Briefings in Bioinformatics, 2017, 1–17

doi: 10.1093/bib/bbx062
Paper

Nanopore sequencing data analysis: state of the art, applications and challenges

Alberto Magi, Roberto Semeraro, Alessandra Mingrino, Betti Giusti, and Romina D'Aurizio

Corresponding author: Alberto Magi, Department of Experimental and Clinical Medicine, University of Florence, Largo Brambilla, 3 - 50134 Florence, Italy. Tel.: +39 055 7948909; Fax: +39 055 7948909; E-mail: albertomagi@gmail.com

Abstract

The nanopore sequencing process is based on the transit of a DNA molecule through a nanoscopic pore, and since the 90s is considered as one of the most promising approaches to detect polymeric molecules. In 2014, Oxford Nanopore Technologies (ONT) launched a beta-testing program that supplied the scientific community with the first prototype of a nanopore sequencer: the MinION. Thanks to this program, several research groups had the opportunity to evaluate the performance of this novel instrument and develop novel computational approaches for analyzing this new generation of data.

Despite the short period of time from the release of the MinION, a large number of algorithms and tools have been developed for base calling, data handling, read mapping, *de novo* assembly and variant discovery. Here, we face the main computational challenges related to the analysis of nanopore data, and we carry out a comprehensive and up-to-date survey of the algorithmic solutions adopted by the bioinformatic community comparing performance and reporting limits and advantages of using this new generation of sequences for genomic analyses.

Our analyses demonstrate that the use of nanopore data dramatically improves the *de novo* assembly of genomes and allows for the exploration of structural variants with an unprecedented accuracy and resolution. However, despite the impressive improvements reached by ONT in the past 2 years, the use of these data for small-variant calling is still challenging, and at present, it needs to be coupled with complementary short sequences for mitigating the intrinsic biases of nanopore sequencing technology.

Key words: nanopore; resequencing; de novo assembly

Introduction

Over the past 10 years, second-generation sequencing (SGS) platforms [1], combined with original computational approaches, have revolutionized biological and biomedical research paving the way for a new era for personal genomics and enabling the sequencing of full human genomes quickly and at affordable prices [2].

Although SGS technology has completely changed our ability to study the genetic variation of any organism, it is apparent that the short reads (100–500 bp) generated by these platforms are insufficient to resolve complex genomic structures such as long repetitive elements, copy number alterations and structural variations that are relevant for studying and understanding evolution, adaptation and disease. The past few years have seen the emergence of a third generation of sequencing (TGS)

Alberto Magi, PhD, is an assistant professor at the University of Florence, Italy. His research interests focus on the development of computational methods for the identification of genomic variants.
Roberto Semeraro, PhD, is a postdoctoral researcher at the University of Florence, Italy. His research is focused on computational methods for the analysis of NGS data.
Alessandra Mingrino, PhD, is a postdoctoral researcher at the University of Florence, Italy. Her research is focused on the development of novel experimental strategies with third-generation sequencing data.
Betti Giusti, PhD, is an associate professor of Clinical Pathology at the University of Florence. Her research activity has focused on genetics of cardiovascular diseases and extracellular matrix disorders.
Romina D'Aurizio, PhD, is a postdoctoral researcher at the Italian National Research Council of Pisa. Her research is focused on computational methods for the analysis of high-throughput sequencing data.
Submitted: 26 February 2017; **Received (in revised form):** 10 April 2017

technologies based on single-molecule real-time (SMRT) [3] and nanopore sequencing [4], which interrogate single molecule of DNA and are capable to produce sequences much longer than those generated by SGS methods. While the SMRT approach is based on directly observing a single molecule of DNA polymerase, as it synthetizes a strand of DNA, the basic principle of nanopore sequencing is the transit of a DNA molecule through a nanoscopic pore and the contemporary measurement of its effect on an electric current.

The idea of using nanopores as biosensors was originally conceived in the 90s by Deamer and Akeson [5], and since the first paper published in Proceedings of the National Academy of Sciences (PNAS) in 1996 [6], the nanopore-based detection of polymeric molecules has become as one of the most promising sequencing approaches. Nanopore biosensors can be classified into two main categories that include biological and solid-state pores, and several papers have demonstrated that both types of nanopores are able to detect biological and chemical molecules at single-molecule level [7–9].

Solid-state nanopores can be fabricated from many materials (i.e. silicon, aluminum, boron, graphene and hybrid materials) by using semiconductor production processes [9], and because of their excellent chemico-physical properties, they can work in a wide variety of experimental conditions and allow for DNA sequencing and protein detection [10].

On the other hand, biological nanopores are transmembrane protein channels embedded in a matrix (i.e. lipid bilayers, liposomes or other polymer films) that are naturally produced by bacteria and can be genetically engineered by using molecular biology techniques that change the amino acid residue at a specific site. At present, four transmembrane protein channels have been widely tested as biosensor, and these include the α-hemolysin (exotoxin secreted by the bacterium *Staphylococcus aureus*), the *Mycobacterium smegmatis* porin A (MspA), the Curlin sigma S-dependent growth (CsgG) and bacteriophage ϕ 29 pores. While the small diameter of α-hemolysin and MspA only permits the sequencing of single-stranded DNA (ssDNA), the ϕ29 pore is capable of sensing larger molecules like double-stranded DNA (dsDNA), complexes of DNA and proteins.

The translocation of DNA through a nanopore is a drift-diffusion process with directed and random motions that needs to be controlled for allowing nucleotide recognition. Between 2005 and 2010, several groups demonstrated that coupling an enzyme motor to the nanopore is a successful strategy for controlling DNA translocation speed that can be lowered to a rate at which single-nucleotide resolution is feasible [11].

Thanks to these proof-of-concept studies, in 2012, the company Oxford Nanopore Technologies (ONT, https://www.nanoporetech.com) announced the first high-throughput (biological) nanopore sequencing platform, the MinION, and in April 2014 launched the MinION Access Programme (MAP), an independent beta-testing program for a developer community made of >1000 laboratories (https://www.nanoporetech.com/community/the-minion-access-programme). Thanks to the MAP, several research groups had the opportunity to evaluate the base throughput, read quality and the global performance of this novel instrument and use it for several applications that range from bacterial/viral genome assembly to cancer variant discovery and transcript isoform identification [12–15]. The MinION is a pocket-sized (90 g and 10 cm in length) device that is able to generate DNA sequences with an average length of 2–10 kb and an error rate in the range 15–40% that challenges the use of available bioinformatics methods originally designed for SGS data [16].

For this reason, the MAP community started to develop and/or adapt computational approaches for managing and analyzing these ultra-long and high error-prone sequences and, despite the short period of time from the beginning of the MAP, a large number of algorithms have been published for base calling, data handling, read mapping, *de novo* assembly and variant discovery.

In this manuscript, we describe the main characteristics of the MinION device and chemistry, and we face the computational challenges related to the analysis of nanopore data by carrying out a comprehensive and up-to-date survey of the algorithmic solutions adopted by the bioinformatic community. Moreover, by combining results from all available studies, we depict a comprehensive review that shows the algorithmic and technological limits and advantages of using this new generation of sequencing data for different genomic applications.

The MinION sequencer

The ONT MinION is the first commercially available device that uses nanopore as biosensor to sequence long (10 kb on average and up to hundreds of kb) ssDNA molecule. Recently, Clive Brown, ONT's Chief Technology Officer, disclosed that the nanopore used in the earliest MinION chemistry was a genetically engineered α-hemolysin, while the latest chemistry is based on mutants of the CsgG nanopore [17].

The current MinION flow cells, the cartridges where DNA libraries are inserted, contain 2048 individual protein nanopores arranged in 512 channels, each containing four pores and sensors. Each channel measures the current signal generated by one of the four pores at once allowing to process up to 512 DNA molecules simultaneously.

To sequence both DNA strands, dsDNA is prepared by adding to its ends two adapters (leader and hairpin adapters) preloaded with motor proteins [18]. The leader adapter guides the dsDNA fragments to the vicinity of pores, and the sequencing process begins when the leader motor protein unzips the dsDNA enabling the first strand (template) to pass through the nanopore one base at a time (Figure 1A). At the end of the template strand, the hairpin motor protein mediates the movement of the complement strand through the pore in a similar fashion.

When the DNA strand passes through the nanopore, a sensor measures ionic current changes with a constant sampling frequency (Figure 1B), and the change between template and complement strand is recognized by the pore via the specific signal generated by an apurinic/apyrimidinic site located in the hairpin [29]. Raw current data are then subjected to base calling by means of a machine learning approach (recurrent neural network, RNN, or hidden Markov model, HMM) to obtain a consensus sequence for template and complement strands separately (Figure 1C). When the two strands of the molecule are read successfully, a consensus is built to obtain a more accurate read (called the 2D read). Otherwise, only the template strand sequence is provided (called the 1D read, Figure 1D).

Over the past 2 years, ONT released five different MinION chemistry versions (R6.0, R7.0, R7.3, R9 and R9.4) and several updates of the base-calling software that strongly improved the performance of the MinION device (Figure 1E–G). While the earliest R6.0 chemistry version allowed to generate a total yield in the order of tens of Mb and 2D reads with 4 kb of median length and 40% of the average error rate, the more recent R7.0 and R7.3 improved the throughput to hundreds of Mb and 2D read size and accuracy to 6 kb and 80–85%, respectively.

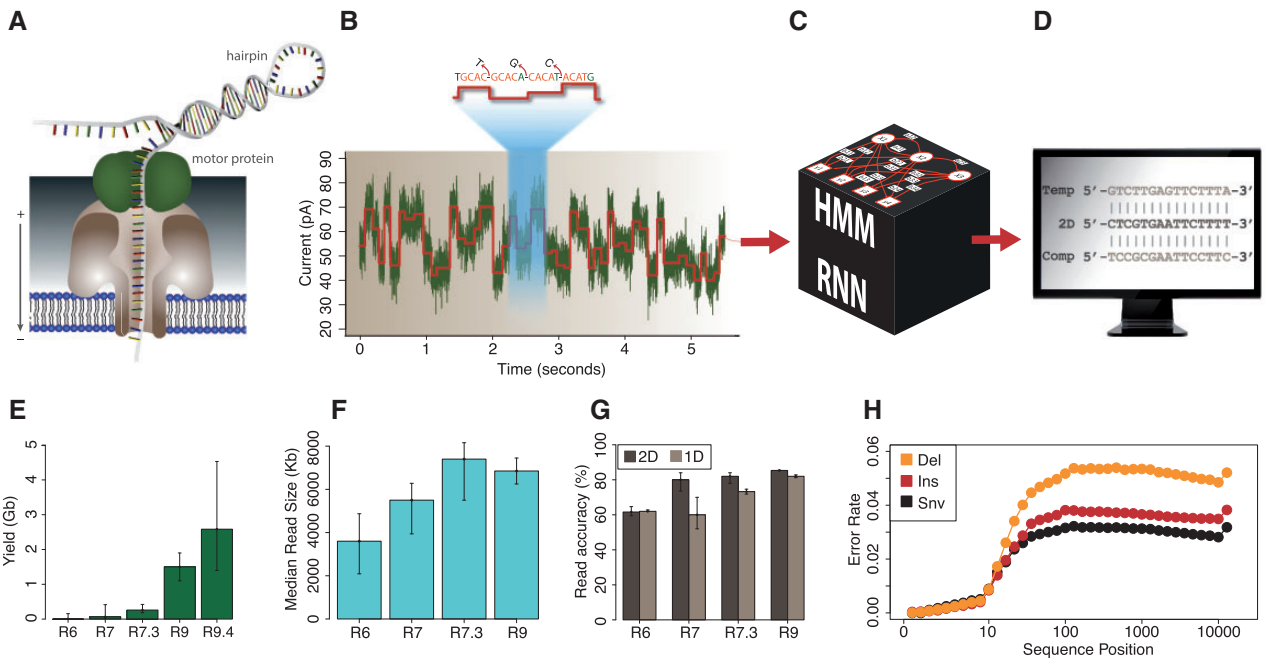


Figure 1. The Oxford Nanopore sequencing process and performance. The motor protein unwinds the dsDNA allowing ssDNA to pass through the pore, while a sensor measures ionic current shifts (**A**) with a constant sampling frequency (at present 5000 Hz). Raw ionic current signals are segmented into discrete 'events' summarized by mean, SD and length (**B**). Segmented events are then analyzed with machine learning approach (black box, **C**) that outputs the sequence of template and complement signals (**D**). Panels E, f and g show performance evolution of ONT chemistry for throughput, read length and accuracy. Panel H shows sequencing error rate as a function of read position as calculated in [16]. All the results reported in Panels E, f and g were manually parsed from [19–28]. Yield statistics for R9.4 flow cells were obtained from https://github.com/nanopore-wgs-consortium/NA12878 and https://github.com/nanoporetech/ONT-HG1.

In May 2016, ONT introduced the new R9 chemistry based on the protein nanopore CsgG, a novel algorithm for base calling that uses deep learning and the 'fast mode' that allows to sequence 250 bases/second (instead of the previous 75 bases/second) with a sampling frequency of 5000 Hz. Although at present few experiments performed with the R9 chemistry are publicly available, the results reported by Istace *et al.* [19] and by Loman (http://lab.loman.net/2016/07/30/nanopore-r9-data-release/) demonstrate an impressive improvement in total throughput (in the order of Gb), and a significant decrease in template sequence error rate that will make 1D read comparable with 2D reads.

From the release of the R9 flow cells, ONT introduced alternative protocols, without hairpin ligation, that allow to sequence high-accuracy 1D read (the template strand) avoiding the complement strand to pass through the pore. Thanks to these technological improvements, kits without hairpin ligation are going to replace 2D kits that will be phased out soon.

Sequencing errors are mainly caused by inserted and deleted bases [16], and slightly depend on read position demonstrating that the DNA strand translocation through the nanopore is not affected by position biases (Figure 1H). This result is of fundamental importance because it suggests that the nanopore sequencing approach can generate high-quality sequences with no theoretical limits on length, except those introduced during sample preparation [16].

In October 2016, new flow cells containing R9.4 chemistry were released by ONT, increasing sequencing speed to 450 bases/second and enabling 5–10 Gb DNA sequencing data to be obtained from a 48 h experiment. R9.4 chemistry was used to generate the first nanopore high-coverage (30×) human genomes (https://github.com/nanopore-wgs-consortium/NA12878; https://github.com/nanoporetech/ONT-HG1) with around 40 flow cells for each genome.

Base calling

When a ssDNA fragment translocates through the nanopore, it causes a drop in the current generating a noisy signal that is characterized by shifts in the mean that reflect DNA bases passing through the pore at that time. Raw current signals are used to infer the exact sequence of the ssDNA by means of a decoding process that is named base calling (Panels A–D of Figure 1).

The first fundamental step of the decoding process is the segmentation of current signal measurements, which consists in detecting the boundaries of each current shift (Figure 1B). The segmentation is performed locally by the ONT software, the MinKNOW, and gives as output a list of segments/events reporting their mean, Standard Deviation (SD) and length. Although current measurements are sampled at constant frequency (at regular time intervals), the translocation of the ssDNA through the pore is a stochastic process governed by enzymes. Ideally, consecutive events should differ by exactly one base, but in practice, consecutive segments differ by more than one base as a consequence of the nonuniform DNA translocation process. Moreover, even segmentation step can introduce noise in the analysis process, and novel and more accurate segmentation algorithms could improve the accuracy of base calling. For this reason, the segment sequences generated during segmentation step need sophisticated machine learning approaches to reconstruct the exact sequence of DNA bases passed through the pore.

ONT developed a proprietary base-caller, named Metrichor, that works in the cloud and is managed by the Metrichor Desktop Agent. As Metrichor is a proprietary software, the computational details at the base of its calling algorithm are not known. It is known (http://www.bio-itworld.com/news/02/17/12/Oxford-strikes-first-in-DNA-sequencing-nanopore-wars.html) that the first versions of the Metrichor algorithm were based on HMM, where hidden states correspond to each

possible k-tuple (sequence made of k bases), and emission probabilities reflect current signals expected for a particular k-mer.

As a result, every base is read as a part of k consecutive events, and transition probabilities take into consideration missed events, falsely split events and other likely errors. The HMM is first trained for parameter estimation, and then, base calling is performed by using the Viterbi algorithm that outputs the best state sequence. The base-calls of 1D sequences are performed separately for template and complement current signals, while the 2D sequences are inferred using both signals as constrains. In the latest implementation of Metrichor, now integrated in the proprietary EPI2ME platform, ONT introduced a novel base-caller based on RNN instead of the previously used HMM caller. From August 2016, EPI2ME allows to perform local base-calling without the need of a constant Internet connection (only for 1D sequencing data).

Recently, Boza *et al.* [30] and David *et al.* [31] have developed two open-source tools, Nanocall and DeepNano, for performing base-calling of MinION data without an Internet connection. Nanocall is based on a HMM similar to that used by Metrichor, where the hidden states are the sequenced k-mers, parameters are trained with an expectation–maximization procedure based on forward–backward algorithms and base-calling is obtained with the Viterbi algorithm.

On the other hand, DeepNano is based on RNNs that take as input the mean, SD and duration of each segmented event and gives as output the probability distribution of called bases. To take into consideration the sequential nature of segmented events, DeepNano exploits a bidirectional neural network that scans data in both directions and concatenates hidden outputs before proceeding to the next layer [30]. To estimate the parameters of the network, the authors use a stochastic gradient descent combined with Nesterov momentum.

To demonstrate the power of their RNN base-caller, Boza *et al.* [30] used the *Escherichia coli* data set generated by Loman *et al.* [20] and a *Klebsiella pneumoniae* data set. The results reported in their paper show that the DeepNano outperforms the Metrichor base-caller in terms of both accuracy (from 70 to 75% sequence identity for 1D read and from 85 to 87% for 2D reads) and computational speed (190 s for a 2D read with Metrichor and 11 s with DeepNano).

Finally, a new RNN base-caller, Nanonet, is now available through the ONT Github (https://github.com/nanoporetech/nanonet). Nanonet is a python-based command line suite to perform event detection step and segmentation for both 1D and 2D base-calling that runs locally using multiple Central Processing Units (CPUs). The package provides also an interface for training networks from FAST5 files by leveraging on CURRENNT, a Computed Unified Device Architecture (CUDA)-enabled machine learning library for RNN, which requires Graphics Processing Units (GPUs), as the training is an intensive process.

MinION data formats and handling

The MinION device is controlled by the MinKNOW software that runs on the host computer to which the MinION is connected and performs several core tasks such as selection of run parameters, data acquisition, real-time signal segmentation and feedback of experimental progression. For each read, the results of signal segmentation (segment mean, variance and duration) and the metadata associated with the sequencing process (channel and well used) are stored by the MinKNOW in FAST5 binary files, a variant of the HDF5 standard (http://www.hdfgroup.org/HDF5/). As explained in previous section, the raw data contained in the HDF5 files are then processed remotely in the cloud by the Metrichor, and the resulting called sequence is stored in other HDF5 files with extension .FAST5 that can contain a template read and a complement read or 2D read.

The pressing need to explore and manipulate the data generated by the MinION device obliged the MAP community to develop a series of software packages for parsing and converting FAST5 files into more conventional FASTA or FASTQ sequence formats (Table 1).

Poretools [32] and poRe [33] were the first software originally devised for exploring and manipulating FAST5 files. They generate basic quality plot such as read-length histograms for template, complement or 2D reads combined with yield-over-time plots and the squiggle plot (sequence of the segmented signals). NanoOK [34] exploits state-of-the-art mapping tool to estimate the three sources of sequencing errors (substitutions, insertions and deletions) and plots errors, coverage and k-mer distribution in a pdf document.

Two recently published tools, npReader [35] and minoTour (http://minotour.nottingham.ac.uk), allow to evaluate the sequencing process in real time by extracting reads, while the samples are being sequenced on the MinION device, and show streaming plots through a graphical user interface with quality statistics of the run. minoTour also provides pore/channel activity and percentage of coverage and average depth in presence of a reference sequence and allows to eject molecules in real time if they are not from the target genome of interest with the 'Read Until' approach [36].

Finally, Tarraga *et al.* [29] introduced a novel tool for exploring and converting FAST5 files, HPG Pore, that can run locally on individual computer or on a computer cluster exploiting the Hadoop distributed computing framework for managing the large amount of data that are expected to be generated by the new ONT instruments, the PromethION and the GridION X5 (see closing remarks section).

De novo assembly

De novo genome assembly consists of computationally reconstructing the entire genome sequence from a collection of sequenced reads much shorter than the genome from which they are generated.

Although the advent of SGS improved the accuracy and completeness of assembled genomes and allowed *de novo* assembly at affordable prices, the use of short reads still generates fragmented assemblies as a consequence of repetitive regions longer than the read length. For this reason, genome assembly obtained from SGS data needs to be refined by using complex approaches that include Sanger sequencing and specially tailored assembly methods. In this scenario, the availability of sequencing technologies that can generate ultra-long reads, much larger than the repetitive regions, is fundamental for reconstructing the complex architecture of genomes.

From an algorithmic point of view, *de novo* assembly methods can be divided into two main classes, those based on the Overlap-Layout-Consensus (OLC) paradigm and those that follow the de Bruijn graph (DBG) approach (see Box 1 for more details). As DBG methods were properly devised to exploit the huge amount of short reads generated by SGS platforms, and OLC algorithms were widely used in the Sanger sequencing era, both approaches are not appropriate to handle the long sequences with high error rates produced by TGS platforms.

To mitigate the high error rate of TGS data, several authors developed novel computational strategies that include hybrid methods that use complementary SGS data to correct long-read

Table 1. Data handling and quality check tools

Tool	Output formats	Output stats	Output plots	Source	Languages	Reference
HPG Pore	FASTQ/FASTA	Text file with:	−read-length/quality/GC histograms[a]	http://github.com/opencb/hpg-pore	Java/ Hadoop	[29]
		Reads number and nucleotide	−Nucleotide/mean quality per position per read			
		−Mean/min/max read length	−Yield over time			
		−Nucleotide distribution	−Number of reads/yield per channel			
		−%GC, mean quality	−Raw signal (squiggle)			
NanoOK	FASTQ/FASTA	−Summary run/	−Read-length histograms[a]	https://github.com/TGAC/NanoOK	Java/R	[34]
		Read statistics	−InDel length histograms			
		Alignment-based QC	−Number of reads per identity %			
		Analysis, including errors, coverage and k-mers stats	−Error, k-mers distribution			
			−Coverage			
npReader	Real-time FASTQ/FASTA generation	Real-time QC metrics	GUI with streaming plots:	https://github.com/mdcao/npReader	Java/R	[35]
			−Real-time read counts			
			−Read-length/quality histograms[a]			
poRe	FASTQ/FASTA organized into run folders	Summary run/read/	−Read-length histograms[a]	http://sourceforge.net/projects/rpore	R	[33]
		Channel statistics	−(Cumulative) yield-over-time			
			−Number of reads/yield per channel			
			−Raw signal (squiggle)			
PoreTools	FASTQ/FASTA	Summary run/	−Read-length histograms[a]	https://www.github.com/arq5x/poretools	Python	[32]
		Read statistics	−Nucleotides/mean quality per position			

Note. In this table are listed the tools properly developed for dealing with the raw data produced by MinION and stored in HDF5/FAST5 binary files. All of the them generate FASTQ/FASTA files and a series of quality control plots (plots column) and statistics (run stats). The source link of each tool is also provided.
[a]Stands for template, complement or 2D reads.

errors and non-hybrid methods in which long reads are self-corrected by exploiting overlaps in high-coverage data. While the first tools, the Hierarchical Genome Assembly Process (HGAP) [37] and the PacBio Corrected Reads (PBcR) [38] pipelines, were designed for PacBio sequences, in the past 2 years, much work has been done to develop novel hybrid and non-hybrid error reduction algorithms properly tailored for ONT data (Table 2).

Goodwin *et al.* [21] developed a hybrid error correction algorithm, Nanocorr, that first aligns MiSeq reads against the nanopore sequences using BLAST, and then selects the optimal set of short read alignments by using a dynamic programming algorithm based on the longest-increasing-subsequence problem. Nanocorr was capable to increase the percent identity of uncorrected reads from 67 (R6.0–R7.3) to 97%.

Madoui *et al.* [39], instead of using SGS data to correct long nanopore reads, developed a hybrid error correction approach, Nanopore Synthetic-long (NaS), in which MinION sequences are used as template to recruit Illumina reads and build synthetic reads by performing local assembly. By combining 50× Illumina 250 bp paired ends and R7.3 MinION nanopore reads at 57×, the NaS approach was able to improve the mean percentage identity of 1D read and 2D reads from 56 and 75% to 99.99% producing synthetic reads at 23× genome coverage.

In 2015, Loman *et al.* [20] were the first to demonstrate that it is possible to assemble an entire bacterial genome (*E. coli* K-12) using solely ONT reads corrected for errors with a novel non-hybrid approach. The error correction method, Nanocorrect, is based on the iterative use of the partial order alignment (POA) graphs on each read and its overlapped reads to determine a consensus sequence. POA uses a directed acyclic graph to compute a multiple alignment and permits a more sensitive reconstruction of consensus sequences in the presence of large number of insertions and deletion (InDels) and is capable to increase the mean percentage identity of 2D reads from 80.5 to 97.7%. However, as POA algorithm is slow, the Nanocorrect tool has been recently deprecated in favor of Nanopolish error correction pipeline http://simpsonlab.github.io/2016/02/25/deprecating-nanocorrect/.

Concurrently, Szalay and Golovchenko [40] proposed a non-hybrid correction approach, PoreSeq, based on HMM that combines Metrichor base-calls and segmented ionic current data from multiple overlapping reads aligned to the same DNA region. PoreSeq first selects an initial best guess sequence among those provided by Metrichor and then takes all the reads that cover the same DNA region. The initial best guess sequence is iteratively improved by introducing random mutations that are generated by a modified single-molecule Viterbi algorithm. Each mutation is tested by recalculating the observation likelihood, and the mutated sequence is kept if this likelihood exceeds the current best. This approach allows to test only for likely mutations, avoiding prohibitively dense search over all possible sequences.

In 2016, Koren *et al.* [43] introduced a novel assembly pipeline, Canu, that implements a non-hybrid error correction method that search for the highest weight path of directed

Table 2. Tools for *de novo* assembly

Tool	Description	Input	Algorithm	Link	Reference
Nanocorr	Error correction	FASTA	Hybrid	https://github.com/jgurtowski/nanocorr	[21]
NaS	Error correction	FASTA FASTQ	Hybrid	https://github.com/institut-de-genomique/NaS	[39]
Nanocorrect	Error correction	FASTA	Non-hybrid	https://github.com/jts/nanocorrect	[20]
PoreSeq	Error correction Polishing	FASTA FASTQ	Non-hybrid	https://github.com/tszalay/poreseq	[40]
Nanopolish	Consensus polishing	FASTA	Non-hybrid	https://github.com/jts/nanopolish	[20]
LQS	Assembly pipeline	FASTA FASTQ	Nanocorrect OLC-Celera Nanopolish	https://github.com/jts/nanopore-paper-analysis	[20]
PBcR	Assembly pipeline	FASTA FASTQ	Hybrid Non-hybrid OLC-Celera	http://wgs-assembler.sourceforge.net	[38]
Falcon	Assembly pipeline	FASTA	Non-hybrid OLC	https://github.com/PacificBiosciences/FALCON	[41]
Miniasm	Assembly algorithm	FASTA FASTQ	No correction OLC	https://github.com/lh3/miniasm	[42]
Canu	Assembly pipeline	FASTA FASTQ	Non-hybrid OLC	https://github.com/marbl/canu	[43]
SMARTdenovo	Assembly pipeline	FASTA FASTQ	No correction OLC	https://github.com/ruanjue/smartdenovo	[44]
SPAdes	Assembly pipeline	FASTA FASTQ	Hybrid DBG	http://bioinf.spbau.ru/spades	[45]
ALLPATHS-LG	Assembly pipeline	FASTA FASTQ	Hybrid DBG	https://software.broadinstitute.org/allpaths-lg/blog/	[46]
ABruijn	Assembly pipeline	FASTA	Non-hybrid DBG	https://github.com/fenderglass/ABruijn	[47]

Note. Table reports software packages tested or properly devised for correcting and assembling nanopore reads. Columns summarize the main computational features of each tool.

acyclic graphs generated from alignments obtained with the Myers' ND algorithm [48].

To evaluate the performance of hybrid and non-hybrid read error correction methods on MinION data, Deschamps *et al.* [49] used reads generated from the multi-chromosome genome of *Agrobacterium tumefaciens* strain LBA4404. Although both approaches were able to correct reads with an average identity >97%, Ilumina-based correction with PBcR and self-correction with Canu provided the best results in terms of percentage of corrected sequences and of reads with at least 99% identity (Figure 2A).

Thanks to their algorithmic structure and to a sensitive overlapping step, OLC methods are generally more suited for assembling long and high error-prone sequences such as those generated by PacBio and ONT platforms. However, owing to the impressive diffusion of SGS data in the past decade, the great majority of state-of-the-art assemblers are based on DBG method, and until recently, the few OLC approaches for long reads (PBcR and HGAP) were mostly based on the Celera Assembler (Table 2), with the only exception of the non-hybrid PacBio tool Falcon. For these reasons, a large number of assembly pipelines developed to date for MinION data are based on the Celera Assembler.

Madoui *et al.* [39] assembled NaS synthetic reads with the Celera algorithm obtaining a 3.6 Mb contig containing complex and repetitive regions and covering 99.8% of the *Acinetobacter baylyi* reference genome with an identity >99.98%. Tests on lower coverage nanopore data sets (14× and 28× of MinION reads) showed that the final assemblies were less fragmented than the Illumina-only assembly.

Loman *et al.* [20], Szalay and Golovchenko [40] and Goodwin [21] combined their error correction methods (Nanocorrect, PoreSeq and Nanocorr) with the Celera Assembler obtaining good results in terms of number/size of contigs and with a per-base identity of 98–99% (Figure 2B–D).

In 2016, Li *et al.* [42] developed Miniasm, a tool to assemble genomes from long reads without performing error-correction, and more recently, the developers of Celera, PBcR and MHAP moved away from their original projects and started to develop Canu [43] that is a complete reworking of the Celera Assembler specifically designed for high error rate sequences and that combines strategies from Celera (http://wgs-assembler.sourceforge.net), Falcon [41] and Pbdagcon (https://github.com/PacificBiosciences/pbdagcon).

In 2016, Sovic *et al.* [51] compared the performance of five non-hybrid (LQS, PBcR, Falcon, Miniasm and Canu) and two hybrid (ALLPATHS-LG [46] and SPAdes [45]) assembly pipelines on publicly available MinION *E. coli* data sets sequenced with different chemistry (R7 and R7.3) and subsampled at various coverages to explore assembler scalability.

They showed that none of the currently available non-hybrid approaches were capable to completely reconstruct a genome with low-quality nanopore reads (mainly 1D read) or when the sequencing coverage is low (20×), while for coverages >20×, all tested tools produce a consistent increase of the quality of assembly in terms of largest contig length and average identity. As expected, hybrid assembly pipelines achieve better results than non-hybrid methods, and this is mainly because of the additional coverage provided by Illumina reads (Figure 2B and C). Remarkably, although hybrid approaches are highly dependent on the quality of SGS data, they are not much influenced by quality and coverage of nanopore sequences and are capable to obtain good assemblies even with low-coverage reads produced by the early R7 chemistry (Figure 2B and C).

Istace *et al.* [19], by sequencing the *Saccharomyces cerevisiae* S288C with 11 MinION runs, tested four non-hybrid assembly pipelines (Canu, Miniasm, SMARTdenovo https://github.com/ruanjue/smartdenovo and ABruijn [47]) with various combinations of R7.3 1D and 2D reads. They found that Canu obtained

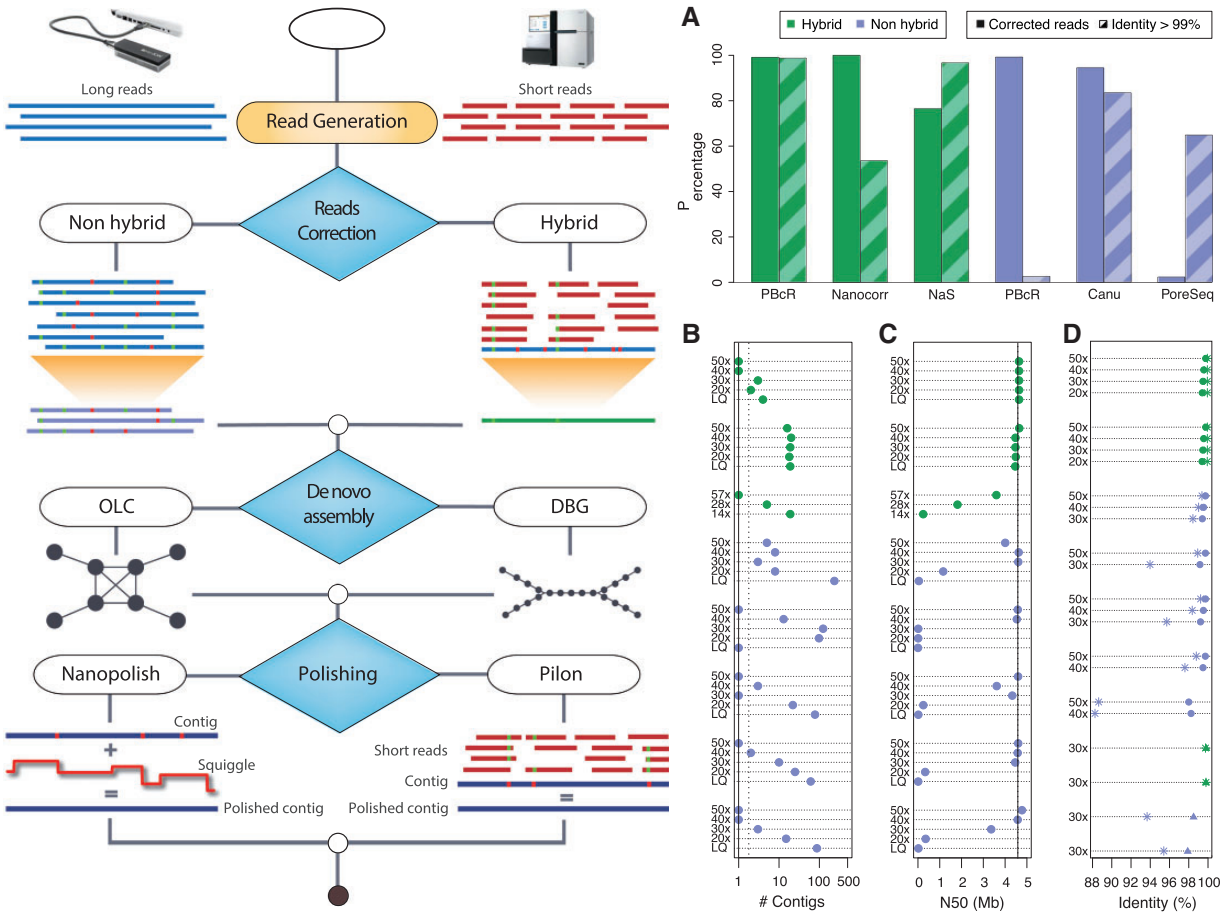


Figure 2. *De novo* assembly. High-level diagram of long-read assembly pipeline. The assembly has three main steps that are error correction, contig assembly and polishing. Error correction approaches belong to two main categories that include hybrid and non-hybrid methods. Hybrid methods use complementary NGS (Illumina) data, while non-hybrid approaches self-correct reads by exploiting overlap of high-coverage data. Contig assembly can be performed with OLC or DBG algorithms. After assembly, contigs are refined by using polishing methods that improve the consensus sequence accuracy by using raw current data or high-quality Illumina reads. Panel (**A**) shows the performance of hybrid and non-hybrid correction methods as a function of fraction of corrected reads and of reads with at least 99% identity between the reference assembly and the aligned portion of the read. The dot chart of Panels (**B**) and (**C**) reports the performance of three hybrid and four non-hybrid assembly methods for different sequencing coverages in terms of N50 and number of contigs. From top to bottom are reported the results of Allpaths-LG, SPAdes, NAS + Celera, LQS, Falcon, PBcR, Canu and Miniasm. Vertical lines of Panels (B) and (C) represent performance statistics for PacBio (dotted) and Illumina (solid) data and were parsed from [50] and [51], respectively. Panel (**D**) shows the effect of polishing strategies on hybrid and non-hybrid assemblies in terms of percentage of identity as a function of sequencing coverages. Asterisks represent the percentage of identity before polishing, while solid circles and triangles after polishing with Nanopolish and Pilon, respectively. Polishing performance was manually parsed from [51] for Nanopolish and from [49] for Pilon. From top to bottom are reported the results of Allpaths-LG, SPAdes, LQS, PBcR (non-hybrid), Canu (non-hybrid), Falcon, Miniasm, PBcR (hybrid), Canu (hybrid), PBcR (non-hybrid) and Canu (non-hybrid).

the best assembly with 67× high-quality 2D pass reads; SMARTdenovo with 30× of the longest 2D reads; ABruijn using all the 2D reads, which represented coverage of ∼120×; and Miniasm using the 2D reads corrected by Canu, which represented coverage of ∼108×. ABruijn, SMARTdenovo and Miniasm generated around 25 contigs and were able to assemble 14 chromosomes in one or two contigs, while the assembly of Canu was composed of 37 contigs, and only seven chromosomes were assembled in one or two contigs.

Once reads have been assembled, the contigs obtained from OLC or DBG algorithms need to be refined by using polishing tools. At present, few computational methods have been tested and developed to make improvements to the quality of draft assemblies obtained from ONT sequences.

Loman *et al.* [20] developed an HMM algorithm, Nanopolish, that takes as input the draft assembly and progressively modifies it by making small localized changes that are accepted if they increase the probability of the electrical current data for a set of reads.

Szalay and Golovchenko [40] used PoreSeq to correct errors in the assembled contigs improving the consensus sequence accuracy from 96 to 98.5%. Goodwin *et al.* calculated the consensus sequence by using the algorithm Pilon [52], which, by exploiting complementary Illumina reads, improved the per-base accuracy of the assembly from 99.78 to 99.88%.

Sovic *et al.* [51] found that the Nanopolish algorithm improves the average identity and contig length of non-hybrid assemblies, while it introduces errors on hybrid assemblies (Figure 2D). This is because of the fact that Nanopolish tries to refine contigs assembled with reads with lower error rate (Illumina reads) using sequences with higher error rate (nanopore 2D reads). On the other hand, the Pilon method, exploiting Illumina reads, is able to improve the identity of the consensus sequence of draft assemblies obtained by both hybrid and non-hybrid approaches [49].

Table 3. Tools for resequencing

Tool	Description	Input	Algorithm	Link	Reference
BLASR	Aligner	FASTA FASTQ	BWT-FM+dynamic programming	https://github.com/PacificBiosciences/blasr	[54]
BWA	Aligner	FASTQ	BWT-FM	http://bio-bwa.sourceforge.net/	[21]
GraphMap	Aligner	FASTA FASTQ	Global alignment	https://github.com/isovic/graphmap	[55]
LAST	Aligner	FASTA FASTQ	Global alignment	http://last.cbrc.jp/	[53]
marginAlign	Re-aligner	BAM	HMM	https://github.com/benedictpaten/marginAlign	[24]
marginCaller	Variant caller	BAM	HMM	https://github.com/benedictpaten/marginAlign	[24]
Nanopolish	Variant caller	BAM FAST5	HMM	https://github.com/jts/nanopolish	[13]
SignalAlign	Methylation mapper	BAM	HMM	https://github.com/ArtRand/signalAlign	[56]
Nanopolish	Methylation mapper	BAM	HMM	https://github.com/jts/nanopolish	[57]

Note. Table reports software packages tested or properly devised for aligning and calling variants with nanopore reads.

Mapping and resequencing

The resequencing approach consists in aligning all the reads produced by a sequencing experiment against a reference genome to find differences between reads and the reference genome (see Box 2).

The alignment of nanopore and, more generally, of TGS data is particularly challenging owing the size (from kb to tens of kb) and the high and nonuniform error profiles of this new generation of sequences. The principal computational problem is how to align long (many kilobase) reads with moderate divergence from the genome (up to 20–30% divergence, concentrated in InDels) with the same speed and sensitivity of SGS alignment methods.

In the past 2 years, since the beginning of the MAP, several authors suggested LAST [53] as one of the best approaches for aligning nanopore reads. LAST (Table 3) was originally designed for comparing large sequence data sets with each other (vertebrate genomes and/or large numbers of DNA reads) and is based on three steps in which it first finds initial matches between reads and genome, then extends them with a gapless X-drop algorithm and finally extends them using a gapped X-drop algorithm [58]. The main difference with respect to other alignment algorithms is that it can find weak similarities between sequences with many mismatches and gaps.

At present, LAST is the most widely used alignment algorithm for nanopore sequences, and it has been successfully exploited for aligning whole-genome sequencing nanopore reads against bacterial [22, 23, 59] and viral genomes [60, 61] and for determining exon connectivity in complex mRNAs [15].

Another widely used mapper for nanopore reads is the Basic Local Alignment with Successive Refinement (BLASR) developed by Chaisson *et al.* [54]. BLASR combines the data structures used in short-read mapping (Burrows–Wheeler Transform Full-text Minute-space index, BWT-FM index) with alignment methods used in whole-genome alignment (dynamic programming alignment) and was originally devised to align sequences generated by the Pacific Bioscience platform (Table 3). The successive refinement approach consists of three phases: (1) detecting candidate intervals by clustering short, exact matches; (2) approximate alignment of reads to candidate intervals using sparse dynamic programming; and (3) detailed banded alignment using the sparse dynamic programming alignment as a guide. BLASR was used to align nanopore reads for resolving the haplotypes of *HLA-A, HLA-B* and *CYP2D6* genes [62] and to identify an *E. coli* sample down to the species level from 16S recombinant DNA amplicons [63].

Burrows–Wheeler Aligner (BWA) [64] is the most used and cited mapper for aligning SGS reads against a large reference sequence allowing mismatches and gaps. BWA was based on backward search with Burrows–Wheeler Transform and was properly devised for the alignment of huge amount of short reads with minimal variation with respect to the reference genome. Li extended the BWA-MEM (Maximal Exact Match) algorithm by combining relaxed scoring of Smith–Waterman with heuristics filtering to support long and high error rate sequences from PacBio and ONT reads. The improved version of BWA-MEM was successfully used to detect complex structural variants (simple interstitial deletions, translocations, inversions and a complex combination of a translocation and an inversion) in cancer samples [14].

In 2016, Sovic *et al.* [55] introduced the first mapping algorithm specifically designed for nanopore reads, GraphMap, which is capable of obtaining BLAST-like sensitivity with high computational speed. GraphMap is based on a five-step procedure that progressively and conservatively reduces the set of candidate mapping positions by redefining the forms of read-to-reference alignment.

To improve the accuracy of nanopore reads alignment, Jain *et al.* [24] proposed a novel approach, named marginAlign, which realigns reads against a reference genome by combining an error model with the alignments generated by LAST and BWA. The proposed error model is a five-state pair HMM that has one state for modeling matches and four states for modeling short/long InDels, and the parameter estimation is performed by means of the Baum–Welch algorithm. The realignment of M13 bacteriophage and *E. coli* nanopore reads, previously aligned with LAST and BWA-MEM, showed an improvement between 2 and 5% in average identity.

Recently, we evaluated the performance [16] of several mapping tools (BWA, BLASR, LAST and marginAlign) by exploiting the nanopore sequences generated by the MinION Analysis and Reference Consortium (MARC) [18]. The MARC experiments were performed by five laboratories that sequenced the same *E. coli* strain, in duplicate, by using the R7.3 flow cells and the Metrichor 1.12 protocol for base-calling.

The results of this comparison show that the five mapping tools align around 99% of the reads with average base quality

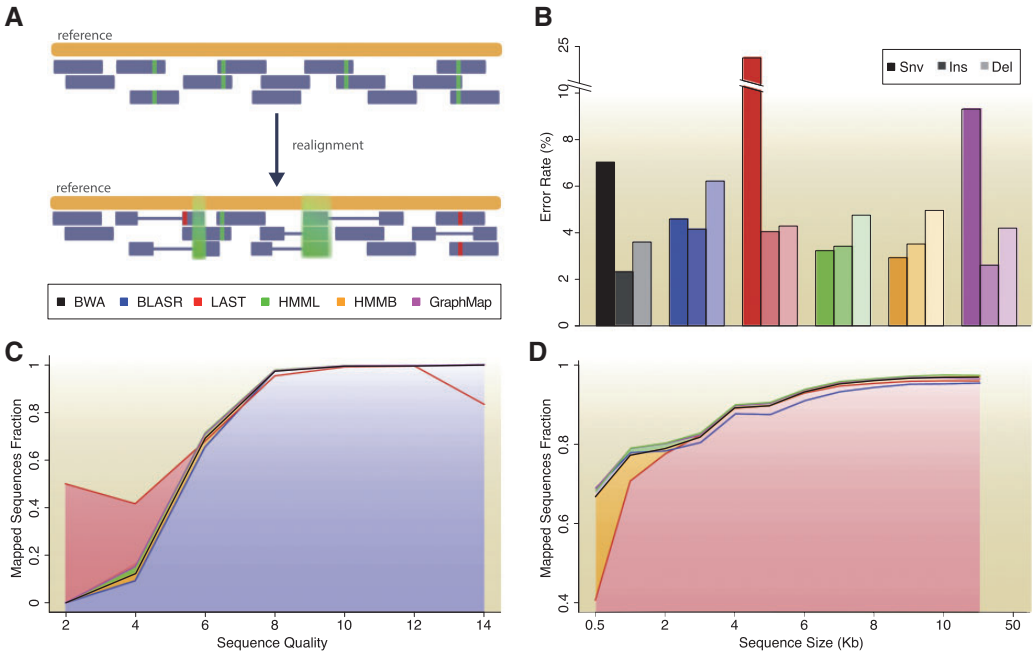


Figure 3. Mapping and resequencing. Panel (**A**) summarizes alignment and realignment approaches. Alignment algorithms are used to find the exact position of each reads with respect to a reference genome sequence. Realignment algorithms (such as MarginAlign) realign reads against a reference genome by combining an error model with the alignments generated by other mapping methods. Panel (**B**) shows the error rate estimated by different alignment algorithms for substituted, inserted and deleted bases. Panels (**C**) and (**D**) report aligners performance as a function of read length and quality. Color legend is referred to Panels (B), (C) and (D). MarginAlign results are reported for LAST (HMML) and BWA (HMMB) alignments. All the results reported in figure are parsed from [16].

≥ 8, and that the alignment performance strongly depends on sequence size, as the longer the reads, the higher the fraction of sequences mapped by each method (Figure 3C and D).

Although the five alignment strategies gave similar results, the LAST algorithm obtained the worst global performance and resulted as the most influenced by sequence length (Figure 3D). Moreover, we also found that the five mapping approaches return different results in handling the three error categories (substitutions, insertions and deletions). In particular, marginAlign was able to mitigate substitution error rate at the expense of InDels, while LAST and BWA gave small InDel errors with higher substitution errors (Figure 3B).

Variant discovery

Genetic variant discovery consists in finding differences between genomes and comprises the identification of single-nucleotide variants (SNVs), small InDels and/or more complex structural variants such as large deletions, duplications, inversions and translocations.

The identification of all these classes of variants can be achieved by either searching for differences between aligned reads and the reference genome (resequencing) or comparing assembled consensus sequences (*de novo* assembly against a reference genome).

In resequencing, SNVs and small InDels are inferred by comparing the number of reads that do not contain the reference allele with the total number of reads aligned to that position, while in *de novo* assembly approach, variants are typically identified by comparing the consensus sequences with the reference genome by using software such as MUMmer [65], Mugsy [66] and Mauve [67].

At present, few papers explored the capability of nanopore data to identify small variants either by resequencing or assembly approach, and all these works demonstrated that performing this task with this new technology is still challenging (Figure 4). By using synthetic reference genomes generated by substituting, inserting and removing bases from the *E. coli* reference genome, we evaluated the capability of aligned nanopore reads to detect SNVs and InDels in monoploid genome [16]. The results show that, although the detection rate for SNVs discovery is around 90% even at low coverages (10×), only 30% of the deletions and <5% of the insertions can be correctly identified with high sequencing coverage (40×).

Jain *et al.* [24] developed marginCaller (Table 3), a tool for SNV discovery in monoploid genomes that starts from alignment produced by marginAlign and computes posterior alignment match probabilities between the bases in the reads and the reference by using a HMM-based realignment strategy. Exploiting altered sequence as reference, they demonstrated that marginCaller is able to identify 97% of the simulated SNVs with 2D read data at 60× sequencing coverage.

More recently, Quick *et al.* [13] introduced a novel SNV caller (integrated in the Nanopolish package) that takes as input data aligned with marginAlign. The Nanopolish variant caller first selects candidate variants corresponding to mismatches between aligned reads and the reference, and then identifies regions with closely spaced variants. Each cluster of variants is used to generate a set of candidate haplotypes from the possible combinations of SNVs and the haplotype that maximizes the probability of the event-level data is called as the sequence for the region (Table 3).

Regarding diploid variant detection, Ammar *et al.* [62], by using the BLASR aligner, demonstrated that currently available state-of-the-art variant caller, such as the Genome Analysis Toolkit's UnifiedGenotyper and HaplotypeCaller, is unable to call any heterozygous variants from ultra-high-coverage (1000×) targeted sequencing data of human genes *HLA-A, HLAB* and *CYP2D6*. They also showed that heterozygous variant calling with threshold-based approach (one-third of the reads to support an allele) is an effective method for SNV detection,

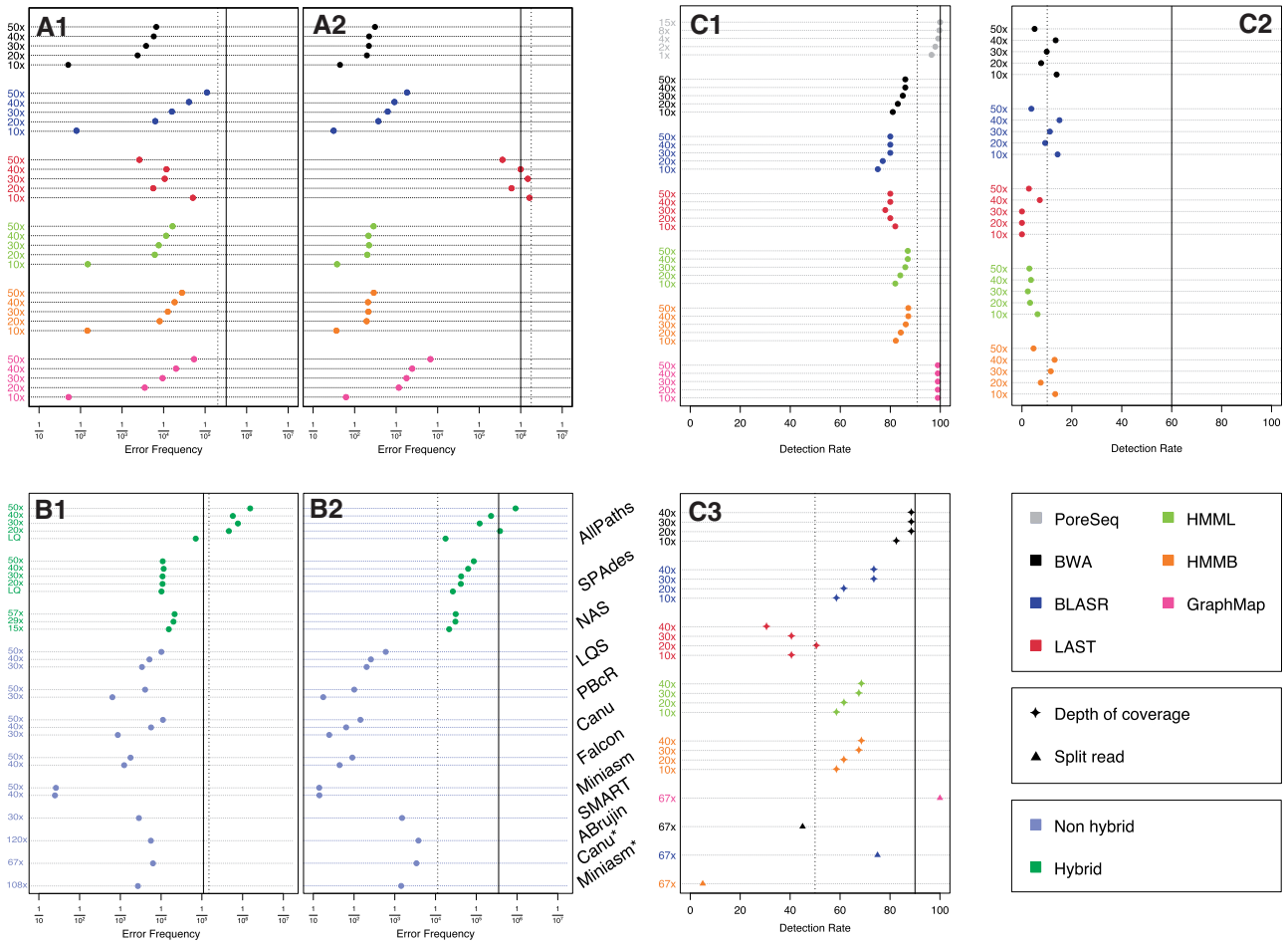


Figure 4. Variants detection performance. The dot charts of Panels A and B report the false-positive frequency of substitutions (**A1** and **B1**) and InDels (**A2** and **B2**) for resequencing (A1 and A2) and *de novo* assembly (B1 and B2) approaches. The dot charts of Panels C1, C2 and C3 report detection accuracy for substitutions (**C1**), Indels (**C2**) and CNVs (**C3**) with different computational approaches. Vertical lines report performance for PacBio (dotted) and Illumina (solid) data and are manually parsed from [50] and [55], respectively. The results plotted in Panels A1 and A2 are parsed from [16], B1 and B2 from [51] and [19], while C1, C2 and C3 from [16, 55] and [40], respectively.

although it leads to the discovery of many false-positive variants, especially deletions, as a consequence of the high error rate.

Recently, Sovic *et al.* [55], reanalyzing the Ammar *et al.* [62] data set by combining the GraphMap aligner with the rare variant caller LoFreq [68], found a significant improvement in terms of both sensitivity and specificity in the identification of SNVs with respect to other mapping methods.

Concerning *de novo* assembly, Szalay and Golovchenko [40] tested the performance of the PoreSeq pipeline in recognizing the original sequence of computationally mutated reference genomes and demonstrated that, even at low coverages (4×), their assembly approach detects small variants with a recall rate >95% demonstrating similar performance to resequencing approach but at an order-of-magnitude lower coverage.

Although these results show that ONT data can be reliably exploited to detect small variants at high recall rate, the precision of both resequencing and *de novo* assembly approaches can be highly influenced by sequencing errors, especially when error profiles are not randomly distributed.

In [16], we demonstrated that the error distribution along the nanopore reads is not completely random, but, on the contrary, errors fall in recurrent positions of the genome and follows specific nucleotide patterns that can be ascribed to intrinsic biases of the nanopore sequencing process. In particular, single-nucleotide errors are enriched of C to G and G to C substitutions, while deleted bases principally involve A and T and mainly occur after A and T nucleotides of the genome. By using the R7.3 data generated by the MARC [18], we found that these patterns generate recurrent errors that, even at high sequencing coverages (≥ 30x), lead to the identification of one false substitution every 10–100 kb and one false InDel every 1–10 kb (Figure 4A), and the aligners that guarantee the smallest false-positive rate for both SNVs and InDels are BLASR and GraphMap.

In *de novo* assembly framework, the results reported by Sovic *et al.* [51] and Istace *et al.* [19] show that the use of non-hybrid assembly approaches produces consensus sequences with error rates similar to those obtained by resequencing approach (one substitution error every 1–10 kb and one InDels every 100–1000 bp), demonstrating that correction methods are not capable to remove the intrinsic biases of the nanopore sequencing process. Conversely, hybrid approaches, thanks to the use of Illumina data, are able to mitigate the error rate of nanopore reads generating a consensus sequences with a false event (SNV or InDel) every 10–100 kb (Figure 4B).

At present, a limited number of papers have faced the problem of detecting structural variants with ultra-long nanopore

reads. The identification of structural variants requires computational methods much more complex than those used for small variant discovery that include split reads (SRs), paired-end mapping (PEM), depth of coverage (DOC) and *de novo* assembly methods. All these methods were originally designed for the analysis of SGS data, and owing the lack of paired end sequences, PEM methods cannot be exploited for long nanopore reads.

SR methods allow for the detection of deletions and insertions on the basis of a split sequence-read signature: the alignment to the genome is broken, and a continuous stretch of gaps in the read indicates a deletion or in the reference indicates an insertion. Norris *et al.* [14] evaluated the capability of nanopore reads to detect previously characterized SVs, including large deletions, inversions and translocations in pancreatic cancer. To this end, they sequenced with R7.3 chemistry the barcoded polymerase chain reaction (PCR) amplicons of 12 regions containing structural variants. Nanopore reads were aligned against the human reference genome with BWA-MEM (ont2d), and SV were searched with the SR approach implemented in the LUMPY package [69]. Although this approach was able to identify almost all SVs, it failed in the detection of the exact break point coordinates because of BWA alignment limits.

By using R7.3 *E. coli* sequences and modified reference genomes with SVs from 100 to 4000 bp, Sovic *et al.* [55] compared the capability of several aligners to produce spanning alignments or split alignments indicative of a structural variation (insertions or deletions). LAST and marginAling revealed poor results, while BLASR and GraphMap arose as the best mapper for producing spanning alignments that accurately identify SV events with an F-measure (Figure 4C).

The DOC approach is based on the simple idea that during the sequencing process, the reads are randomly and independently sequenced from any location of the genome. Under this assumption, the number of reads mapping into a window of the reference genome should be proportional to the number of times the region appears in the DNA sample and follows a Poisson distribution. As a consequence, the copy number of any genomic region can be estimated by calculating the DOC of reads aligned to consecutive and non-overlapping windows of the genome.

In [16], by using the R7.3 data generated by the MARC, we demonstrated that nanopore sequencing is a uniform process that generates sequences randomly and independently without classical sources of bias such as GC content and mappability. Thanks to these properties, nanopore data can be readily used to detect genomic regions involved in copy number variants with high accuracy, outperforming PacBio and even SGS data in terms of both sensitivity and specificity (Figure 4C).

By using the genomes of *S. cerevisiae* natural isolates assembled with solely R7.3 sequences, Istace *et al.* [19] detected 29 translocations and 4 inversions demonstrating that complex SVs can be detected with high accuracy with *de novo* assembly comparison.

DNA methylation

DNA methylation is involved in many biological processes, such as cell differentiation, gene regulation and development and disease [70]. At present, the state-of-the-art method for DNA methylation mapping at a genome-wide level is based on sodium bisulfite treatment followed by SGS experiments. However, these techniques are limited by sequence read length and do not reveal long-range single-read patterns of methylation. Two recently published papers demonstrated that the ONT MinION device is able to directly detect DNA modifications analyzing the changes in ionic current signal without any DNA treatment (Table 3).

Simpson *et al.* [57], by using R7.3 and R9 nanopore chemistry, developed an HMM (https://github.com/jts/nanopolish) that calculates the probability to distinguish 5-methylcytosine (5-mC) from unmethylated cytosine in a CpG context. Their model was able to identify 5-mC at up to 95% accuracy, increasing the stringency for making a call. However, their training set mapped only completely methylated regions reducing the ability to identify heterogeneous methylation regions.

Similarly, Rand *et al.* [56] used a variable-order HMM combined with a hierarchical Dirichlet process to analyze the effects of base methylation on ionic current distribution. To evaluate the performance of their tool (SignalAlign, https://github.com/ArtRand/signalAlign), they applied it to the analysis of the methylation level of three samples sequenced with R7.3 and R9 chemistry: synthetic oligonucleotides, *E. coli* genomic DNA and pUC19 plasmid. SignalAlign was able to recognize methylation with a median accuracy of 80, 96 and 86% for synthetic, *E. coli* and pUC19 plasmid samples, respectively.

Applications

Although the number of papers that make use of nanopore data is still limited, it must be highlighted the marked versatility of MinION use, both in terms of sample types and fields of application (Table 4).

Different studies successfully investigated the potential of the MinION device for characterizing bacterial pathogens. Ashton *et al.* [25] were the first to resolve the structure and chromosomal insertion site of an antibiotic resistance island in *Salmonella typhi* by scaffolding an assembly generated from short-read Illumina data with MinION long reads.

Quick *et al.* [12] showed that MinION can be used to distinguish real-time outbreak from non-outbreak of *Salmonella enterica*. Moreover, the analysis performed by Judge *et al.* [73], on six clinically significant pathogens, proved that, although MinION error rates were higher than those of Illumina, MiSeq and PacBio sequencing, it was able to detect the presence of acquired resistance genes.

MinION was also tested for real-time viral outbreak surveillance in remote and resource-limited area. Indeed, it has been recently used for monitoring ongoing epidemic Ebola outbreaks directly on site in Liberia [71] and Guinea West Africa [13].

Moreover, ONT data were used to obtain a complete influenza virus genome, evidencing 99% identity sequence with data derived from Illumina MiSeq and traditional Sanger sequencing [54], and succeeded in taxonomic classification of high-risk viral pathogens in human clinical samples [72], by fully exploiting its portability.

Furthermore, the investigation of microbial communities also benefits from the use of long-read sequencing, as short reads often limit the microbial composition analysis at the species level because of the high similarity of 16S ribosomal RNA (rRNA) amplicon sequences. Recent studies [63, 74, 77, 78] have shown that reads obtained by Nanopore MinION were sufficiently accurate to identify and differentiate both viral and bacterial species.

Concerning human genomics, Ammar *et al.* [62] used the MinION to sequence the *HLA-A*, *HLA-B* and *CYP2D6* genes,

Table 4. MinION Data applications

Applications	Organism	Sample	Target	Summary	Reference
Pathogen surveillance and bacterial/viral outbreak investigation	Ebola virus	RNA	WGS	142 Ebola-positive samples were sequenced using a target reverse transcriptase polymerase chain reaction (RT-PCR) protocol to isolate only sufficient viral cDNA. Data were analyzed in 24–48 h on site in a resource-limited setting	[13]
	Ebola virus	RNA	WGS	By using a RT-PCR based approach, high-quality complete genome sequences for eight of nine high-virus load samples were obtained	[71]
	Salmonella typhi haplotype 58	DNA	WGS	A hybrid assembly of combined MinION and Illumina HiSeq data allowed to identify the structure and insertion site of a previously uncharacterized antibiotic resistance island in *S. typhi* H58	[25]
	Chikungunya, Ebola and HCV virus	RNA	Viral genome	The strains of high-risk pathogens, such as chikungunya, Ebola and HCV Hepatitis C Virus viruses were identified directly from clinical human samples without culture using MinION	[72]
	Gram bacilli and Methicillin-resistant *S. aureus*	DNA	WGS	MinION error rates were higher than Illumina MiSeq and PacBio RSII platforms, but provided an even coverage across the entire genome length; all of the expected carbapenemases, ESBL genes and the mecA gene have been identified	[73]
	Salmonella enterica	DNA	WGS	Clinical information regarding the *Salmonella* outbreak in hospital could be assessed by using both Illumina MiSeq and MinION with confident assignments of serotypes and clinical information regarding *Salmonella* outbreak in less than half a day	[12]
	Influenza A virus	RNA	Eight viral genes and genome	By using MinION sequencer, a complete influenza virus genome was obtained that shared >99% identity with sequence data obtained from MiSeq and traditional Sanger sequencing	[61]
	Klebsiella pneumonia, *E. coli* and *Enterobacter cloacae*	DNA	Species and resistance genes identification	By using MinION sequencing, it was possible to correctly identify different bacteria and to detect acquired resistance genes, directly from the urine samples	[74]
Haplotypes reconstruction	Human	DNA	*HLA-A*, *HLA-B* and *CYP2D6* genes	Long-read data from a single 24 h nanopore sequencing run were used to reconstruct haplotypes, which were confirmed by HapMap data and statistically phased complete genomics and Sequenom genotypes	[62]
Prenatal diagnosis	Human	DNA	Aneuploidy investigation	MinION technology can be used for rapid real-time acquisition of short DNA reads of sufficient numbers for successful determination of gender and aneuploidy within 2–4 h	[75]
Cancer	Human pancreatic cells	DNA	Structural variations	MinION is able to detect SVs that inactivate oncosuppressor gene such as p16, SMAD4 and TSGs in pancreatic cancer cell lines	[14]
	Lymphocytic leukemia patients	DNA	*TP53* amplicon sequencing	By using a strategy based on long-template PCR, read error correction and post-variant calling filtering, nanopore technology shows to be more sensitive, manageable and less expensive than Sanger sequencing	[76]
Microbial community investigation	20 different bacterial species	DNA	16S rRNA genes	Using R7.3 chemistry and single sequencing run, MinION generated reads were enough to reconstruct >90% of the 16S rRNA gene sequences in species present in a mock reference community	[77]
	Escherichia coli and three poxviruses	DNA	Amplicon sequencing	By using amplicon sequencing, MinION technology can accurately identify and differentiate both viral and bacterial species present within biological samples despite over 98% identity between vaccinia strains	[63]
	Mouse gut microbiome	DNA	16S rRNA genes	At the species level, MinION sequencing allowed identification of more species than short-read sequencing, facilitating the accurate classification of the bacterial community composition	[78]
Transcriptome investigation	*Drosophila melanogaster*	RNA	Rdl, MRP, Mhc and Dscam1 genes	MinION could be used to sequence ‘‘full length’ Dscam1 cDNAs and complex genes with sufficient accuracy to identify isoforms	[15]
	Echis coloratus	RNA	Venom toxin-encoding genes	Both the hybrid (Illumina) and *de novo*-corrected MinION reads provide full-coding sequences and 5′/3′ UTRs for 29 of 33 candidate venom toxins detected, far superior to Illumina data and Sanger-based ESTs	[79]
	ERCC RNA Spike-In mix	RNA	92 polyadenylated transcripts and HEK293 cells	The majority of cDNAs were sequenced as full length by MinION with a good agreement in the measured cDNA abundance with PacBio RS II and Illumina HiSeq 2500 platforms	[80]
Sequencing in microgravity	Mouse bacteria/virus	DNA	WGS	For the first time ever, DNA was successfully sequenced in microgravity as part of the Biomolecule Sequencer experiment performed by NASA	[13]

Note. The table makes a list of all so far published studies involving MinION data, including details about the organism, the type of sample and the target and summarizing the main findings.

demonstrating that both variants and haplotypes can be resolved without the need for statistical phasing, while Minervini *et al.* [76] showed that Nanopore technology can be exploited for *TP53* gene mutations detection, as it correlated with Sanger sequencing but was more sensitive, manageable and less expensive.

Further, Wei *et al.* [75] suggested the potential use of this platform for aneuploidy detection in individuals with trisomies and monosomies, thus supporting its use in prenatal diagnosis and in particular in preimplantation genetic screening of embryos for *in vitro* fertilization.

The emergence of long-read technologies is going to revolutionize also transcriptomic studies, as they have the potential of overcoming assembly limitation of short-read RNA-seq in accurately reconstructing expressed full-length transcripts without the need for an assembly step [81].

Bolisetty *et al.* [15] used MinION data to identify around 8000 isoforms of four complex alternatively spliced genes in *Drosophila*, demonstrating that nanopore reads can be used to deconvolute complex transcriptomes where different but highly similar isoforms of the same gene are expressed, and for genes that have many exons and possible alternative promoters or 3′ ends [15, 79].

Recently, quantification accuracy and coverage performance have been assessed using the benchmarked External RNA Controls Consortium (ERCC) RNA Spike-In mix and a complementary DNA (cDNA) population of Human Embryonic Kidney (HEK)-293 cells [80]. In August 2016, the MinION was successfully used to sequence DNA in microgravity as part of the Biomolecule Sequencer experiment performed by NASA in the International Space Station (ISS). The experiments performed in the ISS demonstrated that DNA sequencing in space gives error profiles similar to those obtained on the Earth, opening a new era of scientific and medical possibilities to protect astronaut health during long-duration missions on the journey to Mars and to identify DNA-based life forms beyond the Earth (https://nanoporetech.com/publications/dna-sequencing-microgravity-international-space-station-iss-using-minion).

Closing remarks

The development of nanopore strand sequencing as a portable device and its commercial diffusion in the scientific community through an independent beta-testing program are going to revolutionize genomics. Thanks to this program, a large community of researchers had the opportunity to test this novel instrument for several applications that range from bacterial to human cancer genomics.

In just 2 years from the beginning of the MAP, the total throughput of a MinION sequencing run has grown from tens of Mb to 1–2 Gb and the read accuracy from 60 to 90% and, despite the short period of time from the beginning of the MAP, a large number of algorithms have been devised for this new generation of sequences. A large part of these methods come from the algorithmic recipes previously developed for PacBio reads and allow to successfully exploit nanopore reads for both resequencing and *de novo* assembly approaches.

The results reviewed here show that the use of nanopore data dramatically improves the *de novo* assembly of genomes in terms of N50 and number of contigs, and owing to the weak effect of classical sequencing bias, such as GC content, it allows for the exploration of structural variants with an unprecedented accuracy and resolution. Thanks to these properties, even small bioinformatic laboratories can now study genome structure all in-house for all those applications, such as structural variants detection, for which exact base variations are not of such high importance. However, despite the impressive improvements reached by ONT in the past 2 years, the use of these data for small variant calling is still challenging and, at present, it needs to be coupled with complementary Illumina sequences for mitigating the intrinsic biases of this technology. Although the R9 chemistry and the RNN base-caller reduced 2D read error rate to almost 7%, the only publicly available data set posted by Nick Loman at http://lab.loman.net/2016/07/30/nanopore-r9-data-release/ shows that errors still fall in recurrent positions of the genome and generate one false variant every 5–10 kb (data not reported).

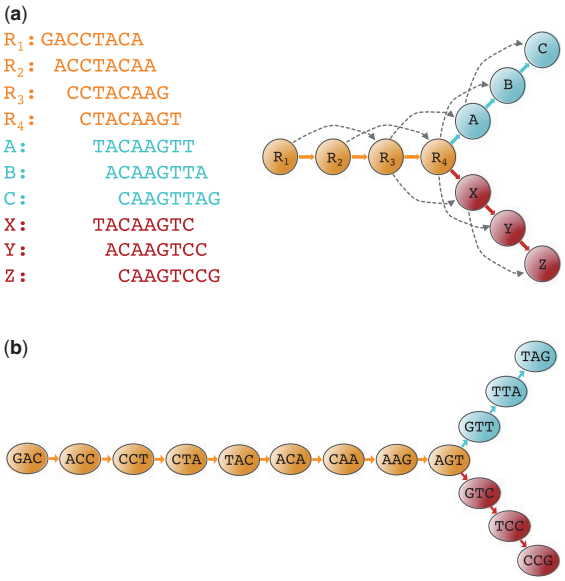
DNA translocation through the pore is a stochastic process in which the intervals between each advance of the strand are variable. These intervals can be too short to be revealed or too long to be indistinguishable from a repetitive sequence, and this is the main reason why the most common errors are deletions. In this scenario, improving the translocation mechanism will be a fundamental step for reaching the sequencing accuracy needed for clinical applications. In the meanwhile, novel base-calling algorithms should take better model event duration, and variant calling methods should include a priori nucleotide information (in a Bayesian framework) to mitigate the high false-positive rate.

At present, the great majority of ONT data applications are focused on small genomes as a consequence of the relatively low throughput offered by the MinION device. Recently, ONT announced the release of two novel high-throughput benchtop systems: the GridION X5 and the PromethION. The GridION X5 will allow to run up to five MinION Flow Cells with a total throughput of 50–100 Gb of data. On the other hand, the PromethION contains docking for 48 flow cells, each with 3000 nanopores (for 144 000 pores on the instrument), that will allow the device to generate a throughput 300 times that of the MinION. Thanks to the throughput of these two novel instruments, nanopore data could be exploited for studying the structure of larger genomes such as the human and other mammalian genomes, for unrevealing the architecture of highly repetitive regions, the detection of complex structural variants and haplotype reconstruction even at chromosomal level.

However, the impressive increase in data generation from a single sequencing run is going to become a serious challenge for the use and diffusion of this technology. At present, an R9.4 experiment can generate up to 1 Tb of HDF5 data, and as eukaryotic genome studies need multiple Flow Cells to gain sufficient coverage, the use of this platform will require computational infrastructures with tens of TB of storage capabilities for the analysis of a single genome. For this reason, the availability of novel tools for the compression of HDF5 files will be of fundamental importance for managing and analyzing the huge amount of data that will be generated by ONT platforms in the near future.

Recently, the Loman and his colleagues (http://lab.loman.net/2017/03/09/ultrareads-for-nanopore/) introduced a novel library preparation protocol that, by involving a modified Sambrook phenol–chloroform extraction/purification and minimal pipetting steps, allows to generate ultra-long reads up to 800 kb. Although ultra-long reads (in the order of 1 Mb) will improve *de novo* assembly and haplotype reconstruction, they pose computational challenges. As reported in Loman blog, the alignment of ultra-long reads is slow with both BWA and GraphMap aligners and will require novel and faster mapping tools.

Box 1-Assembly (Figure 5)



The great majority of *de novo* assembly algorithms are based on the OLC paradigms or on DBG, and both approaches exploit the overlap among sequenced reads to reconstruct an entire genome structure.

The OLC is computational procedure, introduced by Staden [82] and extended by many other scientists, made of three steps that consist in first finding overlap (O) among all reads, then creating a layout (L) of all the reads and their overlap on a graph and finally inferring the consensus (C) sequence. In OLC, the overlap between pairs of reads is calculated explicitly by doing all-against-all pairwise read alignment, and in the resulting graph, two nodes (reads) are linked when two reads overlap larger than a length cutoff.

On the other hand, DBG is an algorithmic approach, originally developed in 1995 by Idury and Waterman [83], in which the reads are chopped into k-mers to construct a graph in which k-mers are nodes and are linked if they are neighbors on the genome.

The algorithmic complexity and computational efficiency of these two approaches differ significantly, as the layout step of OLC is a Hamiltonian path problem, while inferring the contig sequence using the k-mer graph is an Euler path problem that is easier to resolve [84]. In fact, after the layout step, OLC needs to call the consensus sequence from the multiple sequence alignments, whereas the k-mer graph already includes the consensus information. Thanks to their capability to tolerate errors in overlap detection, OLC methods are more suited than DBG to handle high error-prone reads generated by TGS technologies, and several studies demonstrated that the Celera algorithm can produce high-quality assembly with ten times higher N50 and a smaller number of contigs compared with other tools.

The development of computational strategies for *de novo* assembly has accompanied the evolution of sequencing technologies, and while OLC algorithms were widely used in the Sanger sequencing era, the DBG approaches were properly devised to exploit the huge amount of short reads generated by SGS platforms. Publicly available tools based on OLC comprise Arachne [85], Celera Assembler [86], CAP3 [87], PCAP [88] and Phusion [89], whereas Euler-USR [90], Velvet [90], ABySS [91], AllPath-LG [46] and SOAPdenovo [84] are based on DBG.

Box 2-Resequencing

The first fundamental step in a resequencing study is to find the exact position of each read with respect to a reference genome (alignment), accounting for sample variance and sequencing errors.

The algorithmic approaches for sequence alignment fall into two main categories: global and local alignment. Global alignment consists in forcing the match between two sequences to span the entire length of the sequences, while local alignment finds segments of high similarity between a query and a subject sequence, which could be widely divergent. Several algorithms have been developed to face the sequence alignment problem including exact methods like dynamic programming (Smith–Waterman [92] for local and Needleman–Wunsch [44] for global alignment) and heuristic or probabilistic methods designed for large-scale database search, which are faster but do not guarantee to find best matches.

The most suitable algorithmic approach for aligning sequences against a reference genome is the local alignment Smith–Waterman algorithm. However, the use of this method is computationally infeasible for the great majority of real genomes, and for this reason, the alignment strategies developed for rapid sequences mapping are more frequently based on heuristic algorithms that use auxiliary data structures, called indices, to quickly identify matches between the reads and the reference genome. The great majority of indexing algorithms can be grouped into two main classes: algorithms based on hash tables and those based on suffix trees. Reads generated by Sanger sequencing, that are highly accurate and around 1000 bp long, are successfully mapped using methods based on hash table such as MEGABLAST [93] and BLAST [94].

On the other hand, the sequences produced by SGS platforms need approaches that are capable to handle huge amount of data where there is little variation between the read and the genome. The great majority of state-of-the-art methods for SGS reads are based on querying a compressed version of the suffix array, the BWT-FM [95] of a genome. TGS platforms do not have length limits of SGS and Sanger sequencing and generate reads of tens of kb with a high number of errors that are primarily InDels rather than substitutions. In this situation, the computational problem is how to align long (many kilobase) reads with moderate divergence from the genome (up to 20% divergence) and with speed and accuracy similar to SGS aligners.

At present, few approaches have been tested for TGS reads alignment, and these include a method originally designed for comparing large sequences (LAST [53]), an approach that combines the data structures used in short-read mapping with alignment methods used in whole-genome alignment (BLASR [54]), a procedure that progressively reduce the set of candidate mapping position (GraphMap [55]) and an extension of the short reads BWA algorithm that combines relaxed scoring of Smith–Waterman with heuristic filtering.

Key Points

- The Oxford Nanopore MinION is a pocket-sized (90 g and 10 cm in length) device that is able to generate DNA sequences with an average length of 2–10 kb and an error rate in the range 15–40%.
- Nanopore sequencing involves the transit of a ssDNA molecule through a nanoscopic pore and the contemporary measurement of its effect on an electric current. Raw current signals are then used to infer the sequence of the ssDNA by means of machine learning algorithms.
- Recently, various assembly pipelines have been tested for nanopore data: the OLC-based and the hybrid read error correction methods achieve better results than pipelines using the DBG scheme or non-hybrid correction approaches. In addition, currently available mapping algorithms return different results in handling nanopore error profiles (substitutions, insertions and deletions).
- The use of nanopore data for small variant calling is still challenging, and at present, it needs to be coupled with complementary short-read sequences for mitigating the intrinsic biases of this technology.
- Nanopore data are competitive for studying the structure of large genomes, such as the human genome, for unrevealing the architecture of highly repetitive regions, the detection of complex structural variants and haplotype reconstruction, even though they pose new computational challenges.

Funding

The Italian Ministry of Health, Young Investigators Award, Project GR-2011-02352026 'Detecting copy number variants from whole-exome sequencing data applied to acute myeloid leukemias' (to A.M.).

References

1. Metzker ML. Sequencing technologies - the next generation. *Nat Rev Genet* 2010;**11**:31–46.
2. Goodwin S, McPherson JD, McCombie WR. Coming of age: ten years of next-generation sequencing technologies. *Nat Rev Genet* 2016;**17**(6):333–51.
3. Eid J, Fehr A, Gray J, *et al*. Real-time DNA sequencing from single polymerase molecules. *Science* 2009;**323**(5910):133–8.
4. Clarke J, Wu HC, Jayasinghe L, *et al*. Continuous base identification for single-molecule nanopore DNA sequencing. *Nat Nanotechnol* 2009;**4**(4):265–70.
5. Deamer D, Akeson M, Branton D. Three decades of nanopore sequencing. *Nat Biotechnol* 2016;**34**(5):518–24.
6. Kasianowicz JJ, Brandin E, Branton D, *et al*. Characterization of individual polynucleotide molecules using a membrane channel. *Proc Natl Acad Sci USA* 1996;**93**(24):13770–3.
7. Haque F, Li J, Wu HC, *et al*. Solid-state and biological nanopore for real-time sensing of single chemical and sequencing of DNA. *Nano Today* 2013;**8**:56–74.
8. Iqbal SM, Akin D, Bashir R. Solid-state nanopore channels with DNA selectivity. *Nat Nanotechnol* 2007;**2**(4):243–8.
9. Feng Y, Zhang Y, Ying C, *et al*. Nanopore-based fourth-generation DNA sequencing technology. *Genomics Proteomics Bioinformatics* 2015;**13**:4–16.
10. Traversi F, Raillon C, Benameur SM, *et al*. Detecting the translocation of DNA through a nanopore using graphene nanoribbons. *Nat Nanotechnol* 2013;**8**(12):939–45.
11. Wendell D, Jing P, Geng J, *et al*. Translocation of double-stranded DNA through membrane-adapted phi29 motor protein nanopores. *Nat Nanotechnol* 2009;**4**(11):765–72.
12. Quick J, Ashton P, Calus S, *et al*. Rapid draft sequencing and real-time nanopore sequencing in a hospital outbreak of Salmonella. *Genome Biol* 2015;**16**:114.
13. Quick J, Loman NJ, Duraffour S, *et al*. Real-time, portable genome sequencing for Ebola surveillance. *Nature* 2016;**530**(7589):228–32.
14. Norris AL, Workman RE, Fan Y, *et al*. Nanopore sequencing detects structural variants in cancer. *Cancer Biol Ther* 2016;**17**(3):246–53.
15. Bolisetty MT, Rajadinakaran G, Graveley BR. Determining exon connectivity in complex mRNAs by nanopore sequencing. *Genome Biol* 2015;**16**:204.
16. Magi A, Giusti B, Tattini L. Characterization of MinION nanopore data for resequencing analyses. *Brief Bioinform* 2016, doi: 10.1093/bib/bbw077.
17. Goyal P, Krasteva PV, Van Gerven N, *et al*. Structural and mechanistic insights into the bacterial amyloid secretion channel CsgG. *Nature* 2014;**516**(7530):250–3.
18. Ip CLC, Loose M, Tyson JR, *et al*. MinION Analysis and Reference Consortium: Phase 1 data release and analysis. *F1000Res* 2015;**4**:1075.
19. Istace B, Friedrich A, d'Agata L, *et al*. *de novo* assembly and population genomic survey of natural yeast isolates with the Oxford Nanopore MinION sequencer. *Gigascience* 2017; **6**(2): 1–13.
20. Loman NJ, Quick J, Simpson JT. A complete bacterial genome assembled *de novo* using only nanopore sequencing data. *Nat Methods* 2015;**12**(8):733–5.
21. Goodwin S, Gurtowski J, Ethe-Sayers S, *et al*. Oxford Nanopore sequencing, hybrid error correction, and *de novo* assembly of a eukaryotic genome. *Genome Res* 2015;**25**(11):1750–6.
22. Laver T, Harrison J, O'Neill PA, *et al*. Assessing the performance of the Oxford Nanopore Technologies MinION. *Biomol Detect Quantif* 2015;**3**:1–8.
23. Quick J, Quinlan AR, Loman NJ. A reference bacterial genome dataset generated on the MinIONTM portable single-molecule nanopore sequencer. *Gigascience* 2014;**3**:22.
24. Jain M, Fiddes IT, Miga KH, *et al*. Improved data analysis for the MinION nanopore sequencer. *Nat Methods* 2015; **12**(4):351–6.
25. Ashton PM, Nair S, Dallman T, *et al*. MinION nanopore sequencing identifies the position and structure of a bacterial antibiotic resistance island. *Nat Biotechnol* 2015; **33**(3):296–300.
26. Urban J, Bliss J, Lawrence C, *et al*. Sequencing ultra-long DNA molecules with the Oxford Nanopore MinION. bioRxiv preprint first posted online May. 13, 2015 *bioRxiv* 2015, in press.
27. Karlsson E, Lärkeryd A, Sjödin A, *et al*. Scaffolding of a bacterial genome using MinION nanopore sequencing. *Sci Rep* 2015;**5**:11996.
28. Mikheyev AS, Tin MMY. A first look at the Oxford Nanopore MinION sequencer. *Mol Ecol Resour* 2014;**14**(6):1097–102.
29. Tarraga J, Gallego A, Arnau V, *et al*. HPG pore: an efficient and scalable framework for nanopore sequencing data. *BMC Bioinformatics* 2016;**17**:107.
30. Boža V, Brejová B, Vinař T. *DeepNano: Deep Recurrent Neural Networks for Base Calling in MinION Nanopore Reads*. 2016. https://arxiv.org/abs/1603.09195.

31. David M, Dursi LJ, Yao D, *et al*. Nanocall: an open source basecaller for Oxford Nanopore sequencing data. *Bioinformatics* 2017;**33**:49–55.
32. Loman NJ, Quinlan AR. Poretools: a toolkit for analyzing nanopore sequence data. *Bioinformatics* 2014;**30**(23):3399–401.
33. Watson M, Thomson M, Risse J, *et al*. poRe: an R package for the visualization and analysis of nanopore sequencing data. *Bioinformatics* 2015;**31**:114–5.
34. Leggett RM, Heavens D, Caccamo M, *et al*. NanoOK: multi-reference alignment analysis of nanopore sequencing data, quality and error profiles. *Bioinformatics* 2016;**32**:142–4.
35. Cao MD, Ganesamoorthy D, Cooper MA, *et al*. Realtime analysis and visualization of MinION sequencing data with npReader. *Bioinformatics* 2016;**32**(5):764–6.
36. Loose M, Malla S, Stout M. Real-time selective sequencing using nanopore technology. *Nat Methods* 2016;**13**(9):751–4.
37. Chin CS, Alexander DH, Marks P, *et al*. Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data. *Nat Methods* 2013;**10**(6):563–9.
38. Koren S, Schatz MC, Walenz BP, *et al*. Hybrid error correction and *de novo* assembly of single-molecule sequencing reads. *Nat Biotechnol* 2012;**30**(7):693–700.
39. Madoui MA, Engelen S, Cruaud C, *et al*. Genome assembly using Nanopore-guided long and error-free DNA reads. *BMC Genomics* 2015;**16**:327.
40. Szalay T, Golovchenko JA. *De novo* sequencing and variant calling with nanopores using PoreSeq. *Nat Biotechnol* 2015;**33**(10):1087–91.
41. Chin CS, Peluso P, Sedlazeck FJ, *et al*. Phased diploid genome assembly with single-molecule real-time sequencing. *Nat Methods* 2016;**13**(12):1050–4.
42. Li H. Minimap and miniasm: fast mapping and *de novo* assembly for noisy long sequences. *Bioinformatics* 2016;**32**(14):2103–10.
43. Koren S, Walenz B, Berlin K, *et al*. Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. *Genome Res* 2017;**27**(5):722–736.
44. Needleman SB, Wunsch CD. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J Mol Biol* 1970;**48**(3):443–53.
45. Bankevich A, Nurk S, Antipov D, *et al*. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *J Comput Biol* 2012;**19**(5):455–77.
46. Gnerre S, Maccallum I, Przybylski D, *et al*. High-quality draft assemblies of mammalian genomes from massively parallel sequence data. *Proc Natl Acad Sci USA* 2011; **108**(4):1513–8.
47. Lin Y, Yuan J, Kolmogorov M, *et al*. Assembly of long error-prone reads using de Bruijn graphs. *Proc Natl Acad Sci USA* 2016; **113**(52):E8396–E8405.
48. Myers EW. AnO (ND) difference algorithm and its variations. *Algorithmica* 1986;**1**(1–4):251–66.
49. Deschamps S, Mudge J, Cameron C, *et al*. Characterization, correction and *de novo* assembly of an Oxford Nanopore genomic dataset from *Agrobacterium tumefaciens*. *Sci Rep* 2016; **6**:28625.
50. Lu H, Giordano F, Ning Z. Oxford Nanopore MinION sequencing and genome assembly. *Genomics Proteomics Bioinformatics* 2016;**14**(5):265–79.
51. Sović I, Križanović K, Skala K, *et al*. Evaluation of hybrid and non-hybrid methods for *de novo* assembly of nanopore reads. *Bioinformatics* 2016;**32**(17):2582–9.
52. Walker BJ, Abeel T, Shea T, *et al*. Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PLoS One* 2014;**9**(11):e112963.
53. Kiełbasa SM, Wan R, Sato K, *et al*. Adaptive seeds tame genomic sequence comparison. *Genome Res* 2011;**21**(3):487–93.
54. Chaisson MJ, Tesler G. Mapping single molecule sequencing reads using basic local alignment with successive refinement (BLASR): application and theory. *BMC Bioinformatics* 2012;**13**:238.
55. Sović I, Šikić M, Wilm A, *et al*. Fast and sensitive mapping of nanopore sequencing reads with GraphMap. *Nat Commun* 2016;**7**:11307.
56. Rand AC, Jain M, Eizenga JM, *et al*. Mapping DNA methylation with high-throughput nanopore sequencing. *Nat Methods* 2017;**14**(4):411–3.
57. Simpson JT, Workman RE, Zuzarte PC, *et al*. Detecting DNA cytosine methylation using nanopore sequencing. *Nat Methods* 2017;**14**(4):407–10.
58. Altschul SF, Madden TL, Schäffer AA, *et al*. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* 1997;**25**(17):3389–402.
59. Risse J, Thomson M, Patrick S, *et al*. A single chromosome assembly of Bacteroides fragilis strain BE1 from Illumina and MinION nanopore sequencing data. *Gigascience* 2015;**4**:60.
60. Karamitros T, Harrison I, Piorkowska R, *et al*. *De novo* assembly of human herpes virus type 1 (HHV-1) genome, mining of non-canonical structures and detection of novel drug-resistance mutations using short- and long-read next generation sequencing technologies. *PLoS One* 2016;**11**(6):e0157600.
61. Wang J, Moore NE, Deng YM, *et al*. MinION nanopore sequencing of an influenza genome. *Front Microbiol* 2015;**6**:766.
62. Ammar R, Paton TA, Torti D, *et al*. Long read nanopore sequencing for detection of HLA and CYP2D6 variants and haplotypes. *F1000Res* 2015;**4**:17.
63. Kilianski A, Haas JL, Corriveau EJ, *et al*. Bacterial and viral identification and differentiation by amplicon sequencing on the MinION nanopore sequencer. *Gigascience* 2015;**4**:12.
64. Li H, Durbin R. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* 2009;**25**(14):1754–60.
65. Kurtz S, Phillippy A, Delcher AL, *et al*. Versatile and open software for comparing large genomes. *Genome Biol* 2004;**5**(2):R12.
66. Angiuoli SV, Salzberg SL. Mugsy: fast multiple alignment of closely related whole genomes. *Bioinformatics* 2011;**27**(3): 334–42.
67. Darling ACE, Mau B, Blattner FR, *et al*. Mauve: multiple alignment of conserved genomic sequence with rearrangements. *Genome Res* 2004;**14**(7):1394–403.
68. Wilm A, Aw PPK, Bertrand D, *et al*. LoFreq: a sequence-quality aware, ultra-sensitive variant caller for uncovering cell-population heterogeneity from high-throughput sequencing datasets. *Nucleic Acids Res* 2012;**40**(22):11189–201.
69. Layer RM, Chiang C, Quinlan AR, *et al*. LUMPY: a probabilistic framework for structural variant discovery. *Genome Biol* 2014;**15**(6):R84.
70. Schübeler D. Function and information content of DNA methylation. *Nature* 2015;**517**(7534):321–6.
71. Hoenen T, Groseth A, Rosenke K, *et al*. Nanopore Sequencing as a Rapidly Deployable Ebola Outbreak Tool. *Emerg Infect Dis* 2016;**22**(2):331–4.
72. Greninger AL, Naccache SN, Federman S, *et al*. Rapid metagenomic identification of viral pathogens in clinical samples by real-time nanopore sequencing analysis. *Genome Med* 2015;**7**:99.
73. Judge K, Harris SR, Reuter S, *et al*. Early insights into the potential of the Oxford Nanopore MinION for the detection of antimicrobial resistance genes. *J Antimicrob Chemother* 2015;**70**(10):2775–8.

74. Schmidt K, Mwaigwisya S, Crossman LC, *et al*. Identification of bacterial pathogens and antimicrobial resistance directly from clinical urines by nanopore-based metagenomic sequencing. *J Antimicrob Chemother* 2017;**72**:104–14.
75. Wei S, Williams Z. Rapid short-read sequencing and aneuploidy detection using MinION Nanopore technology. *Genetics* 2016;**202**:37–44.
76. Minervini CF, Cumbo C, Orsini P, *et al*. TP53 gene mutation analysis in chronic lymphocytic leukemia by nanopore MinION sequencing. *Diagn Pathol* 2016;**11**:96.
77. Benítez-Páez A, Portune KJ, Sanz Y. Species-level resolution of 16S rRNA gene amplicons sequenced through the MinION™ portable nanopore sequencer. *Gigascience* 2016;**5**:4.
78. Shin J, Lee S, Go MJ, *et al*. Analysis of the mouse gut microbiome using full-length 16S rRNA amplicon sequencing. *Sci Rep* 2016;**6**:29681.
79. Hargreaves AD, Mulley JF. Assessing the utility of the Oxford Nanopore MinION for snake venom gland cDNA sequencing. *PeerJ* 2015;**3**:e1441.
80. Oikonomopoulos S, Wang YC, Djambazian H, *et al*. Benchmarking of the Oxford Nanopore MinION sequencing for quantitative and qualitative assessment of cDNA populations. *Sci Rep* 2016;**6**:31602.
81. Conesa A, Madrigal P, Tarazona S, *et al*. A survey of best practices for RNA-seq data analysis. *Genome Biol* 2016;**17**:13.
82. Staden R. A new computer method for the storage and manipulation of DNA gel reading data. *Nucleic Acids Res* 1980;**8**(16):3673–94.
83. Idury RM, Waterman MS. A new algorithm for DNA sequence assembly. *J Comput Biol* 1995;**2**(2):291–306.
84. Li R, Zhu H, Ruan J, *et al*. *De novo* assembly of human genomes with massively parallel short read sequencing. *Genome Res* 2010;**20**(2):265–72.
85. Batzoglou S, Jaffe DB, Stanley K, *et al*. ARACHNE: a whole-genome shotgun assembler. *Genome Res* 2002;**12**:177–89.
86. Myers EW, Sutton GG, Delcher AL, *et al*. A whole-genome assembly of Drosophila. *Science* 2000;**287**(5461):2196–204.
87. Huang X, Madan A. CAP3: A DNA sequence assembly program. *Genome Res* 1999;**9**(9):868–77.
88. Huang X, Wang J, Aluru S, *et al*. PCAP: a whole-genome assembly program. *Genome Res* 2003;**13**(9):2164–70.
89. Mullikin JC, Ning Z. The phusion assembler. *Genome Res* 2003;**13**:81–90.
90. Chaisson MJ, Brinza D, Pevzner PA. *De novo* fragment assembly with short mate-paired reads: Does the read length matter? *Genome Res* 2009;**19**(2):336–46.
91. Birol I, Jackman SD, Nielsen CB, *et al*. *De novo* transcriptome assembly with ABySS. *Bioinformatics* 2009;**25**(21):2872–7.
92. Smith TF, Waterman MS. Identification of common molecular subsequences. *J Mol Biol* 1981;**147**:195–7.
93. Zhang Z, Schwartz S, Wagner L, *et al*. A greedy algorithm for aligning DNA sequences. *J Comput Biol* 2000;**7**(1-2):203–14.
94. Altschul SF, Gish W, Miller W, *et al*. Basic local alignment search tool. *J Mol Biol* 1990;**215**(3):403–10.
95. Ferragina P, Manzini G. Indexing compressed text. *J. ACM* 2005;**52**(4):552–81. [http://doi.acm.org/10.1145/1082036.1082039].