Chapter 11
Recent Advances in Sequencing Technology

John F. Thompson, Fatih Ozsolak, and Patrice M. Milos

Abstract As we celebrate the tenth anniversary of the sequencing of the first human genome, we recognize the remarkable technological innovation that now provides the ability to resequence thousands of human genomes a year. While the current methods of choice utilize amplification-based methods and the corresponding challenges of sample preparation that accompany these methods, new technologies that do not require amplification have emerged. Single-molecule sequencing methods have the potential to dramatically shape the next 10 years of technological progress driven by the continuing interest of driving the cost of whole genome sequencing below the $1000 cost threshold. Yet while whole genome sequencing remains of interest, sequencing technologies also enable new approaches for genome exploration and experimentation including direct RNA sequencing, complete transcript sequencing and real time methods for both nucleic acid and enzyme kinetics.

11.1 Introduction

The development of Maxam and Gilbert, and separately Sanger, sequencing methodologies in the mid-to late 1970s opened a new era for cataloging the genome of a multitude of organisms [1, 2]. While much of the early sequencing included bacteriophage, cDNA's, ribosomal RNAs and viral genomes using these newly developed methods, the Maxam and Gilbert technology was quickly replaced with

J.F. Thompson
NABsys Inc., Providence, RI, USA

F. Ozsolak
Helicos BioSciences Corporation, Cambridge, MA, USA

P.M. Milos (✉)
Pfizer Center for Therapeutic Innovation, Boston, MA, USA
e-mail: Patrice.M.Milos@pfizer.com

G. Spoto and R. Corradini (eds.), *Detection of Non-Amplified Genomic DNA*, Soft and Biological Matter, DOI 10.1007/978-94-007-1226-3_11,

the safer and more expeditious method of Sanger chain termination chemistries. In Sanger sequencing, a primer is used to hybridize to the template DNA and synthesis of the complementary DNA strand is initiated by the DNA polymerase. Initially, four separate reactions were used to distinguish each dNTP incorporation as well as the chain terminating ddNTP with termination occurring due to the lack of a hydroxyl moiety on the ddNTP. Altering the ratio of dNTP's and ddNTP's results in chain termination along the length of the molecule and the use of radioactive nucleotides allowed the electrophoretic and autoradiographic analysis of the newly formed DNA strands providing the sequence of the synthesized strand in four separate electrophoretic gel lanes. The technical advance of using fluorescently labelled ddNTP's, each A, C, G or T ddNTP containing a unique fluorescent dye with distinguishing spectral properties facilitated single tube chemistry as well as automated sequencing instrumentation using the differential fluorescence of the individual nucleotides to record the incorporation event along the DNA strand through a laser detection system [3].

The commercial opportunity for DNA sequencing was quickly realized when in 1981 two engineers from Hewlett-Packard started Applied Biosystems to manufacture and sell sequencing instrumentation to enable large scale sequencing in genome research laboratories across the globe. Continued improvements in instrumentation allowing ever increasing speed and high-throughput capacity using the same basic Sanger chain terminator chemistry culminated in the release of the ABI Prism® 3700 Genetic Analyzer which became the work horse for the Human Genome Project delivering sequence data at a volume never imagined possible. A consistent theme throughout the revolution of DNA sequencing we are experiencing involves scientists envisioning the need and technologists finding new, innovative solutions to meet those needs.

Initial uses of the sequencing technology focused on cDNAs and the genomes of small organisms, including the first complete 4.6-Mb *Escherichia coli* K-12 genome [4] representing the work of more than 250 people that required a 6 year effort. The draft human genome sequence was published concurrently by the publicly funded Human Genome Project [5] and the commercial venture Celera [6]. The government sponsored project took some 15 years and three Billion dollars yet prompted scientists and technologists alike to begin considering the potential for expanding the number of genomes beyond the initial draft genome sequences. In 2004, following the completion of the human genome sequence [7], the National Human Genome Research Institute under the guidance of Dr. Francis Collins, leader of the HGP, announced a major 70 million dollar funding initiative to fuel development of new DNA sequencing technologies with a goal of achieving the $1000 genome by 2014. In now looking back, the initial round of funding shaped the field of both amplified and unamplified DNA next-generation sequencing as it exists today (Table 11.1). Indeed the goal of achieving the $1000 genome by the year 2014 once again reiterates the notion that clarity of vision allows the developments necessary to achieve the desired objectives, often more quickly than projected.

Today, we have taken the basic principles of DNA sequencing chemistry described above and moved far beyond the throughput and yield envisioned by the early pioneers in sequencing. Further we have extended the bulk DNA methodology

Table 11.1 The initial genome technology grants funded by the National Human Genome Research Institute in 2004

Grantee	Funding amount	Company/Institution	Title
$100,000 Genome Grants: NHGRI's "Near-Term Development for Genome Sequencing"			
Stevan B. Jovanovich, Ph.D.	$6.1 million (3 years)	Microchip Biotechnologies Inc.	Microbead Integrated DNA Sequencer (MINDS) System
Gina L. Costa, Ph.D.	$5.4 million (3 years)	Agencourt Bioscience Corp.	Bead-based Polony Sequencing
Kenton Lohman, Ph.D.	$2 million (2 years)	454 Life Sciences Corp.	Massively Parallel High Throughput, Low Cost Sequencing
Marcel Margulies, Ph.D.	5 million (3 years)	454 Life Sciences Corp.	454 Life Sciences Massively Parallel System DNA Sequencing
John Williams, Ph.D.	$2.5 million (3 years)	Li-COR Inc.	Single-Molecule DNA Sequencing Using Charge-Switch dNTPs
Michael L. Metzker, Ph.D.	$2 million (3 years)	Human Genome Sequencing Center, Baylor College of Medicine	Ultrafast SBS (Sequencing by Synthesis) Method for Large-Scale Human Resequencing
Stephen R. Quake, Ph.D.	$1.8 million (3 years)	Stanford University	High-Throughput, Single-Molecule DNA Sequencing
Mostafa Ronaghi, Ph.D.	$1.8 million (3 years)	Stanford Genome Technology Center	Pyrosequencing Array for DNA Sequencing
Jingyue Ju, Ph.D.	$1.8 million (3 years)	Columbia University	An Integrated System for DNA Sequencing by Synthesis
Peter Williams, Ph.D.	$1.7 million (3 years)	Arizona State University	Multiplexed Reactive Sequencing of DNA
Steven A. Benner, Ph.D.	$800,000 (3 years)	University of Florida	Polymerases for Sequencing by Synthesis
Amit Meller, Ph.D.	$600,000 (2 years)	Rowland Institute at Harvard, Harvard University	Ultra-fast Nanopore Readout Platform for Designed DNA's
"$1000 Genome" Grants NHGRI's "Revolutionary Genome Sequencing Technologies"			
J. Michael Ramsey, Ph.D.	$2 million (2 years)	University of North Carolina, Chapel Hill	Nanotechnology for the Structural Interrogation of DNA
James Weifu Lee, Ph.D.	$700,000 (3 years)	Oak Ridge National Laboratory	"Computational Research & Development for Rapid Sequencing Nanotechnology"
James Weifu Lee, Ph.D.	$750,000 (3 years)	Oak Ridge National Laboratory	"Experimental Research & Development for Rapid Sequencing Nanotechnology"
Scott D. Collins, Ph.D.	$850,000 (2 years)	University of Maine	High-speed Nanopore Gene Sequencing
Steven A. Benner, Ph.D.	$800,000 (3 years)	University of Florida	DNA Sequencing Using Nanopores
Andre Marziali, Ph.D.	$650,000 (3 years)	University of British Columbia, Vancouver	Nanopores for Trans-Membrane Bio-Molecule Detection
Stuart Lindsay, Ph.D.	$550,000 (3 years)	Arizona State University	Molecular Reading Head for Single-Molecule DNA Sequencing
Ronald W. Davis, Ph.D.	$450,000 (2 years)	Stanford University	Single Molecule Nucleic Acid Detection with Nanopipettes

Data Compiled from the National Human Genome Research Institute release announcing these awards (http://www.genome.gov/12513210)

utilized in these early technologies to now allow the detection and sequencing of single molecules of DNA and RNA. This chapter describes the rapid pace of amplified sequencing technology developments and the emergence of non-amplified, single molecule DNA and RNA sequencing, all of which have resulted in dramatic increases in the generation of genomic information including thousands of complete human genome sequences that will help to unravel the complexity of numerous diseases, as well as the use of sequencing for basic biological measurements including cDNA sequencing and quantitation, direct RNA sequencing and quantitation, and new insight into ancient genomes which are also highlighted in this chapter. The field of non-amplified sequencing now offers the opportunity for direct measurements of both DNA and RNA, providing a true measurement of cellular biology.

11.2 Emergence of Amplification Based Short Read Sequencing

Through the recognition of the important role "shot-gun" sequencing played in the commercial venture led by Craig Venter, the important concepts emerged for the "second generation" technologies – if one could simply sequence short fragments of DNA, the problem of assembling the genomes of small organisms with an eye to the whole human genome, seemed entirely possible. 454 Life Sciences emerged as the first entrant into the field of new non-Sanger based sequencing technologies publishing data demonstrating an increase in sequencing throughput enabling some 25 million bases of sequence data or some 100 fold greater than traditional Sanger technology in a period of 4 h [8]. This advance was enabled by the massively parallel synthesis of DNA templates by polymerase chain amplification to provide sufficient substrate for sequencing. This technology relies on the basic principles of sequencing by synthesis applied to pyrosequencing in which the emission of light upon incorporation of the labeled nucleotide for subsequent detection and visualization of the incorporation event, the landscape of DNA sequencing was changed forever. Commercial introduction of the Genome Sequencer 20/FLX in 2005 enabled researchers to achieve the complete sequence of the first human genome subsequent to that published from the Human Genome Project. This genome however was completed in just two months time, using the 454 technology with average read lengths approaching 250 nucleotides, and provided a genome coverage of $\sim$7.5$\times$ allowing redundancy of the reads to ensure both near complete coverage and accuracy of sequence data [9].

Solexa, a company founded in Cambridge, England in the early 2000s was one of the first commercial companies interested in the pursuit of single-molecule sequencing along with Helicos BioSciences founded at about the same time in Cambridge, MA. While both companies started as single-molecule sequencing by synthesis companies, Solexa abandoned the single-molecule approach, was acquired by Illumina in 2006 and their initial commercial platform, the GA1 was introduced in the 2007 timeframe. The GA1, an amplification based second generation system

of short reads, provided the customer with a new level of sequence throughput with 36 nucleotide reads. Today, the sequencing by synthesis approach practiced by Illumina has been continually improved through continued investment in the technology surrounding the chemistry, image analysis, engineering hardware and image analysis software to the point that the current HiSeq instrument allows researchers to sequence two complete human genomes per run at 30× coverage for approximately $10,000, closely approaching the goal of the $1000 genome.

11.2.1 Emerging Low Cost, High Throughput Technologies

While the increasing capacity of instruments, such as the Illumina HiSeq and Life Tech Solid, provide genome centers across the globe with the capacity to sequence hundreds to thousands of genomes per year, access to lower cost platforms to allow genomic scientists at smaller research institutions and translational research centers not well served by the ultra-high-throughput capacity has been limited. Former 454 founder, Jonathan Rothberg meanwhile recognized the limitations of the sequencing-by-synthesis chemistries dependence on ultra-fast imaging requirements. Ion Torrent was founded on the principles of image-independent chemistry in which the nucleotide incorporation event could simply be monitored by local changes in pH evident when a hydrogen atom is released upon nucleotide incorporation, promising a future where scale is only dependent on the ability to create a semiconductor surface which gets smaller and smaller to allow ever increasing numbers of molecular events to be monitored. Ion Torrent, purchased by Life Technologies in 2010, has its eyes set on the ultimate prize of the $1000 genome, however, along that path, the company has introduced the first low cost Personal Genome Analyzer (PGM^{TM}) to address the need of scientific researchers requiring low cost and moderate throughout to enable genomic experimentation. With a throughput today of 100 megabases, and promised improvements of ten-fold every 6 months the Ion Torrent technology appears ready for placing DNA sequencing within the realm of every biological researcher [10]. To exemplify, while this platform today is not able to sequence a complete genome in a single run, this technology enables comprehensive exon sequencing for important genes of biological relevance with the potential for diagnostic applications and when combined with bar-coding for each individual sample, hundreds to thousands of samples could be analyzed at the same time. Like the other sequencing technologies, a broad array of applications are available with the PGM machine.

11.3 Sequencing Applications

While all the technologies mentioned above rely on DNA or cDNA amplification to obtain sequence information, numerous applications are enabled by these methodologies and have revolutionized the manner by which biological measurements are

possible. These applications are briefly highlighted below and will be discussed in more depth as we describe the applications for non-amplified DNA and RNA sequencing later in the Chapter.

11.3.1 Whole Genome Sequencing

Whole genome sequencing costs have dropped dramatically over the last 10 years to the current cost estimate of 12 cents per megabase of DNA sequence as demonstrated in Table 11.2 (www.genome.gov/sequencingcosts). As the technologies have improved read length and paired-read capabilities, the major application driven by the Genome Centers has been genomic sequencing. From viral species and bacterial genome sequencing for purposes of strain identification to the human microbiome to the baboon, chimpanzee and 1,000s of normal human genomes and tumor genomes remains, these technologies have become the workhorse for genome sequencing. While the scope of this chapter is focused on sequencing technology, a brief mention of the informatics developments accompanying the technological advances described herein is appropriate as these developments have been integral in continued sequencing improvements.

Whole genome sequencing in particular has been dramatically improved by new methods and algorithms to enable *de novo* assembly of genomes and are nicely highlighted in recent publications [11, 12].

11.3.2 cDNA Sequencing

For more than a decade much of our exploration of the transcriptome has been conducted using microarrays or hybridization based methods that allow one to reliably detect the relative abundance of the known transcripts which hybridize to known probes on the array surface. This field, much like the field of DNA sequencing, has exploded with the use of next-generation sequencing. Termed RNA-Sequencing or RNA-Seq, scientists have pushed the technology to conduct biological experimentation at a scale not previously imagined and further, in a hypothesis free manner. By generating cDNA molecules through traditional reverse transcriptase methods and including the ligation of adapters that allow for amplification of the corresponding molecules, one has the ability to interrogate the transcriptome to ask important biological questions including information on the relative quantitation of RNA transcripts [13]. Since the initial publications describing the RNA-Seq methods appeared in 2008, important biological insight on stems cells [14], the complexity of the cancer transcriptome [13] and more recently the beginnings of the analysis of single cells [15] has allowed scientists to gain new insight into the RNA which plays such an integral role in cellular biology and disease states.

Table 11.2 The National Human Genome Research Institutes calculated costs for whole genome sequencing

Date	Cost per Mb of DNA sequence	Cost per genome
September-2001	$5,292.39	$95,263,072
March-2002	$3,898.64	$70,175,437
September-2002	$3,413.80	$61,448,422
March-2003	$2,986.20	$53,751,684
October-2003	$2,230.98	$40,157,554
January-2004	$1,598.91	$28,780,376
April-2004	$1,135.70	$20,442,576
July-2004	$1,107.46	$19,934,346
October-2004	$1,028.85	$18,519,312
January-2005	$974.16	$17,534,970
April-2005	$897.76	$16,159,699
July-2005	$898.90	$16,180,224
October-2005	$766.73	$13,801,124
January-2006	$699.20	$12,585,659
April-2006	$651.81	$11,732,535
July-2006	$636.41	$11,455,315
October-2006	$581.92	$10,474,556
January-2007	$522.71	$9,408,739
April-2007	$502.61	$9,047,003
July-2007	$495.96	$8,927,342
October-2007	$397.09	$7,147,571
January-2008	$102.13	$3,063,820
April-2008	$15.03	$1,352,982
July-2008	$8.36	$752,080
October-2008	$3.81	$342,502
January-2009	$2.59	$232,735
April-2009	$1.72	$154,714
July-2009	$1.20	$108,065
October-2009	$0.78	$70,333
January-2010	$0.52	$46,774
April-2010	$0.35	$31,512
July-2010	$0.35	$31,125
October-2010	$0.32	$29,092
January-2011	$0.23	$20,963
April-2011	$0.19	$16,712
July-2011	$0.12	$10,497

Reprinted from Wetterstrand KA. DNA Sequencing Costs: Data from the NHGRI Large-Scale Genome Sequencing Program. Available at: www.genome.gov/sequencingcosts. Accessed 12 Nov 2011

11.3.3 Chromatin Immunoprecipitation Studies

While DNA sequence is clearly the foundation for much of our understanding of the human genome, the additional insight provided by studying the structure of the

DNA in its native chromatin state has been significantly aided by next-generation sequencing. Chromatin immunoprecipitation studies (ChIP) allow one to study the interaction of the DNA sequence with any DNA-binding protein including histone proteins and their various modified forms in the nucleosome structure of the DNA found in the nucleus. Additional proteins including transcription factors which regulate the dynamic state of gene transcription form key interactions with the chromatin. The state of these protein:DNA interactions are studied genome-wide through a snapshot of antibody immunoprecipitation specific to the various proteins followed by isolation and sequencing of the specific DNA region to which the protein-antibody complex had bound. New insight into development [16], hematopoiesis [17] clinical outcome in ovarian cancer [18] has been possible through use of ChIP-Sequencing at the whole genome level.

11.3.4 Limitations of Amplification Technology

While amplification-based sequencing has led to tremendous advances across a range of biological questions, it remains limited or subject to artifacts in a variety of specific areas. Challenges still remain with respect to genomic rearrangements of large scale, accurate quantitation of both DNA and RNA as well as severe bias in regions of the genome with extreme G+C content. The new methods involving non-amplified DNA and RNA sequencing offer significant opportunity in all these areas as the technologies continue to mature [19]. Additionally, the various single-molecule approaches offer other technology-specific advantages including the ability to generate very long reads, to directly detect modified bases, and to directly sequence RNA.

11.4 Emergence of Non-amplified DNA Technologies

Researchers had long been interested in single molecule, non-amplified measurements of DNA. Non-amplified methods offered the potential to eliminate many of the challenges associated with complex sample preparation, PCR amplification, and the ability to interrogate single-cell nucleic acid as well as the potential for real-time interrogation to allow faster and cheaper detection of the sequence information. A landmark publication demonstrating single molecule fluorescence imaging to monitor the turnover of ATP molecules by single muscle myosin molecules using total internal reflection microscope [20] provided insight into the potential for such single-molecule measurements. This was rapidly followed by key developments in this field of non-amplified DNA sequencing which can be divided into three different areas which are sufficiently advanced for the purpose of this review: (1) direct imaging for single DNA molecules to allow both mapping and sequencing of large DNA molecules important for describing higher order structure of the

DNA sequence; (2) Optical sequencing by synthesis technologies and (3) nanopore sequencing technologies. What follows is a summary of these various technologies as well as detailed examples of sequencing applications available today.

11.4.1 Optical Mapping Technologies

As the field of DNA sequencing was maturing and the Human Genome Project was in full swing, many questions remained about the ability of shot-gun sequencing to recapitulate the accurate sequence of a genome the size of humans. Optical mapping techniques emerged in the 1990s as a potential solution to create an ordered structural map of the human genome and was one of the first single molecule methods for visualization of the higher order structure of the genome. Following the isolation of high molecular weight DNA upwards to some 10–20 megabases, DNA molecules are fixed on a glass surface, liquid flow elongates the single DNA molecules and restriction digestion of the DNA is performed directly on the surface. Figure 11.1 illustrates an optical image obtained from human genomic DNA to allow visualization of the single DNA fragments and illustrates a schematic representation of the potential ability to assemble long stretches of DNA along a contiguous stretch thus recapitulating the higher order structure of the human genome [21].

More recently scientists at companies such as Halcyon Molecular (http://halcyonmolecular.com/) and ZS Genetics (http://www.zsgenetics.com/index.html) have turned to the transmission electron microscopy as a new tool to investigate individual DNA molecules at the atomic level to allow visualization of the DNA sequence along the length of the molecule offering the potential to directly image and visualize at the DNA sequence level [22]. These methods while still under development offer the potential to provide sequence data on long stretches of DNA to overcome the limitations of current sequencing technologies which at present are limited by read length and throughput.

11.4.2 Optical Non-amplified Sequencing Technologies and Their Applications

11.4.2.1 Optical Sequencing Technologies

An initial demonstration of single-molecule sequencing-by-synthesis using DNA polymerase and fluorescent nucleotides to monitor the complementary nucleotide incorporation enabled DNA sequence data to be obtained from individual DNA molecules [23]. These initial studies led to the founding of Helicos BioSciences which further developed and commercialized the world's first single-molecule sequencing-by-synthesis technology.

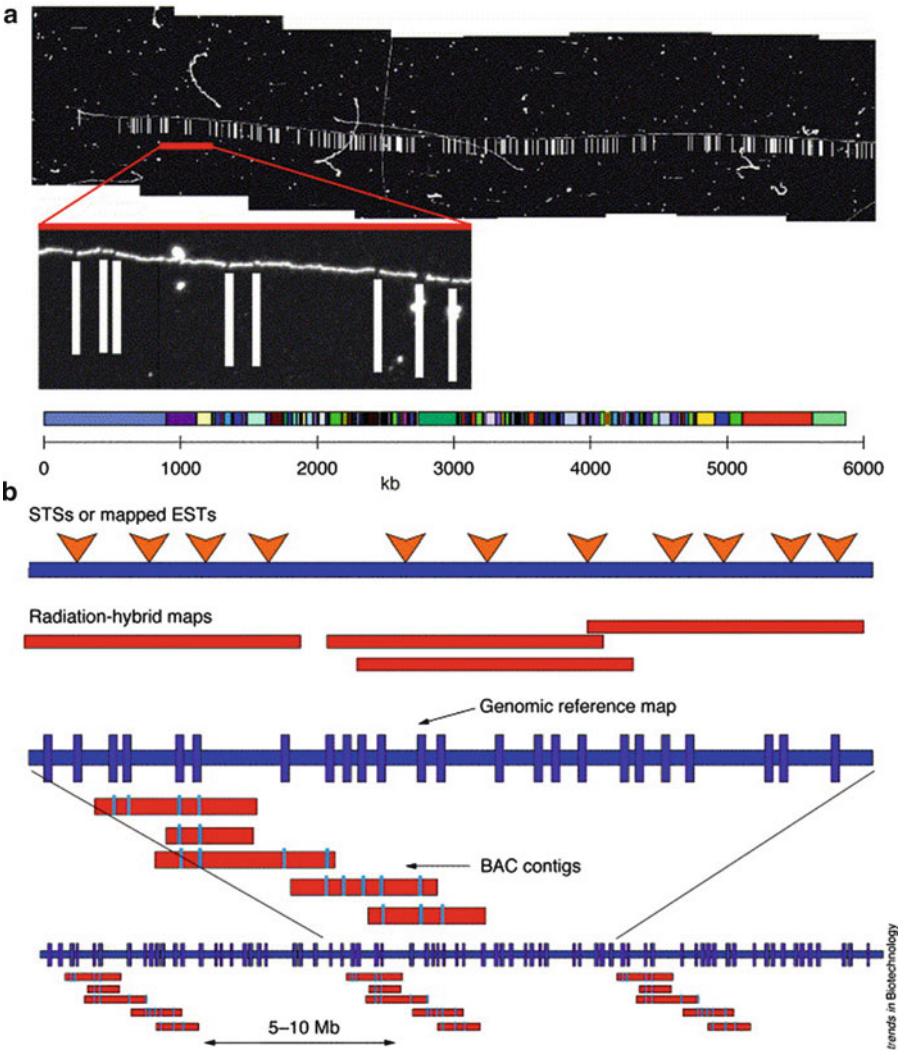
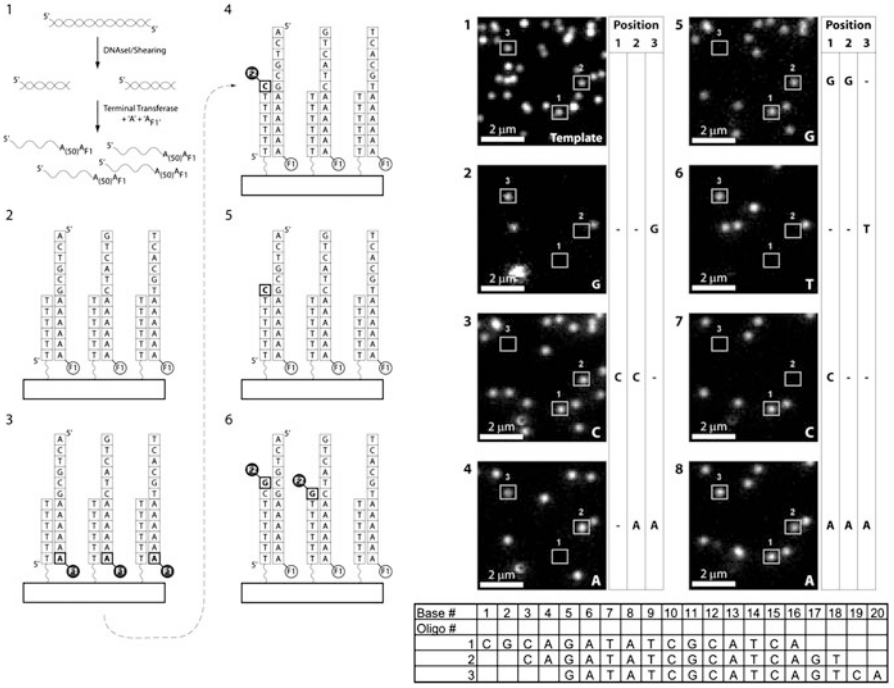


Fig. 11.1 Single molecule optical mapping depiction of the human genome. (**a**) An image of a DNA molecule 2 mm long covering 6 megabases of DNA, digested with PacI represented by overlapping microscope images. *White bars* represent the cutting sites on the DNA molecule. (**b**) Schematic representation depicting the linking of a whole genome optical map with that of physical maps (Reprinted with permission from Trends in Biotechnology, Ref. [21] Copyright 1999, Elsevier Sciences)

The Helicos technology utilizes non-amplified fragments of DNA or RNA for direct capture of the nucleic acid on the glass flow cell surface to which either universal capture primers or gene specific capture primers are covalently affixed. The depiction of the DNA sequencing process, which is also utilized for direct RNA sequencing, as well as the actual images captured during the sequencing process are shown in Fig. 11.2 [24, 25].



Base #	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
Oligo #																				
1	C	G	C	A	G	A	T	A	T	C	G	C	A	T	C	A				
2			C	A	G	A	T	A	T	C	G	C	A	T	C	A	G	T		
3					G	A	T	A	T	C	G	C	A	T	C	A	G	T	C	A

Fig. 11.2 Helicos BioSciences single molecule sequencing sample preparation and imaging of single molecules of DNA (*Left*) Single-molecule sequencing by synthesis process. (**1**) Genomic DNA is fragmented and a 3′ poly(A) tail added, labeled, and blocked using terminal transferase. (**2**) The DNA templates are captured on the surface with covalently bound oligo- dT(50). (**3**) Imaging of the captured templates to template sites on the surface. (**4**) One labeled nucleotide and polymerase mixture are added, followed by rinsing of the synthesis mixture and direct imaging. (**5**) Chemical cleavage of the dye–nucleotide linker to release the dye label. (**6**) Addition of the next nucleotide and polymerase mixture. (*Right*) Image series illustrating template-specific base addition, successful rinsing, and successful linker cleavage. A mix of three templates is used to allow visual sequence assignment. Template complementary sequences are shown in the table (*bottom*). One example of each template is outlined in the figure. Each frame is a 6.6-µm square image of the same sample position, and shows ∼35 of the 1.8×10^6 imaged templates in this experiment. Frame 1 is the image of the template labels. Template activity in three positions is shown in the columns to the right. Frame 2 is after the first synthesis and rinse cycle. Frames 3 to 8 show the effect of six more consecutive cleave, synthesis, and image cycles, using the base identity shown in the lower right corner of the frame (Reprinted with permission from Science, Ref. [24] Copyright 2008, American Association for the Advancement of Sciences)

The universal surface consists of an Oligo-dT50 surface to allow the researcher to add a polyA tail to the individual DNA molecules with terminal transferase for subsequent hybridization on the flow cell surface. When sequencing RNA, the natural polyA tail of the RNA molecules are captured on the surface or alternatively a polyA tail can be added to RNA molecules via polyA polymerase. Following hybridization to the flow cell, fluorescently labeled nucleotides, termed "Virtual

Terminators" and polymerase are added sequentially to the flow cell to allow the incorporation of complementary nucleotides into the growing strands of DNA or RNA. Once incorporated, laser excitation of the fluorophore present on the individual molecules leads to fluorescence that is captured with a charge coupled device (CCD) camera and converted to sequence reads using specialized imaging software. Following 120 cycles of this sequencing-by-synthesis and optical capture of the sequence data, more than one billion individual DNA sequence reads are generated from the DNAs held on the flow cell surface area and available for the variety of research applications of interest to the scientist. The Helicos sequencing technology provides an average read length of 35 nucleotides for each DNA molecule with a raw error profile between 3 and 5% with the predominant error form being a 'dark base' due to the incorporation of a nucleotide which is not visualized during the sequencing by synthesis process. Key is the alignment algorithm, IndexDP Genomic which allows for accurate alignment of DNA sequence data taking into account potential single nucleotide gaps in the DNA sequence [26]. Various applications of the sequencing technology are described in later sections of this Chapter.

Meanwhile, scientists at Cornell University were pursuing the real-time incorporation of fluorescent nucleotides via DNA polymerase into growing strands of DNA using a zero-mode waveguide (ZMW) technology [27]. Commericalized recently by Pacific Biosciences, the sequencing technology utilizes a SMRT™ cell or chip consisting of thousands of ZMW guides which are tiny microwells, nanometers in diameter, created in a metal film on a glass surface (shown in a schematic form in Fig. 11.3). Here the DNA polymerase is affixed to the bottom glass surface of the well. Laser illumination of the bottom 30 nm of the ZMW guide allows detection only of molecules which are near the bottom of the well. Because the DNA polymerase is attached to the bottom of the ZMWs, only labeled molecules bound to the polymerase remain in the illumination region long enough to be detected. DNA molecules are flowed across the surface to allow single molecules of DNA to bind to the polymerase and reside in the ZMW followed by addition of fluorescently-labeled nucleotides. Through diffusion, the nucleotides find their way into the ZMW where incorporation via the DNA polymerase occurs on the growing strands of DNA. Prior to each incorporation event, the fluorescent nucleotide must remain bound to the polymerase prior to incorporation and this results in fluorescence and corresponding detection of the color indicative of the nucleotide fluor. Following incorporation, the signal returns to low background level until the next nucleotide finds its way to the growing DNA strand. With the initial SMRT cell configuration consisting of some 75,000 ZMW, the incorporation events are monitored in real-time allowing the researcher to interrogate thousands of strands of individual DNA molecules in real time. In addition, the natural processivity of the DNA polymerase enzyme as well as the cleavage of the dye molecule attached to the phosphate chain of the nucleotide leaving a natural nucleotide in the growing DNA strand, has the potential to deliver read lengths well exceeding all other amplification and non-amplification based methods. Currently the average read lengths can approach 1,000 bases but continued efforts on both chemistry and detection offer the opportunity to surpass the current read length.

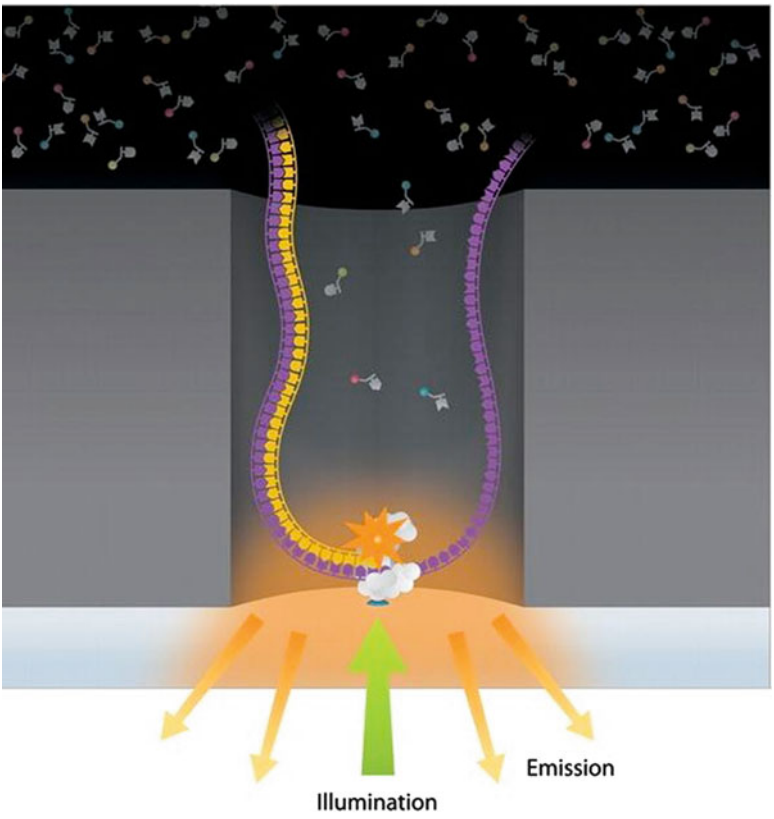


Fig. 11.3 Depiction of Pacific Biosciences Zero Mode Waveguide. The ZMW microwell is depicted here with a polymerase molecule to which DNA is bound. Synthesis of the DNA strand occurs through incorporation of the appropriate labeled nucleotide which will be detected upon laser excitation and emission to the recorded below the ZMW surface (Adapted with Permission from Pacific Biosciences)

Helicos and Pacific Biosciences non-amplified DNA sequencing technology have been used for genome sequencing. The publication describing the first single molecule sequencing of a human genome using the Helicos sequencing technology was indeed a remarkable achievement in the field of non-amplified DNA [28]. Requiring minimal sample preparation and three weeks time for sequencing followed by several weeks of analysis time, the field of non-amplified sequencing was established. Additionally, new insights into ancient genomes have been provided by recent studies of DNA from a Pleistocene-era horse [29] and thus the unique attributes of this ability are highlighted in a later section. Pacific Biosciences and colleagues at Harvard Medical School followed with another single-molecule sequencing study demonstrating the ability of SMRT sequencing to decipher the origin of the cholera strain responsible for the outbreak in Haiti in 2010 [30].

Yet, importantly, these two technologies each provide unique features that differentiate them from the existing amplification-based sequencing technologies. The current ability and future potential of the SMRT sequencing technology to obtain read lengths that far surpass the current technologies combined with the sheer speed of the DNA read-out offer new opportunity for more complete characterization of the genome, allowing us to address the many rearrangements and repetitive regions not possible with current technologies. In addition the ability to obtain full length cDNA transcripts to fully elucidate the complex structure of the transcriptome is entirely within the realm of the technology as throughput improvements occur.

The accuracy of quantitation with the Helicos single molecule sequencing technology is unparalleled as the ability to examine upwards of one billion molecules which have not been amplified or ligated. Thus the inherent bias caused by these molecular manipulations are avoided and the unamplified methodology provides the scientist with the purest quantitative measurement for nucleic acid from a biological source of interest [31] and enables new insight into important new biology revealing new RNA species not previously detected with amplification based technologies [32]. Additionally, the Helicos technology allows the direct sequencing of RNA. Molecules which have not been manipulated via copying with reverse transcriptase can thus be examined in a massively parallel fashion, something never before possible [33]. A more in-depth characterization of these unique applications is detailed below.

11.4.2.2 Gene Expression Measurements: Tag-Based and RNA Seq

Gene expression measurements provide a powerful window into how cells and organisms behave normally as well as in response to various stimuli. Both the number of genes that can be analyzed and the sensitivity with which those genes can be detected has increased substantially as the technology of choice moved from Northern blots and S1 assays to microarrays and qPCR and most recently to RNA Seq. Each of the methodologies has its own strengths and weaknesses with respect to identification of exon structure, crosstalk with similar genes, sensitivity, quantitative accuracy, ability to detect poorly expressed or uncharacterized transcripts and susceptibility to artifacts and errors. No technology is able to provide a complete, quantitative picture of which transcripts are expressed combined with a detailed view of their exon structure and 5′/3′ ends. As such, experimenters need to decide which aspects of the true expression profile are most important for their purposes so the technology can be chosen which is best suited to the needed information.

Most early technologies were able to monitor only a small number of genes for their expression levels so are limited if one wishes to do a complete transcriptome characterization. A more thorough view on transcription became possible with the advent of various types of microarrays and the ability to analyze thousands of different genes simultaneously. However, microarrays suffer from poor sensitivity for genes that are expressed at low levels, can suffer from significant crosstalk with genes that are closely related, has difficulty determining splicing patterns,

and is not generally useful for uncharacterized transcripts. To varying extents, next generation sequencing technologies have overcome all of these difficulties. For NGS technologies with a very high read count, very precise expression patterns can be generated. However, all sequencing methods also have limitations and these must be recognized so that experimental results can be properly interpreted.

Sequencing-based gene expression methods can be divided into two types: tag-based methods and methods that interrogate the entire RNA molecule. Tag-based methods include CAGE (Cap Analysis of Gene Expression) [34], SAGE (Serial Analysis of Gene Expression) [35, 36], PET-Seq (Paired End Tag Seq) [37], DGE (Digital Gene Expression) [38, 39], and DRS (Direct RNA Sequencing) [33]. These techniques capture a specific sequence at either the 5′ end (CAGE and DGE) or 3′ end (SAGE and DRS) or both (PET Seq) and count those molecules for determining gene expression profiles. Little or no information is captured on splicing or the opposite end of the molecule. Thus, these techniques are ideal for assessing and comparing gene expression levels. Additionally, the tag-based systems typically incorporate a selection step in which a specific feature of mRNA (5′ cap or 3′ polyA tail) that allows preferential sequencing of the RNAs of interest with less sequencing of ribosomal RNA that is generally of less importance for expression.

RNA Seq, in contrast to tag-based methods, captures reads from throughout each RNA molecule. This provides information about the entire RNA but, unlike microarrays and tag-based sequencing, this introduces an artifact into RNA Seq data in that the results depend on the length of the RNA being interrogated. If the RNA is long, more reads will arise from that RNA even if expressed at the same number of molecules as a short RNA. Frequently, the raw expression levels are corrected for length but these corrections are imperfect for many reasons [40] and can lead to analysis issues [41]. RNAs with extremes of GC content are less likely to be amplified and thus will appear less often than they should based on actual expression levels [42]. Efforts to eliminate GC bias in library construction have been partially successful [43] but some of these biases remain and are exacerbated by the amplification that is required during sequencing. Even after removing much of the library-induced amplification bias, genomic coverage patterns are still far from what is predicted based on known sequence content [43]. Figure 11.4 demonstrates this principle with data derived from RNA Seq experimentation in which amplification based methods are compared directly to single molecule methods, both involving cDNA analyses. Single molecule sequence data demonstrates more uniform coverage across the gene transcripts depicted [19].

The paired-end protocols used to generate long-range data for splicing analysis can introduce artifacts as well. If the raw reads from RNA Seq libraries are examined, 5–10% of all reads have paired ends that do not match the same gene [44]. Thus, one could expect that some artifacts could also be present in matched pairs, suggesting that rare splice variants should be verified using an independent technology. Additionally, the random hexamers typically used for synthesis of cDNA introduce another level of bias into the process [45], independent of sequencing platform.

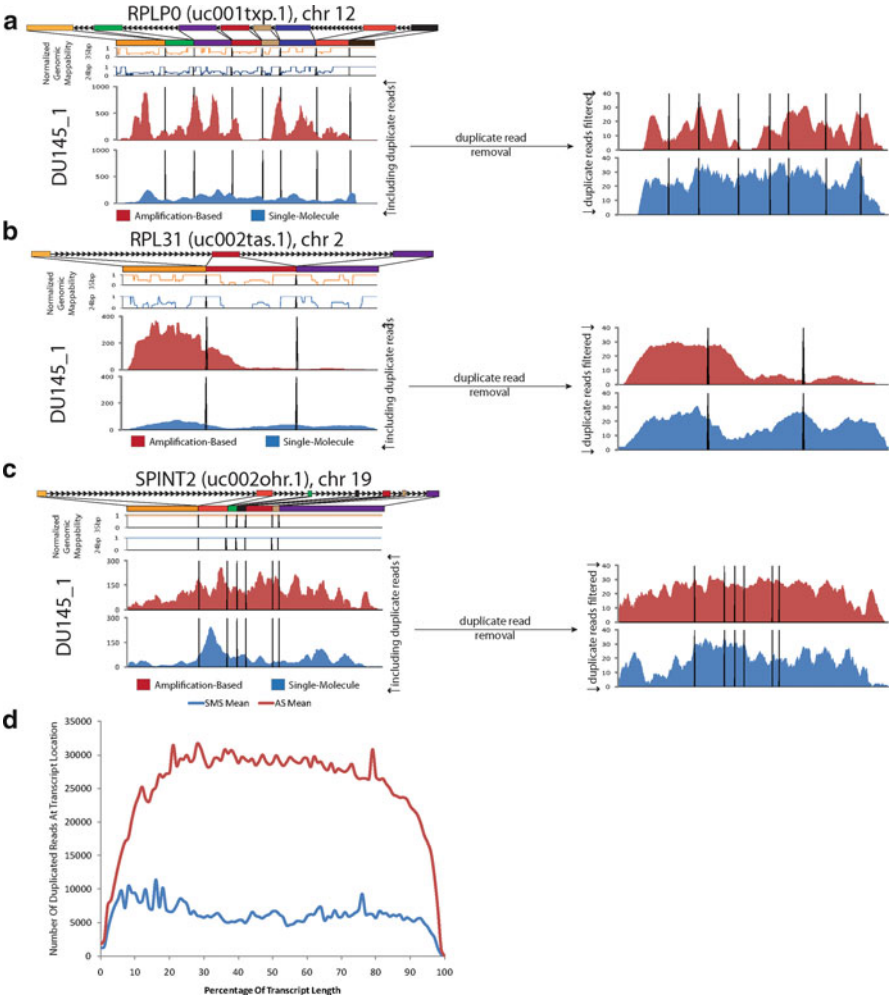


Fig. 11.4 Comparison of amplification and non-amplification RNA Seq data sets derived from cancer cell line. Amplification-based sequencing leads to a bias in high-concentration transcripts. Coverage maps from amplification-based and single molecule sequencing demonstrate significantly greater coverage of (**a**) RPLP0, (**b**) RPL31, and (**c**) SPINT2. Removal of reads with the same start positions significantly reduces the "spikiness" seen in these cases. (**d**) Duplicate reads are relatively evenly distributed along the length of observed transcripts across all samples and artificially inflate the apparent expression of those genes in amplification-based sequencing but not in single-molecule sequencing (Reproduced with Permission from PLoS ONE: Ref. [19])

While amplification-based sequencing has been used extensively for high quality expression profiles [46], there are situations in which single-molecule RNA Seq methods can provide superior results for some aspects of gene expression studies. For example, the Pacific Biosciences system is capable of long reads [47] so offers the potential of a direct view of exon content and start/stop sites for different

transcript isoforms. Unfortunately, the read count with this system is far too low to be generally useful for accurate expression profiles. A reasonable splicing picture of very highly expressed genes probably could be generated but the frequency of reads from most genes will leave them with no coverage or insufficient coverage to assess splicing and start/stop sites. For that reason, no RNA Seq profiles have yet been published with this system.

In contrast to Pacific Biosciences, the Helicos system generates a very high read count and thus can provide a very accurate quantitative assessment of gene expression [38]. Because amplification is not necessary, quantitative accuracy is maintained to a much higher extent than with amplification-based systems.

Any artifacts that arise from reverse transcriptase and cDNA synthesis may still skew the true gene expression measurements unless Direct RNA Sequencing (DRS) is carried out (see below) but biased amplification during library construction and sequencing is not an issue. The Helicos read lengths are generally, though not always, sufficient for unambiguous assignment to a particular gene. With highly homologous genes, some crosstalk may occur so there will be uncertainty in the expression level of such genes. Additionally, though some information can be gleaned about splicing, complex splice patterns are not detectable. Nonetheless, very precise RNA Seq profiles have been obtained [48] and technical replication far superior to amplification-based systems has been described [40].

To a large extent, the differential expression patterns obtained with single-molecule versus amplified sequencing are very similar but there are key differences. In particular, genes that are poorly expressed are found much less often with amplification-based sequencing [19], likely due to issues of limiting library diversity but other issues with PCR cannot be ruled out. In addition to limited visibility of poorly expressed RNAs and those with extreme GC contents, very short RNAs can also be problematic [32]. Many short RNAs are not well amplified during library construction and are too short to be amplified using bridge PCR required in the Illumina platform.

11.4.2.3 Gene Expression Measurements: Direct RNA Sequencing Technology

In addition to the frequent use of amplification, all of the gene expression methods described above rely on conversion of RNA to cDNA prior to sequencing. However, cDNA synthesis is known to be plagued by many artifacts including template switching, primer-independent first and second strand synthesis, and biased cDNA synthesis [49–51]. DRS technology can alleviate many limitations inherent in the transcriptomics methods in use today, and provide new avenues of research and applications in diagnostics. DRS not only eliminates the reverse transcription step, but also the other sample manipulation steps such as ligation and amplification, thus resulting in minimal distortion in the representation of RNA templates. The natural strand-specificity of DRS and its requirement for only femtomole quantities of RNA are advantageous for all aspects of RNA research.

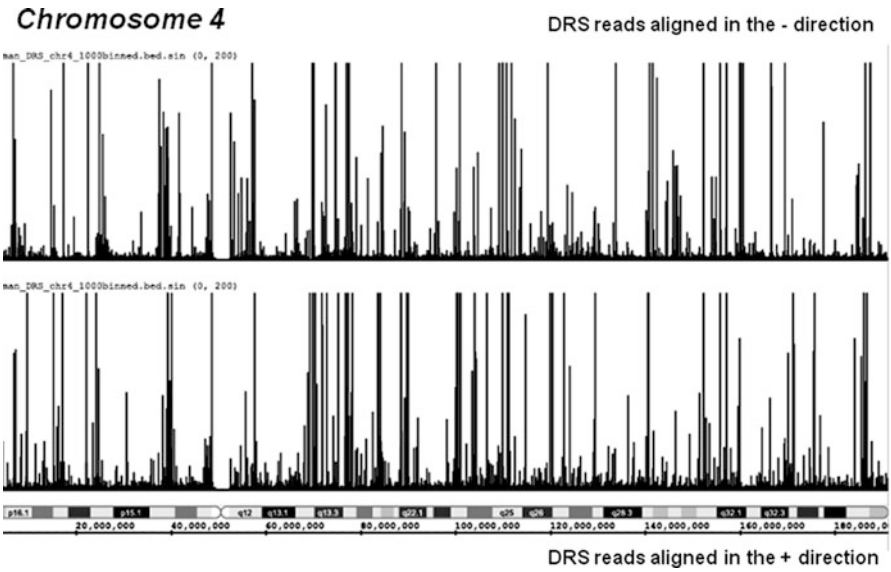


Fig. 11.5 Mapping of liver-derived direct RNA sequencing reads to human Chromosome 4.Total liver RNA was sequencing using the DRS technology. Sequence reads were mapped with high stringency to the human genome reference. Reads mapping to Chromosome 4 are shown (Unpublished data used with Permission from Helicos BioSciences Corporation)

DRS offers a simple route for polyadenylation site mapping [52]. Given its nature of capturing polyadenylated RNAs on poly(dT)-coated surfaces and sequencing after a "fill & lock" step, DRS reads emerge immediately upstream of the polyA-tail. Thus, the 5′ end of DRS reads signify polyadenylation addition locations. The DRS procedure is capable of capturing the polyA + mRNA population from total RNAs or cell lysates directly after blocking only the 3′ hydroxyl. The data generated is quantitative, thus for the first time allows genome-wide study of alternative polyadenylation states in both quantitative and qualitative manner across biological settings of interest. Figure 11.5 demonstrates direct RNA sequence reads obtained from liver total RNA mapped to human chromosome 4. Peaks across the chromosome demonstrate the diversity of RNA reads at low resolution. This data can also be used to generate a tag-based gene expression profile of polyA + mRNAs within cells.

DRS can also be adapted for all RNA analyses being performed today. Whole transcriptome profiling can be done with RNA fragmentation with standard methods, followed by polyadenylation of the RNAs. One advantage of DRS is the universality of sample preparation steps for different applications. In other words, unlike cDNA methods which require different cDNA synthesis and sample manipulation steps for short and long RNAs, DRS requires only 3′ polyadenylated templates. Thus, both short and long RNAs can be sequenced together in a single experiment.

11.4.2.4 Ancient and Degraded DNA

In most situations, long sequence reads are an advantage. However, when the DNA is degraded due to age, chemical fixation, or damaged by other deleterious conditions, it may actually be advantageous to use short read technologies. For example, ancient DNA is frequently contaminated with DNA from other species which is sometimes of more recent origin and hence potentially longer and of higher quality [53]. When such mixtures are amplified, the longer and less damaged modern DNA is preferentially replicated and hence increases its fractional composition and could potentially swamp out the desired signal from the ancient or damaged DNA. This is evident from work with DNA from a Pleistocene-era horse bone in which the same sample sequenced with an amplified versus non-amplified system yielded very different results for the per cent horse versus non-horse sequence reads [29]. On average, $>3\times$ more reads were from horse using the non-amplified sequencing. This difference was substantially increased when the sample preparation for the ancient DNA was modified slightly to remove blocking $3'$ phosphates from the ancient DNA [54]. The extent to which single-molecule sequencing is superior is highly dependent on the quality of the DNA sample of interest. The more degraded the sample, the higher degree of improvement can be obtained. However, it is not necessary for DNA to be Pleistocene era in order to be too degraded for analysis. For example, some remains buried at the National Memorial Cemetery of the Pacific that had been exposed to highly damaging conditions during embalming and which, even after extensive amplification, had not previously provided more than a few dozen base pairs of usable human sequence using amplification-based sequencing was able to be effectively sequenced using the Helicos system [55]. Similarly, many clinical samples are formaldehyde-treated and preserved in paraffin and thus can be significantly damaged or degraded. More recently preserved samples tend to be higher quality as experimenters have realized the importance of mild conditions to allow subsequent sequencing but many samples have not been as carefully handled and thus are problematic with sequencing systems that require longer templates. Again, single-molecule sequencing has been able to provide good sequence data for even RNA samples extracted from FFPE clinical specimens [56].

11.4.3 Nanopore Sequencing Methods

Classical next-generation sequencing techniques have dramatically dropped the price of sequencing while opening up numerous new avenues of scientific investigations. These tremendous advances have served to whet the appetite for even greater capacity at lower costs. However, the current commercial technologies are unlikely to yield orders of magnitude cheaper sequence or provide markedly different capabilities relative to what is already available. Consistent with the initial goals laid out by NHGRI for the $1000 genome and in contrast to the existing technologies, there are a variety of nanopore technologies, though not yet

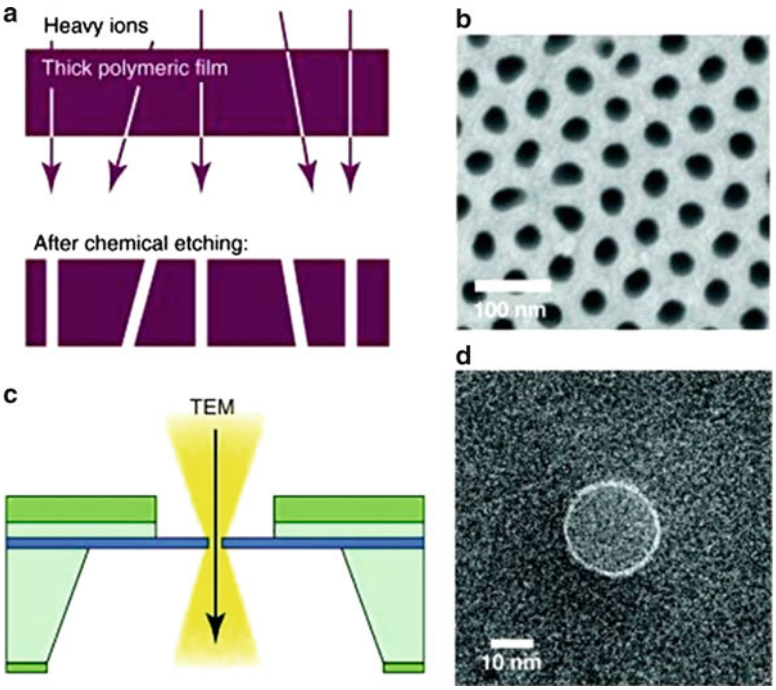


Fig. 11.6 Schematic representation of nanopore fabrication. artificial nanopores and nanochannels can be made by heavy ion and chemical ecthing (**a** and **b**) or by drilling holes through a silicon wafer with a focused electron beam (**c** and **d**). A more detailed description is provided in Ref. [60] (Reprinted with permission from Trends in Biotechnology, Ref. [60] Copyright 2011, Elsevier Sciences)

technically and/or commercially viable, that could provide dramatically different or cheaper output. Nanopore technology is attractive because, in theory, extremely long reads could be generated on single molecules with little or no reagent costs, simple sample preparation in a highly parallel fashion and with a very short time to results. However, realizing all these goals or even some of them, with a single technology will be challenging. To be commercially successful, the output from a technology will either have to provide a qualitatively different result or substantial cost/time savings relative to the already high bar created by existing technologies. Reviews dedicated to nanopores and nanopore sequencing are available and describe the wide variety of approaches being taken in this area [57–61]. Healey [62] has reviewed nanopore sequencing from a more historical perspective.

Nanopore approaches can be categorized in a number of ways based on the nature of the nanopore with the most frequent classification being biological or solid state. Biological nanopores use a protein channel with a pore of the necessary dimensions to allow DNA to pass through [63]. While early biological nanopores were in lipid membranes, some more recent versions are hybrid, using a solid-state scaffold with a protein pore. Solid-state pores (Fig. 11.6) were originally derived from silicon and

its derivatives though recent efforts have also explored graphene as an alternative. In addition to the nature of the pore, another useful categorization is the mode of sequence detection. Most nanopores detect variations in electrical signals (voltage or current) induced by blocking an ionic current between chambers but other methods, including optical detection and differences in transverse electrical currents, can also be used. The theoretical background and modeling of blocking ionic currents have been described in detail [64, 65]. Similarly, the theory and signals expected from transverse currents detected across a tunneling gap have also been described [66, 67] and confirmed experimentally [68–70].

Most efforts directed at biological nanopores have employed either α-hemolysin or MspA as the protein of choice [63, 71, 72] though other proteins may also be used [73]. All of these proteins have a pore with a diameter greater than 1 nm needed to allow single-stranded DNA (ssDNA) to pass through. Some proteins have an even larger pore allowing double-stranded DNA to pass. However, the variation in the lengths, widths and charge distributions within the pores of these proteins is not ideal for sequence detection so efforts have been directed at improving their properties. α-hemolysin and MspA have been successfully mutated to improve a variety of sequencing properties [72, 74, 75]. These mutated pore proteins have been shown to be capable of distinguishing the electrical signals from all four natural nucleosides/nucleotides [76–78]. However, distinguishing free nucleotides is not the same as sequencing extended lengths of DNA. To generate usable sequence data, the DNA must be translocated through the nanopore with sufficient force that movement is unidirectional but not so quickly that the signal from the individual bases cannot be distinguished. The length and width of the pore must be such that the base being sequenced is the primary contributor to the signal or a very complex set of signals will result. These conditions are not easily met so a variety of methods have been used in an attempt to generate sequence data. The speed of DNA translocation through protein pores has been slowed by altering viscosity [79] and by varying salt concentrations [80]. These manipulations have not yet proved sufficient for reading sequence so various proteins have been used to assist in the process.

For example, DNA polymerase has been used to detect sequence incorporations while attached to α-hemolysin [81]. Similarly, the activity of exonuclease I bound to α-hemolysin has been monitored by nanopores [82]. The kinetics of these reactions in pilot studies is not sufficient for effective sequencing but demonstrates the future possibilities if they could be optimized. In addition to detecting the electronic signal directly from nucleotides/nucleobases, there have also been efforts to detect signals optically after converting the naturally occurring sequence to fluorescent emitters [83]. However, methods described thus far introduce far more complexity into sample preparation than is required by other approaches so are not as appealing as the simpler nanopore readouts.

While biological nanopores offer advantages in terms of being easily manipulated with respect to changing the charge and inclusion of complex functional groups in or near the pore, there are only a small number of suitable pore proteins and they provide a limited set of scaffolds with which to work. Also, construction of highly parallel arrays does not have the economies of scale that solid-state

nanopores offer. Solid-state nanopores have the advantage of being much thinner and thus can generate signals arising from a single base more easily. Indeed, graphene offers the thinnest possible nanopore, the thickness of a single carbon atom. Furthermore, nanopores can be constructed with a wide variety of widths and this can be readily changed using well-established techniques [84]. For the most part, solid-state nanopores are made from silicon derivatives. Very long DNAs (97 kb) have been reported to be translocated through silicon at a speed greater than 10 kb/ms [85] and it is likely that much longer DNAs could be used. Even at 97 kb, these DNAs far surpass the lengths that can be interrogated with current sequencing technologies and thus would immediately provide a substantial benefit for genomic analysis if information beyond simply the length could be obtained.

As with biological nanopores, the speed of translocation with current solid-state nanopores is too fast for sufficient signal to be generated for each base to be effectively read. As a result, modifications of the pores and translocation conditions have been carried out to slow the rate of translocation [86]. Capture rate can be adjusted by varying salt concentrations [87] or by altering the manner in which the pores are made [88, 89]. Pores can be chemically modified so that the charge slows translocation speed. Because silicon nanopores can be made wider, ssDNA and dsDNA as well as protein-bound DNA can be successfully translocated through pores. When bound to recA, dsDNA moves through the pores much more slowly and generates a much higher blockade current [90].

While varying translocation conditions can slow the rate of transit through the nanopore that creates new issues with the positioning of the DNA with respect to the nanopore constriction and assuring that the DNA is read sequentially [91]. However, if DNA is translocated through a nanopore at a speed necessary to ensure predictable positioning, it is going too fast to generate sufficient signal. As a result, a variety of methods have been tested in order to provide the optimal mix of speed and sensitivity [92, 93]. Additionally, hybrid pores modified with DNA [94] or protein [95] can also be used. Another approach employs oligonucleotide probes of known sequence to tag regions of interest in DNA [96] and generate position-specific changes in blocking current. By using pools of probes, the entire sequence can be generated *de novo*. This method has the advantage of using groups of nucleotides with each signal and spreading that signal out over a longer physical distance for enhanced signal to noise and ease of detection.

Graphene nanopores also have the subject of much recent interest due to the absolute minimum thickness of their constrictions. Thus far, translocation of DNA has been observed through pores experimentally [97–99] along with theoretical predictions of how sequence signals might be possible [100]. While offering tremendous potential, graphene nanopores are now in their infancy in terms of characterization and their very simplicity will make them difficult to modify as can be accomplished with protein and silicon-based nanopores. Thus, the varying flavors of nanopore sequencing provide the best hope for the next quantum leap in sequencing capacity and capabilities but many issues need to be overcome before they are a commercial reality.

11.5 The Future

The power of amplification-based next generation sequencing has enabled countless new approaches to a host of previously inaccessible biological questions. This treasure trove of new applications and knowledge has obscured the fact that there are severe limitations and artifacts present in this powerful but not omniscient technology. The tremendous variety of single-molecule approaches holds the potential for filling many of those gaps and expanding the reach of massively parallel sequencing ever further. An important example, sequence reads of unprecedented length achieved with orders of magnitude faster speed to results could be generated, all at lower cost than current technologies. Different single-molecule approaches existing today, or in development, offer the potential for simplifying and lowering sample requirements, improved quantitation for gene expression, protein binding, and epigenetics. Thus, we can look to the next 10 year when the promise of the $1000 genome will be realized – and with this new milestone new opportunity for unraveling the complexity of human disease.

References

1. Maxam, A.M., Gilbert, W.: A new method for sequencing DNA. Proc. Natl. Acad. Sci. U.S.A. **74**, 560–564 (1977)
2. Sanger, F., Nicklen, S., Coulson, A.R.: DNA sequencing with chain-terminating inhibitors. Proc. Natl. Acad. Sci. U.S.A. **74**, 463–467 (1977)
3. Prober, J.M., Trainor, G.L., Dam, R.I., Hobbs, F.W., Robertson, C.W., Zagursky, R.I., Cocuzza, A.J., Jensen, M.A., Baumeister, K.: A system for rapid DNA sequencing with fluorescent chain-terminating dideoxynucleotides. Science **238**, 336–341 (1987)
4. Blattner, F.R., et al.: Escherichia coli K-12. Science **277**, 1453–1462 (1997)
5. International Human Genome Sequencing Consortium, Lander, E.S., Linton, L.M., Birren, B., Nusbaum, C., Zody, M.C., et al.: Initial sequencing and analysis of the human genome. Nature **409**, 860–921 (2001)
6. Venter, C.J., Adams, M.D., Myers, E.W., et al.: The sequencing of the human genome. Science **291**, 1304–1351 (1991)
7. International Human Genome Sequencing Consortium: Finishing the euchromatic sequence of the human genome. Nature **431**, 931–945 (2004)
8. Margulies, M., Egholm, M., Altman, W.E., et al.: Genome sequencing in microfabricated high-density picolitre reactors. Science **437**, 376–380 (2005)
9. Wheeler, A., Srinivasan, M., Egholm, M., Shen, Y., Chen, L., McGuire, A., He, W., Chen, Y.-J., Makhijani, V., et al.: The complete genome of an individual by massively parallel DNA sequencing. Nature **452**, 872–876 (2008)
10. Rothberg, J.M., Hinz, W., Rearick, T.M., et al.: An integrated semiconductor device enabling non-optical genome sequencing. Nature **475**, 348–352 (2011)
11. Li, R., Zhu, H., Ruan, J., Qian, W., Fang, X., et al.: de novo assembly of human genomes with massively parallel short read sequencing. Genome Res. **20**, 265–272 (2010)
12. Nagarajan, N., Pop, M.: Sequencing and genome assembly using next-generation technologies. Methods Mol. Biol. **673**, 1–17 (2010)
13. Morin, R., Bainbridge, M., Fejes, A., Hirst, M., Krzywinski, M., Pugh, T., McDonald, H., Varhol, R., Jones, S., Marra, M.: Profiling the HeLa S3 transcriptome using randomly primed cDNA and massively parallel short-read sequencing. Biotechniques **45**, 81–94 (2008)

14. Cloonan, N., Forrest, A.R., Kolle, G., Gardiner, B.B., Faulkner, G.J., Brown, M.K., et al.: Stem cell transcriptome profiling via massive-mRNA sequencing. Nat. Methods **5**, 613–617 (2008)
15. Tang, F., Barbacioru, C., Wang, Y., Nordman, E., Lee, C., Xu, N., Wang, X., Bodeau, J., Tuch, B.B., Siddiqui, A., Lao, K., Surani, M.A.: mRNA-Seq whole-transcriptome analysis of a single cell. Nat. Methods **6**(5), 377–82 (2009)
16. Zhao, X.D., Han, X., Chew, J.L., Liu, J., Chiu, K.P., Choo, A., Orlov, Y.L., Sung, W.K., Shahab, A., Kuznetsov, V.A., Bourque, G., Oh, S., Ruan, Y., Ng, H.H., Wei, C.L.: Whole-genome mapping of histone H3 Lys4 and 27 trimethylations reveals distinct genomic compartments in human embryonic stem cells. Cell Stem Cell **1**, 286–98 (2007)
17. Adli, M., Zhu, J., Bernstein, B.E.: Genome-wide chromatin maps derived from limited numbers of hematopoietic progenitors. Nat. Methods **7**, 615–8 (2010), Epub Jul 11 2010
18. Kennedy, B.A., Deatherage, D.E., Gu, F., Tang, B., Chan, M.W., et al.: ChIP-seq defined genome-wide map of TGFβ/SMAD4 targets: implications with clinical outcome of ovarian cancer. PLoS One **6**, e22806 (2011). E pub 2011 Jul 25 (2011)
19. Sam, L.T., Lipson, D., Raz, T., Cao, X., Thompson, J.F., Milos, P.M., Robinson, D., Chinnaiyan, A.M., Kumar-Sinha, C., Maher, C.A.: A comparison of single molecule and amplification based sequencing of cancer transcriptomes. Plos One **6**, e17305 (2011)
20. Funatsu, T., Harada, Y., Tokunaga, M., Saito, K., Yanagida, T.: Imaging of single fluorescent molecules and individual ATP turnovers by single myosin molecules in aqueous solution. Nature **374**, 555–559 (1994)
21. Aston, C., Mishra, B., Schwartz, D.C.: Optical mapping and its potential for large-scale sequencing projects. Trend. Biotech. **17**, 297–302 (1999)
22. Krivanek, O.L., Chisholm, M.F., Nicolosi, V., Pennycook, T.J., Corbin, G.J., et al.: Atom-by-atom structural and chemical analysis by annular dark-field electron microscopy. Nature **464**, 571–574 (2010)
23. Braslavsky, I., Herbert, B., Kartalov, E., Quake, S.R.: Sequence information can be obtained from single DNA molecules. Proc. Natl. Acad. Sci. U.S.A. **100**, 3960–4 (2003)
24. Harris, T.D., Buzby, P.R., Babcock, H., Beer, E., Bowers, J., Braslavsky, I., et al.: Single-molecule DNA sequencing of a viral genome. Science **320**, 106–109 (2008)
25. Thompson, J.F., Reifenberger, J.G., Giladi, E., Kerouac, K., Gill, J., et al.: Single-step capture and sequencing of natural DNA for detection of *BRCA1* mutations. Genome Res.. doi:10.1101/gr.122192.111. Published in Advance July 15, 2011
26. Giladi, E., Healy, J., Myers, G., Hart, C., Kapranov, P., Lipson, D., et al.: Error tolerant indexing and alignment of short reads with covering template families. J. Comput. Biol. **17**, 1397–1411 (2010)
27. Levene, M.J., Korlach, J., Turner, S.W., et al.: Zero-mode waveguides for single-molecule analysis at high concentrations. Science **299**, 682–6 (2003)
28. Pushkarev, D., Neff, N.F., Quake, S.R.: Single-molecule sequencing of an individual human genome. Nat. Biotechnol. **27**, 847–852 (2009)
29. Orlando, L., Ginolhac, A., Raghavan, M., Vilstrup, J., Rasmussen, M., Magnussen, K., Steinmann, K., Kapranov, P., Thompson, J.F., Zazula, G., Froese, D., Shapiro, B., Hofreiter, M., AL-Rasheid, K.A.S., Mundy, J., Gilbert, M.T.P., Willerslev, E.: True single-molecule DNA sequencing of a pleistocene horse bone. Genome Res. **21**, 1705–1719 (2011)
30. Chin, C.S., et al.: The origin of the Haitian cholera outbreak strain. N. Engl. J. Med. **364**, 33–42 (2011)
31. Kanamori-Katayama, M., Itoh, M., Kawaji, H., Lassmann, T., Katayama, S., Kojima, M., Bertin, N., Kaiho, A., Ninomiya, N., Daub, C.O., Carninci, P., Forrest, A.R., Hayashizaki, Y.: Unamplified cap analysis of gene expression on a single-molecule sequencer. Genome Res. **21**, 1150–9 (2011)
32. Kapranov, P., Ozsolak, F., Kim, S.W., Foissac, S., Lipson, D., Hart, C., Roels, S., Borel, C., Antonarakis, S.E., Monaghan, A.P., John, B., Milos, P.M.: Novel class of gene-termini-associated human RNAs suggests a novel RNA copying mechanism. Nature **466**, 642–646 (2010)

33. Ozsolak, F., Platt, A., Jones, D., Reifenberger, J., Sass, L.E., McInerney, P., Thompson, J.F., Bowers, J., Jarosz, M., Milos, P.: Direct RNA sequencing. Nature **461**, 814–818 (2009)
34. Plessy, C., Bertin, N., Takahashi, H., Simone, R., Salimullah, M., et al.: Linking promoters to functional transcripts in small samples with nanoCAGE and CAGEscan. Nat. Methods **7**, 528–534 (2010)
35. Asmann, Y.W., Klee, E.W., Thompson, E.A., Perez, E.A., Middha, S., Oberg, A.L., Therneau, T.M., Smith, D.I., Poland, G.A., Wieben, E.D., Kocher, J.P.: 3′ tag digital gene expression profiling of human brain and universal reference RNA using Illumina Genome Analyzer. BMC Genomics **10**, 531 (2009)
36. Wu, Z.J., Meyer, C.A., Choudhury, S., Shipitsin, M., Maruyama, R., et al.: Gene expression profiling of human breast tissue samples using SAGE-Seq. Genome Res. **20**, 1730–1739 (2010)
37. Fullwood, M.J., Wei, C.L., Liu, E.T., Ruan, Y.: Next-generation DNA sequencing of paired-end tags (PET) for transcriptome and genome analyses. Genome Res. **19**, 521–532 (2009)
38. Lipson, D., Raz, T., Kieu, A., Jones, D.R., Giladi, E., et al.: Quantification of the yeast transcriptome by single-molecule sequencing. Nat. Biotechnol. **27**, 652–658 (2009)
39. Ozsolak, F., Ting, D.T., Wittner, B.S., Brannigan, B.W., Paul, S., et al.: Amplification-free digital gene expression profiling from minute cell quantities. Nat. Methods **7**, 619–621 (2010)
40. Raz, T., Kapranov, P., Lipson, D., Letovsky, S., Milos, P.M., Thompson, J.F.: Protocol dependence of sequencing-based gene expression measurements. PLoS One **6**, e19287 (2011)
41. Oshlack, A., Wakefield, M.J.: Transcript length bias in RNA-seq data confounds systems biology. Biol. Direct **4**, 14 (2009)
42. Dohm, J.C., Lottaz, C., Borodina, T., Himmelbauer, H.: Substantial biases in ultra- short read data sets from high-throughput DNA sequencing. Nucleic Acids Res. **36**, e105 (2008)
43. Aird, D., Ross, M.G., Chen, W.S., Danielsson, M., Fennell, T., Russ, C., Jaffe, D.B., Nusbaum, C., Gnirke, A.: Analyzing and minimizing PCR amplification bias in Illumina sequencing libraries. Genome Biol. **12**, R18 (2011)
44. Mamanova, L., Andrews, R.M., James, K.D., Sheridan, E.M., Ellis, P.D., et al.: FRT-seq: amplification-free, strand-specific transcriptome sequencing. Nat. Methods **7**, 130–132 (2010)
45. Hansen, K.D., Brenner, S.E., Dudoit, S.: Biases in Illumina transcriptome sequencing caused by random hexamer priming. Nucleic Acids Res. **38**, e131 (2010)
46. Oshlack, A., Robinson, M.D., Young, M.D.: From RNA-seq reads to differential expression results. Genome Biol. **11**, 220 (2010)
47. Eid, J., Fehr, A., Gray, J., Luong, K., Lyle, J., Otto, G., Peluso, P., Rank, D., et al.: Real-time DNA sequencing from single polymerase molecules. Science **323**, 133–138 (2009)
48. Kapranov, P., St Laurent, G., Raz, T., Ozsolak, F., Reynolds, C.P., Sorensen, P.H., Reaman, G., Milos, P., Arceci, R.J., Thompson, J.F., Triche, T.J.: The majority of total nuclear- encoded non-ribosomal RNA in a human cell is 'dark matter' unannotated RNA. BMC Biol. **8**, 149 (2010)
49. Mader, R.M., et al.: Reverse transcriptase template switching during reverse transcriptase-polymerase chain reaction: artificial generation of deletions in ribonucleotide reductase mRNA. J. Lab. Clin. Med. **137**, 422–8 (2001)
50. Cocquet, J., Chong, A., Zhang, G., Veitia, R.A.: Reverse transcriptase template switching and false alternative transcripts. Genomics **88**, 127–31 (2006)
51. Haddad, F., Qin, A.X., Giger, J.M., Guo, H., Baldwin, K.M.: Potential pitfalls in the accuracy of analysis of natural sense-antisense RNA pairs by reverse transcription-PCR. BMC Biotechnol. **7**, 21 (2007)
52. Ozsolak, F., Kapranov, P., Foissac, S., Kim, S.W., Fishilevich, E., Monaghan, A.P., John, B., Milos, P.M.: Comprehensive polyadenylation site maps in yeast and human reveal pervasive alternative polyadenylation. Cell **143**, 1018–29 (2010)
53. Malmström, H., Svensson, E.M., Gilbert, M.T., Willerslev, E., Götherström, A., Holmlund, G.: More on contamination: the use of asymmetric molecular behavior to identify authentic ancient human DNA. Mol. Biol. Evol. **24**, 998–1004 (2007)

54. Ginolhac, A., Vilstrup, J., Stenderup, J., Raghavan, M., Rasmussen, M., Stiller, M., Shapiro, B., Zazula, G., Froese, D., Steinmann, K.E., Thompson, J.F., AL-Rasheid, K.A.S., Gilbert, T., Willerslev, E., Orlando, L.: Improving the performance of true-single molecule sequencing for ancient DNA. (2011) (Submitted)
55. Thompson, J., Lipson, D., Hart, C., Kapranov, P., Letovsky, S., Milos, P., Ozsolak, F., Raz, T., Reifenberger, J., Steinmann, K., Loreille, O., Coble, M.: Sequencing the unsequenceable: applying massively parallel, single-molecule sequencing to badly degraded DNAs. In: Abstracts of the 59th Annual Meeting of The American Society of Human Genetics, Honolulu, 20–20 Oct 2009. http://www.ashg.org/2009meeting/abstracts/fulltext/f21866.htm
56. Yee, A.J., Raz, T., Amzallag, A., Lipson, D., Giladi, E., Lopez, H., Borger, D.R., Mino-Kenudson, M., Thompson, J.F., Iafrate, A.J., Milos, P., Haber, D.A., Ramaswamy, S.: Single molecule RNA sequencing of formalin-fixed paraffin-embedded tissue derived from patients with lung cancer. J. Clin. Oncol. **29**(15_suppl), 10550 (2011), http://www.asco.org/ASCOv2/Meetings/Abstracts?&vmview=abst_detail_view&confID=102&abstractID=78488
57. Branton, D., Deamer, D.W., Marziali, A., Bayley, H., Benner, S.A., et al.: The potential and challenges of nanopore sequencing. Nat. Biotechnol. **26**, 1146–53 (2008)
58. Deamer, D.: Nanopore analysis of nucleic acids bound to exonucleases and polymerases. Annu. Rev. Biophys. **39**, 79–90 (2010)
59. Timp, W., Mirsaidov, U.M., Wang, D., Comer, J., Aksimentiev, A., Timp, G.: Nanopore sequencing: electrical measurements of the code of life. IEEE Trans Nanotechnol. **9**, 281–294 (2010)
60. Kowalczyk, S.W., Blosser, T.R., Dekker, C.: Biomimetic nanopores: learning from and about nature. Trends Biotechnol. **29**(12), 607–614 (2011)
61. Venkatesan, B.M., Bashir, R.: Nanopore sensors for nucleic acid analysis. Nat. Nanotechnol. **6**, 615–24 (2011)
62. Healy, K.: Nanopore-based single-molecule DNA analysis. Nanomedicine **2**, 459–481 (2007)
63. Kasianowicz, J.J., Brandin, E., Branton, D., Deamer, D.W.: Characterization of individual polynucleotide molecules using a membrane channel. Proc. Natl. Acad. Sci. U.S.A. **93**, 13770–3 (1996)
64. Aksimentiev, A.: Deciphering ionic current signatures of DNA transport through a nanopore. Nanoscale **2**, 468–483 (2010)
65. Kowalczyk, S.W., Grosberg, A.Y., Rabin, Y., Dekker, C.: Modeling the conductance and DNA blockade of solid-state nanopores. Nanotechnology **22**, 315101 (2011)
66. Lagerqvist, J., Zwolak, M., Di Ventra, M.: Fast DNA sequencing via transverse electronic transport. Nano Lett. **6**, 779–82 (2006)
67. Krems, M., Zwolak, M., Pershin, Y.V., Di Ventra, M.: Effect of noise on DNA sequencing via transverse electronic transport. Biophys. J. **97**, 1990–6 (2009)
68. Chang, S., Huang, S., He, J., Liang, F., Zhang, P., Li, S., Chen, X., Sankey, O., Lindsay, S.: Electronic signatures of all four DNA nucleosides in a tunneling gap. Nano Lett. **10**, 1070–5 (2010)
69. Tsutsui, M., Taniguchi, M., Yokota, K., Kawai, T.: Identifying single nucleotides by tunnelling current. Nat. Nanotechnol. **5**, 286–90 (2010)
70. Ivanov, A.P., Instuli, E., McGilvery, C.M., Baldwin, G., McComb, D.W., Albrecht, T., Edel, J.B.: DNA tunneling detector embedded in a nanopore. Nano Lett. **11**, 279–85 (2011)
71. Bayley, H., Cremer, P.S.: Stochastic sensors inspired by biology. Nature. **413**, 226–30 (2001)
72. Butler, T.Z., Pavlenok, M., Derrington, I.M., Niederweis, M., Gundlach, J.H.: Single-molecule DNA detection with an engineered MspA protein nanopore. Proc. Natl. Acad. Sci. U.S.A. **105**, 20647–52 (2008)
73. Wendell, D., Jing, P., Geng, J., Subramaniam, V., Lee, T.J., Montemagno, C., Guo, P.: Translocation of double-stranded DNA through membrane-adapted phi29 motor protein nanopores. Nat. Nanotechnol. **4**, 765–772 (2009)
74. Maglia, G., Restrepo, M.R., Mikhailova, E., Bayley, H.: Enhanced translocation of single DNA molecules through alpha-hemolysin nanopores by manipulation of internal charge. Proc. Natl. Acad. Sci. U.S.A. **105**, 19720–5 (2008)

75. Stoddart, D., Heron, A.J., Mikhailova, E., Maglia, G., Bayley, H.: Single-nucleotide discrimination in immobilized DNA oligonucleotides with a biological nanopore. Proc. Natl. Acad. Sci. U.S.A. **106**, 7702–7 (2009)
76. Clarke, J., Wu, H.C., Jayasinghe, L., Patel, A., Reid, S., Bayley, H.: Continuous base identification for single-molecule nanopore DNA sequencing. Nat. Nanotechnol. **4**, 265–70 (2009)
77. Derrington, I.M., Butler, T.Z., Collins, M.D., Manrao, E., Pavlenok, M., Niederweis, M., Gundlach, J.H.: Nanopore DNA sequencing with MspA. Proc. Natl. Acad. Sci. U.S.A. **107**, 16060–5 (2010)
78. Manrao, E.A., Derrington, I.M., Pavlenok, M., Niederweis, M., Gundlach, J.H.: Nucleotide discrimination with DNA immobilized in the MspA nanopore. PLoS One **6**, e25723 (2011)
79. Kawano, R., Schibel, A.E., Cauley, C., White, H.S.: Controlling the translocation of single-stranded DNA through alpha-hemolysin ion channels using viscosity. Langmuir **25**, 1233–7 (2009)
80. de Zoysa, R.S., Jayawardhana, D.A., Zhao, Q., Wang, D., Armstrong, D.W., Guan, X.: Slowing DNA translocation through nanopores using a solution containing organic salts. J. Phys. Chem. B **113**, 13332–6 (2009)
81. Cockroft, S.L., Chu, J., Amorin, M., Ghadiri, M.R.: A single-molecule nanopore device detects DNA polymerase activity with single-nucleotide resolution. J. Am. Chem. Soc. **130**, 818–20 (2008)
82. Hornblower, B., Coombs, A., Whitaker, R.D., Kolomeisky, A., Picone, S.J., et al.: Single-molecule analysis of DNA-protein complexes using nanopores. Nat. Methods **4**, 315–17 (2007)
83. McNally, B., Singer, A., Yu, Z., Sun, Y., Weng, Z., Meller, A.: Optical recognition of converted DNA nucleotides for single-molecule DNA sequencing using nanopore arrays. Nano Lett. **10**, 2237–44 (2010)
84. Healy, K., Schiedt, B., Morrison, A.P.: Solid-state nanopore technologies for nanopore-based DNA analysis. Nanomedicine (Lond.) **2**, 875–97 (2007)
85. Storm, A.J., Storm, C., Chen, J., Zandbergen, H., Joanny, J.-F., Fast, D.C.: Fast DNA translocation through a solid-state nanopore. Nano Lett. **5**, 1193–1197 (2005)
86. He, Y., Tsutsui, M., Fan, C., Taniguchi, M., Kawai, T.: Controlling DNA translocation through gate modulation of nanopore wall surface charges. ACS Nano **5**, 5509–18 (2011)
87. Wanunu, M., Morrison, W., Rabin, Y., Grosberg, A.Y., Meller, A.: Electrostatic focusing of unlabelled DNA into nanoscale pores using a salt gradient. Nat. Nanotechnol. **5**, 160–5 (2009)
88. van den Hout, M., Krudde, V., Janssen, X.J., Dekker, N.H.: Distinguishable populations report on the interactions of single DNA molecules with solid-state nanopores. Biophys. J. **99**, 3840–8 (2010)
89. Luan, B., Aksimentiev, A.: Control and reversal of the electrophoretic force on DNA in a charged nanopore. J. Phys. Condens. Matter **22**, 454123 (2010)
90. Smeets, R.M., Kowalczyk, S.W., Hall, A.R., Dekker, N.H., Dekker, C.: Translocation of RecA-coated double-stranded DNA through solid-state nanopores. Nano Lett. **9**, 3089–96 (2009)
91. Lu, B., Albertorio, F., Hoogerheide, D.P., Golovchenko, J.A.: Origins and consequences of velocity fluctuations during DNA passage through a nanopore. Biophys. J. **101**, 70–9 (2011)
92. Trepagnier, E.H., Radenovic, A., Sivak, D., Geissler, P., Liphardt, J.: Controlling DNA capture and propagation through artificial nanopores. Nano Lett. **7**, 2824–30 (2007)
93. Peng, H., Ling, X.S.: Reverse DNA translocation through a solid-state nanopore by magnetic tweezers. Nanotechnology **20**, 185101 (2009)
94. Iqbal, S.M., Akin, D., Bashir, R.: Solid-state nanopore channels with DNA selectivity. Nat. Nanotechnol. **2**, 243–8 (2007)
95. Hall, A.R., Scott, A., Rotem, D., Mehta, K.K., Bayley, H., Dekker, C.: Hybrid pore formation by directed insertion of α-haemolysin into solid-state nanopores. Nat. Nanotechnol. **5**, 874–7 (2010)

96. Oliver, J., Bready, B., Goldstein, P., Preparata, F.: Biopolymer sequencing by hybridization of probes to form ternary complexes and variable range alignment. US patent application 20090099786 (2008)
97. Garaj, S., Hubbard, W., Reina, A., Kong, J., Branton, D., Golovchenko, J.A.: Graphene as a subnanometre trans-electrode membrane. Nature **467**, 190–3 (2010)
98. Merchant, C.A., Healy, K., Wanunu, M., Ray, V., Peterman, N., Bartel, J., Fischbein, M.D., Venta, K., Luo, Z., Johnson, A.T., Drndić, M.: DNA translocation through graphene nanopores. Nano Lett. **10**, 2915–21 (2010)
99. Schneider, G.F., Kowalczyk, S.W., Calado, V.E., Pandraud, G., Zandbergen, H.W., Vandersypen, L.M., Dekker, C.: DNA translocation through graphene nanopores. Nano Lett. **10**, 3163–7 (2010)
100. Sathe, C., Zou, X., Leburton, J.P., Schulten, K.: Computational investigation of DNA detection using graphene nanopores. ACS Nano. **11**, 8842–51 (2011)