Human Molecular Genetics, 2010, Vol. 19, Review Issue 2 **R227–R240**
doi:10.1093/hmg/ddq416
Advance Access published on September 21, 2010

A window into third-generation sequencing

Eric E. Schadt*, **Steve Turner and Andrew Kasarskis**

Pacific Biosciences, 1380 Willow Road, Menlo Park, CA 94025, USA

Received September 15, 2010; Revised and Accepted September 17, 2010

First- and second-generation sequencing technologies have led the way in revolutionizing the field of genomics and beyond, motivating an astonishing number of scientific advances, including enabling a more complete understanding of whole genome sequences and the information encoded therein, a more complete characterization of the methylome and transcriptome and a better understanding of interactions between proteins and DNA. Nevertheless, there are sequencing applications and aspects of genome biology that are presently beyond the reach of current sequencing technologies, leaving fertile ground for additional innovation in this space. In this review, we describe a new generation of single-molecule sequencing technologies (third-generation sequencing) that is emerging to fill this space, with the potential for dramatically longer read lengths, shorter time to result and lower overall cost.

INTRODUCTION

The genomics community has been enormously enabled by first- and second-generation sequencing (SGS) technologies in comprehensively characterizing DNA sequence variation, *de novo* sequencing of a number of species, sequencing of microbiomes, detecting methylated regions of the genome, quantitating transcript abundances, characterizing different isoforms of genes present in a given sample and identifying the degree to which mRNA transcripts are being actively translated (1–10). One of the hallmark features of the SGS technologies is their massive throughput at a modest cost, with hundreds of gigabases of sequencing now possible in a single run for several thousand dollars (11). Despite the recent and rapid acceptance of SGS technologies, a new generation of single-molecule sequencing (SMS) technologies is emerging (12–15). Unlike major SGS sequencing by synthesis (SBS) technologies that rely on PCR to grow clusters of a given DNA template, attaching the clusters of DNA templates to a solid surface that is then imaged as the clusters are sequenced by synthesis in a phased approach, the new generation of SBS technologies interrogate single molecules of DNA, such that no synchronization is required (a limitation of SGS) (16), thereby overcoming issues related to the biases introduced by PCR amplification and dephasing. More importantly, this new generation of sequencing technologies has the potential to exploit more fully the high catalytic rates and high processivity of DNA polymerase or avoid any biology or chemistry altogether to radically increase read length (from tens of bases to tens of thousands of bases per read) and time to result (from days to hours or minutes). The promises then of this new, third generation of sequencing technologies in offering advantages over current sequencing technologies are (i) higher throughput; (ii) faster turnaround time (e.g. sequencing metazoan genomes at high fold coverage in minutes); (iii) longer read lengths to enhance *de novo* assembly and enable direct detection of haplotypes and even whole chromosome phasing; (iv) higher consensus accuracy to enable rare variant detection; (v) small amounts of starting material (theoretically only a single molecule may be required for sequencing); and (vi) low cost, where sequencing the human genome at high fold coverage for less than $100 is now a reasonable goal for the community.

But how do these next–next-generation technologies work? What scales of data generation will be achieved with these new technologies? What types of 'sequencing' data can be generated? Will they ease analysis issues and/or create new ones? And, most importantly, what are the timelines for these technologies to become available, will they really meet the above promises and what do we need to do to prepare? In this review we will address these questions, providing insights into third-generation sequencing (TGS) that promises to bring sequencing to nearly every aspect of our lives. What will it take to be ready?

RESULTS

A brief history on first-generation sequencing and SGS

The process of sequencing DNA consists of three basic phases comprising sample preparation, physical sequencing and

*To whom correspondence should be addressed. Tel: +1 6505218250; Fax: +1 6503239420; Email: eschadt@pacificbiosciences.com

© *The Author 2010. Published by Oxford University Press.*
This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (http://creativecommons.org/licenses/by-nc/2.5), which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.

Downloaded from https://academic.oup.com/hmg/article/19/R2/R227/641295 by guest on 06 November 2022

Univ. of Washington EX2023
Guardant Health, Inc v. Univ. of Washington
IPR2022-00817

re-assembly. The first step of sample preparation is to break the target genome into multiple small fragments. Depending on the amount of sample DNA, these fragments may be amplified into multiple copies using a variety of molecular methods. In the physical sequencing phase, the individual bases in each fragment are identified in order, creating individual reads. The number of bases identified contiguously is defined as read length. In the re-assembly phase, bioinformatics software is used to align overlapping reads, which allows the original genome to be assembled into contiguous sequences. The longer the read length, the easier it is to reassemble the genome (17).

First-generation sequencing

First-generation sequencing was originally developed by Sanger in 1975 (the chain-termination method) (18,19) and in parallel by Maxam and Gilbert in 1977 (a chemical sequencing method) (20). From these first-generation methods, Sanger sequencing ultimately prevailed given it was less technically complex and more amenable to being scaled up. For Sanger sequencing practiced today, during sample preparation, different-sized fragments of DNA are generated each starting from the same location (Fig. 1A). Each fragment ends with a particular base that is labeled with one of four fluorescent dyes corresponding to that particular base. Then all of the fragments are distributed in the order of their length via capillary electrophoresis. Information regarding the last base is used to determine the original sequence. This method results in a read length that is ~800 bases on average, but may be extended to above 1000 bases (21–23). While fully automated implementations of this approach were the mainstay for the original sequencing of the human genome, their chief limitation was the small amounts of DNA that could be processed per unit time, referred to as throughput, as well as high cost, resulting in it taking roughly 10 years and three billion dollars to sequence the first human genome (11,22) (Table 1).

Second-generation sequencing

Commercial SGS tools emerged in 2005 in response to the low throughput and high cost of first-generation methods. To address this problem, SGS tools achieve much higher throughput by sequencing a large number of DNA molecules in parallel (Fig. 1B). With most SGS technologies, tens of thousands of identical strands are anchored to a given location to be read in a process consisting of successive washing and scanning operations. The 'wash-and-scan' sequencing process involves sequentially flooding in reagents, such as labeled nucleotides, incorporating nucleotides into the DNA strands, stopping the incorporation reaction, washing out the excess reagent, scanning to identify the incorporated bases and finally treating the newly incorporated bases to prepare the DNA templates for the next 'wash-and-scan' cycle (11). This cycle is repeated until the reaction is no longer viable. The array of DNA anchor locations can have a very high density of DNA fragments, leading to extremely high overall throughput and a resultant low cost per identified base when such instruments are run at high capacity. For example, Illumina's HiSeq 2000 instrument can generate upwards of 300 or more gigabases of sequence data in a single run. The time-to-result for these SGS methods is generally long (typically taking many days), due to the large number of scanning and washing cycles required. Furthermore, because step yields for the addition of each base are <100%, a population of molecules becomes more asynchronous as each base is added (16). This loss of synchronicity (called dephasing) causes an increase in noise and sequencing errors as the read extends, effectively limiting the read length produced by the most widely used SGS systems to significantly less than the average read lengths achieved by Sanger sequencing (11,17). Further, in order to generate this large number of DNA molecules, PCR amplification is required. The amplification process can introduce errors in the template sequence as well as amplification bias. The effects of these pathologies are that neither the sequences nor the frequencies with which they appear are always faithfully preserved. In addition, the process of amplification increases the complexity and time associated with sample preparation. Finally, the massively high throughput achieved by SGS technologies per run generates mountains of highly informative data that challenge data storage and informatics operations, especially in light of the shorter reads (compared with Sanger sequencing) that make alignment and assembly processes challenging (17).

First-generation sequencing and SGS technologies have led the way in revolutionizing the field of genomics and beyond, motivating an astonishing number of scientific advances [for a comprehensive review of SGSs, see (11)]. Nevertheless there are sequencing applications and aspects of genome biology that are presently beyond the reach of current sequencing technologies, leaving fertile ground for additional innovation in this space.

Transitioning from SGS to TGS

There may not yet be consensus on what constitutes a third generation, or next–next-generation sequencing instrument, given advances are being made on rapid time scales that do not easily fit into generational time scales. However, for the purposes of this review article, we focus on SMS without the need to halt between read steps (whether enzymatic or otherwise), where reads from SMS instruments represent sequencing of a single molecule of DNA. SMS technologies that do not purposefully pause the sequencing reaction after each base incorporation represent the most thoroughly explored TGS approaches in hopes of increasing sequencing rates, throughput and read lengths, lowering the complexity of sample preparation and ultimately decreasing cost. However, as a result of using these criteria to define TGS, a number of exciting technologies do not fit neatly into this definition, but are nevertheless exciting in terms of how they complement current SGS technologies.

One technology that sits between the SGS and TGS categories is Ion Torrent's (now acquired by Life Technologies) semiconductor sequencer. An interesting facet of Ion Torrent's sequencing instrument is that state-of-the-art semiconductor technology is employed to create a high-density array of micro-machined wells that carry out SBS by sensing the release of hydrogen ions as part of the base incorporation process. This process eliminates the need for light, scanning and cameras to monitor the SBS process, thereby simplifying the overall sequencing process, dramatically accelerating

Downloaded from https://academic.oup.com/hmg/article/19/R2/R227/641295 by guest on 06 November 2022

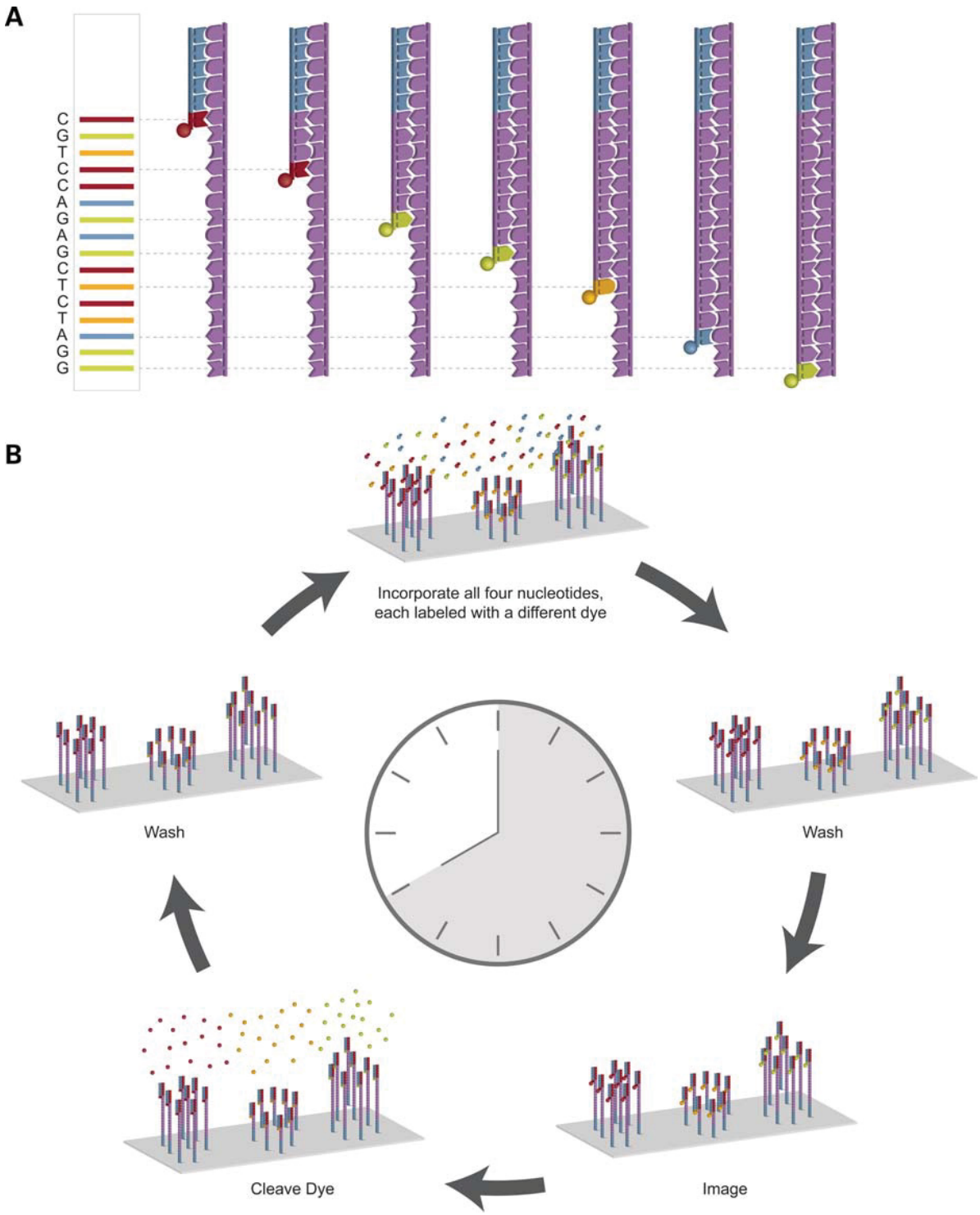

Figure 1. How previous generation DNA-sequencing systems work. (**A**) A modern implementation of Sanger sequencing is shown to illustrate differential labeling and use of terminator chemistry followed by size separation to resolve the sequence. (**B**) The Illumina sequencing process is shown to illustrate the wash-and-scan paradigm common to second-generation DNA-sequencing technologies.

the time to result, reducing the overall footprint of the instrument, and lowering cost to make DNA sequencing more generally accessible to all. However, this technology is still a 'wash-and-scan' system like all current SGS technologies, requiring PCR amplification of the DNA template in each well, as well as termination events typically halting sequencing after each nucleotide incorporation, in order to monitor in succession the incorporation of each of the four bases across all DNA templates. As a result of this process, the overall read length is limited to that of current SGS systems,

Downloaded from https://academic.oup.com/hmg/article/19/R2/R227/641295 by guest on 06 November 2022

Table 1. Comparison of first-generation sequencing, SGS and TGS

	First generation	Second generation[a]	Third generation[a]
Fundamental technology	Size-separation of specifically end-labeled DNA fragments, produced by SBS or degradation	Wash-and-scan SBS	SBS, by degradation, or direct physical inspection of the DNA molecule
Resolution	Averaged across many copies of the DNA molecule being sequenced	Averaged across many copies of the DNA molecule being sequenced	Single-molecule resolution
Current raw read accuracy	High	High	Moderate
Current read length	Moderate (800–1000 bp)	Short, generally much shorter than Sanger sequencing	Long, 1000 bp and longer in commercial systems
Current throughput	Low	High	Moderate
Current cost	High cost per base Low cost per run	Low cost per base High cost per run	Low-to-moderate cost per base Low cost per run
RNA-sequencing method	cDNA sequencing	cDNA sequencing	Direct RNA sequencing and cDNA sequencing
Time from start of sequencing reaction to result	Hours	Days	Hours
Sample preparation	Moderately complex, PCR amplification not required	Complex, PCR amplification required	Ranges from complex to very simple depending on technology
Data analysis	Routine	Complex because of large data volumes and because short reads complicate assembly and alignment algorithms	Complex because of large data volumes and because technologies yield new types of information and new signal processing challenges
Primary results	Base calls with quality values	Base calls with quality values	Base calls with quality values, potentially other base information such as kinetics

[a]There are many TGS technologies in development but few have been reduced to practice. While there is significant potential of TGS to radically improve current throughput and read-length characteristics (among others), the ultimate practical limits of these technologies remain to be explored. Furthermore, there is active development of SGS technologies that will also improve read-length and throughput characteristics.

and ultimately, throughput is limited as well, compared with what SMS platforms will be capable of achieving.

Sitting even closer to the TGS boundary is the Helicos Genetic Analysis Platform, the first commercially available sequencing instrument to carry out SMS (24–27). The Helicos sequencing instrument works by imaging individual DNA molecules affixed to a planar surface as they are extended using a defined primer and a modified polymerase as well as proprietary fluorescently labeled nucleotide analogues, referred to as Virtual Terminator nucleotides, in which the dye is attached to the nucleotide via a chemically cleavable group that allows for step-wise sequencing to be carried out (25). Because halting is still required in this process (similar to SGS technologies), the time to sequence a single nucleotide is high, and the read lengths realized are ~32 nucleotides long. However, given the SMS nature of this technology, no PCR is required for sequencing, a significant advantage over SGS technologies. However, also due to the single-molecule nature of this technology (and all of the SMS technologies), the raw read error rates are generally at or >5%, although the highly parallel nature of this technology can deliver high fold coverage and a consensus or finished read accuracy of >99%. This technology is capable of sequencing an entire human genome, albeit at significant cost by today's standards (roughly $50 000 in reagents) (28). It can follow roughly one billion individual DNA molecules as they are sequenced over the course of many days. Unlike SGS, these many hundreds of millions of sequencing reactions can be carried out asynchronously, a hallmark of TGS. Further, given individual monitoring of templates, the enzymatic incorporation step does not need to be driven to completion, which serves to reduce the overall mis-incorporation error rate. As with the other TGS technologies discussed below, deletions and insertions are a significant issue.

The sample preparation part of this technology involves fragmenting genomic DNA into smaller pieces, adding a 3' poly(A) tail to the fragments, labeling and blocking by terminal transferase. These templates are then captured onto a surface with covalently bound 5' dT(50) oligonucleotides via hybridization (25). The surface is then imaged using charge-coupled device (CCD) sensors, where those templates that have been appropriately captured are identified and then tracked for SBS. The process then resembles the 'wash-and-scan' steps of SGS in which a labeled nucleotide and polymerase mixture are flooded onto the system and incubated for a period of time, the surface is then washed to remove the synthesis mixture and scanned to detect the fluorescent label. The dye–nucleotide linker is then cleaved to release the dye, and this process is repeated.

Not only can this technology be used to sequence DNA, but the DNA polymerase can be replaced with a reverse transcriptase enzyme to sequence RNA directly (29), without requiring the conversion of RNA to cDNA or without the need for ligation/amplification steps, something all existing SGS technologies require for RNA sequencing (5). Instead, each RNA molecule is polyadenylated and 3'-blocked and captured on a surface coated with dT(50) oligonucleotides, similar to the DNA sequencing process. Sequencing is then carried out as described for DNA, but using reverse transcriptase instead of DNA polymerase. In

Downloaded from https://academic.oup.com/hmg/article/19/R2/R227/641295 by guest on 06 November 2022

addition to direct RNA sequencing, the Helicos platform can carry out other sequencing-based assays such as chromatin profiling (30).

While the Helicos SMS technology has been successfully deployed, representing the first example of true SMS, with many significant advantages over SGS technologies, it has many of the characteristics of SGS technologies and so has had a more difficult time clearly differentiating itself from SGS with respect to read lengths, throughput and run times, all of which are similar to leading SGS technologies. When combined with a higher raw read error rate (requiring repetitive sequencing to overcome), the end result is a higher sequencing cost compared with leading SGS technologies. While the Helicos technology may struggle to clearly differentiate itself from SGS in some respects, the direct RNA-sequencing application is the type of advance that will come to clearly distinguish this technology from SGS.

Third-generation DNA sequencing

SMS technologies can roughly be binned into three different categories: (i) SBS technologies in which single molecules of DNA polymerase are observed as they synthesize a single molecule of DNA; (ii) nanopore-sequencing technologies in which single molecules of DNA are threaded through a nanopore or positioned in the vicinity of a nanopore, and individual bases are detected as they pass through the nanopore; and (iii) direct imaging of individual DNA molecules using advanced microscopy techniques. Each of these technologies provides novel approaches to sequencing DNA and has advantages and disadvantages with respect to specific applications. These technologies are at varying stages of development, making the writing of a review on TGS difficult given there is still much to prove regarding the utility of many of the TGS technologies. However, if the full potential of these technologies is realized, in several years time, whole genome sequencing will likely be fast enough and inexpensive enough to resequence genomes as needed for any application. Here we discuss many of the emerging TGS technologies that have the potential to make such stunning advances possible.

SMS sequencing by synthesis

Single-molecule real-time sequencing. The single-molecule real-time (SMRT) sequencing approach developed by Pacific Biosciences is the first TGS approach to directly observe a single molecule of DNA polymerase as it synthesizes a strand of DNA, directly leveraging the speed and processivity of this enzyme to address many of the shortcomings of SGS (14,31). Given that a single DNA polymerase molecule is of the order of 10 nm in diameter, two important obstacles needed to be overcome to enable direct observation of DNA synthesis as it occurs in real time are: (i) confining the enzyme to an observation volume that was small enough to achieve the signal-to-noise ratio needed to accurately call bases as they were incorporated into the template of interest; and (ii) labeling the nucleotides to be incorporated in the synthesis process such that the dye–nucleotide linker is cleaved after completion of the incorporation process so that a natural strand of DNA remains for continued synthesis and so that multiple dyes are not held in the confinement volume at a time (something that would destroy the signal-to-noise ratio).

The problem of observing a DNA polymerase working in real time, detecting the incorporation of a single nucleotide taken from a large pool of potential nucleotides during DNA synthesis, was solved using zero-mode waveguide (ZMW) technology (Fig. 2A) (31). The principle employed is similar to that employed in the protective screen in a microwave oven door. The screen is perforated with holes that are much smaller than the wavelength of the microwaves. Because of their relative size, the holes prevent the much longer microwaves from passing through and penetrating the glass. However, the much smaller wavelengths of visible light are able to pass through the holes in the screen, allowing food to be visible as it is cooked. ZMWs can be made to operate in a similar manner for DNA sequencing.

A ZMW is a hole, tens of nanometers in diameter, fabricated in a 100 nm metal film deposited on a glass substrate. The small size of the ZMW prevents visible laser light, which has a wavelength of ~600 nm, from passing entirely through the ZMW. Rather than passing through, the light exponentially decays as it enters the ZMW. Therefore, by shining laser illumination up through the glass into the ZMW, only the bottom 30 nm of the ZMW becomes illuminated. Within each ZMW, a single DNA polymerase molecule is anchored to the bottom glass surface using biotin/streptavidin interaction (Fig. 2A). Nucleotides, each type labeled with a different colored fluorophore, are then flooded above an array of ZMWs at the required concentration. Diffusion at the nanoscale occurs in microseconds, so that labeled nucleotides travel down into the ZMW, surround the DNA polymerase, then diffuse back up and exit the hole. As no laser light penetrates up through the holes to excite the fluorescent labels, the labeled nucleotides above the ZMWs do not contribute to the measured signals. Only when they diffuse through the bottom 30 nm of the ZMW do they fluoresce. When the correct nucleotide is detected by the polymerase, it is incorporated into the growing DNA strand in a process that takes milliseconds, approximately three orders of magnitude longer than simple diffusion. This difference in time results in higher signal intensity for incorporated versus unincorporated nucleotides, which creates a high signal-to-noise ratio. While held by the polymerase, the fluorescent label emits colored light. The sequencing instrument detects this as a flash whose color corresponds to the base identity. Following incorporation, the signal immediately returns to baseline and the process repeats, with the DNA polymerase continuing to incorporate multiple bases per second. Thus, the ZMW has the ability to detect a single incorporation event against the background of fluorescently labeled nucleotides at biologically relevant concentrations. The first commercial instance of the SMRT sequencing instrument will consist of an array of ~75 000 ZMWs. Each ZMW is capable of containing a DNA polymerase loaded with a different strand of DNA sample. As a result, the array enables the potential detection of ~75 000 SMS reactions in parallel. At present, because the DNA polymerase and DNA template to be sequenced are delivered to ZMWs via a random diffusion process, approximately a third of the ZMWs of a given array are active for a given run.

Downloaded from https://academic.oup.com/hmg/article/19/R2/R227/641295 by guest on 06 November 2022

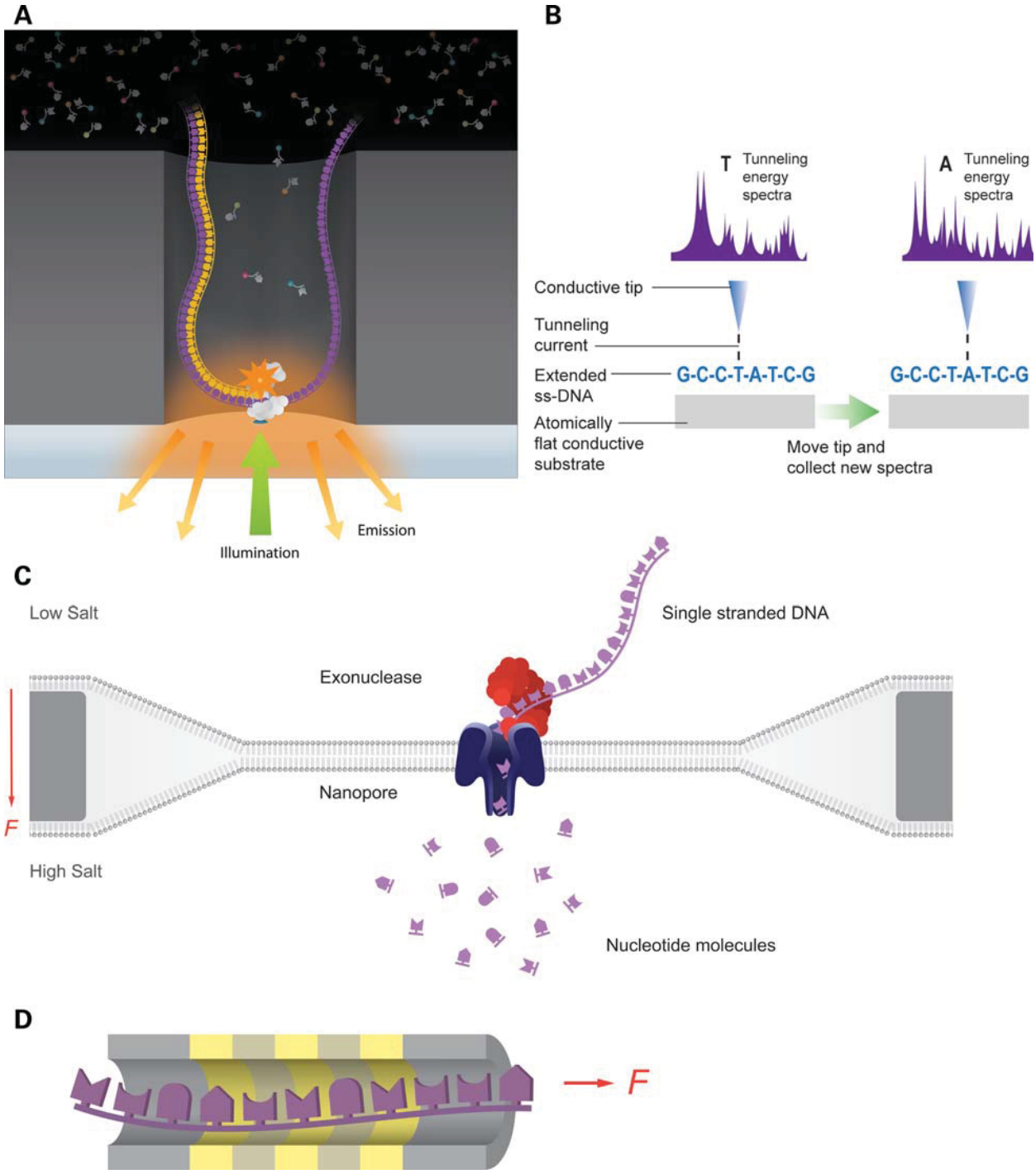

Figure 2. How third-generation DNA-sequencing technologies work. Third-generation DNA-sequencing technologies are distinguished by direct inspection of single molecules with methods that do not require wash steps during DNA synthesis. (**A**) Pacific Biosciences technology for direct observation of DNA synthesis on single DNA molecules in real time. A DNA polymerase is confined in a zero-mode waveguide and base additions measured with florescence detection of gamma-labeled phosphonucleotides. (**B**) Several companies seek to sequence DNA by direct inspection using electron microscopy similar to the Reveo technology pictured here, in which an ssDNA molecule is first stretched and then examined by STM. (**C**) Oxford Nanopore technology for measuring translocation of nucleotides cleaved from a DNA molecule across a pore, driven by the force of differential ion concentrations across the membrane. (**D**) IBM's DNA transistor technology reads individual bases of ssDNA molecules as they pass through a narrow aperture based on the unique electronic signature of each individual nucleotide. Gold bands represent metal and gray bands dielectric layers of the transistor.

ZMWs overcome the first obstacle, but not the second. All SGS technologies directly attach the dye to the base, which is incorporated into the DNA strand. This is problematic for any system attempting to observe DNA synthesis in real time because the dye's large size relative to the DNA can interfere with the activity of the DNA polymerase. Typically, a DNA

Downloaded from https://academic.oup.com/hmg/article/19/R2/R227/641295 by guest on 06 November 2022

polymerase can incorporate only a few base-labeled nucleotides before it halts. The SMRT sequencing approach instead attaches the fluorescent dye to the phosphate chain of the nucleotide rather than to the base. As a natural step in the synthesis process, the phosphate chain is cleaved when the nucleotide is incorporated into the DNA strand. Thus, upon incorporation of a phospholinked nucleotide, the DNA polymerase naturally frees the dye molecule from the nucleotide when it cleaves the phosphate chain. Upon cleaving, the label quickly diffuses away, leaving a completely natural piece of DNA with no evidence of labeling remaining.

The SMRT sequencing platform requires minimal amounts of reagent and sample preparation to carry out a run, and there are no time-consuming scanning and washing steps, enabling time to result in a matter of minutes as opposed to days (14). In addition, SMRT sequencing does not require routine PCR amplification needed by most SGS systems, thereby avoiding systematic amplification bias. Because the processivity of the DNA polymerase is leveraged, SMRT sequencing realizes longer read lengths than any other technology at present, having the potential to produce average read lengths >1000 bp and maximum read lengths in excess of 10 000 bp, enabling *de novo* assembly, direct detection of haplotypes and even providing for the possibility of phasing entire chromosomes. Sample preparation processes for SGS technology often involve costly additional capital equipment, reagents, supplies and physical space. The sample preparation process for SGS can take multiple days. However, with SMRT sequencing, the sample preparation consists of fragmenting the DNA into desired lengths, blunting the ends, ligating hairpin adaptors and then sequencing (32). This provides for considerable flexibility in configuring the system for different applications.

One of the more interesting features of SMRT sequencing is the ability to observe and capture kinetic information. The ability to observe the activity of DNA polymerase in real time allows for the collection, measurement and assessment of the dynamics and timing of enzymatic incorporation, referred to as kinetics. Via the SMRT sequencing process, changes in the kinetics of incorporation associated with chemical modifications to bases, such as methylation, can be detected in the normal course of collecting sequence data (2).

Beyond DNA sequencing, the SMRT sequencing instrument is flexible and should lead to a number of applications that are presently not approachable by any existing technology. For example, one recently published application of the SMRT technology demonstrated direct, real-time observation of the ribosome as it translated mRNA (33). Direct observation of other enzymes, like RNA-dependent polymerases and reverse transcriptase for RNA-sequencing applications, should also be possible.

Despite the many potential advantages of SMRT sequencing, a number of challenges remain. Like the Helicos technology, the raw read error rates can be in excess of 5%, with error rates dominated by insertions and deletions, particularly problematic errors when aligning sequences and assembling genomes. In addition, the throughput of SMRT sequencing will not initially match what can be achieved by SGS. The throughput of SMRT sequencing is a function of the number of ZMWs that can be read at once. While ultimately the potential exists to observe many ZMWs in parallel, the first version released will be capable of only up to 75 000 ZMWs. Finally, as expected to be the case for most TGS technologies, SMRT sequencing data are different in form from SGS data, hence they are amenable to more advanced probabilistic modeling that has the potential to exploit more information about the chemical and structural nature of nucleotide sequences than previous sequencing technologies (as discussed below for most TGS technologies).

Real-time DNA sequencing using fluorescence resonance energy transfer. Other SMS SBS technologies are in development, but little data are available to assess where they are at in development and when they are likely to be released. VisiGen Biotechnologies had one of the more promising approaches to SMS whereby the DNA polymerase is tagged with a fluorophore that when brought into close proximity to a nucleotide, tagged with an acceptor fluorophore, would emit a fluorescence resonance energy transfer (FRET) signal. After incorporation, the fluorophore label on the nucleotide can be released. This type of approach could be considered an improvement over the Helicos technology, and has the potential to move at millions of bases per second, given potential for high multiplex. Visigen Biotechnologies was acquired by Life Technologies recently, and the FRET technology seems to have become one of centerpieces of their SMS efforts, but presently it is hard to gauge progress.

Tunneling and transmission-electron-microscopy-based approaches for DNA sequencing

Direct imaging of DNA using electron microscopy. Halcyon Molecular is pioneering an SMS approach using transmission electron microscopy (TEM) to directly image and chemically detect atoms that would uniquely identify the nucleotides comprising a DNA template. The approach being pursued has been shown to reliably detect atoms in a non-periodic material on a planar surface, using annular dark-field imaging in an aberration-corrected scanning TEM (13). This approach harkens back to a lecture Richard Feynman gave in 1959 at the annual meeting of the American Physical Society at Caltech where he indicated the easiest way to study important biomolecules like DNA, RNA and proteins was to look at them directly. Beyond the TEM technology, Halcyon is developing a number of supporting technologies required to carry out TEM-based DNA sequencing, like the use of functionalized needles to attach stretched molecules of DNA to a substrate for the direct imaging procedure. As of the writing of this review, no publications have appeared demonstrating this procedure for DNA sequencing, but if successfully implemented, the chief advantage of the technology would be very long read lengths (potential into the many millions of bases) at low cost.

ZS genetics is developing another TEM-based DNA sequencing instrument to directly image the sequence. With this technology, labeled atoms within the nucleotides of DNA are imaged using a high-resolution (sub-angstrom) electron microscope where individual bases are detected and identified based on their size and intensity differences between the different labeled bases. While no proof of concept studies have yet been published regarding this technology, ZS Genetics

Downloaded from https://academic.oup.com/hmg/article/19/R2/R227/641295 by guest on 06 November 2022

claims that the technology is capable of producing 10 000–20 000 base reads at a rate of 1.7 billion bases per day. Like most of the other TGS technologies, read length and reduced costs are expected to be the chief advantages.

Direct imaging of DNA sequences using scanning tunneling microscope tips. Reveo is developing a technology related to IBM's DNA transistor approach (see subsequently) in which DNA is placed on a conductive surface to detect bases electronically using scanning tunneling microscope (STM) tips and tunneling current measurements (34). The STM tips are knife-edge shaped and have nanoscale dimensions (Fig. 2B). The aim in applying this technology to SMS is to stretch and confine a molecule of DNA such that tunneling current measurements can be taken to identify individual bases. The procedure for linearizing and depositing DNA sequences on a conductive surface for this application has not yet been described. No proof of concept study for DNA sequencing has been published, but the advantages of this type of technology are expected again to be speed, very long read lengths and a significant reduction in cost, given labeling can be avoided.

DNA sequencing with nanopores

Most nanopore sequencing technologies rely on transit of a DNA molecule or its component bases through a hole and detecting the bases by their effect on an electric current or optical signal. Because this type of technology uses single molecules of unmodified DNA, they have the potential to work quickly on extremely small amounts of input material. Both biological nanopores constructed from engineered proteins and entirely synthetic nanopores are under development. In particular, there is potential to use atomically thin sheets of grapheme as a matrix supporting nanopores (35) and also carbon nanotubes (36).

Direct, electrical detection of single DNA molecules. Oxford Nanopore is commercializing a system for DNA sequencing based on three natural biological molecules that have been engineered to work as a system (Fig. 2C) (37–39). The biological nanopore is constructed from a modified α-hemolysin pore that has an exonuclease attached on the normally extracellular face of the pore. A synthetic cyclodextrin sensor is also covalently attached to the inside surface of the nanopore. This system is contained in a synthetic lipid bilayer so that when DNA is loaded onto its exonuclease-containing face and a voltage is applied across the bilayer by changing the concentration of salt, the exonuclease can cleave off individual nucleotides. The individual nucleotides are detected once they are cleaved based on their characteristic disruption of the ionic current flowing through the pore. Reliable throughput at high multiplex may be difficult to achieve with this system using natural lipid bilayers, but synthetic membranes and solid-state nanopores, if developed, may help overcome this challenge. Like many of the other TGS technologies in this category, the advantages are expected to be long read length and high scalability at low cost, given the technology is driven by electronics, not optics.

Nanopore DNA sequencing with MspA. Another approach aims to use a biological nanopore *directly* on intact DNA. Unlike Oxford Nanopore, which addresses axial resolution limitations in the alpha-haemolysin pore by disassembling the DNA molecule, in this case, the *Mycobacterium smegmatis* Porin A (MspA) protein, which has a shorter blockade region and thus a better resolution, is used as the pore and the effect of a linear molecule of single-stranded DNA (ssDNA) on the current transiting the pore is measured (12). To slow the transit of the ssDNA through the pore to a level allowing detection of individual bases as they interrupt current transiting the pore, a region of double-stranded DNA (dsDNA) is introduced. The ability of this method to directly measure ssDNA in a processive fashion is attractive, but the complexity of introducing the needed dsDNA break on the pore transit velocity appears to be a significant obstacle at this point to an efficient large-scale laboratory workflow for routine DNA sequencing.

Nanopore sequencing with optical readout. One significant challenge in nanopore-based sequencing lies in the need for simultaneously monitoring a large number of nanopores. The first parallel readout of any nanopore-based method has recently been demonstrated through the use of optical multipore detection (40). In this approach, the contrast between the four bases is first increased off-line through a biochemical process that converts each base in the DNA into a specific, ordered pair of concatenated oligonucleotides. Subsequently, two different fluorescently labeled molecular beacons are hybridized to the converted DNA. The beacons are then sequentially unzipped from the DNA molecules as they are translocated through a nanopore. Each unzipping event unquenches a new fluorophore, resulting in a series of dual-color fluorescence pulses that are detected by a high-speed CCD camera with a conventional total internal reflection fluorescence microscopy setup. The unzipping process is slowed down by adjusting the voltage governing DNA translocation through the nanopore to a speed compatible with single-molecule optical detection. With the feasibility of the components of this approach demonstrated, it will be interesting to see whether the potential of extremely high throughput can be achieved through faithful and unbiased biochemical conversion, and accurate, long-read sequencing with high parallelism.

Transistor-mediated DNA sequencing. IBM is developing a nanostructured sequencing device capable of electronically detecting individual bases in a single molecule of DNA (Fig. 2D) (41). The nanostructures are nanometer-sized pores. The surface of the pores consists of axially stratified, alternating layers of metal and dielectric material (like a transistor). Single DNA molecules can then be passed through the pores, controlling the motion of the DNA through the pores by appropriately modulating the current in the electrodes of the transistor. Speed, read length and low cost are again the chief advantages of this type of approach. In fact, the speed of sequencing could be very dramatically increased with this approach, given the theoretical limit has been computed to be 500 000 000 bases read per transistor per second. In addition, like other TGS technologies in this category, the

Downloaded from https://academic.oup.com/hmg/article/19/R2/R227/641295 by guest on 06 November 2022

assay would be label free and require no optics, again greatly diminishing cost.

While the original DNA transistor idea proposed was based on theoretical calculations and molecular dynamic simulations (41), IBM recently published a solution to one of the two technical challenges facing this approach: modulation of the speed with which the DNA molecule is passed through the nanopore to enable optimal base orientation as well as sufficient sampling of a base as it passes through the nanopore (15). The other challenge remaining for this approach is demonstrating that the signal for a single base can be distinguished from the signals of nearby bases. A recent publication related to this challenge indicates via simulation that factors such as ionic motion in the nanopore may not necessarily affect the desired signal of an individual base (42).

Compared with Oxford Nanopore Technologies, the IBM approach would be cheaper and potentially more stable. In particular, IBM's approach will not have the same issues with respect to spatial resolution and sensitivity, which are issues with Oxford Nanopore's approach (43). However, an advantage of Oxford Nanopore over the DNA transistor is that it requires less detection sensitivity, given it is detecting cleaved bases, not intact DNA molecules.

TGS informatics opportunities

The informatics challenges with SGS technologies are largely due to the short reads that are characteristic of these technologies (17). The short-read nature of SGS makes it difficult, even with paired-end reads, to assemble complete genomes *de novo*. Indeed, nearly all human genomes sequenced to date have been assembled using reference-based mapping algorithms (17). While this assembly approach is efficient for accurately identifying SNPs in the human genome (44,45), it does not enable a thorough characterization of structural variations, insertions and deletions. Only *de novo* assembly of individual genomes can accomplish this feat. A number of assemblers for *de novo* assembly have been written for SGS, including overlap graph approaches in which contigs are assembled by looking at sequence overlaps, like Edna (46), VCAKE (47) and SHARCGS (48); those based on the de Bruijn graph data structure (49), like Velvet (50), EULER-SR (51) and ALLPATHS (52); and more recent efforts that use the de Bruijn graph approach but incorporate additional information (e.g. genome repeat structure) to enhance assembly (53). While reasonable assemblies are now feasible using state-of-the-art SGS technologies and algorithms, they are still not capable of achieving the assembly qualities that can be achieved using first-generation Sanger sequencing, with hybrid sequencing approaches that include data generated from multiple technologies now becoming a more standard way to enhance the quality of assemblies (17).

Most of the TGS technologies discussed address (or have the potential to address) the limitations of SGS technologies with respect to assembly quality, given the read lengths and mate-pair distances in TGS are not only significantly beyond those realized with SGS, but with Sanger sequencing as well. Longer reads can span repeat regions that make assembly difficult and can obviate the need for more complex mate-pair strategies required to scaffold SGS reads. As an example, depicted in Figure 3A are seven contigs assembled using Abyss (54) applied to short read data generated from the genome of *Rhodopseudomonas palustris* using the Illumina GA platform. Because the six blue contigs are overlapping, the red contig representing a 1.5 kb repeat region, and because none of the contigs spans the repeat region, the contigs cannot be ordered with respect to each other. However, in Figure 3B, we depict just three molecules of long-range sequencing data (Figure 3, legend) from the TGS SMRT sequencing platform that span the repeat and unambiguously resolve how the contigs should be ordered with respect to one another.

Its strengths notwithstanding, TGS will come with its own set of challenges. Because a TGS system by definition assays a single molecule, there is no longer any safety in numbers to minimize raw read errors. For example, if unlabeled nucleotide were present at 0.001% in a reagent lot for an SBS system, there is certain to be a 0.001% rate of deletions in raw data produced with that reagent unless the labeled base is incorporated more rapidly by the enzyme. Similarly, if a base fails to progress through a nanopore or a DNA transistor as intended and gets counted twice, there will be an insertion in the raw data. Hence, the frequency of errors for raw reads will likely be greater, and the error profile of TGS will certainly differ from that of earlier technologies, so both will need to be accounted for in the algorithms that analyze TGS data. However, because the error profile may be less biased (more uniform), the consensus accuracies have the potential to be significantly higher than that of SGS.

While the longer read lengths of TGS will ease many of the informatics challenges relating to assembly now experienced by those focussed on SGS data, the increased information content will demand new types of mathematical models and algorithms to get the most from the data. For example, real-time monitoring of SMS events can provide kinetic information that transforms one's ability to understand each base as it is incorporated (e.g. the identity of the base, whether the base has been chemically modified, damaged and so on) (2). In addition, because a single molecule at a time is being monitored, the error structure will be significantly different from the ensemble-based approach employed by SGS technologies, with higher error rates for raw reads, but then consensus sequences converging more rapidly to higher quality sequences, given significantly fewer biases in the distribution of errors (14,32). Therefore, because this new generation of sequencing technologies provides for a significant shift in how sequencing is carried out, they demand a new generation of analysis tools to derive maximal information from the raw data.

For example, as described above, Illumina's Genome Analyzer/2 (and HiSeq 2000) sequences clusters of template DNA sequences in a cyclical fashion, where for each cycle a fluorescently labeled complementary base is sequenced for each cluster in which that base represents the next nucleotide in the template sequence (44). As a result, for each cluster, there correspond four images per cycle, and the analysis proceeds by analyzing each of the images and quantitating the intensities for each cluster and selecting the dominant intensity to determine the most likely base for a given cluster at a given cycle. The primary issues relating to analysis of these data

Downloaded from https://academic.oup.com/hmg/article/19/R2/R227/641295 by guest on 06 November 2022

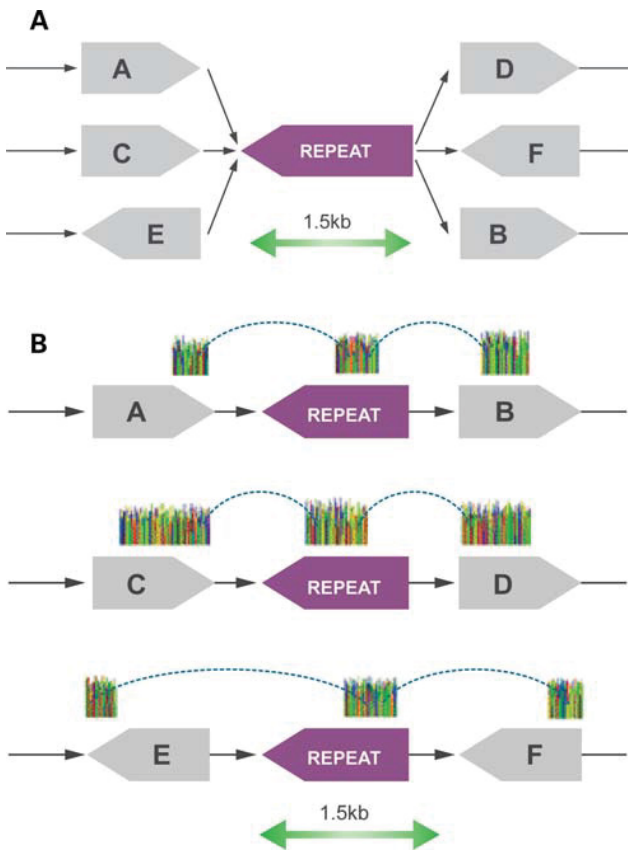

Figure 3. Long reads span long repeats to unambiguously orient contigs. TGS Technologies are capable of generating long reads that are critical for *de novo* assembly of genomes. (**A**) Contigs assembled from short read data alone cannot be unambiguously ordered because they overlap but do not span a repeat region. (**B**) Depicted here by colored traces are individual single-molecule sequences that span several thousand bases, including a copy of the repeat region, overlapping the flanking contigs to unambiguously resolve the contig order.

such that the accuracy of the base calls is optimized are crosstalk, dephasing and chastity filtering (16). On the other hand, SMS SBS technologies involve the real-time monitoring of a single molecule of polymerase interrogating single molecules of DNA at a time. In SMRT sequencing, for example, nucleotides are labeled with four different fluorescent dyes randomly diffusing throughout the sequencing chamber and illuminated by a laser only when the polymerase binds the nucleotide to incorporate it into the sequence being synthesized. A camera monitors illumination events at a rate of 100 frames per second over the course of a sequencing run (typically 15 min), thereby producing a movie comprising 90 000 frames (for a 15 min run). Primary analysis in this instance involves quantifying the intensities for each channel for each sequencing reaction, identifying the illumination events as pulses and then translating the pulses into base calls. While crosstalk between the different dyes is still an issue with this type of sequencing, the phasing correction and chastity filtering are no longer necessary, given the asynchronous nature of SMS. However, given the stochastic fluctuations that result from interrogating a single-template DNA molecule, a number of issues arise that lead to uncertainty around the number and identity of bases read for a given template. For example, a given incorporation event may be missed because the number of photons emitted from the dye attached to the newly incorporated nucleotide could not be distinguished from the background noise, or the polymerase may fail to incorporate a nucleotide and try multiple times before succeeding, creating what appear to be multiple consecutive incorporation events for the same base (14).

There are well-established mathematical frameworks for modeling trace data so that inferences around the interpretation of those traces can be made with respect to the underlying DNA sequence any given trace represents. Most generally, from a given sequencing trace, we detect pulses that represent contiguous time segments in which the intensity for a given channel goes from a background state to significantly above background over the period of time defined by the time segment. A pulse represents an illumination event during the sequencing process that ideally signifies the incorporation of a specific nucleotide into the sequence being synthesized. These types of events represent observations from the sequencing instrument that occur in time over a given sequencing run for a single molecule, which we can represent as the sequence of observations, $O = o_1, \ldots, o_n$. Each o_i can be a measurement or vector of measurements that characterize a given observation event. From these observations O, we want to derive an interpretation that represents the true sequence of nucleotides that were sequenced $T = t_1, \ldots, t_n$. In the first-generation sequencing and SGS contexts, each interpretation t_i of observation o_i represents either a single nucleotide base from the alphabet A, G, C, T or is empty. Because the components of T and O are random variables, the aim is to find the best sequence of interpretations given a sequence of observations. This has classically been represented as the probability of an interpretation t given an observation o (the conditional probability $p(t \mid o)$), where $v(o) = \arg\max_t p(t \mid o)$ then represents the interpretation that gives rise to the maximum probability. From this type of modeling, we can obtain a quality score for the best interpretation, $Q = -10 \log_{10} (1 - p(v(o)|o))$, which is directly analogous to a Phred score for the reliability of the best interpretation (55,56). The complete observation sequence can then be transformed into the sequence of best interpretations $V = v(o_1), \ldots, v(o_n)$ with quality scores $Q(o_1), \ldots, Q(o_n)$, respectively.

While this approach has worked well for FGS and SGS, there are two significant issues relating to the nature of most TGS technologies that necessitate a more advanced formulation of the type of mathematical model just described. First, there are significant stochastic components of SMS that complicate the relationship between observations and interpretations of those observations. For example, pulses observed in SMRT-sequencing trace data will not perfectly convey the sequence of incorporation events—the random and exponentially distributed pulse widths and inter-pulse durations mean that in some instances pulses or gaps between pulses can go undetected. As a result, a final output consisting of a single sequence representation of a given template will not fully reflect the likelihood any given base is correctly positioned, has been called correctly or has been missed or incorrectly inserted. By appropriately characterizing this

Downloaded from https://academic.oup.com/hmg/article/19/R2/R227/641295 by guest on 06 November 2022

uncertainty and incorporating it into the analysis, the ability to appropriately map a given sequence to the correct region of a genome, provide for alternative base calls at any given position or identify more general structural variants is improved. The second major issue relates to the amount of data that will ultimately be achievable with third-generation technologies. With the potential to generate hundreds of billions of reads per day per sequencing instrument, it is not only impractical to store the raw data files, but impractical for most users to store the trace and pulse-level data as well. Therefore, the raw data must be appropriately collapsed to reduce data storage requirements while simultaneously capturing all of the salient features of the pulses from the trace data to enable reliable interpretation of the pulse (observation) events.

Addressing these major issues will be essential to get the most from TGS technologies. It will be important to model the sequence data in the probabilistic sense, as discussed above, where in the context of TGS any given observation may correspond to one or more bases, although we assume that an interpretation will contain at most a few DNA bases (and typically only one). Because it is possible to have multiple different interpretations of the sequence of observations O corresponding to the same underlying read T, one approach would be to connect the multiple interpretations to localized observations via a graphical model for the template distribution, where all different partitions of T are considered: $P(T|O) = \sum_{X \in \text{possible partitions of } T} P(X|O)$. This form is also convenient because the data likelihoods, $P(O|X)$, can be separated from application-specific priors, $P(O)$ (specific to a given TGS technology) via the Bayes Theorem, $P(X|O) = P(O|X)P(X)/P(O)$. It will be critical to develop models from this type of framework or others to best characterize the uncertainty around the identity and number of bases synthesized off of a given template sequence, something that will be important for all downstream applications, including sequence alignment, variant detection and genome assembly.

Beyond the modeling challenges for TGS will come the data management and processing challenges, both demanding access to supercomputing scale resources to handle efficiently. The data from large projects such as 1000 Genomes will collectively approach the petabyte scale just for the raw information. The situation will soon be exacerbated by TGS technologies that will enable scans of entire genomes and microbiomes, transcriptomes and a direct assessment of epigenetic changes, in minutes and for very low cost. Layer on top of this, data from imaging technologies, other high-dimensional sensing technologies, and personal medical records, and the possibility exists to produce terabyte scales of data per individual, and well into the exabyte scales and beyond for populations of individuals. Mining such large high-dimensional datasets poses several challenges for storage and analysis. For biology to accurately model biological systems, advances are needed in data transfer, access control and management, standardization of data formats and integration of data from multiple different dimensions (57).

There are many technologies emerging in the computational space that will make it possible to address our supercomputing needs. Life scientists have begun to borrow solutions from fields such as high-energy particle physics and climatology, which have already passed through similar inflection points. Companies such as Microsoft, Amazon, Google and Facebook have also become masters of petabyte-scale datasets—linking pieces of data distributed over a massively parallel architecture in response to user requests and presenting to the user in a matter of seconds. Users of TGS technologies will need to follow in the footsteps of these others and carve out new paths where needed. For today, the data storage and computational solutions capable of meeting the demands of computing on TGS-scale datasets include cloud-based computing services now available from a number of vendors including Amazon and Microsoft, as well as access to custom high-performance compute clusters (57). In addition, a number of companies like Geospiza are offering services that leverage cloud-based compute resources to enable SGS and TGS users a path to manage and process their raw sequence data. We anticipate that these same resources will become even more critical to TGS users, compared with SGS users, given the scales and diversity of data TGS technologies will generate.

CONCLUSION/PERSPECTIVE

TGS has much to prove in demonstrating that all of the underlying sophisticated machinery upon which these emerging technologies are based can be translated into a true, realized advance over SGS. However, the promise of the dramatic advances the TGS revolution may bring is one of meeting the expectation we have of generating ever higher dimensional data so that we may evolve toward a more complete understanding of living systems and the complex phenotypes (like human disease) that emerge from such systems. The SGS technologies are already having a major impact on the DNA sequencing space, identifying rare variations in tumor tissues associated with different cancer types, for example (58,59). However, TGS promises to deliver entire genomes in less than a day and at reasonable cost (14), increasing the applicability of these technologies in almost every arena in the life and biomedical sciences. Many of the TGS platforms will also have a more general utility beyond DNA sequencing, including identification of patterns of methylation (60), comprehensive characterization of transcriptomes (61) and comprehensive characterization of translation (62). TGS, therefore, stands ready to provide unprecedented snapshots of complex systems that will enable a more accurate network view, which in turn will lead to models of disease that have a greater predictive power.

Ultimately, our ability to construct predictive disease models by integrating very large-scale, high-dimensional data generated by TGS and other technologies will demand that we master the large-scale information being collected on living systems in diverse application areas such as treatment of human diseases, development of alternative biofuels, enhancement of crop yield, ensuring food safety, forensics and beyond. However, without mastering the large-scale molecular data that underlie the broad array of phenotypes linked to each of these areas that TGS and other technologies will generate, without sophisticated mathematical algorithms capable of data integration, and without an appropriate informatics infrastructure to apply these algorithms and translate the results

Downloaded from https://academic.oup.com/hmg/article/19/R2/R227/641295 by guest on 06 November 2022

into manageable bites of information that can be consumed by basic science researchers, clinical researchers, physicians, patients and consumers, efforts to realize the impact TGS can have in areas like medicine, crop and livestock science, and alternative energy will not realize its full potential. Ultimately, through the use of advanced life sciences and informatics technologies, it should be possible for these different communities to become masters of information. Only by marrying information technology to the life sciences and biotechnology will we realize the astonishing potential of the vast amounts of biological data we will be capable of generating with TGS coming on line. Large-scale DNA sequencing, RNA sequencing, translation and related molecular phenotype data, if properly integrated and analyzed, will enable strategies in areas like personalized medicine that would lead to our making better choices that favorably impact human wellbeing.

ACKNOWLEDGEMENTS

We thank J. Korlach for insightful discussion of the manuscript.

Conflict of Interest statement. All authors are employees of Pacific Biosciences and own stock in the company.

FUNDING

Funding to pay the Open Access publication charges for this article was provided by Pacific Biosciences.

GLOSSARY

Amplification bias: Non-uniform amplification of DNA that leads to over-representation of some sequences and under-representation of others.

Assembly: Is a process in which bioinformatics software is used to align overlapping reads, which allows the original genome to be assembled into contiguous sequences. Assembly algorithms can be reference-based and consider a reference sequence as input, or can be *de novo* and blind to any data beyond the sequence reads. Reference-based assembly is an easier computational problem, but has the potential to introduce bias, particularly if a structurally divergent reference sequence is chosen. Longer read length and greater accuracy of each read facilitate both reference-based and *de novo* assembly of genomes.

Consensus reads: If TGS technologies sequence the same template molecule more than once, it is possible to construct a consensus read by aligning all the sequences from each template molecule to reduce stochastic errors in the single-molecule sequence. Consensus reads typically have a greater accuracy than raw reads.

Crosstalk: Overlap between signals for different nucleotides in a sequencing reaction. For instance, the emission spectrum of two fluorophores may overlap somewhat and decrease an instrument's ability to distinguish bases labeled with those fluorophores.

Dephasing: When an ensemble of molecules representing a singleinput sequence are sequenced using wash-and-scan techniques, the sequence reads gradually diverge in length if an extra base is added or a base fails to incorporate.

Direct RNA sequencing: RNA can be sequenced directly by replacing the DNA polymerase in a sequencing reaction with a reverse transcriptase or other RNA-dependent polymerase. Third-generation technologies that dispense with polymerases altogether also have the potential to sequence RNA directly, and in all cases, direct RNA sequencing offers potential decreases in time to result and much more accurate RNA sequencing as cDNA conversion steps are not required prior to sequencing.

Read: A read is the number of bases determined from a single segment of sample nucleic acid by a sequencing instrument.

Read length: The number of individual bases identified contiguously in a read defines its length. Read length can be defined by base-calling software from the instrument output alone, by alignment to a known reference sequence, or by alignment to a *de novo* assembly of sequence.

Sample preparation: All the steps taken to prepare a sample for sequencing after it is taken from a subject or the environment. Sample preparation procedures involve removal of materials other than the type of nucleic acid to be sequenced and purification of that nucleic acid. Depending on the technology, it may also require labeling the nucleic acid, attachment of adapters or amplification. All sample preparation steps have the potential to introduce bias, so technologies that minimize sample preparation have the potential to increase accuracy of and decrease time to a sequencing result.

Second-generation sequencing (SGS): Sequencing of an ensemble of DNA molecules with wash-and-scan techniques.

Sequencing by synthesis (SBS): Sequencing methods that determine the sequence of a DNA template by synthesizing the complementary DNA.

Second-molecule sequencing (SMS): Sequencing of a single DNA or RNA molecules.

Third-generation sequencing (TGS): Sequencing single DNA molecules without the need to halt between read steps (whether enzymatic or otherwise).

Wash-and-scan techniques: These use DNA polymerases just like any other reagent, washing them off after adding a base or an oligonucleotide during an SBS reaction. The many cycles of these approaches necessarily consume a lot of reagents and time.

REFERENCES

1. Drmanac, R., Sparks, A.B., Callow, M.J., Halpern, A.L., Burns, N.L., Kermani, B.G., Carnevali, P., Nazarenko, I., Nilsen, G.B., Yeung, G. *et al.* (2010) Human genome sequencing using unchained base reads on self-assembling DNA nanoarrays. *Science*, **327**, 78–81.
2. Flusberg, B.A., Webster, D.R., Lee, J.H., Travers, K.J., Olivares, E.C., Clark, T.A., Korlach, J. and Turner, S.W. (2010) Direct detection of DNA methylation during single-molecule, real-time sequencing. *Nat. Methods*, **7**, 461–465.
3. Kan, Z., Jaiswal, B.S., Stinson, J., Janakiraman, V., Bhatt, D., Stern, H.M., Yue, P., Haverty, P.M., Bourgon, R., Zheng, J. *et al.* (2010) Diverse somatic mutation patterns and pathway alterations in human cancers. *Nature*, **466**, 869–873.
4. Qin, J., Li, R., Raes, J., Arumugam, M., Burgdorf, K.S., Manichanh, C., Nielsen, T., Pons, N., Levenez, F., Yamada, T. *et al.* (2010) A human gut

Downloaded from https://academic.oup.com/hmg/article/19/R2/R227/641295 by guest on 06 November 2022

microbial gene catalogue established by metagenomic sequencing. *Nature*, **464**, 59–65.
5. Alexander, R.P., Fang, G., Rozowsky, J., Snyder, M. and Gerstein, M.B. (2010) Annotating non-coding regions of the genome. *Nat. Rev. Genet.*, **11**, 559–571.
6. van Bakel, H., Nislow, C., Blencowe, B.J. and Hughes, T.R. (2010) Most 'dark matter' transcripts are associated with known genes. *PLoS Biol.*, **8**, e1000371.
7. Huang, S., Li, R., Zhang, Z., Li, L., Gu, X., Fan, W., Lucas, W.J., Wang, X., Xie, B., Ni, P. *et al.* (2009) The genome of the cucumber, *Cucumis sativus* L. *Nat. Genet.*, **41**, 1275–1281.
8. Wall, P.K., Leebens-Mack, J., Chanderbali, A.S., Barakat, A., Wolcott, E., Liang, H., Landherr, L., Tomsho, L.P., Hu, Y., Carlson, J.E. *et al.* (2009) Comparison of next generation sequencing technologies for transcriptome characterization. *BMC Genomics*, **10**, 347.
9. Wheeler, D.A., Srinivasan, M., Egholm, M., Shen, Y., Chen, L., McGuire, A., He, W., Chen, Y.J., Makhijani, V., Roth, G.T. *et al.* (2008) The complete genome of an individual by massively parallel DNA sequencing. *Nature*, **452**, 872–876.
10. Shendure, J. and Ji, H. (2008) Next-generation DNA sequencing. *Nat. Biotechnol.*, **26**, 1135–1145.
11. Metzker, M.L. (2010) Sequencing technologies—the next generation. *Nat. Rev. Genet.*, **11**, 31–46.
12. Derrington, I.M., Butler, T.Z., Collins, M.D., Manrao, E., Pavlenok, M., Niederweis, M. and Gundlach, J.H. (2010) Nanopore DNA sequencing with Msp.A. *Proc. Natl Acad. Sci. USA*, **107**, 16060–16065.
13. Krivanek, O.L., Chisholm, M.F., Nicolosi, V., Pennycook, T.J., Corbin, G.J., Dellby, N., Murfitt, M.F., Own, C.S., Szilagyi, Z.S., Oxley, M.P. *et al.* (2010) Atom-by-atom structural and chemical analysis by annular dark-field electron microscopy. *Nature*, **464**, 571–574.
14. Eid, J., Fehr, A., Gray, J., Luong, K., Lyle, J., Otto, G., Peluso, P., Rank, D., Baybayan, P., Bettman, B. *et al.* (2009) Real-time DNA sequencing from single polymerase molecules. *Science*, **323**, 133–138.
15. Luan, B., Peng, H., Polonsky, S., Rossnagel, S., Stolovitzky, G. and Martyna, G. (2010) Base-by-base ratcheting of single stranded DNA through a solid-state nanopore. *Phys. Rev. Lett.*, **104**, 8103.
16. Whiteford, N., Skelly, T., Curtis, C., Ritchie, M.E., Lohr, A., Zaranek, A.W., Abnizova, I. and Brown, C. (2009) Swift: primary data analysis for the Illumina Solexa sequencing platform. *Bioinformatics*, **25**, 2194–2199.
17. Schatz, M.C., Delcher, A.L. and Salzberg, S.L. (2010) Assembly of large genomes using second-generation sequencing. *Genome Res.*, **20**, 1165–1173.
18. Sanger, F. and Coulson, A.R. (1975) A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. *J. Mol. Biol.*, **94**, 441–448.
19. Sanger, F., Nicklen, S. and Coulson, A.R. (1977) DNA sequencing with chain-terminating inhibitors. *Proc. Natl Acad. Sci. USA*, **74**, 5463–5467.
20. Maxam, A.M. and Gilbert, W. (1977) A new method for sequencing DNA. *Proc. Natl Acad. Sci. USA*, **74**, 560–564.
21. Hert, D.G., Fredlake, C.P. and Barron, A.E. (2008) Advantages and limitations of next-generation sequencing technologies: a comparison of electrophoresis and non-electrophoresis methods. *Electrophoresis*, **29**, 4618–4626.
22. Schloss, J.A. (2008) How to get genomes at one ten-thousandth the cost. *Nat. Biotechnol.*, **26**, 1113–1115.
23. Venter, J.C., Adams, M.D., Myers, E.W., Li, P.W., Mural, R.J., Sutton, G.G., Smith, H.O., Yandell, M., Evans, C.A., Holt, R.A. *et al.* (2001) The sequence of the human genome. *Science*, **291**, 1304–1351.
24. Bowers, J., Mitchell, J., Beer, E., Buzby, P.R., Causey, M., Efcavitch, J.W., Jarosz, M., Krzymanska-Olejnik, E., Kung, L., Lipson, D. *et al.* (2009) Virtual terminator nucleotides for next-generation DNA sequencing. *Nat. Methods*, **6**, 593–595.
25. Harris, T.D., Buzby, P.R., Babcock, H., Beer, E., Bowers, J., Braslavsky, I., Causey, M., Colonell, J., Dimeo, J., Efcavitch, J.W. *et al.* (2008) Single-molecule DNA sequencing of a viral genome. *Science*, **320**, 106–109.
26. Lipson, D., Raz, T., Kieu, A., Jones, D.R., Giladi, E., Thayer, E., Thompson, J.F., Letovsky, S., Milos, P. and Causey, M. (2009) Quantification of the yeast transcriptome by single-molecule sequencing. *Nat. Biotechnol.*, **27**, 652–658.
27. Tessler, L.A., Reifenberger, J.G. and Mitra, R.D. (2009) Protein quantification in complex mixtures by solid phase single-molecule counting. *Anal. Chem.*, **81**, 7141–7148.
28. Pushkarev, D., Neff, N.F. and Quake, S.R. (2009) Single-molecule sequencing of an individual human genome. *Nat. Biotechnol.*, **27**, 847–852.
29. Ozsolak, F., Platt, A.R., Jones, D.R., Reifenberger, J.G., Sass, L.E., McInerney, P., Thompson, J.F., Bowers, J., Jarosz, M. and Milos, P.M. (2009) Direct RNA sequencing. *Nature*, **461**, 814–818.
30. Goren, A., Ozsolak, F., Shoresh, N., Ku, M., Adli, M., Hart, C., Gymrek, M., Zuk, O., Regev, A., Milos, P.M. *et al.* (2010) Chromatin profiling by directly sequencing small quantities of immunoprecipitated DNA. *Nat. Methods*, **7**, 47–49.
31. Levene, M.J., Korlach, J., Turner, S.W., Foquet, M., Craighead, H.G. and Webb, W.W. (2003) Zero-mode waveguides for single-molecule analysis at high concentrations. *Science*, **299**, 682–686.
32. Travers, K.J., Chin, C.S., Rank, D.R., Eid, J.S. and Turner, S.W. (2010) A flexible and efficient template format for circular consensus sequencing and SNP detection. *Nucleic Acids Res.*, **38**, e159.
33. Uemura, S., Aitken, C.E., Korlach, J., Flusberg, B.A., Turner, S.W. and Puglisi, J.D. (2010) Real-time tRNA transit on single translating ribosomes at codon resolution. *Nature*, **464**, 1012–1017.
34. Blow, N. (2008) DNA sequencing: generation next-next. *Nat. Methods*, **5**, 267–274.
35. Bayley, H. (2010) Nanotechnology: holes with an edge. *Nature*, **467**, 164–165.
36. Liu, H., He, J., Tang, J., Pang, P., Cao, D., Krstic, P., Joseph, S., Lindsay, S. and Nuckolls, C. (2010) Translocation of single-stranded DNA through single-walled carbon nanotubes. *Science*, **327**, 64–67.
37. Clarke, J., Wu, H.C., Jayasinghe, L., Patel, A., Reid, S. and Bayley, H. (2009) Continuous base identification for single-molecule nanopore DNA sequencing. *Nat. Nanotechnol.*, **4**, 265–270.
38. Howorka, S., Cheley, S. and Bayley, H. (2001) Sequence-specific detection of individual DNA strands using engineered nanopores. *Nat. Biotechnol.*, **19**, 636–639.
39. Stoddart, D., Heron, A.J., Mikhailova, E., Maglia, G. and Bayley, H. (2009) Single-nucleotide discrimination in immobilized DNA oligonucleotides with a biological nanopore. *Proc. Natl Acad. Sci USA*, **106**, 7702–7707.
40. McNally, B., Singer, A., Yu, Z., Sun, Y., Weng, Z. and Meller, A. (2000) Optical recognition of converted DNA nucleotides for single-molecule DNA sequencing using nanopore arrays. *Nano Lett.*, **10**, 2237–2244.
41. Polonsky, S., Rossnagel, S. and Stolovitzky, G. (2007) Nanopore in metal–dieletric sandwich for DNA position control. *Appl. Phys. Lett.*, **91**, 153103.
42. Krems, M., Zwolak, M., Pershin, Y.V. and Di Ventra, M. (2009) Effect of noise on DNA sequencing via transverse electronic transport. *Biophys. J.*, **97**, 1990–1996.
43. Branton, D., Deamer, D.W., Marziali, A., Bayley, H., Benner, S.A., Butler, T., Di Ventra, M., Garaj, S., Hibbs, A., Huang, X. *et al.* (2008) The potential and challenges of nanopore sequencing. *Nat. Biotechnol.*, **26**, 1146–1153.
44. Bentley, D.R., Balasubramanian, S., Swerdlow, H.P., Smith, G.P., Milton, J., Brown, C.G., Hall, K.P., Evers, D.J., Barnes, C.L., Bignell, H.R. *et al.* (2008) Accurate whole human genome sequencing using reversible terminator chemistry. *Nature*, **456**, 53–59.
45. Wang, J., Wang, W., Li, R., Li, Y., Tian, G., Goodman, L., Fan, W., Zhang, J., Li, J., Guo, Y. *et al.* (2008) The diploid genome sequence of an Asian individual. *Nature*, **456**, 60–65.
46. Hernandez, D., Francois, P., Farinelli, L., Osteras, M. and Schrenzel, J. (2008) *De novo* bacterial genome sequencing: millions of very short reads assembled on a desktop computer. *Genome Res.*, **18**, 802–809.
47. Jeck, W.R., Reinhardt, J.A., Baltrus, D.A., Hickenbotham, M.T., Magrini, V., Mardis, E.R., Dangl, J.L. and Jones, C.D. (2007) Extending assembly of short DNA sequences to handle error. *Bioinformatics*, **23**, 2942–2944.
48. Dohm, J.C., Lottaz, C., Borodina, T. and Himmelbauer, H. (2007) SHARCGS, a fast and highly accurate short-read assembly algorithm for *de novo* genomic sequencing. *Genome Res.*, **17**, 1697–1706.
49. Pevzner, P.A., Tang, H. and Waterman, M.S. (2001) An Eulerian path approach to DNA fragment assembly. *Proc. Natl Acad. Sci. USA*, **98**, 9748–9753.
50. Zerbino, D.R. and Birney, E. (2008) Velvet: algorithms for *de novo* short read assembly using de Bruijn graphs. *Genome Res.*, **18**, 821–829.
51. Chaisson, M.J. and Pevzner, P.A. (2008) Short read fragment assembly of bacterial genomes. *Genome Res.*, **18**, 324–330.

Downloaded from https://academic.oup.com/hmg/article/19/R2/R227/641295 by guest on 06 November 2022

52. Butler, J., MacCallum, I., Kleber, M., Shlyakhter, I.A., Belmonte, M.K., Lander, E.S., Nusbaum, C. and Jaffe, D.B. (2008) ALLPATHS: *de novo* assembly of whole-genome shotgun microreads. *Genome Res.*, **18**, 810–820.
53. Li, R., Zhu, H., Ruan, J., Qian, W., Fang, X., Shi, Z., Li, Y., Li, S., Shan, G., Kristiansen, K. *et al.* (2010) *De novo* assembly of human genomes with massively parallel short read sequencing. *Genome Res.*, **20**, 265–272.
54. Simpson, J.T., Wong, K., Jackman, S.D., Schein, J.E., Jones, S.J. and Birol, I. (2009) ABySS: a parallel assembler for short read sequence data. *Genome Res.*, **19**, 1117–1123.
55. Ewing, B. and Green, P. (1998) Base-calling of automated sequencer traces using phred. II. Error probabilities. *Genome Res.*, **8**, 186–194.
56. Ewing, B., Hillier, L., Wendl, M.C. and Green, P. (1998) Base-calling of automated sequencer traces using phred. I. Accuracy assessment. *Genome Res.*, **8**, 175–185.
57. Schadt, E.E., Linderman, M.D., Sorenson, J., Lee, L. and Nolan, G.P. (2009) Computational solutions to large-scale data management and analysis. *Nat. Rev. Genet.*, **11**, 647–657.
58. Jones, S., Zhang, X., Parsons, D.W., Lin, J.C., Leary, R.J., Angenendt, P., Mankoo, P., Carter, H., Kamiyama, H., Jimeno, A. *et al.* (2008) Core signaling pathways in human pancreatic cancers revealed by global genomic analyses. *Science*, **321**, 1801–1806.
59. Parsons, D.W., Jones, S., Zhang, X., Lin, J.C., Leary, R.J., Angenendt, P., Mankoo, P., Carter, H., Siu, I.M., Gallia, G.L. *et al.* (2008) An integrated genomic analysis of human glioblastoma multiforme. *Science*, **321**, 1807–1812.
60. Cokus, S.J., Feng, S., Zhang, X., Chen, Z., Merriman, B., Haudenschild, C.D., Pradhan, S., Nelson, S.F., Pellegrini, M. and Jacobsen, S.E. (2008) Shotgun bisulphite sequencing of the Arabidopsis genome reveals DNA methylation patterning. *Nature*, **452**, 215–219.
61. Wang, Z., Gerstein, M. and Snyder, M. (2009) RNA-Seq: a revolutionary tool for transcriptomics. *Nat. Rev. Genet*, **10**, 57–63.
62. Ingolia, N.T., Ghaemmaghami, S., Newman, J.R. and Weissman, J.S. (2009) Genome-wide analysis *in vivo* of translation with nucleotide resolution using ribosome profiling. *Science*, **324**, 218–223.

Downloaded from https://academic.oup.com/hmg/article/19/R2/R227/641295 by guest on 06 November 2022