Thompson and Milos *Genome Biology* 2011, **12**:217
http://genomebiology.com/2011/12/2/217

Genome Biology

REVIEW

The properties and applications of single-molecule DNA sequencing

John F Thompson* and Patrice M Milos

Abstract

Single-molecule sequencing enables DNA or RNA to be sequenced directly from biological samples, making it well-suited for diagnostic and clinical applications. Here we review the properties and applications of this rapidly evolving and promising technology.

Classical DNA sequencing (sometimes referred to as first generation sequencing) was developed in the late 1970s and evolved from a low-throughput, almost 'artisan' approach, in which the same radiolabeled DNA sample was run on a gel with one lane for each nucleotide [1,2], to an automated method in which all four fluorescently labeled dye terminators for a single sample [3] were loaded onto individual capillaries. These capillary-based instruments, introduced in 1998, could handle hundreds of individual samples per week, in a manner sufficiently powerful that the first draft sequence of a human genome was finished in 2001 using this technology. In the intervening years, incremental improvements have been made in dye chemistry, DNA polymerases, and electrophoresis conditions, pushing read lengths up to 1,000 bp; however, the underlying technology has remained the same, sequencing individual clones or samples.

After more than 25 years of steady improvements in first generation sequencing technology, the next generation of sequencing technology (now called second generation; see Box 1 for discussion of third generation sequencing terminology) emerged in 2005 with an immediate 100-fold increase in sequencing throughput using the 454 pyrosequencing approach [4]. This advance was followed by introductions of other technologies (such as Solexa/Illumina and ABI SOLiD) that varied in their technological details but increased sequencing throughput and reduced costs by additional orders of magnitude (reviewed in [5-7]). These second generation technologies drastically increased throughput because the sequencing target had changed from single clones or samples to many independent DNA fragments, enabling large sets of DNAs to be sequenced in parallel. Until recently, all second generation technologies achieved massively parallel sequencing by imaging light emission from the sequenced DNA, although the new sequencing system from Ion Torrent will probably be the first commercial system to change that paradigm by detecting hydrogen ions instead of light [8]. However, the key advance in all second generation technologies has been the avoidance of the bottleneck that resulted from the individual preparation of DNA templates that first generation approaches required. When coupled with powerful new bioinformatic tools and computational capabilities optimized for these new technologies, a prodigious increase in data output has resulted. This is highlighted in Figure 1, where the accumulation of sequence in classical GenBank from its inception in 1982 is compared with data in the Sequence Read Archive (originally known as the Short Read Archive, both abbreviated SRA). Less than a year after its initiation, the SRA had already surpassed classical GenBank and it now accounts for over 95% of all new sequence deposits. Furthermore, this is likely to be an under-representation of the level of new sequencing results because of the challenges of incorporating the new data types and difficulties in transferring the large volume of data.

Although the second generation technologies were initially inferior to classical sequencing in terms of read length (about 35 nucleotides (nt) for Illumina versus about 700 nt for classical sequencing) and single-read error rate (about 2% versus less than 0.1%), these shortcomings could be overcome by the sheer volume of data. Furthermore, continuous improvements in sequencing chemistry have narrowed the gap with respect to read length and errors, as exemplified by Roche 454 now routinely achieving read lengths of 400 nt at >99% accuracy [9] and Illumina moving from an initial read length of 36 nt to the current 76 nt or more and raw error rates well below 1%. These technologies have allowed DNA sequencing to move beyond a method for

*Correspondence: jthompson@helicosbio.com
Helicos BioSciences Corporation, Building 200LL, One Kendall Square, Cambridge, MA 02139, USA

Box 1

The logical term for the next round of sequencing technology advances would be 'third generation sequencing' and this has frequently been used to describe single-molecule sequencing. However, third generation sequencing has also been defined by some as real-time sequencing or solid-state sequencing, so the term has achieved Alice in Wonderland status of meaning whatever its user wants it to mean, and it will therefore not be used here. Instead, the more precise term of single-molecule sequencing will be used and only for those technologies that actually generate a sequencing signal from a single nucleic acid molecule. The definition of single-molecule sequencing has been stretched by some to include systems that start with a single molecule but then make multiple copies of the DNA before sequencing or detection [67]. However, the properties of any sequencing system could be stretched to assert that the sequencing process actually started with a single molecule, even though the unique advantages of single-molecule sequencing would be lost. On the other hand, there are also technologies that do not strive to generate sequence information from every nucleotide but only from a subset of positions. When such partial sequences are generated from a single molecule, this single-molecule mapping data can be combined with sequence data for a complete genomic view. Indeed, no current technology can provide individual read lengths sufficient for whole genome coverage of even the smallest organisms, so methods for combining partial reads into a complete coverage map are an important component of the overall sequencing process for whole genomes.

accumulating genomic information to another level at which sequencing has become the digital measuring stick for a host of important biological processes, including gene expression, splicing, characterizing complex mixed populations of organisms, detecting protein binding, and defining genome methylation sites [10].

Single-molecule sequencing provides solutions to some of the most vexing problems that face second generation sequencing by simplifying sample preparation, reducing sample mass requirements, and eliminating amplification of DNA templates. Every sample manipulation and especially amplification can cause quantitative and qualitative artifacts [11]; these have especially detrimental impacts on quantitative applications, such as chromatin immunoprecipitation sequencing (ChIP-Seq) and RNA/cDNA sequencing. Amplification also places limitations on the size of the DNA being sequenced because molecules that are too short or too long will not be amplified well. The simplified sample preparation and higher consistency caused by eliminating amplification makes single-molecule sequencing well suited for diagnostic and clinical applications [12]. Thus, the need for continuing advances in sequencing technology is apparent.

Because the properties of different single-molecule sequencing technologies vary so much from each other and from other generations of sequencing technologies, it is important to understand those properties and their impact on experimental design and output. Some advantages of single-molecule sequencing may be universal, such as the ability to resequence the same molecule multiple times for improved accuracy and the ability to sequence molecules that cannot be readily amplified because of extremes of GC content, secondary structure, or other reasons. The ability to make use of some of these advantages, such as long read length and quantitative superiority, depends on the details of the technology, as not all single-molecule technologies have the same performance characteristics - in terms of long read length or the throughput in terms of read count - to be appropriate for all applications. Other reviews have examined various aspects of different single-molecule technologies [5-7,13-17]. Here, we focus on the unique attributes of each technology (Figure 2) and how each might be best used to answer questions of biological interest that are not well addressed by current first and second generation sequencing.

Single-molecule sequencing technologies

When considering the properties of single-molecule sequencing technologies, the focus is most frequently on read length, error rate, and throughput (Figure 3); however, input sample quantity and quality requirements, simplicity and parallelizability of sample preparation, and data analysis are also important components that must be factored in when considering whether a technology, single-molecule or otherwise, is appropriate for a given problem. Some of the applications frequently undertaken with current sequencing technologies and the relative importance of various properties of different sequencing methods are shown in Table 1. Important properties of single-molecule technologies that relate to these various applications are discussed below.

Sequencing by synthesis

The first commercially available single-molecule sequencing system was developed by our colleagues at Helicos BioSciences [18]. In this system, individual molecules are hybridized to a flow cell surface containing covalently attached oligonucleotides. Fluorescently labeled nucleotides and a DNA polymerase are added sequentially and incorporation events detected by laser excitation and recording with a charge coupled device (CCD) camera. The fluorescent 'Virtual Terminator' nucleotide prevents the incorporation of any subsequent nucleotide until the nucleotide dye moiety is cleaved [19]. The images from each cycle are assembled to generate an overall set of sequence reads. On a standard run, 120 cycles of nucleotide addition and detection are carried out. Well over a billion molecules can be followed simultaneously in this

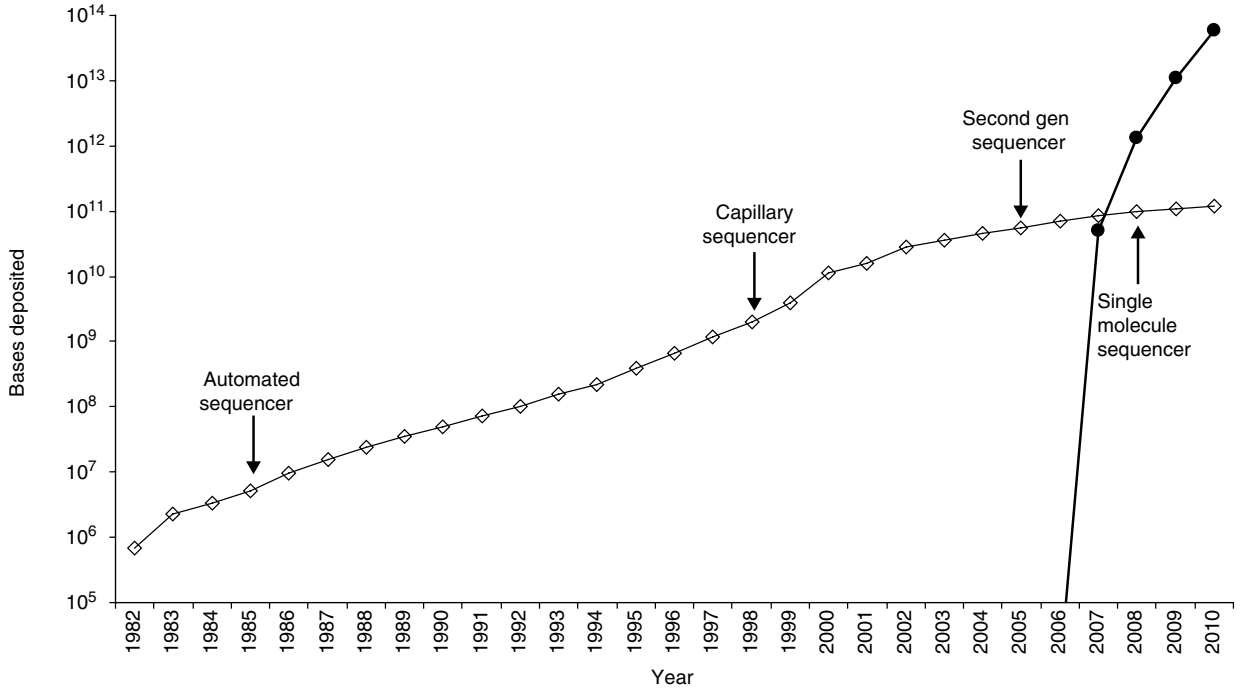

Figure 1. Sequence database submissions from 1982 to 2010. Nucleotides submitted to the classical version of GenBank (diamonds, thin line) and to the Sequence Read Archive (circles, thick line) are shown as a function of time. Data for GenBank up to 2008 were obtained from the NCBI website [68] and subsequent years were obtained from GenBank publications [69,70]. Data for SRA was obtained from publications for 2008 to 2010 [71-73] and estimated for 2007 on the basis of 44 projects being in the database at the end of the year [74] and using February 2008 data from NCBI [75] to estimate the approximate number of bases likely to have been submitted from that spectrum of projects. Key advances in sequencing technology are shown with arrows. The development of second generation sequencing technologies and single-molecule sequencing has had a dramatic increase in the number of sequences deposited in public databases. Less than a year after its initiation, the SRA had already surpassed classical GenBank and it now accounts for over 95% of all new sequence deposits.

approach. Because there are two 25-channel flow cells in a standard run, 50 different samples can be sequenced simultaneously, with the additional possibility of significantly greater throughput of samples through multiplexing. Sample requirements are the simplest of all technologies: sub-nanogram amounts are necessary and very poor quality DNA, including degraded or modified DNA, can be sequenced [20,21]. Average read lengths are relatively short (about 35 nt) with raw individual nucleotide error rates currently about 3 to 5%, occurring randomly throughout the sequence reads and predominantly in the form of a 'dark base' or deletion error, which is accounted for in the alignment algorithm [22]. This error rate is not an issue when detecting polymorphisms because 30x coverage is typically used for diploid genomes with second generation systems to overcome the uneven coverage induced by amplification. Oversampling is needed to overcome the stochastic nature of heterozygote detection, with 30x coverage advisable to ensure that nearly all heterozygotes are called correctly. At this coverage level, accurate consensus sequences are generated regardless of error rates within this range. Single-molecule systems have a much more even coverage and thus do not require as much depth for complete detection of heterozygotes. The even coverage relative to second generation systems was shown with ChIP experiments, in which sequence reads were relatively constant with respect to GC content with single-molecule sequencing, whereas significant deviations were observed at both high and low GC content with amplification-based sequencing [23] and with whole-genome sequencing of a human sample [24].

The Helicos Sequencer system can also sequence RNA molecules directly, thus avoiding the many artifacts associated with reverse transcriptase and providing unparalleled quantitative accuracy for RNA expression measurements [25]. The very high read count per sample allows precise expression measurements to be made with either RNA or cDNA [26-29], a feature not yet possible with other single-molecule technologies. Indeed, whole classes of RNA molecules that cannot be visualized using other technologies can be detected using a single-molecule approach [30,31]. As with many single-molecule systems, repeated reads of the same molecule can

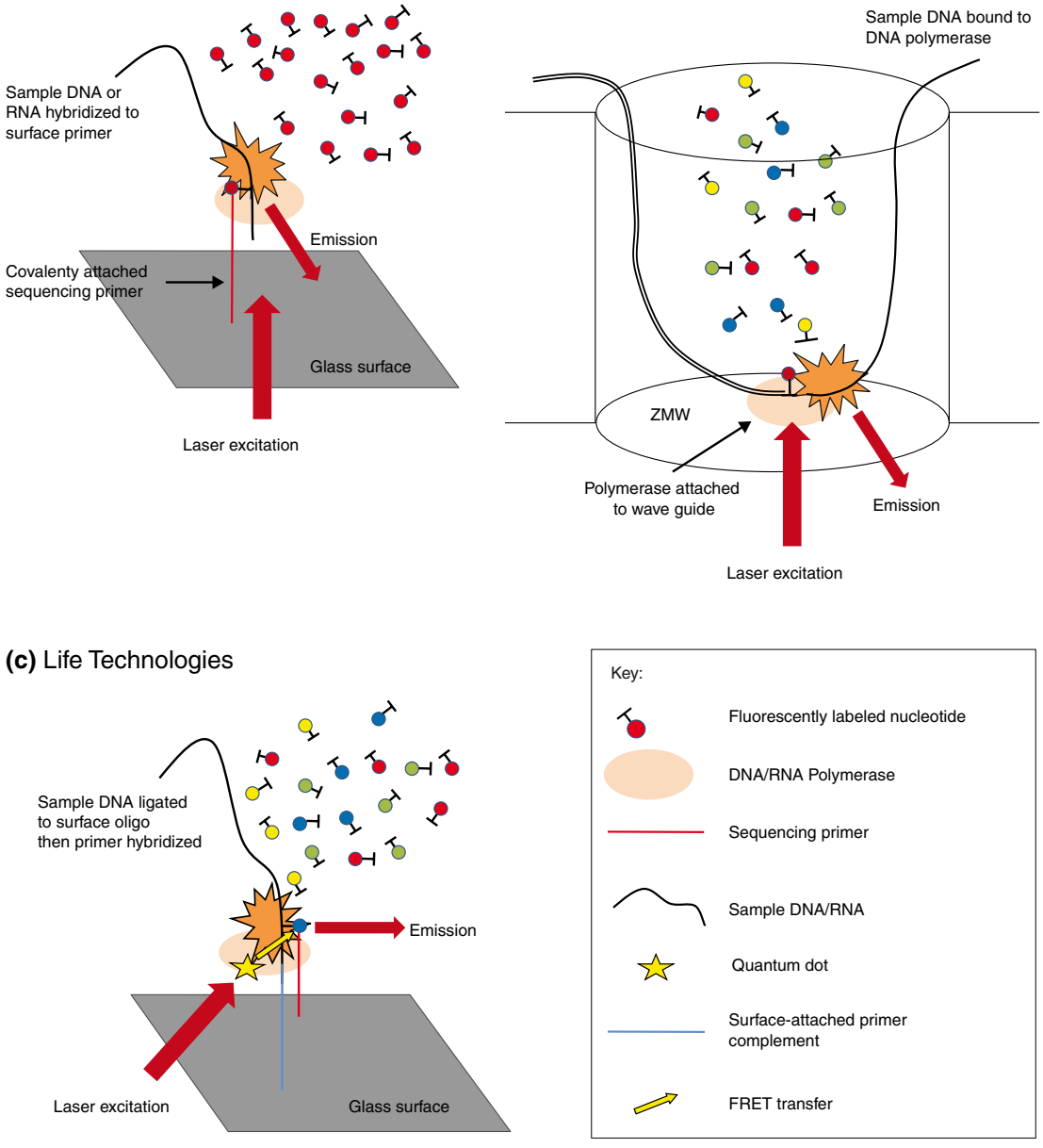

Figure 2. Overview of single-molecule sequencers. The three most advanced single-molecule sequencing systems all carry out sequencing-by-synthesis using laser excitation to generate a fluorescent signal from labeled nucleotides, which is then detected using a camera. **(a)** In the Helicos BioSciences system [18], single nucleotides, each with a fluorescent dye attached to the base, are sequentially added. **(b,c)** In the Pacific Biosciences [35] and Life Technologies [41] systems, four different nucleotides, each with a different color dye attached to the phosphates, are continuously added. Background fluorescence is minimized differently in the three systems. (a) Helicos uses total internal reflectance fluorescence (TIRF) to create a narrow evanescent field of light in which the intensity of the light decays exponentially away from the glass surface. Only dyes that are in the TIRF evanescent field can fluoresce. (b) Pacific Biosciences uses a zero mode waveguide (ZMW), which limits illumination to a narrow region near the bottom of the well containing the polymerase. Only dyes near the opening of the ZMW can fluoresce. (c) Life Technologies uses fluorescence resonance energy transfer (FRET) between the initially absorbing quantum dot on the polymerase and the emitting dye on the nucleotide. Only dyes close to the polymerase-attached quantum dot can be excited by FRET and then fluoresce. For the three systems, DNA is immobilized for viewing over time by a surface-attached sequencing primer (Helicos (a)), by interaction with a surface-bound polymerase (Pacific Biosciences (b)), or by ligating to a surface-attached oligonucleotide (Life Technologies (c)). For Helicos (a), the polymerase is replaced after every cycle of nucleotide addition. For Life Technologies (c), the polymerase can be replaced on a given DNA molecule after each read is completed. For Pacific Biosciences (b), the polymerase cannot be replaced.

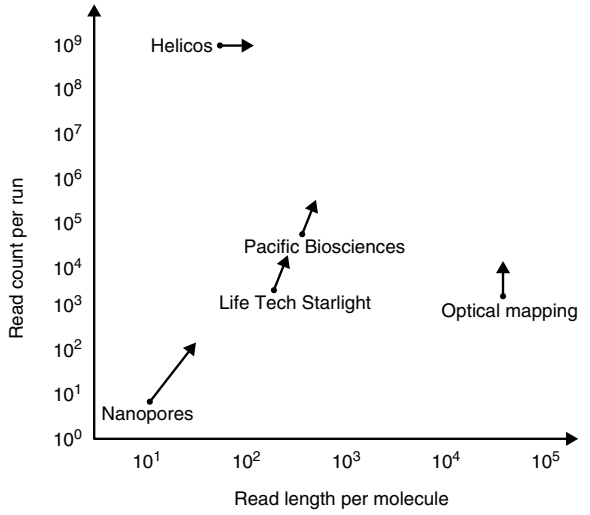

Figure 3. The attributes of single-molecule sequencing technology. The current read counts and read lengths for single-molecule sequencing technologies are shown by the dots. Each technology is striving for improvements in their key attributes with the research aimed in the directions shown by the arrow.

markedly improve the error rate and also allow detection of very rare variants in a mixed sample. For example, a rare variant in a sample containing a mixture of few tumor cells among many normal cells might not be detectable with amplified DNA. With repeat sequencing of the same molecule, the error rate can be driven sufficiently low that mutations in heterogeneous samples such as tumors can be readily detected. Because of the minimal sample preparation needs, the ability to use exceptionally small starting quantities, and the high read count, this technology is ideal for quantitative applications such as ChIP, RNA expression, and copy number variation, and situations in which sample quantity is limiting or degraded [20,23]. Standard, whole human genome resequencing is readily accomplished [24], but it is currently less expensive on second generation systems.

Pacific Biosciences has developed another sequencing-by-synthesis approach using fluorescently labeled nucleotides. In this system, DNA is constrained to a very small volume in a zero-mode wave guide [32] and the presence of a fluorescently labeled cognate nucleotide near the DNA polymerase is measured. The dimensions of the wave guide are so small that light can penetrate only the region very close to the edge, where the polymerase used for sequencing is constrained. Only nucleotides in that small volume near the polymerase can be illuminated and fluoresce for detection. Because the nucleotide that is being incorporated in the extending DNA strand spends a longer time near the polymerase, it can, to a large extent, be distinguished from non-cognate nucleotides. All four potential nucleotides are included in the reaction, each labeled with a different color fluorescent dye so that they can be distinguished from each other. Each nucleotide has a characteristic incorporation time that can further aid in improving base calls. Sequence reads of up to thousands of bases, longer than possible with second generation systems, are obtained in real time for each individual molecule [33-36]. However, the current throughput is less than 100,000 reads per run, so the overall sequence yield is much lower than second generation systems and the Helicos system. In addition, the raw error rate, currently 15 to 20% [37,38], is significantly higher than with any other current sequencing technology, creating challenges in using the data for some applications, such as variant detection.

Much longer reads, referred to as 'strobe reads' [39], can be generated by turning off the laser for periods of time during sequencing, which prevents premature termination caused by laser-induced photodamage to the polymerase and nucleotides. If long reads are not necessary, the high raw error rate can be overcome by ligating a hairpin oligonucleotide to each end of the DNA, creating a circular template (called SMRTbell for single molecule real time), and then repeatedly sequencing the same molecule [37]. This procedure works when the molecules are relatively short but it cannot be used with long reads, so those retain the high raw error rate. Even with a high error rate, the very long reads can be productively used for joining sequence contigs. An additional benefit for this system is the ability to potentially detect modified bases. It is possible to detect 5-methylcytosine [40], although the role of sequence context and other factors in affecting the accuracy of such assignments remains to be clarified. In principle, direct RNA sequencing should also be possible with this system, but this has not been reported yet for natural RNA molecules because nucleotides bind repeatedly to the reverse transcriptase before nucleotide incorporation, thereby giving false signals with multiple insertions that prevent determination of a meaningful sequence. In addition, the low read count of this system will limit it to the identification of common mRNA isoforms rather than quantitative expression profiling or complete transcriptome coverage, both of which require a much higher read count than possible in the foreseeable future. In general, the long reads and short turnaround time make this system most useful for helping to assemble genomes, assessing the analysis of structural variation, haplotyping, metagenomics, and identification of splicing isoforms.

Life Technologies, a major provider of both first and second generation sequencing systems, is developing the fluorescence resonance energy transfer (FRET)-based single-molecule sequencing-by-synthesis technology initially introduced by Visigen [41]. Substantial advances have been made, with commercial release of the

Table 1. Which sequencing technology to use and when?[a]

	Read length	Read count	Sequence throughput[b]	Quantitative accuracy	Single pass error rate	Multiple pass error rate	Consensus error rate	Sample manipulations or perturbations	Sample preparation costs	Informatics costs	Optimal single-molecule technology
Genomics											
Variant detection			High				High				Helicos
Rare variant detection			High		Moderate	High					Helicos
Whole genome assembly	High		High							High	Mix
Metagenomics	High		High		Moderate					High	PacBio/Starlight
Degraded samples								High			Helicos
Copy number variation		High		High							Helicos
Large structural variations	High										Optical mapping
Transcriptomics											
Gene expression		High		High	Moderate			Moderate	High		Helicos
Splicing patterns	High	Moderate		Moderate							PacBio/Starlight
Small RNA quantification		High		High	Moderate			High	High		Helicos
Novel RNA discovery					Moderate		High	High			Helicos

[a]The characteristic features of sequencing technologies are shown, along with a qualitative assessment of how each of those features affect the ease with which an application can be carried out. For example, 'High' indicates that the application requires a high level of the particular feature. This is a general evaluation and particular experiments may vary with respect to the impact of each attribute. The choice of which method to use for a given application depends on the properties of that technology. [b]Sequence throughput is defined as read length multiplied by read count.

'Starlight' system expected in the near future. The current technology consists of a quantum-dot-labeled polymerase that synthesizes DNA using four distinctly labeled nucleotides in a real-time system [42]. Quantum dots, which are fluorescent semiconducting nanoparticles, have an advantage over fluorescent dyes in that they are much brighter and less susceptible to bleaching, although they are also much larger and more susceptible to blinking. The genomic sample to be sequenced is ligated to a surface-attached oligonucleotide of defined sequence and then read by extension of a primer complementary to the surface oligonucleotide. When a fluorescently labeled nucleotide binds to the polymerase, it interacts with the quantum dot, causing an alteration in the fluorescence of both the nucleotide and the quantum dot. The quantum dot signal drops, whereas a signal from the dye-labeled phosphate on each nucleotide rises at a characteristic wavelength. The real-time sequence is captured for each extending primer. Because each sequence is bound to the surface, it can be reprimed and sequenced again for improved accuracy. It is not clear what the sequence specifications will be but its similarity to the Pacific Biosciences technology make that a likely reference point. If so, it will have the same strengths in terms of applications (genome assembly, structural variation, haplotyping, metagenomics) whereas potentially being challenged with quantitative applications requiring a high read count (such as ChIP or RNA expression).

Optical sequencing and mapping

There are other technologies that enable very long reads to be produced but at the cost of significantly lower throughput. For example, it is possible to adhere very long DNA molecules, up to hundreds of kilobases long, to surfaces and interrogate them for particular sequences by cutting them with various restriction enzymes or labeling them after treatment with sequence-specific nicking enzymes. The lengths of the examined molecules are dependent on the ability to handle such long DNA without mechanically shearing it. Complete restriction digests that allow ordering of sequence contigs have been generated for human and other genomes from collections of single molecules spanning entire genomes [43]. Highly repetitive and duplicated genomes, such as maize, are particularly difficult to assemble with traditional sequencing but have been successfully analyzed with this single-molecule system [44]. The restriction sites provide sequence landmarks on the DNA and thus long repeat regions and other intricate structural variations can be assigned in an unambiguous manner. Specialized applications such as genome-wide methylation mapping can also be undertaken [45].

Similarly, DNA molecules can be constrained to nanotubes and specifically labeled for viewing [46]. Single molecules of RNA have been visualized using scanning tip Raman spectroscopy [47]. In an alternative method also using adsorption of long DNA molecules to a surface, guanines could be distinguished from all other bases and the partial sequence read with a scanning electron microscope [48]. Possibilities for reading other bases through insertion of heavy atoms such as bromine or iodine on particular nucleotides have been suggested by ZS Genetics [49]. Although the low strand throughput and incomplete sequence reading are currently limiting, there is potential for reads that are hundreds of kilobases long, again limited primarily by the ability to handle the DNA without shearing it. Other technologies using direct reading of stretched DNA have been reviewed elsewhere [7]. These optical sequencing technologies provide a powerful view of genome structure, but they cannot provide the detailed sequence data or access to many other sequencing applications that require high read counts, such as gene expression measurements.

Nanopores

All of the sequencing techniques described so far require some kind of label on the DNA or nucleotide substrates to detect the individual base for sequencing. However, nanopore approaches generally do not require an exogenous label but rely instead on the electronic or chemical structure of the different nucleotides for discrimination. The advantages and potential means of using nanopores have been reviewed [14,50]. Nanopores of greatest interest thus far include those assembled with solid-state systems constructed of materials such as carbon nanotubes or thin films [51-54] and the biologically based α-hemolysin [55-59] or MspA [60,61]. These bacterial pore proteins have been extensively studied and engineered to optimize the detection of specific bases and the translocation rate of DNA through the pore. Although sequencing native DNA based on its natural properties would eliminate the labeling step and potentially allow very long reads with minimal sample preparation, thus reducing costs, the differences among nucleotides are very modest and their detection is compounded by difficulties in controlling the pace and directionality of the DNA through the nanopore. Specific detection and unidirectional flow are required for high accuracy sequencing.

A variety of methods have been used to slow the pace of DNA through nanopores, including attachment of polystyrene beads [53], salt concentrations [62], viscosity [63], magnetic fields [64], and the introduction of regions of double-stranded DNA on a single-stranded target [54,58]. At the high translocation speeds typically found (potentially millions of bases per second), detecting a signal over background noise from each nucleotide can be a challenge, and this has been

overcome in some cases by reading groups of nucleotides (such as by using hybridization of known sequences as is being developed by NabSys [53]) or encoding the original sequence in a more complex manner by converting the nucleotide sequence using a binary code of molecular beacons (as is being developed by NobleGen [65]). Maintaining a unidirectional flow of DNA has been enhanced by coupling an exonuclease to the process and reading the cleaved nucleotides (as developed by Oxford Nanopore [66]).

Although nanopore sequencing technologies continue to advance, simply showing the ability to sequence DNA, something not yet demonstrated by nanopores with natural DNA, is not sufficient. There needs to be a path to lower costs, longer reads, or higher accuracy relative to other technologies that will provide nanopores with a unique advantage relative to other methods. Even if reagent costs can be significantly reduced, sample preparation and informatic costs remain and these may become the dominant costs of sequencing and will vary depending on the technology being used. The ever-rising hurdles created by extant technology will not be easy to overcome. With the variety of second generation and single-molecule technologies already commercialized and others on the horizon, there will need to be substantial advances on many fronts to make these technologies commercially viable.

Conclusions

The development of single-molecule sequencing in which individual DNA or RNA molecules, derived directly from biological samples, are sequenced not only in a massively parallel manner but also without any type of amplification before or during the sequencing reaction promises yet another inflection point in terms of technology. It offers the potential for lower costs, higher throughput, improved quantitative accuracy, increased read lengths, and the ability to directly sequence RNA and detect methylation and other nucleotide modifications. Just as second generation sequencing did not completely displace first generation sequencing, single-molecule sequencing will not immediately displace earlier technologies. First generation, second generation, and single-molecule sequencing will each be used for the biological problems to which they are most suited, with that range of problems changing over time as each technology is improved.

Although no one mode of single-molecule sequencing can yet provide all the advantages potentially attainable, rapid progress is being made to achieve these goals. Technologies such as the Helicos system using fluorescent detection are now available commercially, and others such as Pacific Biosciences and Life Technologies Starlight will be available soon. New methods that rely on the natural chemical properties of native DNA are still in their infancy but raise the hope that sequencing might at some point become even simpler and cheaper by avoiding the need for labeling DNA and the use of enzymatic activities. Because experimenters have such widely varying requirements for the problems that need to be addressed, each should consider whether the best and most complete answer can be generated with older, sequencing-by-committee approaches or whether a true understanding of their biological questions requires the exquisite quantitative accuracy and in-depth sequencing power of a single-molecule approach. Single-molecule sequencing will continue to advance and offer researchers a variety of options, especially in the diagnostic and clinical fields.

Acknowledgements

This publication was made possible by grants R01 HG004144 and RC2 HG005598 from the National Human Genetics Research Institute (NHGRI) at the National Institutes of Health. Its contents are solely the responsibility of the authors and do not necessarily represent the official views of the NHGRI.

Published: 24 February 2011

References

1. Maxam AM, Gilbert W: **A new method for sequencing DNA.** *Proc Natl Acad Sci U S A* 1977, **74:**560-564.
2. Sanger F, Nicklen S, Coulson AR: **DNA sequencing with chain-terminating inhibitors.** *Proc Natl Acad Sci U S A* 1977, **74:**5463-5467.
3. Prober JM, Trainor GL, Dam RJ, Hobbs FW, Robertson CW, Zagursky RJ, Cocuzza AJ, Jensen MA, Baumeister K: **A system for rapid DNA sequencing with fluorescent chain-terminating dideoxynucleotides.** *Science* 1987, **238:**336-341.
4. Margulies M, Egholm M, Altman WE, Attiya S, Bader JS, Bemben LA, Berka J, Braverman MS, Chen YJ, Chen Z, Dewell SB, Du L, Fierro JM, Gomes XV, Godwin BC, He W, Helgesen S, Ho CH, Irzyk GP, Jando SC, Alenquer ML, Jarvie TP, Jirage KB, Kim JB, Knight JR, Lanza JR, Leamon JH, Lefkowitz SM, Lei M, Li J, *et al.*: **Genome sequencing in microfabricated high-density picolitre reactors.** *Nature* 2005, **437:**376-380.
5. Voelkerding KV, Dames SA, Durtschi JD: **Next-generation sequencing: from basic research to diagnostics.** *Clin Chem* 2009, **55:**641-658.
6. Metzker ML: **Sequencing technologies - the next generation.** *Nat Rev Genet* 2010, **11:**31-46.
7. Pettersson E, Lundeberg J, Ahmadian A: **Generations of sequencing technologies.** *Genomics* 2009, **93:**105-111.
8. Delseny M, Han B, Hsing YL: **High throughput DNA sequencing: the new sequencing revolution** *Plant Sci* 2010, **179:**407-422.
9. **454 Sequencing: Products & Solutions** [http://454.com/products-solutions/system-benefits.asp]
10. Kahvejian A, Quackenbush J, Thompson JF: **What would you do if you could sequence everything?** *Nat Biotechnol* 2008, **26:**1125-1133.
11. Dohm JC, Lottaz C, Borodina T, Himmelbauer H: **Substantial biases in ultra-short read data sets from high-throughput DNA sequencing.** *Nucleic Acids Res* 2008, **36:**e105.
12. Milos PM: **Emergence of single-molecule sequencing and potential for molecular diagnostic applications.** *Expert Rev Mol Diagn* 2009, **9:**659-666.
13. Gupta PK: **Single-molecule DNA sequencing technologies for future genomics research.** *Trends Biotechnol* 2008, **26:**602-611.
14. Branton D, Deamer DW, Marziali A, Bayley H, Benner SA, Butler T, Di Ventra M, Garaj S, Hibbs A, Huang X, Jovanovich SB, Krstic PS, Lindsay S, Ling XS, Mastrangelo CH, Meller A, Oliver JS, Pershin YV, Ramsey JM, Riehn R, Soni GV, Tabard-Cossa V, Wanunu M, Wiggin M, Schloss JA: **The potential and challenges of nanopore sequencing.** *Nat Biotechnol* 2008, **26:**1146-1153.
15. Xu M, Fujita D, Hanagata N: **Perspectives and challenges of emerging single-molecule DNA sequencing technologies.** *Small* 2009, **5:**2638-2649.
16. Efcavitch JW, Thompson JF: **Single-molecule DNA analysis.** *Annu Rev Anal Chem (Palo Alto Calif)* 2010, **3:**109-128.

17. Treffer R, Deckert V: **Recent advances in single-molecule sequencing.** *Curr Opin Biotechnol* 2010, **21:**4-11.
18. Harris TD, Buzby PR, Babcock H, Beer E, Bowers J, Braslavsky I, Causey M, Colonell J, Dimeo J, Efcavitch JW, Giladi E, Gill J, Healy J, Jarosz M, Lapen D, Moulton K, Quake SR, Steinmann K, Thayer E, Tyurina A, Ward R, Weiss H, Xie Z: **Single-molecule DNA sequencing of a viral genome.** *Science* 2008, **320:**106-109.
19. Bowers J, Mitchell J, Beer E, Buzby PR, Causey M, Efcavitch JW, Jarosz M, Krzymanska-Olejnik E, Kung L, Lipson D, Lowman GM, Marappan S, McInerney P, Platt A, Roy A, Siddiqi SM, Steinmann K, Thompson JF: **Virtual terminator nucleotides for next-generation DNA sequencing.** *Nat Methods* 2009, **6:**593-595.
20. Hart C, Lipson D, Ozsolak F, Raz T, Steinmann K, Thompson J, Milos PM: **Single-molecule sequencing: sequence methods to enable accurate quantitation.** *Methods Enzymol* 2010, **472:**407-430.
21. Thompson JF, Steinmann KE: **Single molecule sequencing with a HeliScope genetic analysis system.** *Curr Protoc Mol Biol* 2010, **Chapter 7, Unit7:**10.
22. Giladi E, Healy J, Myers G, Hart C, Kapranov P, Lipson D, Roels S, Thayer E, Letovsky S: **Error tolerant indexing and alignment of short reads with covering template families.** *J Comput Biol* 2010, **17:**1397-1411.
23. Goren A, Ozsolak F, Shoresh N, Ku M, Adli M, Hart C, Gymrek M, Zuk O, Regev A, Milos PM, Bernstein BE: **Chromatin profiling by directly sequencing small quantities of immunoprecipitated DNA.** *Nat Methods* 2010, **7:**47-49.
24. Pushkarev D, Neff NF, Quake SR: **Single-molecule sequencing of an individual human genome.** *Nat Biotechnol* 2009, **27:**847-852.
25. Ozsolak F, Platt AR, Jones DR, Reifenberger JG, Sass LE, McInerney P, Thompson JF, Bowers J, Jarosz M, Milos PM: **Direct RNA sequencing.** *Nature* 2009, **461:**814-818.
26. Lipson D, Raz T, Kieu A, Jones DR, Giladi E, Thayer E, Thompson JF, Letovsky S, Milos P, Causey M: **Quantification of the yeast transcriptome by single-molecule sequencing.** *Nat Biotechnol* 2009, **27:**652-658.
27. Ozsolak F, Goren A, Gymrek M, Guttman M, Regev A, Bernstein BE, Milos PM: **Digital transcriptome profiling from attomole-level RNA samples.** *Genome Res* 2010, **20:**519-525.
28. Ozsolak F, Ting DT, Wittner BS, Brannigan BW, Paul S, Bardeesy N, Ramaswamy S, Milos PM, Haber DA: **Amplification-free digital gene expression profiling from minute cell quantities.** *Nat Methods* 2010, **7:**619-621.
29. Ozsolak F, Kapranov P, Foissac S, Kim SW, Fishilevich E, Monaghan AP, John B, Milos PM: **Comprehensive polyadenylation site maps in yeast and human reveal pervasive alternative polyadenylation.** *Cell* 2010, **143:**1018-1029.
30. Kapranov P, Ozsolak F, Kim SW, Foissac S, Lipson D, Hart C, Roels S, Borel C, Antonarakis SE, Monaghan AP, John B, Milos PM: **New class of gene-termini-associated human RNAs suggests a novel RNA copying mechanism.** *Nature* 2010, **466:**642-646.
31. Kapranov P, St Laurent G, Raz T, Ozsolak F, Reynolds CP, Sorensen PH, Reaman G, Milos P, Arceci RJ, Thompson JF, Triche TJ: **The majority of total nuclear-encoded non-ribosomal RNA in a human cell is 'dark matter' unannotated RNA.** *BMC Biol* 2010, **8:**149.
32. Korlach J, Marks PJ, Cicero RL, Gray JJ, Murphy DL, Roitman DB, Pham TT, Otto GA, Foquet M, Turner SW: **Selective aluminum passivation for targeted immobilization of single DNA polymerase molecules in zero-mode waveguide nanostructures.** *Proc Natl Acad Sci U S A* 2008, **105:**1176-1181.
33. Levene MJ, Korlach J, Turner SW, Foquet M, Craighead HG, Webb WW: **Zero-mode waveguides for single-molecule analysis at high concentrations.** *Science* 2003, **299:**682-686.
34. Korlach J, Bjornson KP, Chaudhuri BP, Cicero RL, Flusberg BA, Gray JJ, Holden D, Saxena R, Wegener J, Turner SW: **Real-time DNA sequencing from single polymerase molecules.** *Methods Enzymol* 2010, **472:**431-455.
35. Eid J, Fehr A, Gray J, Luong K, Lyle J, Otto G, Peluso P, Rank D, Baybayan P, Bettman B, Bibillo A, Bjornson K, Chaudhuri B, Christians F, Cicero R, Clark S, Dalal R, Dewinter A, Dixon J, Foquet M, Gaertner A, Hardenbol P, Heiner C, Hester K, Holden D, Kearns G, Kong X, Kuse R, Lacroix Y, Lin S, *et al.*: **Real-time DNA sequencing from single polymerase molecules.** *Science* 2009, **323:**133-138.
36. Korlach J, Bibillo A, Wegener J, Peluso P, Pham TT, Park I, Clark S, Otto GA, Turner SW: **Long, processive enzymatic DNA synthesis using 100% dye-labeled terminal phosphate-linked nucleotides.** *Nucleosides Nucleotides Nucleic Acids* 2008, **27:**1072-1083.
37. Travers KJ, Chin CS, Rank DR, Eid JS, Turner SW: **A flexible and efficient template format for circular consensus sequencing and SNP detection.** *Nucleic Acids Res* 2010, **38:**e159.
38. Chin CS, Sorenson J, Harris JB, Robins WP, Charles RC, Jean-Charles RR, Bullard J, Webster DR, Kasarskis A, Peluso P, Paxinos EE, Yamaichi Y, Calderwood SB, Mekalanos JJ, Schadt EE, Waldor MK: **The origin of the Haitian cholera outbreak strain.** *N Engl J Med* 2010, **364:**33-42.
39. Ritz A, Bashir A, Raphael BJ: **Structural variation analysis with strobe reads.** *Bioinformatics* 2010, **26:**1291-1298.
40. Flusberg BA, Webster DR, Lee JH, Travers KJ, Olivares EC, Clark TA, Korlach J, Turner SW: **Direct detection of DNA methylation during single-molecule, real-time sequencing.** *Nat Methods* 2010, **7:**461-465.
41. Hardin SH: **Real-time DNA sequencing.** In *Next-Generation Genome Sequencing: Towards Personalized Medicine.* Edited by Janitz M. Weinheim: Wiley-VCH; 2008:97-102.
42. Pennisi E: **Genomics. Semiconductors inspire new sequencing technologies.** *Science* 2010, **327:**1190.
43. Teague B, Waterman MS, Goldstein S, Potamousis K, Zhou S, Reslewic S, Sarkar D, Valouev A, Churas C, Kidd JM, Kohn S, Runnheim R, Lamers C, Forrest D, Newton MA, Eichler EE, Kent-First M, Surti U, Livny M, Schwartz DC: **High-resolution human genome structure by single-molecule analysis.** *Proc Natl Acad Sci U S A* 2010, **107:**10848-10853.
44. Zhou S, Wei F, Nguyen J, Bechner M, Potamousis K, Goldstein S, Pape L, Mehan MR, Churas C, Pasternak S, Forrest DK, Wise R, Ware D, Wing RA, Waterman MS, Livny M, Schwartz DC: **A single molecule scaffold for the maize genome.** *PLoS Genet* 2009, **5:**e1000711.
45. Ananiev GE, Goldstein S, Runnheim R, Forrest DK, Zhou S, Potamousis K, Churas CP, Bergendahl V, Thomson JA, Schwartz DC: **Optical mapping discerns genome wide DNA methylation profiles.** *BMC Mol Biol* 2008, **9:**68.
46. Das SK, Austin MD, Akana MC, Deshpande P, Cao H, Xiao M: **Single molecule linear analysis of DNA in nano-channel labeled with sequence specific fluorescent probes.** *Nucleic Acids Res* 2010, **38:**e177.
47. Bailo E, Deckert V: **Tip-enhanced Raman spectroscopy of single RNA strands: towards a novel direct-sequencing method.** *Angew Chem Int Ed Engl* 2008, **47:**1658-1661.
48. Tanaka H, Kawai T: **Partial sequencing of a single DNA molecule with a scanning tunnelling microscope.** *Nat Nanotechnol* 2009, **4:**518-522.
49. **ZS Genetics, Inc.** [http://www.zsgenetics.com/thetech/prep/index.html]
50. Griffiths J: **The realm of the nanopore. Interest in nanoscale research has skyrocketed, and the humble pore has become a king.** *Anal Chem* 2008, **80:**23-27.
51. Schneider GF, Kowalczyk SW, Calado VE, Pandraud G, Zandbergen HW, Vandersypen LM, Dekker C: **DNA translocation through graphene nanopores.** *Nano Lett* 2010, **10:**3163-3167.
52. Merchant CA, Healy K, Wanunu M, Ray V, Peterman N, Bartel J, Fischbein MD, Venta K, Luo Z, Johnson AT, Drndic M: **DNA translocation through graphene nanopores.** *Nano Lett* 2010, **10:**2915-2921.
53. Balagurusamy VS, Weinger P, Ling XS: **Detection of DNA hybridizations using solid-state nanopores.** *Nanotechnology* 2010, **21:**335102.
54. Garaj S, Hubbard W, Reina A, Kong J, Branton D, Golovchenko JA: **Graphene as a subnanometre trans-electrode membrane.** *Nature* 2010, **467:**190-193.
55. Cockroft SL, Chu J, Amorin M, Ghadiri MR: **A single-molecule nanopore device detects DNA polymerase activity with single-nucleotide resolution.** *J Am Chem Soc* 2008, **130:**818-820.
56. Stoddart D, Heron AJ, Mikhailova E, Maglia G, Bayley H: **Single-nucleotide discrimination in immobilized DNA oligonucleotides with a biological nanopore.** *Proc Natl Acad Sci U S A* 2009, **106:**7702-7707.
57. Purnell RF, Schmidt JJ: **Discrimination of single base substitutions in a DNA strand immobilized in a biological nanopore.** *ACS Nano* 2009, **3:**2533-2538.
58. Stoddart D, Heron AJ, Klingelhoefer J, Mikhailova E, Maglia G, Bayley H: **Nucleobase recognition in ssDNA at the central constriction of the alpha-hemolysin pore.** *Nano Lett* 2010, **10:**3633-3637.
59. Lathrop DK, Ervin EN, Barrall GA, Keehan MG, Kawano R, Krupka MA, White HS, Hibbs AH: **Monitoring the escape of DNA from a nanopore using an alternating current signal.** *J Am Chem Soc* 2010, **132:**1878-1885.
60. Butler TZ, Pavlenok M, Derrington IM, Niederweis M, Gundlach JH: **Single-molecule DNA detection with an engineered MspA protein nanopore.** *Proc Natl Acad Sci U S A* 2008, **105:**20647-20652.
61. Derrington IM, Butler TZ, Collins MD, Manrao E, Pavlenok M, Niederweis M, Gundlach JH: **Nanopore DNA sequencing with MspA.** *Proc Natl Acad Sci U S A* 2010, **107:**16060-16065.
62. de Zoysa RS, Jayawardhana DA, Zhao Q, Wang D, Armstrong DW, Guan X: **Slowing DNA translocation through nanopores using a solution containing organic salts.** *J Phys Chem B* 2009, **113:**13332-13336.

63. Kawano R, Schibel AE, Cauley C, White HS: **Controlling the translocation of single-stranded DNA through alpha-hemolysin ion channels using viscosity.** *Langmuir* 2009, **25:**1233-1237.
64. Peng H, Ling XS: **Reverse DNA translocation through a solid-state nanopore by magnetic tweezers.** *Nanotechnology* 2009, **20:**185101.
65. 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* 2010, **10:**2237-2244.
66. Clarke J, Wu HC, Jayasinghe L, Patel A, Reid S, Bayley H: **Continuous base identification for single-molecule nanopore DNA sequencing.** *Nat Nanotechnol* 2009, **4:**265-270.
67. McCaughan F, Dear PH: **Single-molecule genomics.** *J Pathol* 2010, **220:**297-306.
68. **GenBank Statistics** [http://www.ncbi.nlm.nih.gov/genbank/genbankstats.html]
69. Benson DA, Karsch-Mizrachi I, Lipman DJ, Ostell J, Sayers EW: **GenBank.** *Nucleic Acids Res* 2011 **39:**D32-D37.
70. Benson DA, Karsch-Mizrachi I, Lipman DJ, Ostell J, Sayers EW: **GenBank.** *Nucleic Acids Res* 2010, **38:**D46-D51.
71. Shumway M, Cochrane G, Sugawara H: **Archiving next generation sequencing data.** *Nucleic Acids Res* 2010, **38:**D870-D871.
72. Leinonen R, Sugawara H, Shumway M; International Nucleotide Sequence Database Collaboration: **The sequence read archive.** *Nucleic Acids Res* 2010, **39(Database issue):**D19-21.
73. Sayers EW, Barrett T, Benson DA, Bryant SH, Canese K, Chetvernin V, Church DM, DiCuccio M, Edgar R, Federhen S, Feolo M, Geer LY, Helmberg W, Kapustin Y, Landsman D, Lipman DJ, Madden TL, Maglott DR, Miller V, Mizrachi I, Ostell J, Pruitt KD, Schuler GD, Sequeira E, Sherry ST, Shumway M, Sirotkin K, Souvorov A, Starchenko G, Tatusova TA, *et al.*: **Database resources of the National Center for Biotechnology Information.** *Nucleic Acids Res* 2009, **37:**D5-D15.
74. Wheeler DL, Barrett T, Benson DA, Bryant SH, Canese K, Chetvernin V, Church DM, Dicuccio M, Edgar R, Federhen S, Feolo M, Geer LY, Helmberg W, Kapustin Y, Khovayko O, Landsman D, Lipman DJ, Madden TL, Maglott DR, Miller V, Ostell J, Pruitt KD, Schuler GD, Shumway M, Sequeira E, Sherry ST, Sirotkin K, Souvorov A, Starchenko G, Tatusov RL, *et al.*: **Database resources of the National Center for Biotechnology Information.** *Nucleic Acids Res* 2008, **36:**D13-D21.
75. **The NCBI Short Read Archive (SRA), a New Primary Data Archive Resource** [http://www.ncbi.nlm.nih.gov/Traces/sra/static/AGBT_Poster_Final.pdf]

doi:10.1186/gb-2011-12-2-217
Cite this article as: Thompson JF, Milos PM: **The properties and applications of single-molecule DNA sequencing.** *Genome Biology* 2011, **12:**217.