cleavage of the dyes may allow cleavage in the next ligation reaction, which then allows for the extension in the next but one cycle. This causes a different phasing effect and additional noise from the previous cycle's dyes in the dye identification process.

The SOLiD system currently allows sequencing of more than 300 million beads in parallel, with a typical read length of between 25 and 75 nt. At the time of writing, the ABI SOLiD system is therefore comparable to the Illumina GA

Illumina Exhibit 1010, Page 6 of 13

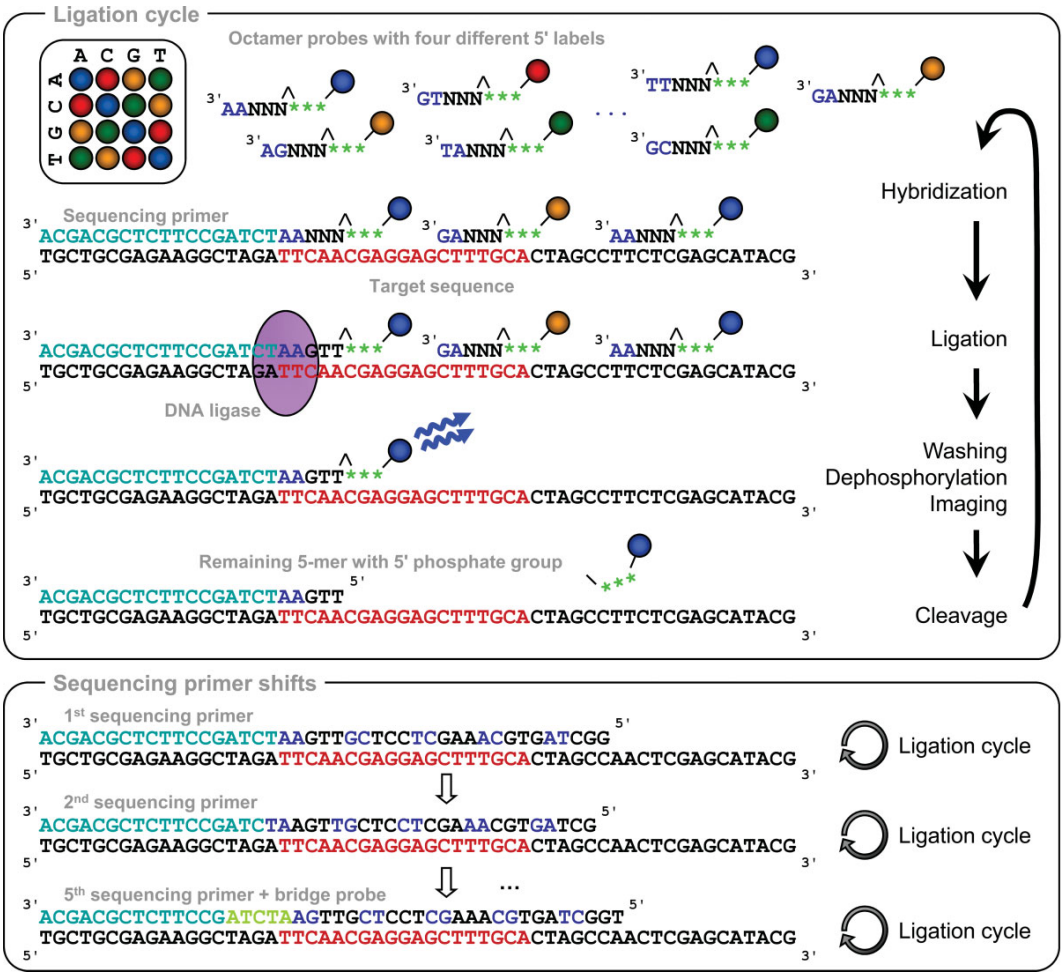

Figure 4. Applied Biosystem's SOLiD sequencing by ligation. A sequencing primer is annealed to single-stranded copies of sequences to be determined. Octamer probes are hybridized, ligated to the sequencing primer, and a fluorescent dye at the 5′ end of the ligated 8-mer probes, encoding the two 3′-most nucleotides of the probe, is read out. Non-extended primers are dephosphorylated. Three nucleotides of the probe including the dye are cleaved, creating a free 5′ phosphate for further ligations. After multiple ligations, the synthesized strands are melted and the ligation product is washed away before a new, by-one-nucleotide-shifted sequencing primer is annealed. Starting from the new sequencing primer the ligation reaction is repeated. The same is done for three other primers, allowing the read out of the dinucleotide label for every position in the sequence.

system in terms of throughput and price per million nucleotides (~5,000 Mb/day, ~0.50$/Mb). Average error rates are, however, dependent on the availability of a reference genome for error correction (10^{-3}–10^{-4} *vs.* 10^{-2}–10^{-3}). In the absence of a reference genome, assembly and consensus calling may be performed based on dye read outs (so-called color space sequences) to reduce the errors before conversion to the nucleotide sequence. If no reference genome is available for error correction, and no assembly and consensus calling is performed, then the average error rate is higher than for the Illumina GA.

Helicos HeliScope

Helicos is the first company to sell a sequencer able to sequence individual molecules instead of molecule ensembles created by an amplification process. Single molecule sequencing has the advantage that it is not affected by biases or errors introduced in a library preparation or amplification step, and may facilitate sequencing of minimal amounts of input DNA. Using methods able to detect non-standard nucleotides, it could also allow for the identification of DNA modifications, commonly lost in the *in vitro* amplification process.

The HeliScope, as the Helicos sequencer is called, was first sold in March 2008, and by the end of the first quarter of 2009 only four machines have been installed worldwide. This might be surprising given the advantages of single molecule sequencing, but probably reflects both the specific limitations of this platform, the price (about one million dollars), and a relatively small market that has already invested extensively in new sequencing technologies.

The technology applied (Fig. 5) could be termed asynchronous virtual terminator chemistry [15]. Input DNA is fragmented and melted before a

Illumina Exhibit 1010, Page 7 of 13

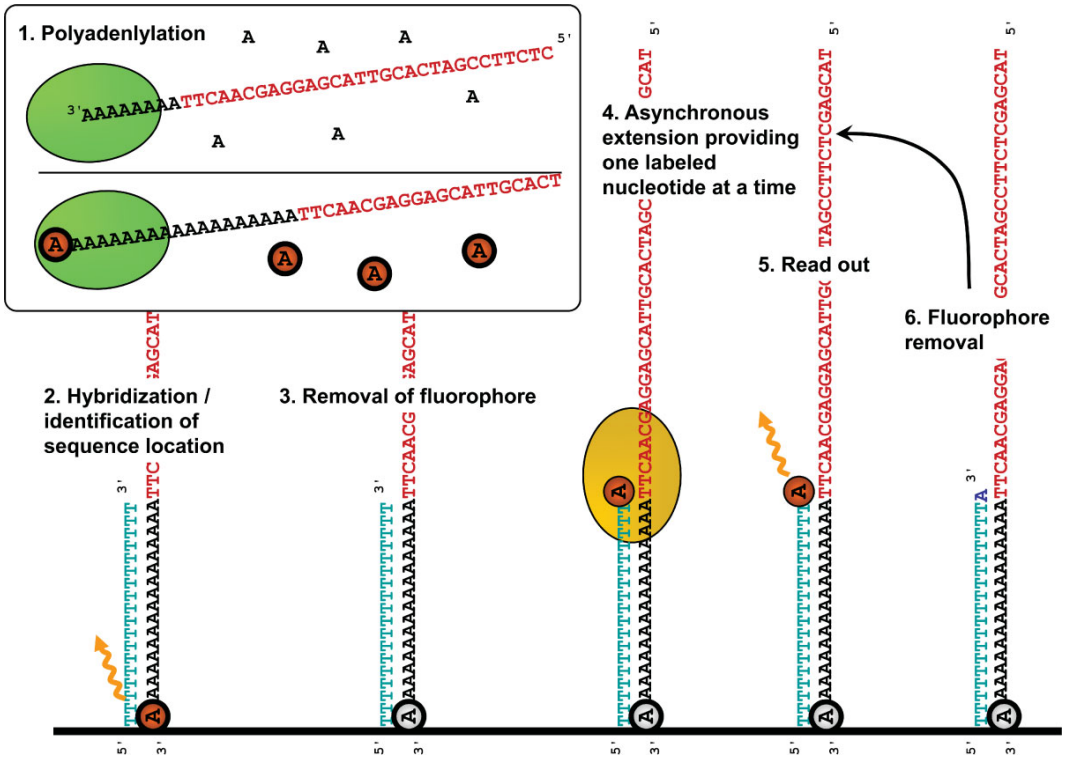

Figure 5. Asynchronous virtual terminator chemistry performed by the HeliScope. Input DNA is fragmented, melted, and polyadenylated. A fluorescently labeled adenine is added in the last step. This single-stranded DNA is washed over a flow cell with poly-T oligonucleotides allowing hybridization. The bound coordinates on the flow cell are determined using the fluorescently labeled adenines. Having the coordinates identified, the fluorescent label of the 3′ adenines is removed. Polymerases are washed through with one type of fluorescently labeled nucleotides (A, C, G, and T) at a time, and the polymerases extend the reverse strand of the sequences starting from the poly-T oligonucleotides. The nucleotide incorporation of the polymerases is slowed down by the fluorescent labeling and allows for at most one incorporation before the polymerase is washed away. The flow cell is then imaged, the fluorescent dyes removed, and the reaction continued with another nucleotide.

poly-A-tail is synthesized onto each single-stranded molecule using a polyadenylate polymerase. In the last step of polyadenylation, a fluorescently labeled adenine is added. The library is washed over a flow cell where the poly-A tails bind to poly-T oligonucleotides. The bound coordinates on the flow cell are determined using a fluorescence-based read out of the flow cell. Having these coordinates identified, the fluorescent label of the 3′ adenine is removed and the sequencing reaction started. Polymerases are washed through the flow cell with one type of fluorescently labeled nucleotide (A, C, G, or T) at a time and the polymerases extend the reverse strand of the sequences starting from the poly-T oligonucleotides. The nucleotide incorporation of the polymerases is slowed down by the fluorescent labeling and allows for at most one incorporation before the polymerase is washed away together with the non-incorporated nucleotides (termed virtual termination [48, 49]). The flow cell is then imaged again, the fluorescent dyes are removed, and the reaction continued with another nucleotide. By this process not every molecule is extended in every cycle, which is why it is an asynchronous sequencing process resulting in sequences of different length (as is the case for the 454/Roche platform).

Since single molecules are sequenced, the signals being measured are weak, and there is no possibility that misincorporation errors can be corrected by an ensemble effect. Due to the fact that molecules are attached to the flow cell by hybridization only, there is a chance that template molecules can be lost in the wash steps. In addition, molecules may be irreversibly terminated by the incorporation of incorrectly synthesized nucleotides. Overall, reads are between 24 and 70 nt long (average 32 nt) [50] and thus shorter than for the other platforms. Due to the higher number of sequences determined in parallel, the total throughput per day (4150 Mb/day with a cost of ~0.33$/Mb [50]) is in the same range as for the GA and SOLiD systems. The average error rate, which is in the range of a few percent, is slightly higher than for all other instruments and biased toward InDels rather than substitutions.

Applications and general considerations

All current high-throughput technologies have an average error rate that

Illumina Exhibit 1010, Page 8 of 13

is considerably higher than the typical 1/10,000 to 1/100,000 observed for high-quality Sanger sequences. Further, the GS FLX Titanium, GA, SOLiD, and HeliScope platforms each have very specific biases and limitations, making it necessary to choose a platform appropriate for a specific project or application (for a summary see Table 1). A combination of technologies [51–54] and experimental protocols [55–57] may also be appropriate, and even complementary, for specific projects.

High-quality Sanger sequencing is now commonly used to generate low-coverage sequencing of individual positions and regions (*e.g.*, diagnostic genotyping) or the sequencing of virus- and phage-sized whole genomes. As the Sanger sequence length is longer than most abundant short repeat classes, it allows the unambiguous assembly of most genomic regions – something that is generally not possible using the shorter read platforms. However, the technology is expensive and too slow for sequencing large samples, extended genomic regions, or the many molecules required for quantitative applications [*e.g.*, gene expression quantification; chromatin immunoprecipitation sequencing (ChIP-Seq); and methylation-dependent immunoprecipitation sequencing (MeDip-Seq)]. For quantitative applications the HeliScope provides the highest throughput in terms of sequence number and has the advantage of not requiring a multistep library preparation protocol. On the other hand, the HeliScope provides the lowest resolution in mapping accuracy for complex genomes due its short read length and error profile. The GA or SOLiD platforms may thus provide equivalent results for quantitative applications, while providing fewer but longer reads and requiring a more elaborate library preparation.

While it has not yet been fully analyzed, it is possible (and even likely) that library preparation protocols could bias the sequence representation in a sample [42, 58, 59], making the replacement of this step an important goal. Further, multistep library preparation protocols require higher amounts of input material, limiting their general application. However, protocols for library construction from limited sample amounts are available or being developed for each of the platforms, and publications demonstrate that, while vendor protocols indicate the need for higher sample quantities (microgram range), many users are proceeding successfully with low input DNA amounts (nanogram to picogram range), as, for example, from ancient DNA specimens [60–62].

Like Sanger sequencing, the GS FLX Titanium provides a read length spanning many of the short repeat sequences – an important feature for accurate sequence mapping and assembly of genomes [63]. Despite the InDel errors, this technology has very low rates of misidentifying individual bases, making it perfectly suited for the identification of single nucleotide polymorphisms (SNPs). Also geared to the identification of SNPs, at least for samples with an existing reference genome, is the SOLiD instrument with its dinucleotide encoding scheme [46]. Considerably higher coverage is needed to perform SNP calling with similar accuracy using the Illumina GA [64]. Neither the Illumina GA nor the ABI SOLiD sequencing systems are prone to generate high rates of small InDels, making them well suited for studying InDel variation.

As mentioned earlier, the drawback of short reads (below about 75 nt) obtained from Helicos, SOLiD, or GA instruments is in genome assembly and mapping applications, where the placement of repeated or very similar sequences cannot be resolved unambiguously. The correct placement is further complicated by high error rates introducing a requirement for a minimum sequence distance of an unambiguous placement. Paired-end or mate-pair protocols help to overcome some of these limitations of short reads [65] by providing information about relative location and orientation of a pair of reads. Currently a paired-end protocol is only commonly applied on the GA, while mate-pair protocols are available for SOLiD, GS FLX Titanium, and GA. In paired-end sequencing the actual ends of rather short DNA molecules (<1 kb) are determined, while mate-pair sequencing requires the preparation of special libraries. In these protocols, the ends of longer, size-selected molecules (*e.g.*, 8, 12, or 20 kb) are connected with an internal adapter sequence in a circularization reaction. The circular molecule is then processed using restriction enzymes or fragmentation before outer library adapters are added around the two combined molecule ends. The internal adapter can then be used as a second priming site for an additional sequencing reaction on the same immobilized molecules. Thus, mate-pair sequencing provides distance information useful for assembly, but does not allow the merging of the two overlapping end reads, since by design the molecules will not overlap in sequencing. However, merging of two overlapping forward and reverse paired end reads from short insert libraries allows the reconstruction of a complete consecutive molecule sequence, longer than the individual read length, and with reduced average error rates in the overlapping sequence part [60, 66].

Due to the large amounts of sequences created, there is interest in sequencing targeted regions (*e.g.* a genomic locus, from sequence capture experiments [67–69]) in multiple individuals/samples instead of sequencing one sample in excessive depth. All technologies therefore provide a separation of their sequencing plate into defined regions or channels. However, at most, 16 such regions/channels are available (GS FLX Titanium and HeliScope plates), which may not be sufficient for some applications. Using different library construction protocols, some platforms allow addition of sample specific barcode (sometimes called "index") sequences to the library molecules. These molecules can then be sequenced in the same region/channel, and later separated (computationally) based on their barcode sequence [70–73]. This facilitates highly parallel sequencing of a large number of samples beyond that possible using the physical lane/channel separation. Currently such protocols (mostly non-vendor protocols) are available for the GS FLX Titanium, GA, and SOLiD instrument.

Although sequencing prices per gigabase have fallen considerably in recent years, making projects like the 1000 Human Genome Variation Project, 1001 *Arabidopsis thaliana* Genomes Project, the Mammalian Genome Project, or the International Cancer Genome Consortium possible, high-throughput sequencing still has high acquisition, running and maintenance costs, which

Illumina Exhibit 1010, Page 9 of 13

Table 1. Comparison of high-throughput sequencing technologies available

	Throughput	Length	Quality	Costs	Applications	Main sources of errors
Sanger	6 Mb/day	800 nt	10^{-4}–10^{-5}	~500$/Mb	Small sample sizes, genomes/scaffolds, InDels/SNPs, long haplotypes, low complexity regions, etc.	Polymerase/amplification, low intensities/missing termination variants, contaminant sequences
454/Roche	750 Mb/day	400 nt	10^{-3}–10^{-4}	~20$/Mb	Complex genomes, SNPs, structural variation, indexed samples, small RNA^{+}, mRNAs^{+}, etc.	Amplification, mixed beads, intensity thresholding, homopolymers, phasing, neighbor interference
Illumina	5,000 Mb/day	100 nt	10^{-2}–10^{-3}	~0.50$/Mb	Complex genomes, counting (SAGE, CNV ChIP, small RNA), mRNAs, InDels/homopolymers, structural variation, bisulfite data, indexing, SNPs^{+}, etc.	Amplification, mixed clusters/neighbor interference, phasing, base labeling
SOLiD	5,000 Mb/day	50 nt	10^{-2}–10^{-3}	~0.50$/Mb	Complex small genomes, counting (SAGE, ChIP, small RNA, CNV), SNPs, mRNAs, structural variation, indexing, etc.	Amplification, mixed beads, phasing, signal decline, neighbor interference
Helicos	5,000 Mb/day	32 nt	10^{-2}	<0.50$/Mb	Non-amplifiable samples, counting (SAGE, ChIP, small RNA), etc.	Polymerase, low intensities/thresholding, molecule loss/termination

The table summarizes throughput, length, quality, and costs for the current versions of the mentioned technologies. These approximate numbers are constantly improving and based on figures available in January 2010. Costs do not include instrument acquisition and maintenance; further they may be affected by discounts and scale effects for multiple instruments. Where numbers are very similar, colors ranging from red (low performance) to green (good performance) indicate a general trend. In the last column, example applications fitting the throughput and error profiles of each of the platforms are given. Typically, this does not mean that the technology is limited to these applications, but that it is currently best suited to such applications.

$^{+}$ High sequencing depth/number of runs required.

are not included in Table 1. Further, each of these platforms requires a substantial investment in data management and analysis, time, and personnel [74–77]. Smaller research groups may still find prohibitive the costs of the infrastructure needed for storing, handling, and analyzing several tens of gigabytes of pure sequence data and terabytes of several thousand intermediate files generated by these instruments each week. Even for larger, experienced genome centers this aspect remains an ever-increasing challenge for the ongoing use of these platforms.

Upcoming developments

Motivated by the goal of a $1,000-genome set by NIH/NHGRI to enable personalized medicine, the throughput of all systems described is constantly increasing and the numbers given here are rapidly outdated. However, in addition to the improvements of current technologies, including the January 2010 announcement of the Illumina HiSeq 2000 system, which determines sequences of clusters on bottom and top of the flow cell and processes two flow cells in parallel, a new generation of sequencers is already on the horizon.

What started with the Helicos system – the sequencing of single molecules without prior library preparation or amplification – will likely become a popular paradigm. Specifically, three other systems have captured media and scientific attention well in advance of their actual availability: Pacific Bioscience's Single Molecule Real Time (SMRT) sequencing technology [18], Oxford Nanopore's BASE technology [14] and, recently, IBM's proposal of silicon-based nanopores [78].

Pacific Biosciences' SMRT technology performs the sequencing reaction on silicon dioxide chips with a 100 nm metal film containing thousands of tens-of-nanometer diameter holes, so-called zero-mode waveguides (ZMWs) [79]. Each ZMW is used as a nano-visualization chamber, providing a detection volume of 20 zeptoliters (10^{-21} l). At this volume, a single molecule can be illuminated while excluding other labeled nucleotides in the background – saving time and sequencing chemistry by omitting wash steps. A single DNA polymerase is fixed to the bottom of the surface within the detection volume, and nucleotides, with different dyes attached to the phosphate chain, are used in concentrations allowing normal enzyme processivity. As the polymerase incorporates complementary nucleotides, the nucleotide is held within the detection volume for tens

Illumina Exhibit 1010, Page 10 of 13

of milliseconds, orders of magnitude longer than for unspecific diffusion events. This way the fluorescent dye of the incorporated nucleotide can be identified during normal speed reverse strand synthesis [79]. In pilot experiments, Pacific Biosciences has shown that its technology allows for direct sequencing of a few thousand bases before the polymerase is denatured due to the laser read out of the dyes. The SMRT technology is intended for release in 2010. Even though further development is needed to create a more robust system, the omission of library preparation and amplification as well as the long sequences generated will undoubtedly provide an advantage over the current systems for many applications.

Oxford Nanopore's BASE technology is unlikely to be released as soon as the SMRT technology. BASE offers the potential to identify individual nucleotide modifications (*e.g.* 5-methylcytosine *vs.* cytosine) during the sequencing process [14]. The idea behind this technology is the identification of individual nucleotides using a change in the membrane potential as they pass through a modified α-hemolysin membrane pore with a cyclodextrin sensor [14, 80]. However, to apply this technology for sequencing, the pore has to be fused to an exonuclease, which degrades single-stranded DNA sequences and releases individual nucleotides into the pore. In addition, the technology needs to be parallelized in array format, before its release as a high-throughput sequencing platform. While the sensitivity for individual nucleotide modifications seems to be a major advantage, the destructive fashion of the outlined sequencing process might be considered a hindrance for applications with precious samples, and it does not allow a second read cycle for error reduction.

In early October 2009, IBM issued a press release [78] describing a method to slow down the speed of an individual DNA strand passing through a nanopore. For this purpose they developed a multilayer metal/dielectric nanopore device that utilizes the interaction of the DNA backbone charges with a modulated electric field to trap and slowly releases an individual DNA molecule. The technology described could theoretically be combined with, for example, the Nanopore technologies developed at Harvard University [81] or the previously described BASE technology where it may overcome the destructive approach followed so far.

Conclusion

Current high-throughput sequencing technologies provide a huge variety of sequencing applications to many researchers and projects. Given the immense diversity, we have not discussed these applications in depth here; other reviews with a stronger focus on specific applications and data analysis are available [24, 82–88]. The discussed technologies make it possible for even single research groups to generate large amounts of sequence data very rapidly and at substantially lower costs than traditional Sanger sequencing. While costs have been reduced to less than 4–0.1% and time has been shortened by a factor of 100–1,000 based on daily throughput, the error profiles and limitations observed for the new platforms differ significantly from Sanger sequencing and between approaches. Further, each of these new sequencing platforms requires substantial additional investments – factors that have often not be sufficiently stressed in research publications describing a specific application. Some vendors have recently started to offer budget versions of their instruments (*e.g.* Illumina GA IIe or 454/Roche GS Junior) with lower sequencing capacity. However, while the instrument price is lower, the financial investment remains high. Costs per base are generally higher than for the standard instrument, and very similar overall infrastructure is still required. Often the choice of an appropriate sequencing platform is project specific and sometimes combinations can be advantageous. This may open the market further to companies providing sequencing-on-demand services, but will not replace the need for laboratories to invest considerable time and expertise in both the production of libraries and analysis of the vast quantities of data that will be generated.

New technologies on the horizon, SMRT by Pacific Biosciences, BASE by Oxford Nanopore, and other technologies such as that suggested by IBM, demonstrate the major future directions in the field of DNA sequencing: the ability to use individual molecules without any library preparation or amplification, the identification of specific nucleotide modifications, and the ability to generate longer sequence reads. These developments will facilitate future research in many fields, make data analysis easier, and further reduce sequencing costs, hopefully achieving the aim of a $1,000 human genome suggested by NIH/NHGRI to be required for personalized medicine.

Acknowledgments

We thank the members of the Department of Evolutionary Genetics, and particularly members of the sequencing group, for providing sequencing data from multiple platforms, as well as interesting discussions and useful insights. We are also indebted to A. Wilkins and the three anonymous reviewers for critical reading of the manuscript and thoughtful comments. This work was supported by the Max Planck Society.

References

1. **Sanger F**, **Air GM**, **Barrell BG**, *et al*. 1977. Nucleotide sequence of bacteriophage phi X174 DNA. *Nature* **265**: 687–95.
2. **Gilbert W**, **Maxam A.** 1973. The nucleotide sequence of the lac operator. *Proc Natl Acad Sci USA* **70**: 3581–4.
3. **Sanger F**, **Nicklen S**, **Coulson AR.** 1977. DNA sequencing with chain-terminating inhibitors. *Proc Natl Acad Sci USA* **74**: 5463–7.
4. **Sanger F**, **Coulson AR.** 1975. A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. *J Mol Biol* **94**: 441–8.
5. **Wu R**, **Kaiser AD.** 1968. Structure and base sequence in the cohesive ends of bacteriophage lambda DNA. *J Mol Biol* **35**: 523–37.
6. **Smith LM**, **Sanders JZ**, **Kaiser RJ**, *et al*. 1986. Fluorescence detection in automated DNA sequence analysis. *Nature* **321**: 674–9.
7. **Swerdlow H**, **Gesteland R.** 1990. Capillary gel electrophoresis for rapid, high resolution DNA sequencing. *Nucleic Acids Res* **18**: 1415–9.
8. **Zagursky RJ**, **McCormick RM.** 1990. DNA sequencing separations in capillary gels on a modified commercial DNA sequencing instrument. *Biotechniques* **9**: 74–9.
9. **Huang XC**, **Quesada MA**, **Mathies RA.** 1992. DNA sequencing using capillary array electrophoresis. *Anal Chem* **64**: 2149–54.
10. **Kambara H**, **Takahashi S.** 1993. Multiple-sheathflow capillary array DNA analyser. *Nature* **361**: 565–6.

Illumina Exhibit 1010, Page 11 of 13

11. **Ueno K**, **Yeung ES.** 1994. Simultaneous monitoring of DNA fragments separated by electrophoresis in a multiplexed array of 100 capillaries. *Anal Chem* **66**: 1424–31.
12. **Kim S**, **Yoo HJ**, **Hahn JH.** 1996. Postelectrophoresis capillary scanning method for DNA sequencing. *Anal Chem* **68**: 936–9.
13. **Bentley DR**, **Balasubramanian S**, **Swerdlow HP**, *et al*. 2008. Accurate whole human genome sequencing using reversible terminator chemistry. *Nature* **456**: 53–9.
14. **Clarke J**, **Wu HC**, **Jayasinghe L**, *et al*. 2009. Continuous base identification for single-molecule nanopore DNA sequencing. *Nat Nanotechnol* **4**: 265–70.
15. **Harris TD**, **Buzby PR**, **Babcock H**, *et al*. 2008. Single-molecule DNA sequencing of a viral genome. *Science* **320**: 106–9.
16. **Margulies M**, **Egholm M**, **Altman WE**, *et al*. 2005. Genome sequencing in microfabricated high-density picolitre reactors. *Nature* **437**: 376–80.
17. **Shendure J**, **Porreca GJ**, **Reppas NB**, *et al*. 2005. Accurate multiplex polony sequencing of an evolved bacterial genome. *Science* **309**: 1728–32.
18. **Korlach J**, **Marks PJ**, **Cicero RL**, *et al*. 2008. Selective aluminum passivation for targeted immobilization of single DNA polymerase molecules in zero-mode waveguide nanostructures. *Proc Natl Acad Sci USA* **105**: 1176–81.
19. **Ansorge WJ.** 2009. Next-generation, DNA sequencing techniques. *Nat Biotechnol* **25**: 195–203.
20. **Mardis ER.** 2008. Next-generation, DNA sequencing methods. *Annu Rev Genomics Hum Genet* **9**: 387–402.
21. **Schuster SC.** 2008. Next-generation sequencing transforms today's biology. *Nat Methods* **5**: 16–8.
22. **Shendure J**, **Ji H.** 2008. Next-generation, DNA sequencing. *Nat Biotechnol* **26**: 1135–45.
23. **Shendure JA**, **Porreca GJ**, **Church GM.** 2008. Overview of DNA sequencing strategies. *Curr Protoc Mol Biol* Chapter 7: Unit 7.1.
24. **Metzker ML.** 2010. Sequencing technologies – the next generation. *Nat Rev Genet* **11**: 31–46.
25. **George KS**, **Zhao X**, **Gallahan D**, *et al*. 1997. Capillary electrophoresis methodology for identification of cancer related gene expression patterns of fluorescent differential display polymerase chain reaction. *J Chromatogr B Biomed Sci Appl* **695**: 93–102.
26. **Blazej RG**, **Kumaresan P**, **Mathies RA.** 2006. Microfabricated bioprocessor for integrated nanoliter-scale Sanger DNA sequencing. *Proc Natl Acad Sci USA* **103**: 7240–5.
27. **Mariella R Jr.** 2008. Sample preparation: the weak link in microfluidics-based biodetection. *Biomed Microdevices* **10**: 777–84.
28. **Roper MG**, **Easley CJ**, **Legendre LA**, *et al*. 2007. Infrared temperature control system for a completely noncontact polymerase chain reaction in microfluidic chips. *Anal Chem* **79**: 1294–1300.
29. **Emrich CA**, **Tian H**, **Medintz IL**, *et al*. 2002. Microfabricated 384-lane capillary array electrophoresis bioanalyzer for ultrahigh-throughput genetic analysis. *Anal Chem* **74**: 5076–83.
30. **Shibata K**, **Itoh M**, **Aizawa K**, *et al*. 2000. RIKEN integrated sequence analysis (RISA) system – 384-format sequencing pipeline with 384 multicapillary sequencer. *Genome Res* **10**: 1757–71.
31. **Hert DG**, **Fredlake CP**, **Barron AE.** 2008. Advantages and limitations of next-generation sequencing technologies: a comparison of electrophoresis and non-electrophoresis methods. *Electrophoresis* **29**: 4618–26.
32. **Ewing B**, **Hillier L**, **Wendl MC**, *et al*. 1998. Base-calling of automated sequencer traces using phred. I. Accuracy assessment. *Genome Res* **8**: 175–85.
33. **Ewing B**, **Green P.** 1998. Base-calling of automated sequencer traces using phred. II. Error probabilities. *Genome Res* **8**: 186–94.
34. **Ronaghi M**, **Karamohamed S**, **Pettersson B**, *et al*. 1996. Real-time DNA sequencing using detection of pyrophosphate release. *Anal Biochem* **242**: 84–9.
35. **Quinlan AR**, **Stewart DA**, **Stromberg MP**, *et al*. 2008. Pyrobayes: an improved base caller for SNP discovery in pyrosequences. *Nat Methods* **5**: 179–81.
36. **Wicker T**, **Schlagenhauf E**, **Graner A**, *et al*. 2006. 454 sequencing put to the test using the complex genome of barley. *BMC Genomics* **7**: 275.
37. **Green RE**, **Malaspinas AS**, **Krause J**, *et al*. 2008. A complete Neandertal mitochondrial genome sequence determined by high-throughput sequencing. *Cell* **134**: 416–26.
38. **Green RE**, **Krause J**, **Ptak SE**, *et al*. 2006. Analysis of one million base pairs of Neanderthal DNA. *Nature* **444**: 330–6.
39. **Turcatti G**, **Romieu A**, **Fedurco M**, *et al*. 2008. A new class of cleavable fluorescent nucleotides: synthesis and optimization as reversible terminators for DNA sequencing by synthesis. *Nucleic Acids Res* **36**: e25.
40. **Fedurco M**, **Romieu A**, **Williams S**, *et al*. 2006. BTA, a novel reagent for DNA attachment on glass and efficient generation of solid-phase amplified DNA colonies. *Nucleic Acids Res* **34**: e22.
41. **Kircher M**, **Stenzel U**, **Kelso J.** 2009. Improved base calling for the Illumina Genome Analyzer using machine learning strategies. *Genome Biol* **10**: R83.
42. **Dohm JC**, **Lottaz C**, **Borodina T**, *et al*. 2008. Substantial biases in ultra-short read data sets from high-throughput DNA sequencing. *Nucleic Acids Res* **36**: e105.
43. **Erlich Y**, **Mitra PP**, **delaBastide M**, *et al*. 2008. Alta-Cyclic: a self-optimizing base caller for next-generation sequencing. *Nat Methods* **5**: 679–82.
44. **Rougemont J**, **Amzallag A**, **Iseli C**, *et al*. 2008. Probabilistic base calling of Solexa sequencing data. *BMC Bioinf.* **9**: 431.
45. **Drmanac R**, **Sparks AB**, **Callow MJ**, *et al*. 2010. Human genome sequencing using unchained base reads on self-assembling DNA nanoarrays. *Science* **327**: 78–81.
46. Applied Biosystems. A Theoretical Understanding of 2 Base Color Codes and Its Application to Annotation, Error Detection, and Error Correction. White Paper SOLiD™ System; 2008.
47. **Dimalanta ET**, **Zhang L**, **Hendrickson CL**, *et al*. 2009. Increased Read Length on the SOLiD™ Sequencing Platform. Poster SOLiD™ System.
48. **Zhu Z**, **Waggoner AS.** 1997. Molecular mechanism controlling the incorporation of fluorescent nucleotides into DNA by PCR. *Cytometry* **28**: 206–11.
49. **Bowers J**, **Mitchell J**, **Beer E**, *et al*. 2009. Virtual terminator nucleotides for next-generation DNA sequencing. *Nat Methods* **6**: 593–5.
50. **Pushkarev D**, **Neff NF**, **Quake SR.** 2009. Single-molecule sequencing of an individual human genome. *Nat Biotechnol* **27**: 847–52.
51. **Reinhardt JA**, **Baltrus DA**, **Nishimura MT**, *et al*. 2009. De novo assembly using low-coverage short read sequence data from the rice pathogen *Pseudomonas syringae* pv. *oryzae*. *Genome Res* **19**: 294–305.
52. **Diguistini S**, **Liao NY**, **Platt D**, *et al*. 2009. De novo genome sequence assembly of a filamentous fungus using Sanger, 454 and Illumina sequence data. *Genome Biol* **10**: R94.
53. **Miller JR**, **Delcher AL**, **Koren S**, *et al*. 2008. Aggressive assembly of pyrosequencing reads with mates. *Bioinformatics* **24**: 2818–24.
54. **Chen W**, **Ullmann R**, **Langnick C**, *et al*. 2009. Breakpoint analysis of balanced chromosome rearrangements by next-generation paired-end sequencing. *Eur J Hum Genet* DOI: 10.1038/ejhg .2009.21118 [Epub ahead of print].
55. **Zimin AV**, **Delcher AL**, **Florea L**, *et al*. 2009. A whole-genome assembly of the domestic cow, *Bos taurus*. *Genome Biol* **10**: R42.
56. **Zhou X**, **Su Z**, **Sammons RD**, *et al*. 2009. Novel software package for cross-platform transcriptome analysis (CPTRA). *BMC Bioinf.* **11**: S16.
57. **Kim JI**, **Ju YS**, **Park H**, *et al*. 2009. A highly annotated whole-genome sequence of a Korean individual. *Nature* **460**: 1011–5.
58. **Linsen SE**, **de Wit E**, **Janssens G**, *et al*. 2009. Limitations and possibilities of small RNA digital gene expression profiling. *Nat Methods* **6**: 474–6.
59. **Quail MA**, **Swerdlow H**, **Turner DJ.** 2009. Improved protocols for the Illumina Genome Analyzer sequencing system. *Curr Protoc Hum Genet* Chapter 18: Unit 18.2.
60. **Briggs AW**, **Stenzel U**, **Meyer M**, *et al*. 2009. Removal of deaminated cytosines and detection of *in vivo* methylation in ancient DNA. *Nucleic Acids Res* **38**(6): e87 [Epub ahead of print].
61. **Maricic T**, **Paabo S.** 2009. Optimization of 454 sequencing library preparation from small amounts of DNA permits sequence determination of both DNA strands. *Biotechniques* **46**: 51–2, 54–7.
62. **Rohland N**, **Hofreiter M.** 2007. Comparison and optimization of ancient DNA extraction. *Biotechniques* **42**: 343–52.
63. **Wheeler DA**, **Srinivasan M**, **Egholm M**, *et al*. 2008. The complete genome of an individual by massively parallel DNA sequencing. *Nature* **452**: 872–6.
64. **Harismendy O**, **Ng PC**, **Strausberg RL**, *et al*. 2009. Evaluation of next generation sequencing platforms for population targeted sequencing studies. *Genome Biol* **10**: R32.
65. **Chaisson MJ**, **Brinza D**, **Pevzner PA.** 2009. De novo fragment assembly with short mate-paired reads: Does the read length matter? *Genome Res* **19**: 336–46.
66. **Krause J**, **Briggs AW**, **Kircher M**, *et al*. 2009. A complete mtDNA genome of an early modern human from Kostenki, Russia. *Curr Biol* **20**: 231–6.
67. **Gnirke A**, **Melnikov A**, **Maguire J**, *et al*. 2009. Solution hybrid selection with ultra-long oligonucleotides for massively parallel targeted sequencing. *Nat Biotechnol* **27**: 182–9.

Illumina Exhibit 1010, Page 12 of 13

68. **Hodges E**, **Rooks M**, **Xuan Z**, *et al*. 2009. Hybrid selection of discrete genomic intervals on custom-designed microarrays for massively parallel sequencing. *Nat Protoc* **4**: 960–74.
69. **Briggs AW**, **Good JM**, **Green RE**, *et al*. 2009. Targeted retrieval and analysis of five Neandertal mtDNA genomes. *Science* **325**: 318–21.
70. **Meyer M**, **Stenzel U**, **Hofreiter M.** 2008. Parallel tagged sequencing on the 454 platform. *Nat Protoc* **3**: 267–78.
71. **Meyer M**, **Stenzel U**, **Myles S**, *et al*. 2007. Targeted high-throughput sequencing of tagged nucleic acid samples. *Nucleic Acids Res* **35**: e97.
72. **Erlich Y**, **Chang K**, **Gordon A**, *et al*. 2009. DNA Sudoku – harnessing high-throughput sequencing for multiplexed specimen analysis. *Genome Res* **19**: 1243–53.
73. **Meyer M**, **Kircher M.** 2010. Illumina sequencing library preparation for highly multiplexed target capture and sequencing. *Cold Spring Harb Protoc* DOI: 10.1101/pdb.prot5448.
74. **Pop M**, **Salzberg SL.** 2008. Bioinformatics challenges of new sequencing technology. *Trends Genet* **24**: 142–9.
75. **Richter BG**, **Sexton DP.** 2009. Managing and analyzing next-generation sequence data. *PLoS Comput Biol* **5**: e1000369.
76. **Quail MA**, **Kozarewa I**, **Smith F**, *et al*. 2008. A large genome center's improvements to the Illumina sequencing system. *Nat Methods* **5**: 1005–10.
77. **Batley J**, **Edwards D.** 2009. Genome sequence data: management, storage, and visualization. *Biotechniques* **46**: 333–4, 336.
78. IBM Research. 2009. IBM research aims to build nanoscale DNA sequencer to help drive down cost of personalized genetic analysis. In Loughran M, ed.; *Press Releases*, Vol. **2009**. New York: IBM.
79. **Eid J**, **Fehr A**, **Gray J**, *et al*. 2009. Real-time DNA sequencing from single polymerase molecules. *Science* **323**: 133–8.
80. **Astier Y**, **Braha O**, **Bayley H.** 2006. Toward single molecule DNA sequencing: direct identification of ribonucleoside and deoxyribonucleoside 5′-monophosphates by using an engineered protein nanopore equipped with a molecular adapter. *J Am Chem Soc* **128**: 1705–10.
81. **Albertorio F**, **Hughes ME**, **Golovchenko JA**, *et al*. 2009. Base dependent DNA-carbon nanotube interactions: activation enthalpies and assembly-disassembly control. *Nanotechnology* **20**: 395101.
82. **Medvedev P**, **Stanciu M**, **Brudno M.** 2009. Computational methods for discovering structural variation with next-generation sequencing. *Nat Methods* **6**: S13–20.
83. **Pepke S**, **Wold B**, **Mortazavi A.** 2009. Computation for ChIP-seq and RNA-seq studies. *Nat Methods* **6**: S22–32.
84. **Flicek P**, **Birney E.** 2009. Sense from sequence reads: methods for alignment and assembly. *Nat Methods* **6**: S6–12.
85. **Park PJ.** 2009. ChIP-seq: advantages and challenges of a maturing technology. *Nat Rev Genet* **10**: 669–80.
86. **Wall PK**, **Leebens-Mack J**, **Chanderbali AS**, *et al*. 2009. Comparison of next generation sequencing technologies for transcriptome characterization. *BMC Genomics* **10**: 347.
87. **Holt RA**, **Jones SJ.** 2008. The new paradigm of flow cell sequencing. *Genome Res* **18**: 839–46.
88. **Dalca AV**, **Brudno M.** 2010. Genome variation discovery with high-throughput sequencing data. *Brief Bioinf.* **11**: 3–14.

Illumina Exhibit 1010, Page 13 of 13