Trends in Genetics

CellPress REVIEWS

Review

The Third Revolution in Sequencing Technology

Erwin L. van Dijk,[1,*] Yan Jaszczyszyn,[1] Delphine Naquin,[1] and Claude Thermes[1]

Forty years ago the advent of Sanger sequencing was revolutionary as it allowed complete genome sequences to be deciphered for the first time. A second revolution came when next-generation sequencing (NGS) technologies appeared, which made genome sequencing much cheaper and faster. However, NGS methods have several drawbacks and pitfalls, most notably their short reads. Recently, third-generation/long-read methods appeared, which can produce genome assemblies of unprecedented quality. Moreover, these technologies can directly detect epigenetic modifications on native DNA and allow whole-transcript sequencing without the need for assembly. This marks the third revolution in sequencing technology. Here we review and compare the various long-read methods. We discuss their applications and their respective strengths and weaknesses and provide future perspectives.

Highlights

Long-read/third-generation sequencing technologies are causing a new revolution in genomics as they provide a way to study genomes, transcriptomes, and metagenomes at an unprecedented resolution.

SMRT and nanopore sequencing allow for the first time the direct study of different types of DNA base modifications.

Moreover, nanopore technology can sequence directly RNA and identify RNA base modifications.

Owing to the portability of the MinION and the existence of extremely simple library preparation methods, nanopore technology allows the performance of high-throughput sequencing for the first time in the field and at remote places. This is of tremendous importance for the survey of outbreaks in developing countries.

A Brief History of Sequencing Technology

The introduction of Maxam and Gilbert's chemical chain termination method for DNA sequencing in 1977 [1] closely followed by **Sanger's 'dideoxy method'** (see Glossary) the same year [2] caused a revolution in biology. These methods led to ever larger genomes being sequenced, culminating with the Human Genome Project [3,4]. The Human Genome Project has been the world's largest collaborative biological project to date and has taken 13 years to complete at a cost of almost US$3 billion.

As a next step, large-scale sequencing projects were undertaken to study human sequence variation. However, for these types of projects Sanger sequencing was too labor intensive, time consuming, and expensive. In 2004 the National Human Genome Research Institute (NHGRI) initiated a program to bring the cost of whole-genome sequencing down to US$1000 in 10 years [5]. This accelerated the development of cheaper and faster methods and in the years that followed **NGS technologies** generating thousands to many millions of sequencing reactions per run appeared. Major advantages of these NGS technologies were that they did not require bacterial cloning of DNA fragments and electrophoretic separation of sequencing products. Various NGS technologies coexisted for a number of years; today, the market is largely dominated by Illumina. Through a spectacular price reduction, NGS rapidly brought genome sequencing within reach of small laboratories, and the original goal of sequencing a human genome for less than US$1000 was achieved a few years ago [6]. Today, NGS has become a standard tool for many applications in basic biology as well as for clinical and agronomical research.

Drawbacks of NGS Methods

While NGS technologies are extremely powerful, they also have some drawbacks. One major limitation is their relatively short reads. Genomes often contain numerous repeated sequences that are longer than the NGS reads, which may lead to misassemblies and gaps [7,8] (Figure 1,

[1]Institute for Integrative Biology of the Cell, UMR9198, CNRS CEA Université Paris-Sud, Université Paris-Saclay, 9198 Gif sur Yvette Cedex, France

*Correspondence: erwin.vandijk@i2bc.paris-saclay.fr (E.L. van Dijk).

Trends in Genetics

CellPress REVIEWS

Key Figure

Comparison of the Performance of Next-Generation Sequencing (NGS) Short-Read and Long-Read Methods

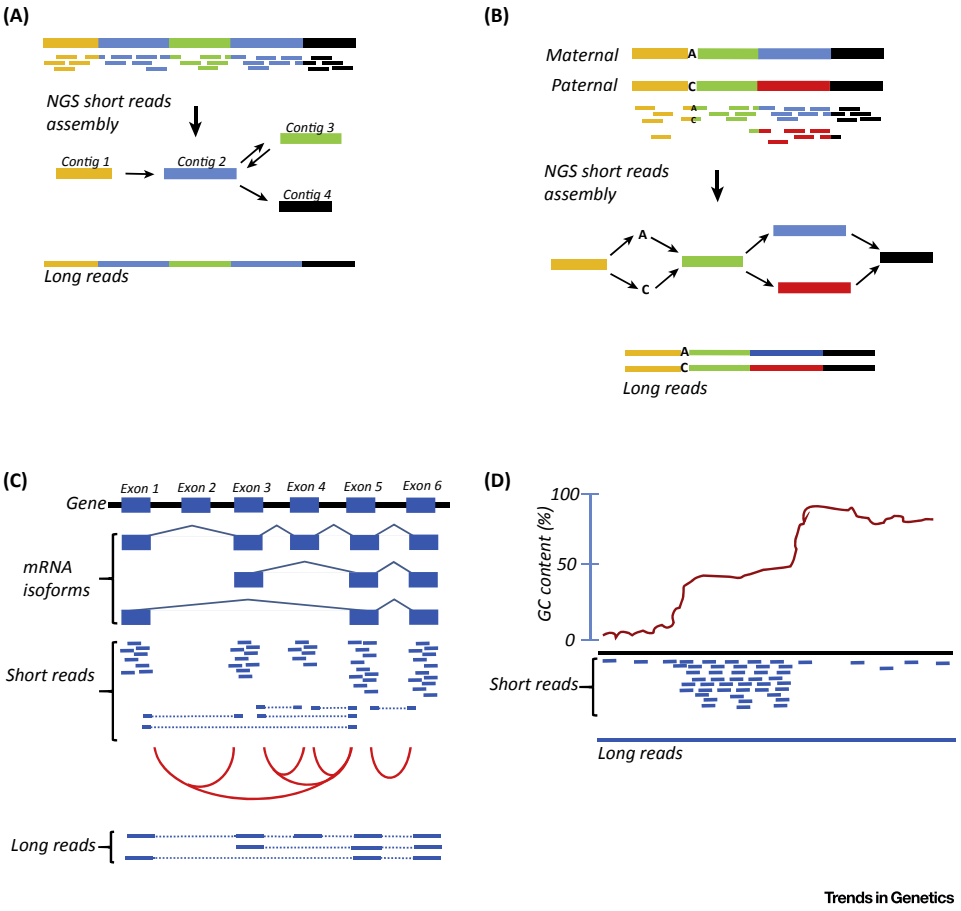

Figure 1. (A) Example showing the resolution of a repeated genome region by short-read assembly or long reads. Sequencing a region with two nearly identical repeats (blue) separated by a unique sequence will generate reads corresponding to the upstream region (yellow), the repeats, the sequence between (green), and the downstream region (black), and some reads will overlap the boundaries. Assembly programs cannot assign reads falling in the repeats to unique positions and will assemble those reads into a single contig. The sequence between the repeats cannot be assigned to a unique position either as it can be placed either upstream or downstream of the 'blue' region. Due to this ambiguity, the sequences upstream of, between, and downstream of the repeats will be assembled into separate contigs. Similar problems arise with structural variants that involve repetitive regions. (B) Haplotype phasing. SNPs (single nucleotide: A or C) or larger variations (red or blue) between maternal and paternal alleles located too far apart to be covered by a single read will be difficult to phase to the parental allele of origin. This will lead to ambiguous trajectories that result in fragmented assemblies. (C) A multitude of mRNA transcript isoforms can be generated from a single gene through alternative intron splicing. Short-read sequencing of those isoforms will produce reads falling within the exons present in the pool (unbroken lines) and there will be reads that overlap the various exon junctions (broken lines). Thus, the alternative splicing events will be detected; the exon–exon junctions detected in the pool are indicated by red lines. However, information about the combination of exon junctions in the individual transcripts is lacking. Long-read sequencing covers the entire transcript and will thus provide this additional information. (D) NGS methods depend on PCR amplification during library preparation and/or on the flow cells. PCR is a bias-prone process that tends to be inefficient at extreme GC content. As a result, regions with extreme GC content will often be poorly covered. Single-molecule real-time sequencing (SMRT) and nanopore long-read sequencing technologies do not require PCR amplification and suffer much less of this problem (although SMRT sequencing uses a polymerase).

Glossary

BAC-by-BAC sequencing: a sequencing method in which a physical map of the target genome, or chromosome, is established using a set of overlapping bacterial artificial chromosome (BAC) clones. The individual clones are subsequently fragmented and subjected to shotgun sequencing.

Chromosome conformation capture: a set of molecular biology methods used to analyze the spatial organization of chromatin in a cell.

Circular consensus sequencing (CCS): in PacBio CCS, the DNA polymerase reads a ligated circular DNA template multiple times, generating a consensus sequence with a high level of accuracy.

Contig: from contiguous; a set of overlapping DNA segments that together represent a consensus region of DNA.

Dideoxy or Sanger sequencing: a method of DNA sequencing based on the selective incorporation of chain-terminating dideoxynucleotides by DNA polymerase. The resulting DNA fragments are heat denatured and separated by size using gel electrophoresis.

Genome phasing: diploid genomes have two copies of each chromosome that differ at various loci along each chromosome. Genome phasing (also called haplotyping or haplotype estimation) allows the determination of which chromosome such heterozygous variants are derived from. Assembly of reads that share the same variation enables reconstruction of the parental homologs (haplotype reconstruction).

Indel: a common class of mutations comprising an insertion or deletion of one or more DNA bases into a genome.

N50: a statistical measure of the average length of a set of sequences; used widely in genomics, especially in reference to read, contig, or **scaffold** lengths in a draft assembly. For reads it indicates the length such that reads of this length or greater sum to half of the total number of bases. For contigs or scaffolds it indicates the size such that contigs or scaffolds of this length or greater sum to at least half of the haploid genome size.

Key Figure). As a result, many available genomes are heavily fragmented into hundreds or thousands of **contigs**.

In addition, while small variants such as single-nucleotide variations (SNVs) and short **indels** can be accurately detected using short reads, larger structural variations (SVs) are more challenging to detect and characterize. This is an important issue, given that SVs are implicated in a number of diseases [9]. Moreover, short reads have a limited capacity to link (even short) independent variations on the same nucleic acid molecule. As a result, NGS methods are not well suited to discriminate and phase alleles to their respective parental homolog, which is important for many aspects of human genetics [10], and have a limited capacity to characterize transcript isoforms generated by alternative splicing.

In addition to the abovementioned limitations due to the short reads, the fact that NGS methods rely on PCR causes difficulties with regions of extreme GC%, as these are inefficiently amplified by PCR.

Therefore, while NGS methods have revolutionized biology, there has been the need to develop methods better capable of dealing with the abovementioned issues.

The Advent of Third-Generation Sequencing (TGS)/Long-Read Sequencing

Shortly after the appearance of NGS, TGS technologies emerged. Distinguishing features of TGS are single-molecule sequencing (SMS) and sequencing in real time (as opposed to NGS, where sequencing is paused after each base incorporation) [11]. The first SMS technology, commercialized by Helicos Biosciences, resembled Illumina sequencing but without any bridge amplification [12]. As the method was relatively slow, expensive, and produced short reads (~32 bp), it did not prove viable. The first 'true' TGS technology was released on the market in 2011 by Pacific Biosciences (PacBio) and is termed 'single-molecule real-time' (SMRT) sequencing [13]. More recently (2014), Oxford Nanopore Technologies (ONT) introduced nanopore sequencing [14]. Besides the absence of PCR amplification and the real-time sequencing process, an important feature of SMRT and nanopore sequencing is the production of long reads. As an alternative approach, Illumina introduced a library preparation kit for 'synthetic long reads' (SLRs) in 2014 (formerly Moleculo [15]). One year later 10X Genomics introduced a microfluidics variant of SLR with much higher partitioning capacity [16]. Note that SLR technologies are not TGS methods as they are based on classical Illumina sequencing.

These long-read technologies are now revolutionizing genomics research as they enable researchers to explore genomes at an unprecedented resolution. In the subsequent sections we examine in more detail these new methodologies. Due to length limitations we do not discuss in detail the analysis of long-read sequence data. Excellent recent reviews focusing on long-read bioinformatics tools can be found elsewhere [17,18].

Long-Read Technologies

SMRT Sequencing: PacBio

In early 2011, PacBio released their PacBio RS sequencer, which uses SMRT technology (Box 1). While initially average read lengths were relatively short (~1.5 kb) and average error rates were high (~13%) [19], the technology has strongly improved over recent years. Average read lengths have increased more than tenfold and the throughput per run has increased by about 100-fold owing to the development of improved **sequencing chemistries** and the release of a new sequencer, the Sequel. This machine generates about tenfold more sequence

Nanochannel genome mapping: high-throughput (optical) genome mapping technology commercialized by BioNano Genomics, also referred to as next-generation mapping (NGM). Long DNA molecules are nick labeled at specific sites and linearized in nanochannel arrays. The length of the DNA molecules and the positions of nick labels are determined after automated image capture.

Next-generation sequencing (NGS) technologies: methods based on of massive parallel sequencing via spatially separated, clonally amplified DNA templates in a flow cell. Typically, reads of up to several hundreds of base pairs are produced.

Quality value (QV): also referred to as the Phred quality score; indicates the probability that a given base is called incorrectly by the sequencer. QVs are logarithmically related to the base-calling error probability (P)[2], $Q = -10\log_{10}P$. For example, QV30 is equivalent to the probability of an incorrect base call 1 in 1000 times.

Scaffold: a noncontiguous series of genomic sequences is linked together into a scaffold comprising sequences separated by gaps of known length. The sequences that are linked are typically contiguous sequences corresponding to read overlaps.

Sequencing chemistry: the molecular mechanism of a given sequencing method. Several technologies are based on 'sequencing by synthesis' in which sequence information is generated by a polymerase that copies a DNA strand. By contrast, nanopore sequencing directly 'reads' the original DNA or RNA molecule.

Short tandem repeats (STPs): or microsatellites; comprise a unit of 2–13 nucleotides repeated many times (up to hundreds or thousands) in a row on a DNA strand.

Structural variations (SVs): genomic rearrangements affecting more than 50 bp. SVs are often multiple kilobases or even megabases in size and include deletions, insertions, inversions, mobile-element transpositions, translocations, tandem repeats, and copy number variants (CNVs).

data than the upgraded PacBio RS (RSII) and is twofold less expensive (Table 1). The 'single-pass' error rate has remained roughly the same since the beginning (~13%), but molecules of up to ~1–2 kb can now be sequenced many times owing to the circular templates [20] and increased polymerase processivity, strongly improving overall accuracy (see Figure ID in Box 1). Moreover, increased throughput has led to a sharp reduction in cost per base ([19]; http://allseq.com/knowledge-bank/sequencing-platforms/pacific-biosciences/).

For genomic DNA library preparation, PacBio commercialized a 'SMRTbell template prep kit' and an 'express' variant thereof for rapid library preparation with an approximately 3-h workflow. For transcriptome analysis an 'isoform sequencing' protocol is available (https://www.pacb.com/wp-content/uploads/ Procedure-Checklist-20-kb-Template-Preparation-Using-BluePippin-Size-Selection-System-15-20-kb-Cutoff-Sequel-Systems.pdf).

Nanopore Sequencing: ONT

The idea of using nanopores in a membrane to sequence single-stranded (ss) DNA or RNA molecules originated at the end of the 1980s [21]. However, due to technical obstacles the first successful sequencing results were reported only in 2012 [22]. Two years later, ONT released their first nanopore sequencer, the pocket-sized MinION, in a large-scale collaborative MinION Access Program (MAP). For technical details, see Box 2.

Through rapid evolution of chemistries, a significant increase in throughput has been achieved. While the early chemistries produced ~184–450 million bases of sequence data per 48-h run [14], today's R9.4 flow cells in combination with the latest library preparation kit versions can produce up to 20 Gb of sequence data (Table 1). Also, the translocation speed has increased; from 30 bases per second (bps) with R7.3 flow cells to 450 bps using today's 9.4 chemistry [23].

For greater flexibility in throughput, ONT introduced several new sequencers. The PromethION was introduced through a new access program in July 2015. This machine can hold up to 48 flow cells with 3000 channels each, making a total of 144 000 channels available per run,

Box 1. SMRT Sequencing Technology

The technology used by PacBio is called SMRT sequencing [12]. This technology uses a closed, circular ssDNA template called a SMRTbell, which is created by ligating hairpin adaptors to both ends of a target dsDNA molecule [15] (Figure IA).

A primer and a polymerase are annealed to the adapter, followed by library loading onto a specialized flow cell with 150 000 picolitre wells called zero mode waveguides (ZMWs), for the PacBio RSII or 1 million for the newer Sequel platform. In each ZMW, a single polymerase is immobilized at the (transparent) bottom, where it replicates a target DNA molecule. During the replication process, the incorporation of fluorescently labeled nucleotides produces fluorescence signals on excitation by a laser and a camera system records in real time (a 'movie') the color and duration of emitted light. The diameter of the ZMWs is smaller than the excitation light's wavelength, which causes it to decay exponentially, exclusively illuminating the very bottom of the wells. This reduces interference from other labeled dNTPs in solution (Figure IB). The time between nucleotide incorporations is called the 'interpulse duration' (IPD). This feature allows the detection of base modifications such as 6 mA as these influence the speed of base incorporation (Figure IC).

Because the SMRTbell forms a closed circle, after the polymerase replicates one strand of the target dsDNA it can continue using the adaptor and then the other strand as a template. If the lifetime of the polymerase is long enough, both strands can be sequenced multiple times (called 'passes') in a single continuous long read (CLR). The CLR can then be split to multiple 'subreads' by recognizing and cutting out the adaptor sequences. The consensus sequence of multiple subreads in a single ZMW yields a CCS with higher accuracy. Figure ID shows the relation between the coverage of a given sequence (i.e., number of passes) and the accuracy, expressed as **quality value (QV)**. Note that there is a tradeoff between molecule length and sequencing accuracy due to the fact that for longer molecules a lower number of passes will be generated.

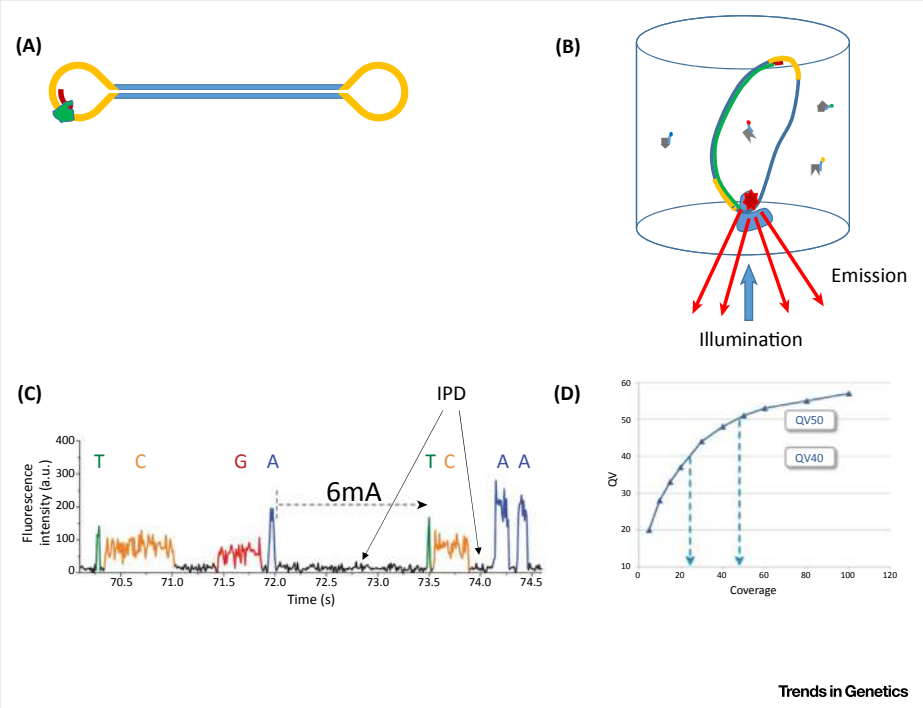

Figure I. Overview of Pacific Biosciences' Single-Molecule Real-Time Sequencing (SMRT) Technology. (A) Library preparation comprises the ligation of hairpin adapters (yellow) to double-stranded DNA molecules (blue), thereby creating circular molecules called 'SMRTbells'. Next, a primer (red) and a polymerase (green) are annealed to the adapter. (B) Schematic representation of a zero mode waveguide (ZMW), a nanoscale observation chamber. The polymerase–primer–SMRTbell complex binds to the bottom of the ZMW through biotin–streptavidin chemistry. Note, however, that not all ZMWs will contain a DNA molecule because the library is loaded by diffusion. The polymerase incorporates fluorescently labeled nucleotides emitting a fluorescent signal on illumination from below. These signals are recorded by a camera in real time in a process called a 'movie'. (C) In a movie, not only the fluorescence color is registered, but also the time between nucleotide incorporations, called the interpulse duration (IPD) (black). The presence of an epigenetic modification, such as 6-methyladenosine (6 mA), results in a delayed IPD. Adapted with permission from Pacific Biosciences. (D) Multiple 'passes' of the circular library can be combined into a circular consensus sequence (CCS) that increases in accuracy as the number of passes increases. Accuracy is expressed as the quality value (QV). Note that at ~25 passes, the accuracy reaches 99.999% (QV40), which is similar to the accuracy of Illumina sequencing. At ~50 passes, accuracy can even reach 99.9999% (QV50). Adapted with permission from Pacific Biosciences. The data indicated in the figure are based on a bacterial genome run on the Sequel system with 2.1 chemistry and 5.1 Sequel software.

against 512 per MinION run. As a result, a stunning theoretical maximum of 15 Tb (best real result obtained: ~6 Tb) of sequence data can be generated per 48-h run. This makes the PromethION a serious competitor for Illumina's HiSeq X Ten, which generates a maximum of 16–18 Tb per run (total of a cluster of ten instruments; http://www.illumina.com). In early 2017, ONT released the GridION X5, which can hold up to five MinION flow cells and can generate up to 100 Gb of data per run. ONT also announced the release of the SmidgION, which is even smaller than a MinION and can be controlled by a smartphone. Together these sequencers allow maximum scalability, from extremely small and portable to extremely powerful and high throughput.

In contrast to SMRT sequencing, nanopore read length is not limited by the technology itself but rather by the length of the DNA molecules to be sequenced. Therefore, provided that the DNA is

Table 1. Comparison of the Various Long-Read Methods

Platform	Instrument	Average read length (kb)	Maximum read length (kb)	Instrument cost (US$)	Cost per run (US$)	Cost per Gb (US$)	Input requirement	Throughput per run	Run time	Refs
PacBio	PacBio RSII	10–15	>80	700 000	400	400[g]	>1 μg DNA[h]	0.5–1 Gb	Up to 4 h	http://allseq.com/knowledge-bank/sequencing-platforms/pacific-biosciences/ http://www.pacb.com/blog/new-chemistry-software-sequel-system-improve-read-length-lower-project-costs/ http://dnatech.genomecenter.ucdavis.edu/wp-content/uploads/2014/07/Pacbio-Guidelines-SMRTbell-Libraries-v1.0.pdf
	Sequel	10–15	>80	350 000	850	85[g]	>1 μg DNA[h]	5–10 Gb	Up to 4 h	
ONT	MinION	Variable[a]	Variable[a]	1000	475–900[e]	24[g]	~1 μg DNA	Up to 20 Gb	Up to 48 h	[25] https://nanoporetech.com/products#modal=comparison
	GridION	Variable[a]	Variable[a]	49 955[b] 125 000[c] Free (U$142 500 for reagents)[d]	475–900[e] per flow cell	24[g]	~1 μg DNA	Up to 100 Gb (five flow cells)	Up to 48 h	
	PromethION	Variable[a]	Variable[a]	135 000	625–2000[e] per flow cell	5[g]	~1 μg DNA	Up to 125 Gb (one flow cell) Up to 6 Tb (48 flow cells)	Up to 64 h	
Illumina SLR	Illumina sequencer: no additional equipment	NA	~10 kb	No additional instrument[f]	17 602	12–27	~500 ng DNA	650 Gb to 1.5 Tb (Illumina HiSeq3000/4000, paired end 150 bp)	~3.5 days	https://www.illumina.com/documents/products/datasheets/datasheet-truseq-synthetic-kit-assembly.pdf https://support.illumina.com/content/dam/illumina-support/documents/documentation/chemistry_documentation/samplepreps_truseq/truseqsynthlongreaddna/truseq-synth-long-read-dna-library-prep-guide-15047264-b.pdf

Table 1. (continued)

Platform	Instrument	Average read length (kb)	Maximum read length (kb)	Instrument cost (US$)	Cost per run (US$)	Cost per Gb (US$)	Input requirement	Throughput per run	Run time	Refs
10X Genomics SLR	Chromium + Illumina sequencer	NA	Up to 100 kb	125 000[f]	6705–19 700[h]	8–11	~1 ng DNA	800 Gb to 1.8 Tb[i] (Illumina HiSeq X, paired end 150 bp)	~3 days	https://support.10xgenomics.com/de-novo-assembly/sequencing/doc/specifications-sequencing-requirements-for-de-novo-assembly https://emea.illumina.com/systems/sequencing-platforms/hiseq-x/specifications.html

[a]Read lengths are limited only by the molecule lengths in the sample.
[b]'Starter pack': allows use of GridION for 60 flow cells over 6 months.
[c]'CapEx purchase of device': includes right to purchase up to 450 flow cells at US$299 each over 18-month period.
[d]'OpEx option': the sum of US$142 500 is a commitment for reagent use (at least 300 flow cells per year).
[e]Reverse correlation between flow cell price and order volume.
[f]To be added: price of Illumina sequencer (http://www.illumina.com).
[g]Based on lowest possible flow cell price and highest output.
[h]Amount required for 1–2-kb insert libraries; larger-insert libraries require more material.
[i]Depending on whether HiSeq X Five or HiSeq X Ten and a single- or dual-flow cell run is used.

of sufficient quality, extremely long reads can be obtained; recently, reads of up to ~1 Mb have been reported [24].

A drawback is the high error rate (~15%) [25]. Nanopore technology does not have the possibility of sequencing the same strand multiple times, as with SMRT sequencing. For higher accuracy, ONT developed a method to sequence both strands of a double-stranded (ds) DNA molecule. First, a hairpin adapter was used on one extremity such that the second strand would be sequenced after the first. This system was called 'two-directional' (2D) sequencing as opposed to 1D sequencing where only one strand is read. This system has recently been replaced by the '$1D^2$' system, using a normal adapter with a specialized sequence that promotes entry of the second strand into the pore after the first strand has passed through. ONT claims that this reduces the error rate to about 3% (https://store.nanoporetech.com/1d-2-sequencing-kit.html), but throughput is reduced as both strands of each molecule are sequenced, doubling the time of passage through the pore. It is interesting to note here that an approach to mimic PacBio **circular consensus sequencing (CCS)** has been reported; it uses the phi29 polymerase to produce a tandem array of copies of the original DNA molecule [26].

For library preparation, a wide diversity of kits is available, including a 'rapid sequencing kit' that has an extremely fast (less than 10 min) and simple workflow with a minimal requirement for equipment. This kit is therefore particularly suitable for use in the field.

SLR: Illumina/10X Genomics

As an alternative to the methods developed by PacBio and ONT, SLR technologies partition large DNA fragments into microtiter wells or an emulsion so that very few fragments exist in each partition. In each partition the template fragments are sheared and barcoded. This approach uses classical short-read sequencing, after which reads with the same barcode are assembled locally as they must be derived from the same original large fragment [15].

Box 2. Nanopore Sequencing: ONT

Nanopore sequencing occurs in a flow cell in which two ionic solution-filled compartments are separated by a membrane with 2048 (MinION) or 12 000 (PromethION) individual nanopores incorporated. A constant voltage bias produces an ionic current through the nanopore and on translocation of a DNA or RNA molecule, a change in the ionic current can be observed and characterized (Figure IA). The current in the nanopore is measured by a sensor several thousand times per second and is graphically represented in a 'squiggle plot'. Finally, data processing is performed by the minKNOW software, which deals with data acquisition and analysis.

Nanopore sequencing requires library preparation, in which DNA fragments, sheared or not, are end-repaired followed by adapter ligation (Figure IB. The adapters are DNA–protein complexes with a tightly bound polymerase or helicase enzyme that ensures stepwise movement of the DNA through the pore by a ratcheting mechanism. dsDNA is unwound at the pore after which one strand passes through. For increased read accuracy, mechanisms have been developed to sequence the second strand after the first strand has finished passing through the pore. To this end, the '$1D^2$' system has been introduced. Specialized adapter sequences are used to increase the likelihood that the second strand will follow after the first has passed through the pore. This allows base calling using information from both strands. The first results demonstrating the feasibility of nanopore sequencing were obtained using α-hemolysin pores (Figure IC). However, for accurate sequencing, pores with shorter sensing regions had to be developed. The first real nanopore sequencing results were obtained using the MspA pore and it currently ONT uses the CsgG pore.

ONT sequencing chemistries use the following naming scheme: versions of pores, motor enzymes, and membranes are indicated by R, E, and M numbers, respectively. Commonly however, flow cell/chemistry versions are indicated by R numbers only, without mentioning the motor enzyme and membrane version; the first chemistry released was R6 and the current chemistry is R9 (with R8 never having been released).

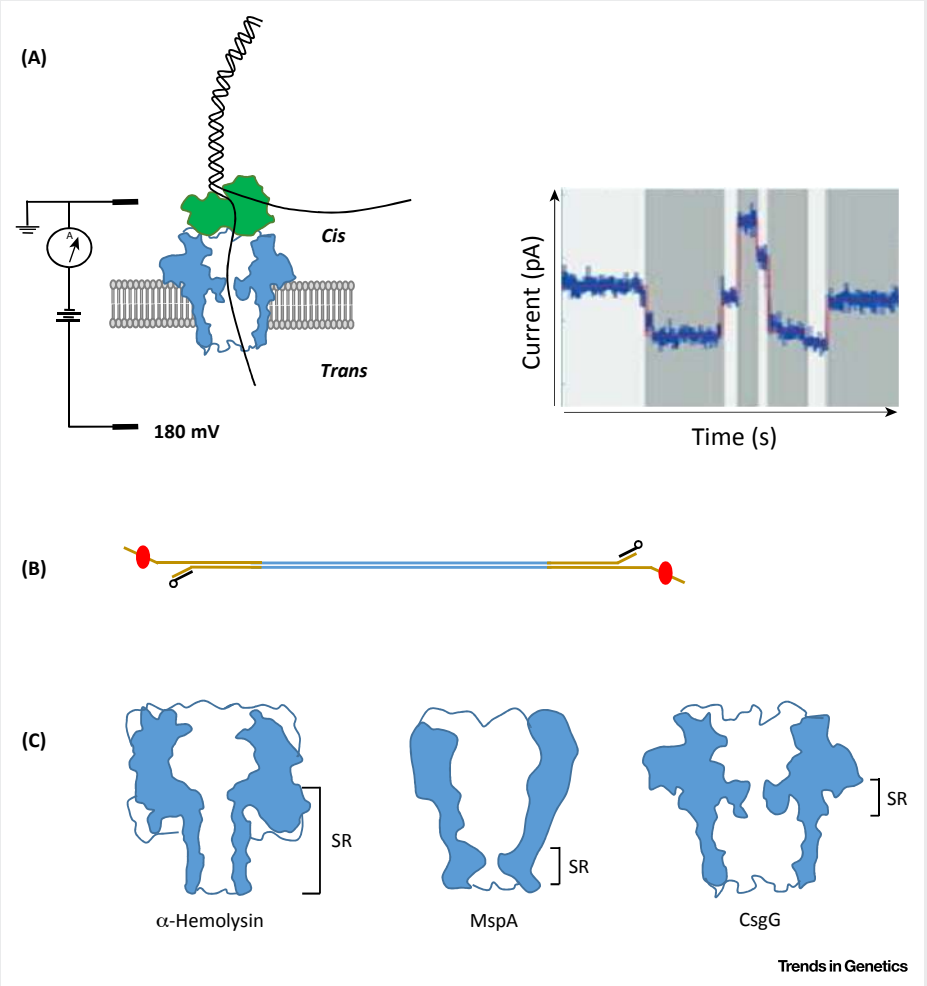

Figure I. Schematic Representation of Nanopore Technology. (A) In an Oxford Nanopore Technologies (ONT) flow cell, two chambers (*cis* and *trans*) filled with ionic solutions are separated by a membrane containing a CsgG nanopore (blue; R9 chemistry). A nucleic acid (black) is electrophoretically driven through the pore in a controlled manner owing to the presence of a 'motor' protein (green). Note that the nucleic acid is unwound on translocation and only one strand passes through the pore. As the DNA or RNA translocates through the pore, current shifts are recorded in real time and are characteristic for particular *k*-mer sequences. The current shifts are graphically represented in a 'squiggle plot'. (B) Typical nanopore library. Double-stranded DNA fragments (blue) often undergo an optional DNA repair step, as single-stranded nicks will lead to premature termination of nanopore sequencing. Then, the extremities are processed to create suitable substrates for ligation of adapters (brown). The adapters have 5′ protruding ends to which a 'motor' protein is bound (red); this extremity will enter the pore first and thus sequencing occurs in the 5′-to-3′ direction. To the other strand of the adapter, an oligonucleotide with a cholesterol moiety (black) is hybridized, which will tether the library molecules to the membrane and increase the efficiency of nanopore sequencing. (C) Different types of nanopores. The α-hemolysin pore, the MspA pore, and the CsgG pore, which is currently being used by ONT (R9 chemistry). The narrow 'sensing regions' (SRs) of the different pores are indicated; note that the MspA and CsgG pores have shorter SRs than α-hemolysin. As a result, a smaller number of nucleotides contribute to the signal, leading to more accurate base determination.

Advantages of this method are low error rates and high throughput. However, SLR library preparation requires PCR and epigenetic modifications cannot be detected directly.

SLR methods show similarities to an earlier technology, **'BAC-by-BAC' sequencing**, in which a set of overlapping bacterial artificial chromosome (BAC) clones is ordered along the chromosomes of a target genome followed by shotgun sequencing of each clone individually. There are currently two SLR systems available: the Illumina SLR platform and the 10X Genomics emulsion-based system (Box 3). While the latter system requires the purchase of additional equipment, it has two major advantages. First, it requires much less starting material (~1 ng, against 500 ng for Illumina SLR). Second, it allows a much higher level of partitioning into barcoded pools. The original GemCode instrument, introduced in 2015, already enabled partitioning into 100 000 pools, with 737 000 different barcodes. One year later, the Chromium device was released, with 1 million partitions and 4 million barcodes [27]. For comparison, Illumina SLR sequencing has only 384 partitions (wells) and indexes. As a result, local complexity is much higher, requiring much deeper sequencing. Another difference is that while DNA is sheared into ~10-kb fragments with Illumina SLR, the 10X Genomics system uses natural fragments of arbitrary size (up to ~100 kb; Table 1).

A summary of the strengths and weaknesses of the various technologies is presented in Table 2.

Long-Read Sequencing Methods Are Causing a New Revolution

While NGS methods can produce reads of up to several hundreds of base pairs (2 × 300 bp maximum read length of Illumina's MiSeq platform), long-read technologies generate reads of up to several tens of kilobases or even up to 1 Mb (nanopore). In addition, the lack of PCR amplification allows less bias and more homogeneous genome coverage. This is enormous technological progress, leading to superior performance in the analysis of repeated regions and SVs, haplotype phasing, and transcriptome analysis (Figure 1). For the first time, genome regions that remained ambiguous to date can now be resolved, and the complexity of transcriptomes can be explored in unprecedented detail. Long-read methods are thus causing a new revolution in genomics research and are leading to important new discoveries in many areas. Below we discuss some applications in more detail.

Specific Applications of Long-Read Technologies

(Whole-) Genome Sequencing

Long-read sequencing methods are frequently used to finish previous short-read assemblies. One example is the human genome, which is considered to be one of the most complete mammalian reference assemblies. However, more than 160 euchromatic gaps remain [28] that are often enriched for repeated sequences and high GC content [29]. A majority of these gaps were closed or extended using SMRT sequencing, adding more than 1 Mb of novel sequence,

Box 3. SLR Technologies

Unlike true long-read sequencing platforms, SLR technology relies on a system of barcoding to associate fragments that are sequenced on existing short-read sequencers. Currently two systems exist: the Illumina SLR sequencing platform (former Moleculo; Figure IA) and the 10X Genomics emulsion-based system (Figure Ib). With the Illumina system, genomic DNA is sheared to 8–10-kb fragments, followed by 'long-fragment' adapter ligation, which can be used to denote the extremities of contigs during downstream short-read assembly. These large fragments are then partitioned into a microtiter plate (~3000 fragments per well) and undergo further shearing and adapter addition through a tagmentation process. Each well contains a single barcode. The DNA is then pooled and subjected to classical Illumina sequencing followed by local assembly to reconstruct the original long fragments.

With 10X Genomics emulsion-based sequencing, natural DNA fragments of up to ~100 kb are mixed into micelles in an emulsion ('GEMs') together with gel beads containing adapter and barcode sequences (Figure Ib). Owing to microfluidics technology, the 10X Genomics instruments require very small amounts of starting material (~1 ng). In each GEM the gel bead dissolves and smaller fragments of DNA are amplified from the original large fragments, each with a barcode identifying the source GEM. Barcoded fragments are then pooled followed by classical Illumina library preparation and sequencing. The obtained reads are assembled to form a series of anchored fragments across a span of ~50 kb.

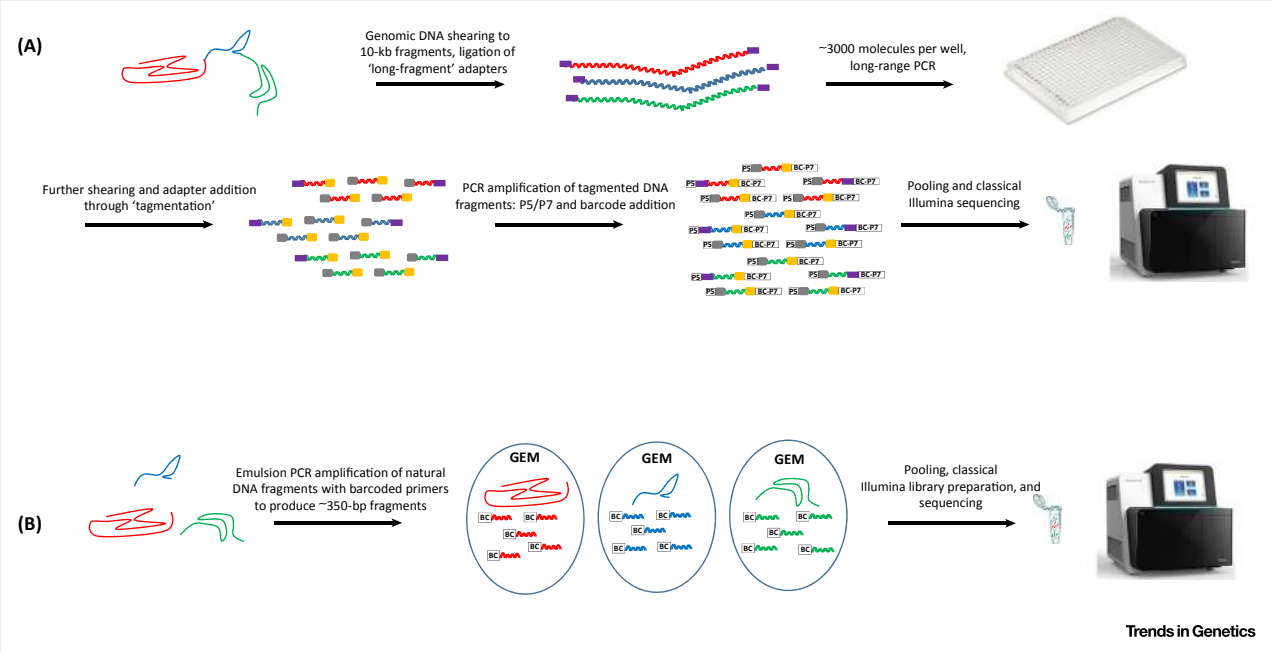

Figure I. Schematic Outline of Synthetic Long-Read Methods. (A) Workflow of Illumina's Truseq synthetic long-read technology. Genomic DNA is sheared into ~10-kb fragments, followed by ligation with 'long-fragment' adapters. The resulting fragments are diluted in a 384-well plate to about 3000 molecules per well followed by long-range PCR. In a subsequent 'tagmentation' process (Illumina's Nextera technology), the PCR products are further fragmented and adapter sequences are added simultaneously. Next, the tagmented DNA is again amplified by PCR and a unique index and the P5 and P7 adapter sequences are added in each well of the 384-well plate. These final products are pooled and sequenced as a classical Illumina library. (B) 10X Genomics' emulsion-based sequencing. The Chromium machine can partition natural DNA fragments in up to 1 million micelles ('GEMs') along with barcoded primers (4 million barcodes available) and PCR reagents. In each GEM, smaller fragments of DNA are amplified from the original larger fragment, each with a barcode identifying the source GEM. The resulting smaller fragments are pooled and sequenced as a classical Illumina library. After sequencing, the reads are aligned and linked together to form a series of anchored fragments across a span of ~50 kb ('read clouds').

and tens of thousands of structural variants were resolved [30,31]. Thus, SMRT is a great tool to resolve extremely repetitive and GC-rich regions, which is also illustrated by the fact that more than 2.25-kb-long stretches of CGG **short tandem repeats (STPs)** implicated in fragile X syndrome (FXS) have been resolved using SMRT sequencing [29].

Initially, the low throughput of the MinION limited its use to the sequencing and assembly of small bacterial genomes [32,33]. More recently, assemblies of larger genomes have been reported, such as those of *Caenorhabditis elegans* [34] and human [24]. In the latter study, extremely long reads of up to 882 kb were obtained. Comparative studies suggest that SMRT and nanopore technologies perform similarly well for *de novo* genome assembly [35,36]. However, SMRT sequencing of the human genome in two independent studies allowed the closure of 50 [30] and 105 [31] euchromatic gaps, both adding more than 1 Mb of novel sequence at 40- and 101-fold coverage, respectively. By contrast, nanopore sequencing closed 12 gaps, adding ~84 kb of novel sequence. It should be noted, however, that this was at lower coverage (30-fold), with an additional fivefold coverage with ultralong reads. It should be noted also that the ultralong nanopore reads enabled measurement of telomere repeat length, which is not possible with the shorter SMRT reads.

Table 2. Pros and Cons of the Various Long-Read Platforms

Platform	Pros	Cons
PacBio	Real long reads Extremely high accuracy with CCS (>99.999% at 20 passes) Direct detection of epigenetic modifications; also here, high level of accuracy with CCS No problem with repeats, low/high %GC	Expensive sequencer (Sequel list price: US$350 000) and relatively high cost per Gb Large amounts of starting material required for library preparation High error rate at single pass (~15%) Only one sequencer available (Sequel) with limited throughput per SMRT Cell (~10 Gb) Maximum read length limited by polymerase processivity (~80 kb)
ONT	Real (ultra-) long reads; in principle no upper limit to read length (~1-Mb reads have been obtained) Cost-effective sequencers (MinION, GridION) Direct detection of epigenetic modifications Extremely fast library preparation ('rapid sequencing kit') Portability (MinION and SmidgION) Scalability; from extremely small and portable (SmidgION, MinION) to extremely powerful and high throughput (PromethION) Direct sequencing of RNA and detection of RNA modifications	High overall error rate (1D: ~15%, $1D^2$: ~3%[a]) and systematic errors with homopolymers Large amount of starting material required for library preparation Frequent changes of software versions, flow cells, and kits
Illumina/10X Genomics SLR	Based on classical Illumina sequencing; low error rates, high throughput Relatively low cost per Gb Illumina SLR: no specialized equipment required 10X Genomics: small amounts of material required for library preparation (~1 ng)	No real long reads PCR amplification required for library preparation No direct detection of epigenetic modifications Illumina SLR: limited partitioning capacity (384 wells and indexes) Illumina SLR: 10 kb maximum SLR length

[a]Although $1D^2$ reduces error rates, it also reduces throughput and it is not very efficient.

Thus, a particular strength of nanopore ultralong reads is the resolution of extremely long repeats that cannot be resolved with any other technology. As an example, even the most complete human chromosome assemblies to date lack centromeric sequences that comprise essentially hundreds or thousands of repeats of AT-rich ~171-bp 'alpha satellite' DNA monomers, with more than 99% sequence conservation [37]. Recently, Jain *et al.* succeeded in producing nanopore reads long enough to cover the hundreds-of-kilobase-long centromeric sequences of the Y chromosome [38]. While individual reads had too many errors to resolve the centromeric sequences, 60-fold coverage with full-length 1D reads followed by polishing steps with Illumina data resulted in a high level of accuracy (>99% alignment identity).

Illumina SLR can resolve certain types of repetitive elements [39] but have difficulties in resolving more tandemly arranged repetitive sequences [40]. The 10X Genomics system is more powerful in resolving complex genomes [16,41,42] but this system remains limited by the fact that it relies on the (local) assembly of short reads and the requirement for PCR. As a result, this technique is likely to be less suitable for sequencing highly repetitive regions, especially those of

extreme GC content. This technique has thus mainly been applied for **genome phasing**, where the high level of accuracy is clearly an advantage in phasing SNPs [16,42,43]. Also, the higher throughput and cost-effectiveness may make this method suitable for large-scale genome variation studies [27].

Other genomic methods that are often used to complement or validate sequencing results in genome assembly are discussed in Box 4.

RNA-Seq

Short-read RNA-seq methods are frequently used for gene expression profiling. However, due to the short reads it is difficult to infer the combinations of splice site usage in individual transcripts. Reconstructing transcripts from short reads often leads to inaccurate annotations that may lack terminal exons or splice junctions between exons [48]. This is a major limitation given the importance of alternative splicing in many physiological and developmental processes [49]. By producing reads that cover entire transcripts, SMRT sequencing provides superior evidence for alternative splicing, can improve the accuracy of existing gene models, and enables phasing of transcripts to the alleles from which they were transcribed [50–52]. The PacBio workflow to study alternative splicing is commonly referred to as 'Isoseq' [53].

A comparative study suggests that, despite its higher error rates, nanopore sequencing may perform as well as SMRT [54]. In addition, it offers the unique possibility of directly sequencing RNA [55]. Recently, direct RNA-seq was used to detect modified nucleotides in bacterial 16S ribosomal RNA, which could be used for rapid identification of microbes in environmental and clinical settings [56].

Tilgner *et al.* described an RNA-seq method called SLR-RNA-seq [57] based on Illumina SLR. A potential concern with this approach is that, for highly expressed genes, different RNA isoforms might be present in the same microtiter well and eventually be assembled into a false-positive novel isoform. In addition, there is the need for PCR amplification, which may introduce biases. More recently, a more powerful SLR-RNA-seq method based on 10X Genomics microfluidics was described [58].

Detection of Epigenetic Modifications

DNA base modifications are of major biological significance [59] but have been difficult to study. A common NGS approach, bisulfite sequencing, can detect C modification only, on treatment of DNA with bisulfite; it cannot discriminate between 5mC and 5hmC and requires a reference

Box 4. Other Genomics Technologies

Other genomic methods are often used to complement or validate sequencing results in genome assembly. One example is **nanochannel genome mapping**. This technology, commercialized by BioNano Genomics, uses fluorescent markers to tag particular sequences in long (>100 kb) DNA fragments. The results are imaged and aligned to each other, and/or to a reference, to map the locations of the fluorescent signals relative to each other [44]. Such optical maps improve *de novo* genome assemblies and/or genome phasing by providing a long-range scaffold on which to align sequencing reads. This technology was used to anchor contigs obtained with SMRT sequencing (**N50** length of 17.9 Mb) into scaffolds with an N50 size of 44.8 Mb [31].

Another example is **chromosome conformation capture** [45]. This technique provides 3D-proximity information of genomic loci through chromosome conformation capture sequencing (Hi-C). As the spatial proximity of genomic loci is highly dependent on their distances in the linear genome [31], this method can be used to order and orient contigs and scaffolds obtained by (long read) sequencing. Recent examples are a study by Mascher *et al.* [46], who used Hi-C to generate an improved reference sequence for the barley genome, and a study that used both optical mapping and Hi-C to improve genome assembly [47].

genome [60]. By contrast, SMRT and nanopore sequencing directly detect base modifications on native DNA without the need for a reference genome. In SMRT sequencing, the kinetics of base addition is measured during sequencing (Box 1). These kinetic measurements present characteristic patterns ('fingerprint') in response to over 25 different types of base modification [60]. Importantly, SMRT sequencing led to the recent discovery of 6 mA in *C. elegans* and in mouse embryonic stem cells [61,62]. The m5C modification, ubiquitous in mammals, has a less pronounced effect on polymerase kinetics [63] but can nevertheless be accurately detected, provided coverage is high enough [64].

Recently, it was shown that nanopore sequencing can also detect 5mC and 6 mA [65,66]. In contrast to SMRT sequencing, it has higher sensitivity for 5mC than for 6 mA [65]. Euskirchen *et al.* demonstrated the potential of 5mC methylation profiling by nanopore sequencing in molecular diagnostics of brain tumors [67]. This study illustrates the power of this technology for rapid classification of tumors with minimal capital cost to inform diagnosis, prognosis, and treatment decisions.

Concluding Remarks and Future Perspectives

Over recent years, long-read sequencing methods have strongly improved. These technologies now enable the study of genomes and transcriptomes at an unprecedented resolution. Also, metagenomics analyses benefit from long-read sequencing, which allows for the first time the resolution of microbial communities at the species level [68–70]. Long-read sequencing is likely to become a standard medical diagnostic tool in the near future, as exemplified by a recent SMRT sequencing study of a patient's genome revealing a SV that could not be detected despite extensive genetic testing with other methods [71].

In particular, nanopore sequencing has improved rapidly. A theoretical 1× coverage of the *Escherichia coli* genome was obtained with just seven ultralong reads (http://lab.loman.net/2017/03/09/ultrareads-for-nanopore/) and a human genome has been assembled using nanopore reads alone [24]. Ultralong nanopore reads may allow complete, gapless assembly of human genomes in the near future, which will further boost human genetics research and personalized medicine. The portability of the MinION allows for the first time sequencing in the field, which is of great importance for the survey of outbreaks in developing countries [72,73].

However, there remains room for improvement. A weakness of nanopore sequencing is the high error rate. In 2010, Stoddart *et al.* proposed the development of nanopores with multiple recognition points for DNA sequence determination [74]. This would provide a proofreading mechanism improving the overall quality of sequencing. As an alternative solution to reduce error rates, a method resembling PacBio CCS has been proposed [26]. On the other hand, to keep up with nanopore technology it will be important for PacBio to increase overall read length and throughput. Current loading methods depend on passive diffusion and are biased towards shorter fragments. A novel, voltage-induced loading technique increases the efficiency of loading long DNA molecules [75]. However, it seems unlikely that SMRT sequencing will approach the ultralong reads currently obtained with nanopores, due to the limitation of polymerase processivity. Thus, SMRT, nanopore, and SLR sequencing methods each have their particular strengths and weaknesses (Table 2), and depending on the specific application either one technology or another may be preferred.

It is worth mentioning here that various other companies are also investing in novel methods for rapid, cost-effective, and portable sequencing and it will be interesting to see whether any of these technologies will see light in the near future (see Outstanding Questions).

Outstanding Questions

While nanopore technology has several advantages over SMRT sequencing (ultralong reads, cost-effectiveness, portability), a major drawback is the higher overall error rate due to the lack of CCS. Will ONT succeed in solving this problem?

Although PacBio's Sequel has a 100-fold-higher throughput per run than the first PacBio RSI (10 Gb versus 0.1 Gb), a single MinION run can now produce twice this amount of data. To keep up with ONT, it will thus be important for PacBio to significantly further increase its throughput in the near future. Will they succeed in doing so or will this technology reach its limits?

Currently, the short-read NGS market is dominated by Illumina and the long-read market is dominated by PacBio, ONT, and 10X Genomics. However, various other companies are developing novel methods for rapid, cost-effective, and portable sequencing. For example, Genapsys is developing an iPad-sized portable sequencer (http://www.genapsys.com), Roche is investing in the development of nanopore sequencing (http://sequencing.roche.com/en/technology-research/technology/nanopore-sequencing.html), Stratos Genomics is developing a 'fourth-generation' sequencing method called 'sequencing by expansion' (SBX) (http://www.stratosgenomics.com/), and GnuBio (Bio Rad) is developing a sequencer based on emulsion microfluidics (http://gnubio.com/). Will any of these technologies see light in the coming years?

While rapid and powerful methods now exist for the direct single-molecule analysis of genomes and transcriptomes, high-throughput proteomics still relies on mass spectrometry methods that require large amounts of material and are currently limited to short proteins (<30 kDa). In principle, nanopores could be used to sequence denatured proteins of any length by translocation. Will methods appear in the near future that allow single-molecule protein sequencing?

Recently, SMRT technology was used to sequence the genome of a patient who suffered recurrent benign tumors

Last, an exciting possibility of nanopore technology is the sequencing of denatured peptide chains, and recent results confirm its feasibility [76]. It will be interesting to see whether further progress will be made in the future to make single-molecule protein sequencing a reality.

In any case, we are at only the beginning of the third revolution in sequencing technology and the coming years promise to bring exciting new developments and discoveries.

Acknowledgments

This work was supported by the National Center for Scientific Research (CNRS), the French Alternative Energies and Atomic Energy Commission (CEA), and Paris-Sud University.

References

1. Maxam, A.M. and Gilbert, W. (1977) A new method for sequencing DNA. *Proc. Natl. Acad. Sci. U. S. A.* 74, 560–564
2. Sanger, F. *et al.* (1977) DNA sequencing with chain-terminating inhibitors. *Proc. Natl. Acad. Sci. U. S. A.* 74, 5463–5467
3. Venter, J.C. *et al.* (2001) The sequence of the human genome. *Science* 291, 1304–1351
4. Lander, E.S. *et al.* (2001) Initial sequencing and analysis of the human genome. *Nature* 409, 860–921
5. Schloss, J.A. (2008) How to get genomes at one ten-thousandth the cost. *Nat. Biotechnol.* 26, 1113–1115
6. van Dijk, E.L. *et al.* (2014) Ten years of next-generation sequencing technology. *Trends Genet.* 30, 418–426
7. Goodwin, S. *et al.* (2016) Coming of age: ten years of next-generation sequencing technologies. *Nat. Rev. Genet.* 17, 333–351
8. Salzberg, S.L. and Yorke, J.A. (2005) Beware of mis-assembled genomes. *Bioinformatics* 21, 4320–4321
9. Weischenfeldt, J. *et al.* (2013) Phenotypic impact of genomic structural variation: insights from and for human disease. *Nat. Rev. Genet.* 14, 125–138
10. Tewhey, R. *et al.* (2011) The importance of phase information for human genomics. *Nat. Rev. Genet.* 12, 215–223
11. Schadt, E.E. *et al.* (2010) A window into third-generation sequencing. *Hum. Mol. Genet.* 19, R227–R240
12. Pushkarev, D. *et al.* (2009) Single-molecule sequencing of an individual human genome. *Nat. Biotechnol.* 27, 847–850
13. Eid, J. *et al.* (2009) Real-time DNA sequencing from single polymerase molecules. *Science* 323, 133–138
14. Jain, M. *et al.* (2015) Improved data analysis for the MinION nanopore sequencer. *Nat. Methods* 12, 351–356
15. Voskoboynik, A. *et al.* (2013) The genome sequence of the colonial chordate, *Botryllus schlosseri*. *Elife* 2, e00569
16. Zheng, G. *et al.* (2016) Haplotyping germline and cancer genomes with high-throughput linked-read sequencing. *Nat. Biotechnol.* 34, 303–311
17. Sedlazeck, F.J. *et al.* (2018) Piercing the dark matter: bioinformatics of long-range sequencing and mapping. *Nat. Rev. Genet.* 19, 329–346
18. Chu, J. *et al.* (2017) Innovations and challenges in detecting long read overlaps: an evaluation of the state-of-the-art. *Bioinformatics* 33, 1261–1270
19. Quail, M.A. *et al.* (2012) A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers. *BMC Genomics* 13, 341
20. Travers, K.J. *et al.* (2010) A flexible and efficient template format for circular consensus sequencing and SNP detection. *Nucleic Acids Res.* 38, e159
21. Deamer, D. *et al.* (2016) Three decades of nanopore sequencing. *Nat. Biotechnol.* 34, 518–524
22. Manrao, E.A. *et al.* (2012) Reading DNA at single-nucleotide resolution with a mutant MspA nanopore and phi29 DNA polymerase. *Nat. Biotechnol.* 30, 349–353
23. Brown, C.G. and Clarke, J. (2016) Nanopore development at Oxford Nanopore. *Nat. Biotechnol.* 34, 810–811
24. Jain, M. *et al.* (2018) Nanopore sequencing and assembly of a human genome with ultra-long reads. *Nat. Biotechnol.* 36, 338–345
25. Jain, M. *et al.* (2017) MinION Analysis and Reference Consortium: phase 2 data release and analysis of R9.0 chemistry. *F1000Res.* 6, 760
26. Li, C. *et al.* (2016) INC-seq: accurate single molecule reads using nanopore sequencing. *Gigascience* 5, 34
27. Marks, P. *et al.* (2017) Resolving the full spectrum of human genome variation using linked-reads. *BioRxiv* Published online December 8, 2017. http://dx.doi.org/10.1101/230946
28. Genovese, G. *et al.* (2013) Using population admixture to help complete maps of the human genome. *Nat. Genet.* 45, 406–414
29. Schmidt, M.H. and Pearson, C.E. (2016) Disease-associated repeat instability and mismatch repair. *DNA Repair (Amst.)* 38, 117–126
30. Chaisson, M.J. *et al.* (2015) Resolving the complexity of the human genome using single-molecule sequencing. *Nature* 517, 608–611
31. Seo, J.S. *et al.* (2016) *De novo* assembly and phasing of a Korean human genome. *Nature* 538, 243–247
32. Quick, J. *et al.* (2015) Rapid draft sequencing and real-time nanopore sequencing in a hospital outbreak of *Salmonella*. *Genome Biol.* 16, 114
33. Loman, N.J. *et al.* (2015) A complete bacterial genome assembled de novo using only nanopore sequencing data. *Nat. Methods* 12, 733–735
34. Tyson, J.R. (2018) MinION-based long-read sequencing and assembly extends the *Caenorhabditis elegans* reference genome. *Genome Res.* 28, 266–274
35. Michael, T.P. *et al.* (2018) High contiguity *Arabidopsis thaliana* genome assembly with a single nanopore flow cell. *Nat. Commun.* 9, 541
36. Giordano, F. *et al.* (2017) *De novo* yeast genome assemblies from MinION, PacBio and MiSeq platforms. *Sci. Rep.* 7, 3935
37. Rudd, M.K. and Willard, H.F. (2004) Analysis of the centromeric regions of the human genome assembly. *Trends Genet.* 20, 529–533
38. Jain, M. *et al.* (2017) Linear assembly of a human Y centromere using nanopore long reads. *BioRxiv* Published online July 21, 2017. http://dx.doi.org/10.1101/170373
39. McCoy, R.C. *et al.* (2014) Illumina TruSeq synthetic long-reads empower de novo assembly and resolve complex, highly-repetitive transposable elements. *PLoS One* 9, e106689
40. Li, R. *et al.* (2015) Illumina synthetic long read sequencing allows recovery of missing sequences even in the "finished" *C. elegans* genome. *Sci. Rep.* 5, 10814
41. Hulse-Kemp, A.M. *et al.* (2018) Reference quality assembly of the 3.5-Gb genome of *Capsicum annuum* from a single linked-read library. *Hortic. Res.* 5, 4

in his heart and glands. The individual satisfied the clinical criteria for the Carney complex, but after 8 years of genetic evaluation, including whole-genome short-read sequencing, experts were still unable to pinpoint the underlying genetic mutation and confirm a diagnosis. SMRT sequencing identified a 2.2-kb deletion affecting *PRKAR1A*, the gene involved in the Carney complex. This case highlights the potential of the application of long-read sequencing to precision medicine. Will this become routine in the clinic in the near future?

42. Porubsky, D. *et al.* (2017) Dense and accurate whole-chromosome haplotyping of individual genomes. *Nat. Commun.* 8, 129
43. Hui, W.W. *et al.* (2017) Universal haplotype-based noninvasive prenatal testing for single gene diseases. *Clin. Chem.* 63, 513–524
44. Lam, E.T. *et al.* (2012) Genome mapping on nanochannel arrays for structural variation analysis and sequence assembly. *Nat. Biotechnol.* 30, 771–776
45. Burton, J.N. *et al.* (2013) Chromosome-scale scaffolding of *de novo* genome assemblies based on chromatin interactions. *Nat. Biotechnol.* 31, 1119–1125
46. Mascher, M. *et al.* (2017) A chromosome conformation capture ordered sequence of the barley genome. *Nature* 544, 427–433
47. Bickhart, D.M. *et al.* (2017) Single-molecule sequencing and chromatin conformation capture enable *de novo* reference assembly of the domestic goat genome. *Nat. Genet.* 49, 643–650
48. Steijger, T. (2013) Assessment of transcript reconstruction methods for RNA-seq. *Nat. Methods* 10, 1177–1184
49. Kalsotra, A. and Cooper, T.A. (2011) Functional consequences of developmentally regulated alternative splicing. *Nat. Rev. Genet.* 12, 715–729
50. Tilgner, H. *et al.* (2014) Defining a personal, allele-specific, and single-molecule long-read transcriptome. *Proc. Natl. Acad. Sci. U. S. A.* 111, 9869–9874
51. Wang, B. *et al.* (2016) Unveiling the complexity of the maize transcriptome by single-molecule long-read sequencing. *Nat. Commun.* 7, 11708
52. Lagarde, J. *et al.* (2017) High-throughput annotation of full-length long noncoding RNAs with capture long-read sequencing. *Nat. Genet.* 49, 1731–1740
53. Sharon, D. *et al.* (2013) A single-molecule long-read survey of the human transcriptome. *Nat. Biotechnol.* 31, 1009–1014
54. Weirather, J.L. *et al.* (2017) Comprehensive comparison of Pacific Biosciences and Oxford Nanopore Technologies and their applications to transcriptome analysis. *F1000Res.* 6, 100
55. Garalde, D.R. *et al.* (2018) Highly parallel direct RNA sequencing on an array of nanopores. *Nat. Methods* 15, 201–206
56. Smith, A.M. *et al.* (2017) Reading canonical and modified nucleotides in 16S ribosomal RNA using nanopore direct RNA sequencing. *BioRxiv* Published online April 29, 2017. http://dx.doi.org/10.1101/132274
57. Tilgner, H. *et al.* (2015) Comprehensive transcriptome analysis using synthetic long-read sequencing reveals molecular co-association of distant splicing events. *Nat. Biotechnol.* 33, 736–742
58. Tilgner, H. *et al.* (2018) Microfluidic isoform sequencing shows widespread splicing coordination in the human transcriptome. *Genome Res.* 28, 231–242
59. Bierne, H. *et al.* (2012) Epigenetics and bacterial infections. *Cold Spring Harb. Perspect. Med.* 2, a010272
60. Flusberg, B.A. *et al.* (2010) Direct detection of DNA methylation during single-molecule, real-time sequencing. *Nat. Methods* 7, 461–465
61. Greer, E.L. *et al.* (2015) DNA methylation on N^6-adenine in *C. elegans*. *Cell* 161, 868–878
62. Wu, T.P. *et al.* (2016) DNA methylation on N^6-adenine in mammalian embryonic stem cells. *Nature* 532, 329–333
63. Murray, I.A. *et al.* (2012) The methylomes of six bacteria. *Nucleic Acids Res.* 40, 11450–11462
64. Ichikawa, K. *et al.* (2017) Centromere evolution and CpG methylation during vertebrate speciation. *Nat. Commun.* 8, 1833
65. Rand, A.C. *et al.* (2017) Mapping DNA methylation with high-throughput nanopore sequencing. *Nat. Methods* 14, 411–413
66. Simpson, J.T. *et al.* (2017) Detecting DNA cytosine methylation using nanopore sequencing. *Nat. Methods* 14, 407–410
67. Euskirchen, P. *et al.* (2017) Same-day genomic and epigenomic diagnosis of brain tumors using real-time nanopore sequencing. *Acta Neuropathol.* 134, 691–703
68. Pootakham, W. *et al.* (2017) High resolution profiling of coral-associated bacterial communities using full-length 16S rRNA sequence data from PacBio SMRT sequencing system. *Sci. Rep.* 7, 2774
69. Shin, J. *et al.* (2016) Analysis of the mouse gut microbiome using full-length 16S rRNA amplicon sequencing. *Sci. Rep.* 6, 29681
70. Kuleshov, V. *et al.* (2016) Synthetic long-read sequencing reveals intraspecies diversity in the human microbiome. *Nat. Biotechnol.* 34, 64–69
71. Merker, J.D. *et al.* (2018) Long-read genome sequencing identifies causal structural variation in a Mendelian disease. *Genet. Med.* 20, 159–163
72. Quick, J. *et al.* (2016) Real-time, portable genome sequencing for Ebola surveillance. *Nature* 530, 228–232
73. Faria, N.R. *et al.* (2016) Mobile real-time surveillance of Zika virus in Brazil. *Genome Med.* 8, 97
74. Stoddart, D. *et al.* (2010) Multiple base-recognition sites in a biological nanopore: two heads are better than one. *Angew. Chem. Int. Ed. Engl.* 49, 556–559
75. Larkin, J. *et al.* (2017) Length-independent DNA packing into nanopore zero-mode waveguides for low-input DNA sequencing. *Nat. Nanotechnol.* 12, 1169–1175
76. Kolmogorov, M. *et al.* (2017) Single-molecule protein identification by sub-nanopore sensors. *PLoS Comput. Biol.* 13, e1005356