Review

Single-molecule DNA sequencing technologies for future genomics research

Pushpendra K. Gupta

Molecular Biology Laboratory, Department of Genetics and Plant Breeding, Chaudhary Charan Singh University, Meerut 250004, India

During the current genomics revolution, the genomes of a large number of living organisms have been fully sequenced. However, with the advent of new sequencing technologies, genomics research is now at the threshold of a second revolution. Several second-generation sequencing platforms became available in 2007, but a further revolution in DNA resequencing technologies is being witnessed in 2008, with the launch of the first single-molecule DNA sequencer (Helicos Biosciences), which has already been used to resequence the genome of the M13 virus. This review discusses several single-molecule sequencing technologies that are expected to become available during the next few years and explains how they might impact on genomics research.

Introduction

Since the onset of genomics research in the mid-1990s, whole-genome sequencing has been undertaken for a large number of prokaryotes and eukaryotes. However, the initial enthusiasm and euphoria for whole-genome sequencing activity has now given way to a demand for large-scale whole-genome resequencing (see Glossary) or the sequencing of target regions and/or metagenomes and pan-genomes (see Glossary), which require an increased sequencing speed and reduced costs [1,2]. With this in mind, in 2004, the National Human Genome Research Institute of the National Institutes of Health (NIH–NHGRI) announced a total of US$70 million in grant awards for the development of DNA sequencing (see Glossary) technologies that would reduce the cost of sequencing the human genome from US$ 3×10^9, the amount spent on the public Human Genome Project, to US$$10^3$ by 2014 (www.genome.gov/12513210). In October 2006, the X Prize Foundation (Santa Monica, CA, USA) announced a US$10 million 'Archon X Prize for Genomics' to the first private effort that could sequence 100 human genomes in 10 days for less than US$10 000 per genome (http://genomics.xprize.org/genomics/archon-x-prize-for-genomics). These incentives have contributed to an explosion of research activity in the development of new DNA sequencing technologies.

Further impetus for developing new DNA sequencing technologies came from the emerging field of personal genomics (see Glossary), which aims to study variations

Glossary

Bioinformatics: the application of molecular biology as an information science, especially involving the use of computers in genomics research.
DNA resequencing: sequencing an individual's specific DNA segment, for which sequence information is already available from one or more other individuals.
DNA sequencing: this term encompasses biochemical methods for determining the order of the nucleotide bases, adenine, guanine, cytosine and thymine, in a DNA oligonucleotide.
Epigenome: a form of the genome that has been modified, without alteration in DNA sequence, by changes like DNA methylation and/or histone modifications. The epigenome can differ among different cell types of the same individual. Thus, an organism can have several epigenomes, but only one genome. The epigenome controls the differential expression of genes in specific cells.
EST (expressed sequence tag): a short sub-sequence of a transcribed, spliced, nucleotide sequence, produced by one-shot sequencing of a cloned mRNA (i.e. sequencing several hundred base pairs from an end of a cDNA clone taken from a cDNA library).
Exon: a sequence of DNA that codes information for protein synthesis that is transcribed into spliced messenger RNA.
Flow cell: a reaction chamber containing templates tethered to a solid surface, to which nucleotides and reagents are iteratively applied and removed by washing for 'sequencing by synthesis' (SBS); most second- and third-generation sequencers use flow cells.
Genomics: the study of the entire genome of an organism; structural genomics includes whole-genome sequencing, whereas functional genomics aims to determine the functions of all genes.
Metagenome: the genetic material present in an environmental sample that consists of the genomes of the many individual organisms.
MicroRNAs (miRNA): single-stranded RNA molecules of 21–23 nucleotides in length. miRNAs are encoded by genes that are transcribed into untranslatable non-coding RNAs (ncRNA), which regulate gene expression at the mRNA level.
Non-Sanger sequencing systems: systems that do not make use of Sanger's dideoxy termination method, which up until a few years ago was the only DNA sequencing method used in all sequencing systems.
Pan-genome: the total gene repertoire in a given species; it includes the 'core genome', which is shared by all individuals, the 'dispensible genome', which is shared by some individuals, and the 'unique genome', which is unique to an individual.
Personal Genome Project (PGP): an initiative aiming to publish complete genomes and medical records of several volunteers, to be used for personalized medicine.
SAGE (serial analysis of gene expression): a technique used to produce a snapshot of the mRNA population in a sample of interest.
Single-pass sequencing: sequencing each individual DNA template only once, resulting in error-prone sequencing reads, depending on the type of sequencer used. However, if the single-pass sequencing covers the given genome or a part of the genome many times due to the presence of multiple copies of the sequence in the library, the sequences might overlap. Thus, when assembled to generate a contiguous sequence, most errors are removed.
Transcriptome: the set of all mRNA molecules or 'transcripts' that are produced either in all cells or in a particular cell type of an organism.
Two-pass sequencing: sequencing each individual template twice, giving two reads from the same position on the same strand to reduce the error rate. In the first pass, a template is sequenced as usual (pass 1); the primers are then melted off and the same template is sequenced a second time (pass 2). It requires covalent attachment of template strands to the surface in a stable and biocompatible fashion, and enables high accuracy in sequencing, because only sequences occurring in both reads are accepted as correct.
Zero-mode waveguides (ZMWs): sub-wavelength optical nanostructures, 50–200 nm in size, used as devices for focal volume confinement.

Corresponding author: Gupta, P.K. (pkgupta36@gmail.com).

0167-7799/$ – see front matter doi:10.1016/j.tibtech.2008.07.003 Available online 21 August 2008

Box 1. Personal genome projects launched during 2005–2008

Personal Genome Project (PGP)
In January 2006, George Church of Harvard Medical School launched PGP (see Glossary) to develop a context for genome–phenotype relationships. In the first instance, the genomes of 10 volunteers are being sequenced under 'Genome 10' or 'PGP10'. This project will expand to ~100 000 volunteers in 2008.

Yanhuang Project
In April 2007, a large-scale whole-genome sequencing project, described as 'The Yanhuang Project', was inaugurated in China. This project will involve sequencing the entire genomes of 100 Chinese individuals over three years. Once this project is completed, the Beijing Genomics Institute (BGI) aims to sequence the genomes of 1000 more people, including members of ethnic groups from other Asian countries.

1000 Genomes Project
A large multigenome international project, which aims to sequence the genomes of ~1000 individuals, was formally unveiled in January 2008. This project is a collaboration among the National Human Genome Research Institute of the US National Institutes of Health (NIH–NHGRI; Bethesda, MD USA), the Wellcome Trust Sanger Institute (Hinxton, UK) and the BGI (Shenzhen, China). It has been informally called the '1000 Genomes Project' and will most probably include the hundreds of individuals who also participated in the International HapMap Project – an ongoing study of genetic diversity – in addition to other individuals.

in the genomes of individual humans. To date, only a few personal genomes have been fully sequenced, and the genome sequences of Craig Venter and James Watson have been the only ones published [3,4]. Nevertheless, initiatives are underway to further increase the freely available sequence data for a large number of individual human genomes, as proposed in the Personal Genome Project and other similar projects launched recently (Box 1) [5,6].

To meet the increased sequencing demands, several non-Sanger ultra-high-throughput sequencing systems became commercially available in 2007 [7–12] (for non-Sanger sequencing systems, see Glossary). These were described as 'second generation' or 'next generation' sequencing systems, and included the following: Genome Sequencer 20/FLX (commercialized by 454/Roche); 'Solexa 1G' (later named 'Genome Analyzer' and commercialized by Illiumina/Solexa); SOLiD™ system (commercialized by Applied Biosystems); and Polonator G.007 (commercialized by Dover Systems). These developments have significantly reduced the cost of sequencing – from 1 cent for 10 bases to 1 cent for 1000 bases – and have simultaneously yielded an increase in DNA sequencing speed. As a result, several commercial genotyping services [e.g. Knome, DeCode, 23andMe and Navigenics (http://www.technologyreview.com/Biotech/20926/)] have also appeared on the market. These companies provide services to examine the genome of a person at as many as one million sites for US$1,000 to US$2,500. Future services might include the sequencing of whole genomes of individual humans, if there is a demand. However, there is also an attempt to regulate these gene tests, particularly in the states of New York and California, where gene test firms are being asked to obtain permits and conduct these tests only on the advice of a physician (http://tinyurl.com/55zzk8; http://tinyurl.com/5qgnr9).

Another non-Sanger DNA sequencing approach is the use of single-molecule sequencing (SMS), which only became available in 2008. This approach has been described as a 'third generation' or 'next-next generation' sequencing technology. It is anticipated that SMS will be much faster and cheaper, so that researchers in the near future should be able to pursue new scientific enquiries that are currently not possible owing to the prohibitive cost of sequencing. Whether or not SMS will fulfill this promise is debatable; compared with the earlier methods mentioned above, a significant reduction in sequencing costs has not yet been demonstrated for the only commercially available SMS technology, which is provided by Helicos Biosciences (Table 1).

Nevertheless, additional major efforts are underway to develop novel SMS technologies, which will be discussed below. In addition, much work is being done to address the outstanding issues that have become apparent during the development of these second- and third- generation sequencing technologies. Most of these outstanding issues (e.g. short read-lengths, higher error-rates, and the difficulty of managing massive amounts of data) are actually common to both second- and third-generation sequencing technologies, and bioinformatics (see Glossary) tools are being developed to deal with them. However, this article focuses on the third-generation SMS technologies, the so-called 'next-next generation'. These systems will probably constitute a significant fraction of future genomics research efforts, mainly because they will significantly reduce the cost and effort of sequencing in comparison with second-generation technologies – despite the fact that cost reductions continue to be made for second-generation approaches. SMS technology should not necessarily be considered as a panacea that can overcome all limitations associated with earlier technologies, and it is unlikely that it will completely replace all earlier sequencing technologies. Rather, it should be seen as another new and promising technology; the actual benefits and limitations of will only become fully known after it has been used by a large number of researchers. It currently appears that, in future, SMS technologies will either be the dominant DNA sequencing methods or co-exist with other technologies.

Table 1. A comparison of new-generation DNA sequencing platform

Features	Second-generation sequencers			Third-generation sequencers (single molecule-SBS)			
	454-FLX	Solexa	SOLiD	Helicos tSMS	PacBio SMRT	Nanopore and modified forms	ZS Genetics TEM
Read-length (bp)	240–400	35	35	30	100 000	Potentially unlimited?	Potentially unlimited?
Cost/human genome (US$)	1 000 000	60 000	60 000	70 000	Low	Low	Low
Run time (h/Gb)	75	56	42	~12	<1	>20	~14
Ease of use	Difficult	Difficult	Difficult	Easy	Easy	Easy	Easy

'State of the art' of single-molecule sequencing

The amplification of the target DNA by polymerase chain reaction (PCR) is an integral step in all second-generation sequencing technologies. However, it creates several problems, including the introduction of a bias in template representation, and the introduction of errors during amplification. Another problem associated with second-generation sequencing technologies is 'dephasing' of the DNA strands due to loss of synchronicity in synthesis (i.e. different strands being sequenced in parallel). This can lead to errors during sequencing. These limitations can be largely overcome by SMS. This method was proposed as early as 1989 by Keller and coworkers [13–15], and it has since been realized in the laboratory through several approaches, such as scanning probe microscopy, exonuclease sequencing, and sequencing by synthesis (SBS), among others (Figure 1) [16–20]. However, it was only in 2008 that the first commercial SMS instrument was launched. This system, developed by Helicos Biosciences, is based on SBS. Other promising technologies, also using the SBS principle, are likely to be commercially realized in the near future. Two other novel and promising SMS technologies include nanopore sequencing and transmission electron microscopy (TEM). In contrast to SBS, which relies on DNA synthesis, these systems involve no chemistry,. The following section discusses some of these SMS technologies.

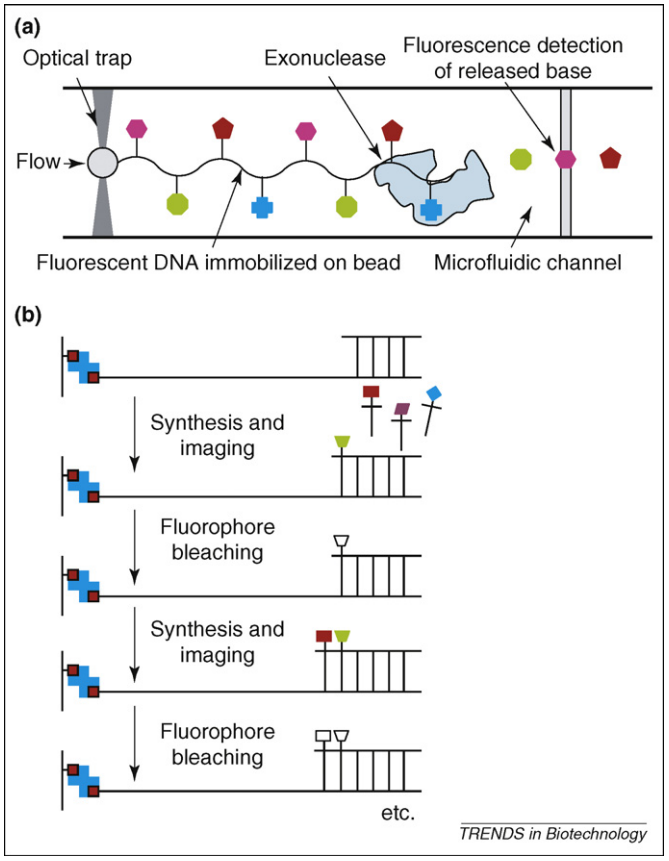

Figure 1. Two different approaches for SMS. **(a)** Exonuclease sequencing. Enzymatic digestion is used to cleave one base at a time from a transcript that is obtained by incorporation of fluorescent nucleotides; each base is cleaved by the exonuclease and identified by single-molecule fluorescence spectroscopy. **(b)** Sequencing by synthesis (SBS). The enzyme-catalyzed addition of each base is monitored by a fluorescence signal (e.g. by using nucleotides with fluorescent bases); in this method, each incorporated, labeled base is identified with the help of a label – several methods for identification are available – and the label is removed before another labeled base is incorporated. The four nucleotides are either labeled with the same dye (e.g. Cy5 in the tSMS method of Helicos Biosciences) or differently labeled (as in several other SBS methods). Reproduced with permission from [20].

Sequencing by synthesis

SBS on a solid surface in a flow cell (see Glossary) is the DNA sequencing technique most commonly used in the newest second- and third-generation sequencing systems. The methods used in these systems are based on the detection of individual nucleotides that are incorporated during DNA synthesis. However, the techniques used for labeling the nucleotides needed during DNA synthesis differ from those used to determine the identity of incorporated, labeled nucleotides. Below, we discuss three SMS methods, based on SBS, that are likely to be commercialized within the next few years, although several other methods are also being tried.

True single-molecule sequencing (*tSMS*^{TM}). This proprietary method, which is commercialized by Helicos Biosciences (Cambridge, MA, USA; http://www.helicosbio.com/), has recently been used to resequence the entire genome of the virus M13, which is approximately one million times smaller than the human genome [21]. In this technology, the target DNA is used for the construction of a library of poly(dA)-tailed templates, which pair with millions of poly(dT)-oligonucleotides that are anchored to a glass cover-slip. The positions of each of these individual poly(dT) oligos – and hence those of the respective paired poly(dA)oligos – on the cover slip are determined by camera imaging. The sequence of each poly(dA)-tailed fragment is determined by adding nucleotides – each labeled with the same cyanine dye Cy5 (a non-radioactive fluorescent dye) – in a cyclic manner, one at a time. The incorporation of nucleotides to each poly(dT) – or, indeed, the lack of incorporation, depending upon complementarity – enables faithful copying of the paired poly(dA)-tailed templates for sequencing. The events of nucleotide incorporation are imaged with a camera and used to obtain ~30-base-long sequences for each paired poly(dA)-tailed fragment (Figure 2).

During the resequencing of the M13 genome with tSMS, the error rate was reduced by performing a so-called 'two-pass' sequencing (see Glossary). Misincorporations arising from dephasing, which results from asynchronous synthesis of fragments, were circumvented by monitoring each template molecule individually, thereby obviating the need for synchronization between different template molecules. This also enabled homopolymers to be read accurately. By contrast, in other sequencing methods, this stage often leads to errors, because several molecules of a particular nucleotide need to be incorporated in the same cycle, when the enzyme encounters a homopolymer segment in a particular template. The other available templates with no homopolymers at that point of the sequence use only a single molecule of the nucleotide during this cycle, thus causing dephasing as a result of asynchronous synthesis.

FRET-based approach. VisiGen Biotechnologies (Houston, TX, USA; http://www.visigenbio.com) is developing an approach that uses fluorescence resonance energy transfer

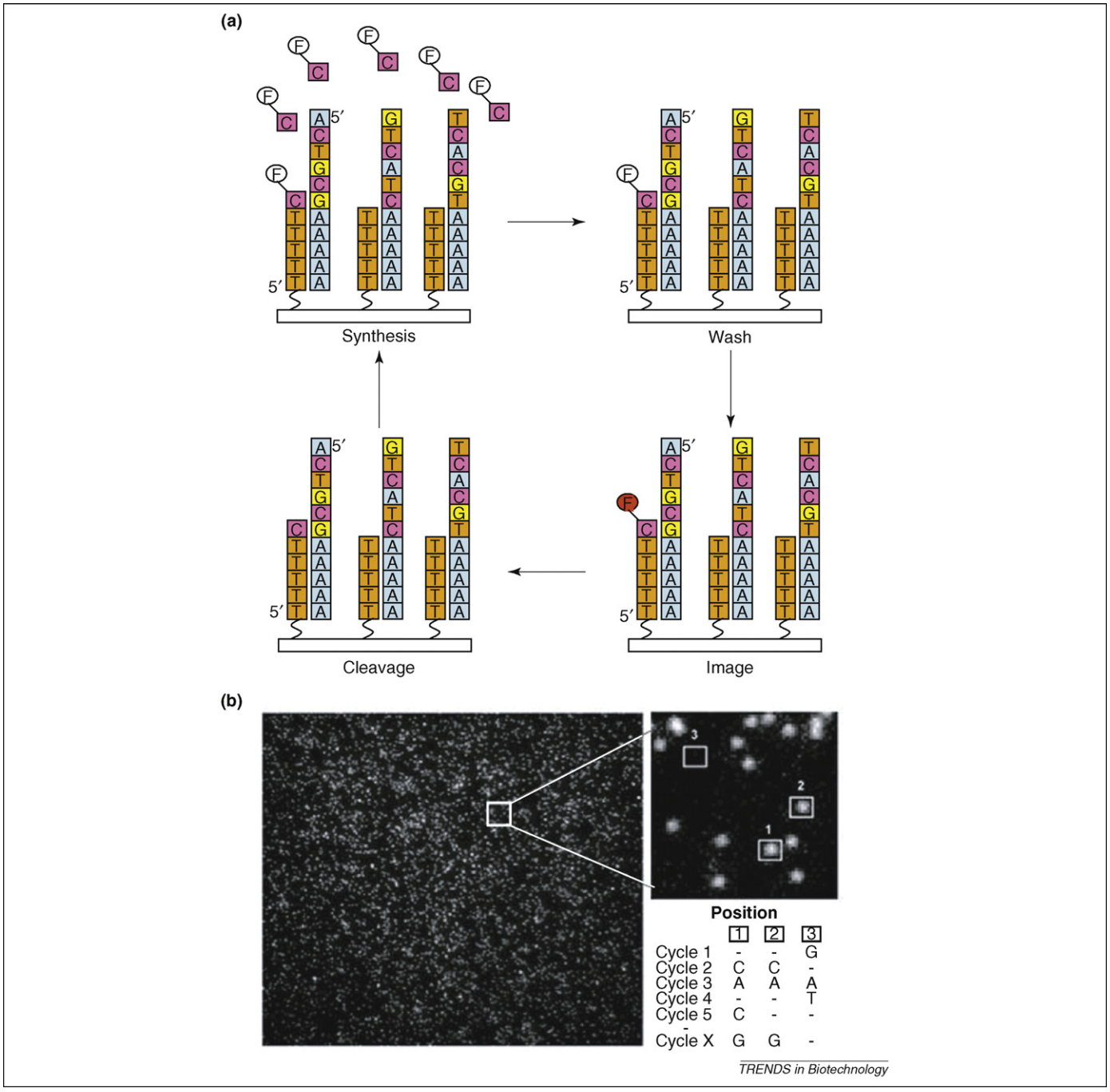

Figure 2. The true single-molecule sequencing (tSMS) technology for DNA sequencing (Helicos Biosciences): **(a)** Schematic illustration showing the different steps involved in tSMS (adapted from [65]). Four steps are shown in a clockwise order starting from top left: (i) incubation of flow cell (glass surface) with a dye-labeled nucleotide and its incorporation; (ii) washing step to remove all unincorporated, labeled nucleotide molecules; (iii) camera imaging of the incorporated nucleotide through the attached Cy5 label (exciting at 647 nm); (iv) cleavage of dye–nucleotide linker to release the dye label. In the example shown, the incorporated nucleotide is cytosine (C) and the label is cyanine Cy5 (F); **(b)** example for an image, and raw data obtained with tSMS technology (from http://www.helicosbio.com/Technology/TrueSingleMoleculeSequencing/tabid/64/Default.aspx). The left picture shows an image of poly(dT) templates after incubation with a dye-labeled nucleotide. The inset on the right side shows a close-up view of individual, single molecules, and the positions of three templates are marked (1–3) after the second cycle of incubation with labeled cytosine. Cytosine has been incorporated at positions 1 and 2, but not at position 3. Figures reproduced with expressed permission of Helicos Biosciences Corporation.

(FRET). In this system, the DNA sequence is read in real time by monitoring a polymerase as it incorporates bases into a DNA strand. The polymerase that is used contains a donor fluorophore, and each nucleotide carries one of four differently colored acceptor fluorophores. When a nucleotide is incorporated, the proximity of donor and acceptor fluorophores results in a FRET signal. This signal is specific for the particular nucleotide incorporated at this position, owing to its particular fluorophore label. After the nucleotide has been incorporated, the pyrophosphate containing the fluorophore is released, thereby quenching the signal and preparing the reaction for the next incorporation step. In this respect, the FRET-based SMS–SBS approach might be regarded as an improvement on the Helicos Biosciences tSMS technology, which requires the cyclic addition of reagents, increasing the time and cost of sequencing. VisiGen envisages that this approach will reach the market in 2009, and proposes a sequencing speed

of one million bases per second, which would mean that an individual human genome might be fully sequenced within an hour.

SMRT sequencing. Single-molecule real-time sequencing (SMRT™) is another proprietary approach that is being pursued by Pacific Biosciences (PacBio, Menlo Park, CA, USA; http://www.pacificbiosciences.com). It involves the use of so-called SMRT chips, each made up of a 100-nm-thick metal film and containing thousands of zero-mode waveguides (ZMWs; see Glossary), which are cavities of 10–50 nm in diameter. Within each ZMW, a DNA polymerase molecule is attached at the bottom and is visualized as DNA is synthesized from a single-stranded DNA molecule template (Figure 3) [22]. This reaction uses fluorophore-labeled nucleotides, but the label is attached to the phosphate group rather than to the base, and each nucleotide is labeled with a different fluorophore. When a nucleotide is incorporated during DNA synthesis, the attached fluorophore lights up owing to laser-beam-mediated illumination of a small detection volume (20 zeptoliters = 20×10^{-21} liters). This allows identification of each incorporated nucleotide. During formation of the phosphodiester bond, a nucleotide is held up in the detection volume for a much longer time (milliseconds) than the time (microseconds) needed for a nucleotide to diffuse in and out of the detection volume. This increase in time facilitates detection. During this detection process, the other unincorporated nucleotides float in the dark in the un-illuminated volume of ZMW and do not light up. As the fluoropohore is attached to the phosphate group, it has to be cleaved and released before the next nucleotide is incorporated; the phosphate–dye complex is actually released and quickly diffused out of the detection volume after incorporation of the base. This set-up enables nucleotides to be incorporated at a speed of ten bases per second, giving rise to a chain of thousands of nucleotides in length within minutes [23]. Importantly, simultaneous and continuous detection occurs across all of the thousands of ZMWs that are located on the SMRT chip in real-time, which facilitates the determination of thousands of sequences, each sequence thousands of bases long. The proof-of-concept for this technique has already been provided with the use of synthetic DNA templates of known sequences.

PacBio anticipates selling their instruments in 2010 or 2011 at a price comparable to that of the ABI SOLiD or

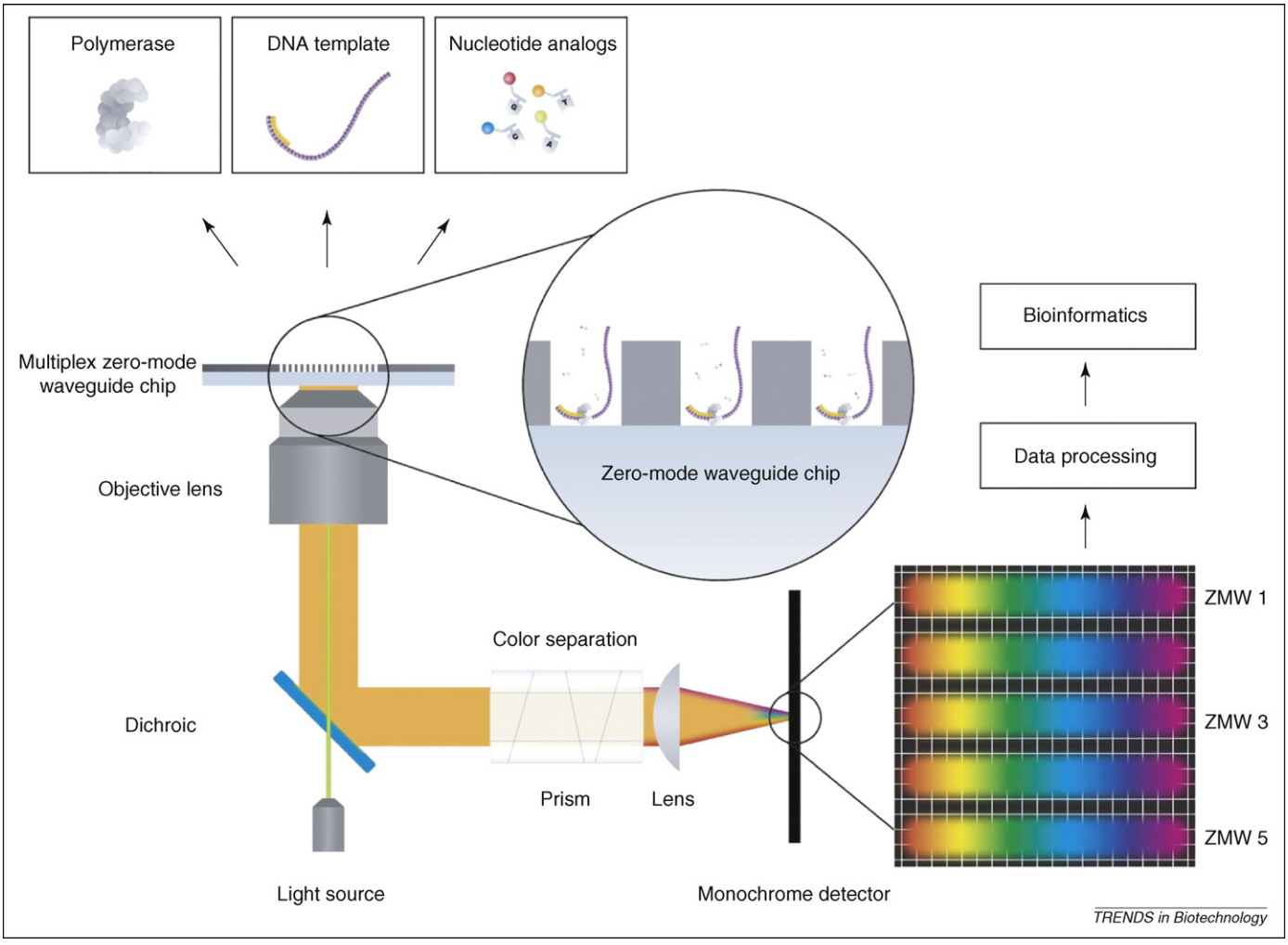

Figure 3. A schematic illustration representing the highly parallel optic system used in Single Molecule Real-Time (SMRT) DNA sequencing technology (Pacific Biosciences). This method uses SMRT chips that contain thousands of zero-mode waveguides (ZMWs). The presence of a fluorescent dye within the detection volume (within the ZMWs) indicates nucleotide incorporation. This leads to a light flash that is separated into a spatial array, from which the identity of the incorporated base can be determined. Figure reproduced from the document 'Pacific Biosciences Technology Backgrounder' (dated 2/2/2008), with expressed permission of Pacific Biosciences.

Illumina 'Genome Analyzer'. PacBio claims that, by 2013, the technology will be able to give a 'raw' human genome sequence in less than 3 min, and a complete high-quality sequence in just 15 min (http://www.bio-itworld.com/BioIT_Content.aspx?id=71746andamp;terms=Feb+12+2008+Pacific+Biosciences).

Nanopore sequencing

The idea of nanopore sequencing was first conceived in 1989 by David Deamer of the University of California (Santa Cruz, CA, USA) (http://www.futurepundit.com/archives/000017.html). It is believed to have great promise, and many researchers in the field are eagerly waiting to see the method working at a commercial scale. The basic outline of nanopore sequencing and some of its modifications are discussed below.

Nanopore sequencing technology involves the use of a very thin membrane that contains nanopores of ~1.5–2 nm in diameter. The target single-stranded DNA is placed to one side of the membrane, and a current is applied across the nanopore [24]. The negatively charged DNA traverses through the channel (the nanopore), towards the positive charge, thus blocking the channel and generating a change in the electrical conductance of the membrane. This, in turn, leads to alternations in the current – in the range of picoamperes (pA) – and these changes can be measured in an electric circuit. This enables researchers to discriminate between DNA molecules with different sequences (Figure 4). It has been shown that the duration of the translocation of a polynucleotide through a nanopore channel depends on its sequence rather than on its length. Also, the nanopore approach is able to distinguish between polynucleotides that are similar in length and composition (GC:AT ratio) but which differ in sequence, even if at only a single position (see [25,26] and references therein). However, despite enormous activity in this research area, the nanopore sequencing approach has not been successfully used for the sequencing of any significant number of nucleotides, and most applications are at the initial proof-of-concept stage. It is estimated that the basic technology might take at least another five years to become commercially available.

A major drawback of the nanopore technology is that the DNA molecule might pass through the nanopore at a speed that is too fast to enable the resolution of individual bases. Therefore, efforts have been made to overcome this problem through modifications to the basic nanopore technology desribed. Two of these specific modified nanopore-sequencing methods will be discussed below.

'*Hybridization*'-*assisted nanopore sequencing* (*HANS*). This proprietary technique of the company NABsys (http://nabsys.com/) was developed as a joint venture with Brown University (both at Providence, RI, USA). It combines nanopore sequencing with sequencing by hybridization (SBH). The approach involves hybridizing each molecule being sequenced with individual 6-mers from a library that consists of all possible 6-mer nucleotide probes. The resulting hybrid duplex DNA will pass through a nanopore, and the resulting changes in current can be measured and used for the detection of sites of hybridization for each probe, because the effect of duplex DNA hybrid segments on the current differs from that of a single-stranded DNA molecule. This approach should give a full-length map for each probe, such that sites of a large number of 6-mer probes – each with known sequence – encompassing the entire genome will be known and can be used to determine the DNA sequence.

A technique similar to HANS has been developed by Complete Genomics (Mountain View, CA, USA) in collaboration with BioNanomatrix (Philadelphia, PA, USA; http://www.technologyreview.com/Biotech/20640). This technology involves hybridization of 5-mer probes – in groups of several differently labeled 5-mer probes at a time – with the target DNA molecule, which is threaded into a microfluidic device with channels of 100 nm in diameter. The fluorescent label signals arising from the hybrid duplex DNA segments are recorded using a special camera – unlike HANS, which measures changes in current – and provide the hybridization sites of all 5-mers used. The process is repeated with different sets of 5-mers until the entire target DNA molecule is covered. From these sites, the complete target sequence is inferred. This can be several thousand bases long, which is the major advantage of this technique. The company claims that this technology should become available within five years and that it should cost only US$100 to sequence a human genome.

'*Design polymer*'-*assisted nanopore sequencing*. This technology has been developed by LingVitae (Oslo, Norway; http://www.lingvitae.com/) and involves the conversion of target DNA into a magnified form, the so-called 'design polymer'. In this design polymer, each nucleotide of the original DNA segment is replaced, one by one, by a block of 20 nucleotides (using a proprietary technology). The block itself consists of two types of 10-mers. These blocks constitute a two-unit code (e.g. 01, in which 0 represents one 10-mer, and 1 represents the other 10-mer within the 20 nucleotide block). Using this binary code, the four bases of DNA can be encoded as 00, 01, 10 and 11. The 10-mers are optically detected with the use of fluorescently labeled probes that hybridize with the codes and which are unzipped during translocation through the nanopores. The use of 'design polymers' and the 'unzipping' of probes will overcome the main problem of nanopore sequencing – that a single 0.34-nm-long nucleotide passes too fast through a nanopore – because the 20-nucleotide-long coding units measure ~7 nm, which is long enough to allow for their discrimination. In this approach, the size of the coding unit can range from 4 to 40 nucleotides.

Transmission electron microscopy for DNA sequencing

In this novel SMS platform, DNA sequences are read directly with the help of a specialized transmission electron microscope (TEM). This approach is being developed by ZS Genetics (ZSG; North Reading, MA, USA; http://zsgenetics.com/application/GenSeq/index.html), which has been accepted as one of the competitors for the previously mentioned 'Archon X Prize'. The details of this process were presented for the first time at the recent Cambridge Healthtech Institute (CHI) Sequencing Conference in San Diego, held from 23–24 April, 2008 (http://www.healthtech.com/Conferences_Overview.aspx?id=60306andc=542). The technology involves the linearization of the target DNA molecule, followed by synthesis of a complimentary strand,

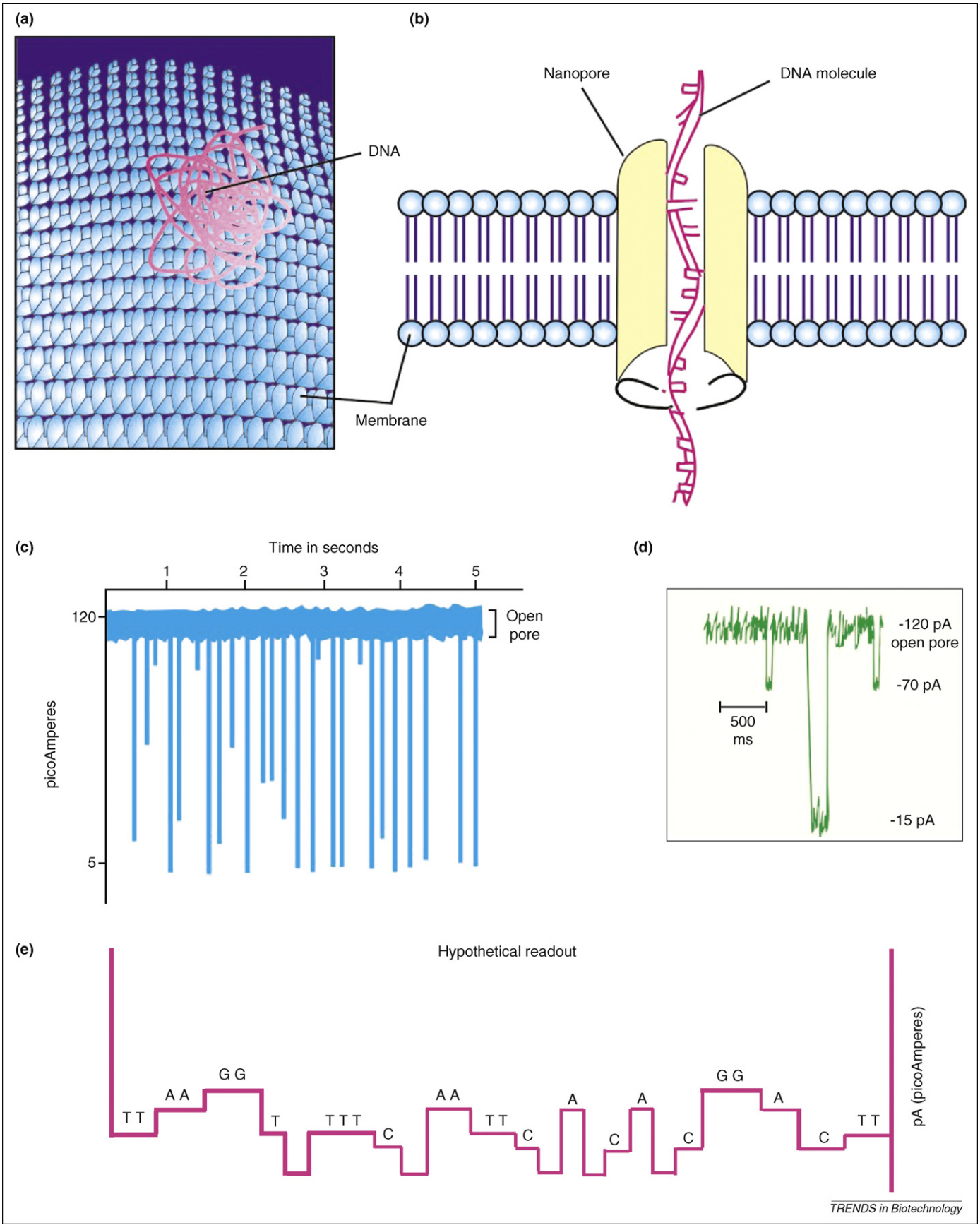

Figure 4. A schematic representation of the basic principle involved in nanopore sequencing. **(a)** Shown here is a surface view of a membrane with numerous nanopores. **(b)** Shown is a cross-section of a part of the membrane, with a single-stranded DNA molecule traversing through a nanopore (not to scale). **(c)** Example of a peak profile obtained as a result of current alterations (in pA) that occur when DNA is passing through the nanopore. **(d)** A hypothetical close-up of peaks that show expected changes in current that would arise from a DNA with homopolymer segments, which are indicated by the presence of three distinct peaks. **(e)** A hypothetical example of a sequence read-out that demonstrates single-base resolution of the technique based on minor but nevertheless detectable differences in current alterations. The extent of an observed alteration in current can be correlated to a particular nucleotide passing through the nanopore.

whereby three of the four bases are labeled with heavy atoms, and the fourth base remains unlabeled. Given that atoms such as C, O, N, H and P present in DNA have low atomic number (Z = 1–15), natural DNA is transparent when viewed with TEM. However, bases labeled with heavier elements, with high Z values (e.g. iodine with Z = 53; bromine with Z = 35), make the DNA heavier and, therefore, visible under TEM. Thus, when the resulting complementary strand is observed under TEM, the four bases can be discriminated by the size and intensity of dots representing the four bases. ZSG claims to be able to generate read lengths of 10 000 to 20 000 bases, with a rate of 1.7 billion base pairs (equivalent to 400 million base pairs, with 4x coverage) per day, and has already released images of a 23-kb piece of DNA. The company claims that it will be able to produce several-fold increases in the sequencing potential with future improved versions of this technology.

Table 1 demonstrates a comparison of these recent SMS technologies with the second-generation technologies, which emerged in 2007 and still remain highly popular among users. The table also demonstrates the advantages that future SMS technologies promise to offer with regards to time and cost.

Outstanding crucial issues

Despite these advantages, the SMS technologies share with other recent technologies some of the limitations and outstanding problems. These problems have been widely discussed, and they need to be addressed before any of these technologies can be extensively used in genomics research in a user-friendly and cost-effective manner.

Bioinformatics: Assembly of short read-lengths, read quality and the ability to produce paired-end reads

There has been much debate on the short read-lengths that are typically generated by new sequencing technologies. These cannot be easily used to reconstruct the much longer *de novo* sequences needed for whole-genome sequencing ([27] and references therein). However, all the novel SMS technologies – with the exception of the tSMS technology – seem, at least in theory, to be able to yield much longer read-lengths. Therefore, although several algorithms have been developed in attempts to overcome the limitations associated with short sequence read-lengths [28], short read-lengths might not be a crucial problem for most of the novel SMS technologies. However, some additional problems arise, and these are receiving the increased attention of bioinformaticians [27]. For instance, the sequence read-quality, which causes errors, needs to be improved (besides other reasons, kinetics of the enzyme could be one possible reason for high error-rate, and efforts are being made to improve the enzyme). Another problem is the inability to generate paired-end sequence reads, which are needed to determine the orientation and relative positions of contigs during the assembly process. Finally, the management of the enormous amount of data generated by the new sequencing technologies is becoming a serious problem.

Enrichment of selected targets for resequencing

It has been recognized that, in most cases, resequencing the whole genome might be unnecessary; rather, sequencing of only selected regions of the genome might be adequate. For instance, in some cases, researchers might need to sequence only the exons (see Glossary) in single-copy genes or only the specific regions that are known to be associated with certain diseases. In other cases, one might be interested in studying genomic DNA sequences that are methylated or associated with histone modifications, or which are bound to proteins and/or RNA involved in regulation of gene expression, or even those that are actively involved in transcription at a particular time or in a given space. Therefore, techniques are being developed to enrich the target DNA in a desired manner, before it is used for sequencing [29–32].

Conclusions and perspectives

The demand for large-scale DNA sequencing has dramatically increased in recent years. As a result, we are witnessing the development of several mutually competitive DNA sequencing systems that are much faster and cheaper and which have a higher level of precision. These new systems involve ultra-high-throughput sequencing of a large number of DNA fragments in parallel and include the following three classes of sequencing systems: the first-generation systems, which are based on the old, Sanger dideoxy method, but which include improved technologies; the second-generation systems, which are mainly based on amplified SMS; and the third-generation systems, which are based on true SMS (e.g. the tSMS technology of Helicos Biosciences) and which do not involve amplification steps. These ultra-high-throughput DNA sequencing technologies are already producing massive amounts of data at an unprecedented rate, thus also making it necessary to develop bioinformatics tools and hardware for storage, retrieval and maintenance of data. Of the above methods, the SMS technology has attracted particular attention and excitement. However, it remains to be seen whether this technology will become the dominant technology in the future or whether it will co-exist with other second-generation technologies, because these latter methods have only recently taken over from the first-generation technologies as the market leaders.

Among the SMS technologies discussed in this review, the tSMS technology developed by Helicos Biosciences is the only technology that has already been commercialized and which, through resequencing of the M13 genome, has provided proof-of-concept. However, currently, this technology seems to be equally expensive, promises no more than 30-base read-lengths, and has yet to be used to sequence a complex genome of gigabase size. By contrast, several other SMS technologies that are still under development (e.g. SMRT, HANS, Design Polymer and TEM) are more promising with regards to speed and cost and also have the potential to yield longer read-lengths. Thus, short read-length, which is a serious limitation of the second-generation sequencing instruments, might no longer be an issue with future SMS instruments. The projected cost and sequencing time in these technologies also appears to be much lower. For instance, Complete Genomics projects a human genome sequence for a mere US$100 (http://www.technologyreview.com/Biotech/20640/), and PacBio promises it within 3–15 min. Eminent

scientists, such as Craig Venter, the founder of 'The Institute for Genomic Research' (TIGR) and Celera Genomics, as well as George Church of Harvard Medical School, have expressed the view that such claims might indeed be realistic (www.bio-itworld.com).

If the promises described in this review are fulfilled, one can look forward to yet another genomics revolution, in which individual scientists in small laboratories can undertake genomics research. This is a stark contrast to the multinational, multi-billion dollar projects of the past. Consequently, many more microbial, animal and plant genomes could be sequenced to investigate how genomes evolved [33–39]. The sequencing of increasing numbers of metagenomes can lead to environmental clues in terms of diverse and novel DNA sequences occurring in each individual environment [40,41]. Also, sequencing of pan-genomes (see Glossary) will allow identification of 'unique sequences', 'core sequences' and 'dispensable sequences' in the genome of an individual species [42,43]. Similarly, sequencing of genomes of different species within a genus will allow a study of variation among genomes within a genus [44,45], and resequencing the human genome will permit study of copy number variations (CNVs) and structural variations (SVs) associated with several human diseases [46,47]. Enormous activity would also be expected in the areas of functional genomics; for example, the development of additional EST–SAGE (expressed sequence tags–serial analysis of gene expression; see Glossary) data, and the analysis of transcriptomes, epigenomes and microRNA (see Glossary), as witnessed recently [48–56]. Several recent studies on transcriptome analysis in yeasts, *Arabiopsis* and mouse used the so-called 'mRNA-Seq' method of analysis using Solexa/SOLiD systems. These analyses indicated that the new sequencing technologies might perhaps replace microarray technology, which until now has been considered extremely powerful for expression studies [57–61]. Future SMS technologies will certainly also have an impact on this area of genomics research.

However, it should be noted that, although SMS is a promising technology that draws a lot of attention, various groups are simultaneously developing other ultra-high-throughout, low-cost sequencing technologies that use both the Sanger method and amplicative non-Sanger methods [62–64]. It is therefore possible that SMS technology will co-exist with other sequencing technologies – including even the Sanger method. The above developments in sequencing technologies will, however, create work for bioinformaticians, because data management will be a challenge in any case. They are already aware of this problem and are busy in dealing with it. We look forward to another exciting period of genomics research.

Acknowledgements

The Indian National Science Academy (INSA) awarded the position of INSA Honorary Scientist to P.K.G.; the Head of the Department of Genetics and Plant Breeding, Chaudhary Charan Singh University, Meerut, India provided the facilities; Ajay Kumar helped in various ways during the preparation of this manuscript; Sachin Rustgi helped in improving the quality of the figures; and the Editor subjected the manuscript to several rounds of critical reading, which led to significant improvement of this review.

References

1 Bentley, D.R. (2006) Whole genome resequencing. *Curr. Opin. Genet. Dev.* 16, 545–552
2 Olson, M. (2007) Enrichment of super-sized resequencing targets from the human genome. *Nat. Methods* 4, 891–892
3 Levy, S. *et al.* (2007) The diploid genome sequence of an individual human. *PloS Biol.* 5, e254
4 Wheeler, D.A. *et al.* (2008) The complete genome of an individual by massively parallel DNA sequencing. *Nature* 452, 872–876
5 Kaiser, J. (2008) A plan to capture human diversity in 1000 genomes. *Science* 319, 395
6 Church, G.M. (2006) Genomes for all. *Sci. Am.* 294, 46–54
7 Shendure, J. *et al.* (2004) Advanced sequencing technologies. *Nat. Rev. Genet.* 5, 335–344
8 Metzker, M.L. (2005) Emerging technologies in DNA sequencing. *Genome Res.* 15, 1767–1776
9 Chan, E.Y. (2005) Advances in sequencing technology. *Mutat. Res.* 573, 13–40
10 Mitchelson, K. (ed.) (2007) *New High Throughput Technologies for DNA Sequencing and Genomics* (Vol. 2), Elsevier
11 Mardis, E.R. (2008) The impact of next generation sequencing technology on the genetics. *Trends Genet.* 24, 133–141
12 Gupta, P.K. (2008) Ultrafast and low cost sequencing methods for applied genomics research. *Proc. Natl. Acad. Sci. India* 78, 91–102
13 Jett, J.H. *et al.* (1989) High-speed DNA sequencing: an approach based upon fluorescence detection of single molecules. *J. Biomol. Struct. Dyn.* 7, 301–309
14 Davis, L.M. *et al.* (1991) Rapid DNA sequencing based upon single molecule detection. *Genet. Anal. Tech. Appl.* 8, 1–7
15 Harding, J.D. and Keller, R.A. (1992) Single-molecule detection as an approach to rapid DNA sequencing. *Trends Biotechnol.* 10, 55–57
16 Brakmann, S. *et al.* (2002) A further step towards single-molecule sequencing: *Escherichia coli* exonuclease III degrades DNA that is fluorescently labeled at each base pair. *Angew. Chem. Int. Ed. Engl.* 41, 3215–3217
17 Werner, J.H. *et al.* (2003) Progress towards single-molecule DNA sequencing: a one color demonstration. *J. Biotechnol.* 102, 1–14
18 Crut, A. *et al.* (2005) Detection of single DNA molecules by multicolor quantum-dot end-labeling. *Nucleic Acids Res.* 33, e98
19 Braslavsky, I. *et al.* (2003) Sequence information can be obtained from single DNA molecules. *Proc. Natl. Acad. Sci. U. S. A.* 100, 3960–3964
20 Bayley, H. (2006) Sequencing single molecules of DNA. *Curr. Opin. Chem. Biol.* 10, 628–637
21 Harris, T.D. *et al.* (2008) Single molecule DNA sequencing of a viral genome. *Science* 320, 106–109
22 Levene, M.J. *et al.* (2003) Zero-mode waveguides for single molecule analysis at high concentrations. *Science* 299, 682–686
23 Korlach, J. *et al.* (2008) Selective aluminum passivation for targeted immobilization of single DNA polymerase molecules in zero-mode waveguide nanostructures. *Proc. Natl. Acad. Sci. U. S. A.* 105, 1176–1181
24 Kasianowicz, J.J. *et al.* (1996) Characterization of individual polynucleotide molecules using a membrane channel. *Proc. Natl. Acad. Sci. U. S. A.* 93, 13770–13773
25 Rhee, M. and Burns, M.A. (2006) Nanopore sequencing technology: research trends and applications. *Trends Biotechnol.* 24, 580–586
26 Ryan, D. *et al.* (2007) Towards nanoscale genome sequencing. *Trends Biotechnol.* 25, 385–389
27 Pop, M. and Salzberg, S.L. (2008) Bioinformatics challenges of new sequencing technology. *Trends Genet.* 24, 142–149
28 Karow, J. (2008) Moving from simulations to real data, short read assemblers. In Sequence, 2008 March 18, Vol. 2, Iss. 12
29 Albert, T.J. *et al.* (2007) Direct selection of human genomic loci by microarray hybridization. *Nat. Methods* 4, 903–905
30 Okou, D.T. *et al.* (2007) Microarray-based genomic selection for high-throughput resequencing. *Nat. Methods* 4, 907–909
31 Porreca, G.J. *et al.* (2007) Multiplex amplification of large sets of human exons. *Nat. Methods* 4, 931–936
32 Hodges, E. *et al.* (2007) Genome-wide *in situ* exon capture for selective resequencing. *Nat. Genet.* 39, 1522–1527
33 Jackson, S. *et al.* (2006) Comparative sequencing of plant genomes: choices to make. *Plant Cell* 18, 1100–1104
34 Pennisi, E. (2007) The greening of plant genomics. *Science* 317, 317

35 Paterson, A.H. (2006) Leafing through the genomes of our major crop plants: strategies for capturing unique information. *Nat. Rev. Genet.* 7, 174–184
36 Paterson, A.H. *et al.* (2008) The fruits of tropical plant genomics. *Trop. Plant Biol.* 1, 3–19
37 McNally, K.L. *et al.* (2006) Sequencing multiple and diverse rice varieties, connecting whole-genome variation with phenotypes. *Plant Physiol.* 141, 26–31
38 Negrão, S. *et al.* (2008) Integration of genomic tools to assist breeding in the *japonica* subspecies of rice. *Mol. Breeding* 22, 151–168
39 Medini, D. *et al.* (2008) Microbiology in the post-genomic era. *Nat. Rev. Microbiol.* 6, 419–430
40 Schloss, P.D. and Handelsman, J. (2005) Metagenomics for studying unculturable microorganisms: cutting the Gordian knot. *Genome Biol.* 6, 229–233
41 Nealson, K.H. and Venter, J.C. (2007) Metagenomics and the global ocean survey: what's in it for us, and why should we care. *ISME Journal* 1, 185–187
42 Vernikos, G.S. (2008) Genome watch: overtake in reverse gear. *Nat. Rev. Microbiol.* 6, 334–335
43 Morgante, M. *et al.* (2007) Transposable elements and the plant pan genome. *Curr. Opin. Plant Biol.* 10, 149–155
44 *Drosophila* 12 Genomes Consortium (2007) Evolution of genes and genomes on the *Drosophila* phylogeny. *Nature* 450, 203–218.
45 Kim, H. *et al.* (2008) Construction, alignment and analysis of twelve framework physical maps that represent the ten genome types of the genus *Oryza*. *Genome Biol.* 9, R45
46 Bennett, S.T. *et al.* (2005) Toward the 1.000 dollars human genome. *Pharmacogenomics* 6, 373–382
47 Lupski, J.R. (2007) Structural variation in the human genome. *N. Engl. J. Med.* 356, 1169–1171
48 Wold, B. and Myers, R.M. (2008) Sequence census methods for functional genomics. *Nat. Methods* 5, 19–21
49 Emrich, S.J. (2007) Gene discovery and annotation using LCM-454 transcriptome sequencing. *Genome Res.* 17, 69–73
50 Warthmann, N. *et al.* (2008) Highly specific gene silencing by artificial miRNA in rice. *PLoS ONE* 3, e1829
51 Lu, C. *et al.* (2008) Genome-wide analysis for discovery of rice microRNAs reveals natural antisense microRNAs (nat-miRNAs). *Proc. Natl. Acad. Sci. U. S. A.* 105, 4951–4956
52 Mardis, E.R. (2007) ChIP-seq: welcome to the new frontier. *Nat. Methods* 4, 613–614
53 Cokus, S.J. (2008) Shotgun bisulphate sequencing of the *Arabidopsis* genome reveals DNA methylation patterning. *Nature* 452, 215–219
54 Li, X. (2008) High-resolution mapping of epigenetic modifications of the rice genome uncovers interplay between DNA methylation, histone methylation, and gene expression. *Plant Cell* 20, 259–276
55 Lister, R. *et al.* (2008) Highly integrated single-base resolution maps of the epigenome in *Arabidopsis*. *Cell* 133, 523–536
56 Zhang, X. (2008) The epigenetic landscape of plants. *Science* 320, 489–492
57 Graveley, B.R. (2008) Power sequencing. *Nature* 453, 1197–1198
58 Wilhelm, B.T. *et al.* (2008) Dynamic repertoire of a eukaryotic transcriptome surveyed at single-nucleotide resolution. *Nature* 453, 1239–1243
59 Nagalakshmi, U. *et al.* (2008) The transcriptional landscape of the yeast genome defined by RNA sequencing. *Science* 320, 1344–1349
60 Mortazavi, A. *et al.* (2008) Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat. Methods* 5, 621–628
61 Cloonan, N. *et al.* (2008) Stem cell transcriptome profiling via massive-scale mRNA sequencing. *Nat. Methods* 5, 613–619
62 Fredlake, C.P. *et al.* (2008) Ultrafast DNA sequencing on a microchip by a hybrid separation mechanism that gives 600 bases in 6.5 minutes. *Proc. Natl. Acad. Sci. U. S. A.* 105, 476–481
63 Pihlak, A. *et al.* (2008) Rapid genome sequencing with short universal tiling probes. *Nat. Biotechnol.* 26, 676–684
64 Guo, J. *et al.* (2008) Four colour DNA sequencing with 3′-O-modified nucleotide reversible terminators and chemically cleavable dideoxynucleotides. *Proc. Natl. Acad. Sci. U. S. A.* 105, 9145–9150
65 Blow, N. (2008) DNA sequencing: generation next-next. *Nat. Methods* 5, 267–274