Oxford NANOPORE Technologies

WHITE PAPER

Nanopore sequencing

The advantages of long reads for genome assembly

SEPTEMBER 2017

Contents

Introduction

Over the last decade, improvements in next generation DNA sequencing technology have transformed the field of genomics, making it an essential tool in modern genetic and clinical research laboratories. The facility to sequence whole genomes or specific genomic regions of interest is delivering new insights into a variety of applications such as human health and disease, metagenomics, antimicrobial resistance, evolutionary biology and crop breeding.

For applications such as the analysis of larger structural variation, or *de novo* assembly, whole genome sequencing (WGS) is typically the technique of choice.

One of the key steps of WGS is the accurate assembly of the vast amount of data generated into a contiguous stretch of DNA sequence. This review provides a background to the DNA assembly process and the associated advantages of long or ultra-long DNA reads, as provided by nanopore sequencing technology.

Whole genome assembly – solving the puzzle

Traditional technologies have required users to sequence short lengths of DNA, which must then be reassembled back into their original order as accurately as possible. Such short-read sequencing technologies, however, present a number of challenges, particularly the difficulty of accurately analysing repetitive regions and large structural variations.[1]

This means that many reference genomes that were created using short-read sequencing are highly fragmented, which in turn introduces bias into any alignments made against that reference[2]. This review shows how these challenges are now being met, by the emergence of long-read nanopore sequencing[3].

1

De novo assembly

De novo assembly aims to reconstruct the original genome sequence from the set of reads, for example when there is no reference genome available or a scientist prefers to avoid potential bias that could arise from using an imperfect reference. Generally, the first step is to find overlaps between the reads and build a graph to describe their relationship.

Several approaches have been used for this purpose, including Greedy extension, de Bruijn graphs and Overlap Layout Consensus (OLC) (Figure 1). The graph is then simplified and a consensus is extracted. A key step in this process is the assembly of contiguous, uninterrupted stretches of overlapping DNA, so-called contigs. As detailed in the next section, long-read sequencing technology offers many advantages for both *de novo* and alignment-based genome assembly.

Long-read sequencing technology offers simplified and less ambiguous genome assembly.

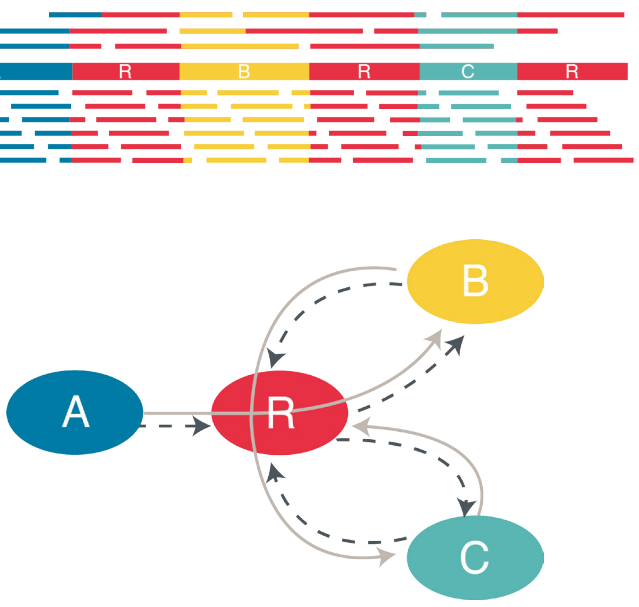


Figure 1
A schematic showing how long-read sequencing can deliver simplified, less ambiguous genome assembly. Long reads (solid arrows) have greater overlap with other reads than is provided by short reads (dashed arrows), allowing more accurate assemblies, especially in repeat regions (R). Image adapted from Schatz (2014)[4].

2

Advantages of long-read sequencing for genome assembly

Until recently, cost-effective generation of large amounts of sequence data could only be performed using short-read (<300 bp) sequencing technologies; however, nanopore-based sequencing can process very long DNA fragments (currently up to 950 kb)[5] to create long reads, which deliver a number of significant benefits:

Long-read nanopore sequencing offers easier assembly and the ability to span repetitive genomic regions.

Ease of assembly

The longer a sequencing read, the more overlap it will have with other reads. As such, it is much easier to assemble the DNA fragments back into the correct order. This can be visualised much like a jigsaw; the larger the pieces, the easier the puzzle (Figure 2).

Facility to span repetitive genomic regions

Most genomes contain significant amounts of repetitive DNA (e.g. transposons, satellites, gene duplications). As the short reads produced by traditional next generation sequencing (NGS) technology may not span each given repetitive region, the resulting genome assemblies can be highly fragmented[6]. Long-read sequencing technologies have a significant advantage here as the reads generated are more likely to span the full repetitive region, allowing the creation of accurate genome assemblies with minimal gaps (see Figure 3 and Case Study 1, page 7).

~ 50 base read
~ 92,000 "pieces"

~ 500,000 base read
~ 9 "pieces"

Figure 2
Like a jigsaw puzzle with large pieces, long-read DNA is much easier to assemble than short-read DNA. The *Escherichia coli* genome comprises 4.6 million bases, which would equate to 92,000 fragments of 50 bp or just 9 fragments of 500 kb in length.

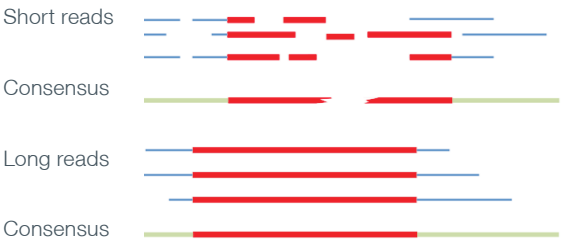


Figure 3
A schematic highlighting the advantages of long reads in *de novo* assembly of repetitive regions. Long read lengths are more likely to incorporate the whole repetitive region (shown in red) allowing more accurate assembly with fewer gaps. Image adapted from Sam Demharter[7].

Identification of large structural variation

The term structural variation (SV) covers a range of genetic alterations, including copy number variation (CNV), duplications, translocations and inversions (Table 1).

Traditionally, structural variants were defined as spanning over 1,000 base pairs; however, with the advent of higher resolution genome-analysis techniques, some researchers have now revised this definition to spanning over 50 base pairs[8]. Such structural variants have been associated with a number of diseases (including autism, schizophrenia and cancer) and, for this reason, have become increasingly important areas for research[9].

Most existing genome assemblies have been created using short-read sequencing technology, which is limited in its ability to capture large structural variation. Even the well-studied human genome has large gaps in its assembly[6].As CNV alone is estimated to comprise 4.8–9.5% of the human genome[10], it is clear that accurate analysis of such regions is of significant importance.

Unlike short-read technology, longer nanopore-sequencing reads can cover the whole structure variant in one read. This results in more accurate genome assemblies facilitating better understanding of genome architecture in genomic diseases (see Figure 4 and Case Studies 2 and 3, pages 8, 9 and 10)[11,12,13].

Structural variants	Definition
Copy number variation (includes insertions and deletions)	A segment of DNA which has greater (insertion) or fewer (deletion) copies than expected
Duplications	A segment of DNA that occurs in two or more copies per haploid genome
Translocation	A change in position of a region of DNA which involves no change to the total DNA content
Inversions	A segment of DNA that is reversed end to end

Table 1
Definition of common large structural variants.

Figure 4
A schematic highlighting the advantages of long reads in *de novo* assembly of copy number variations (CNVs). Long read lengths are more likely to contain the whole CNV, providing less potential for error in the consensus sequence. Image adapted from Sam Demharter[7].

Assembly quality

The quality of genome assemblies is often assessed by the number of contigs required to represent the genome, the length of these contigs relative to the size of the genome and the proportion of reads that can be assigned to the contigs. One of the most commonly used assembly metrics is N50.

The N50 value is calculated by sorting all contigs by length, calculating the cumulative assembly length (i.e. adding the length of all contigs together) and identifying the length of the contig which is situated in the middle of the cumulative length (Figure 5). The higher the N50 value, the more contiguous the assembly.

The much higher N50 values generated using long-read nanopore sequencing, underlines the superior quality and completeness of genome assembly when compared with short-read sequencing technology (Table 2 and Case Study 1).

Figure 5
The N50 metric is one measure of the quality of a genome assembly. Other metrics such as N75 and N25 are also sometimes used to further assess assembly quality. Image adapted from Jansen (2016)[14].

Table 2
Comparison between genome assemblies of *Rhizoctonia solani* using long- and short-read sequencing platforms. The higher N50 value coupled with the fewer number of contigs, obtained using long-read length nanopore technology is indicative of superior genome assembly quality. Table adapted from Datema *et al* (2016)[15].

Isolate	Technology	Number of contigs	N50 length
AG1-IA	Short read	6,452	20 kb
AG3	Short read	6,040	26 kb
AG8	Short read	7,606	7 kb
AG1-IB	Short read	3,793	35 kb
AGX	Nanopore long read	606	199 kb

3

Genome assembly tools

Many software tools are now available for genome assembly utilising long-read sequencing data. Tools differ in what they offer and the extent of their use and should be assessed according to the specific project requirements (Table 3). Assembly tools often include, or are coupled with, error-correction methods, that utilise information in the assembled reads, or couple information from other reads to produce high consensus accuracy.

A recent study by Cherukuri and Janga (2016) comparing a number of assembly algorithms found the OLC approach to be most favourable for long-read nanopore sequencing – delivering higher genome coverage combined with significantly larger average contig lengths and fewer overall contig numbers[27]. However, other studies utilising combined de Bruijn and OLC approaches also report accurate genome assemblies[25,28].

Table 3
List of assemblers designed or adapted for Oxford Nanopore long reads. Adapted from Lu *et al* (2016)[26]. For the latest information on genome assemblers, visit www.nanoporetech.com/publications

Assembler name	Algorithms	Error correction	Link	Reference
LQS	DALIGNER, Celera OLC	Nanocorrect, Nanopolish	https://github.com/jts/nanopolish	Loman (2015)[16]
Canu	MHAP, Celera OLC	Canu	https://github.com/marbl/canu	Berlin (2015)[17]
Canu	MHAP, Celera OLC	Racon, Pilon	https://github.com/nanoporetech/ont-assembly-polish	nanoporetech[18]
Miniasm	OLC		https://github.com/lh3/minia	Li (2016)[19]
Miniasm	OLC	Racon	https://github.com/isovic/racon	Vaser (2017)[20]
Ra-integrate	OLC		https://github.com/mariokostelac/ ra-integrate/	Sovic (2016)[21]
ALLPATHS-LG	de Bruijn graph	ALLPATHS-LG	https://www.broadinstitute.org/software/allpaths-lg/blog/?page_id	Gnerrea (2011)[22]
SPAdes	de Bruijn graph	SPAdes	http://bioinf.spbau.ru/spades	Bankevich (2012)[23]
SMART denovo	Smith-Waterm, dot matrix		https://github.com/ruanjue/smart denovo	Ruan[24]
ABruijn	de Bruijn graph		https://github.com/fenderglass/ABruijn	Lin (2016)[25]

LQS: Loman, Quick and Simpson; MHAP: MinHash alignment process; OLC: Overlap Layout Consensus.

4

Case study 1
The need for accurate plant and plant pathogen genomes

KeyGene, the crop innovation company with headquarters in the Netherlands focusses on enhancing both yield and quality aspects in vegetable and field crops. Plant and plant pathogen genomes offer many challenges, including large polyploid genome sizes and high levels of repeat content (Figure 6). To address these issues Dr Alexander Wittenberg's team at KeyGene are utilising long read-length sequencing technologies such as the MinION™ to deliver high-quality reference genomes.

KeyGene were able to assemble the genome with almost 10-fold fewer contigs than when using short-read sequencing, enabling a much more complete and contiguous genome assembly.

The team chose the prevalent plant pathogen *Rhizoctonia solani* with its highly repetitive genome as a suitable test candidate for genome sequencing. While existing reference genomes are available for this species, they are highly fragmented, reflecting the short-read sequencing technology used to generate them[30].

Utilising the MinION, KeyGene were able to assemble the genome with almost 10-fold fewer contigs than when using short-read sequencing, enabling a much more complete and contiguous genome assembly. An N50 of 199 kb for the long-read sequencing compared favourably against an N50 size of 7–35 kb obtained with short-read approaches (Table 2).[15]

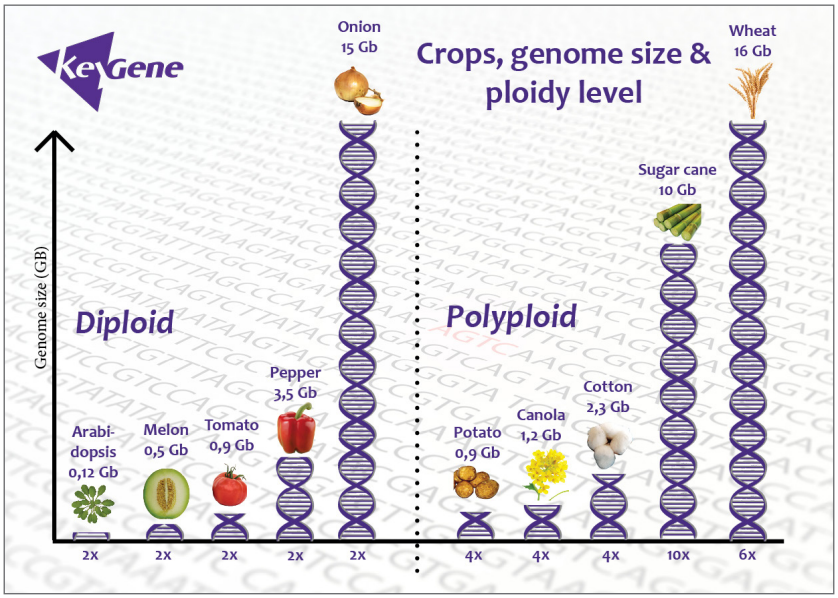


Figure 6
Polyploidy, where there are more than two paired chromosomes, is especially common in plants and can result in large genome sizes. For reference, the human genome is 3.3 Gb. Image courtesy of KeyGene, The Netherlands.

Case study 2
Long-read applications for human genome analysis

While the human genome has been extensively studied, there are still gaps in the coverage that cannot be easily resolved using short-read sequencing[6].
As a consequence, a number of researchers are now applying long-read sequencing technology to improve the human genome assembly and accurately resolve complex structural genomic variation. Recently, two independent groups have utilised the MinION to accurately resolve haplotype phasing, allowing the correct determination of which alleles are on the same chromosome.

Dr. Wigard Kloosterman at UMC Utrecht demonstrated that the long-read MinION platform could accurately determine haplotype phasing for chromothripsis – a highly clustered rearrangement of chromosomal DNA which has been shown to occur in both cancerous and congenital disease samples[9] (Figure 7). In addition to the detection of 40 chromothripsis breakpoints previously identified by short-read sequencing, the MinION platform identified many additional breakpoints with a high degree of accuracy.

In addition to the detection of 40 chromothripsis breakpoints previously identified by short-read sequencing, the MinION platform identified many additional breakpoints with a high degree of accuracy.

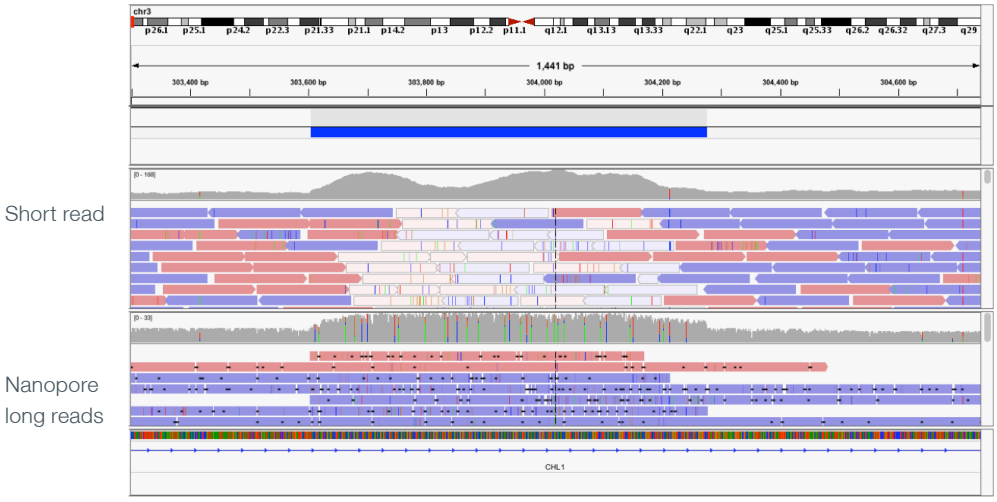


Figure 7
Nanopore long-read sequencing allowed identification of a tandem duplication, which was not evident when using short-read traditional sequencing technology. Figure courtesy of Dr. Wigard Kloosterman, UMC Utrecht, The Netherlands.

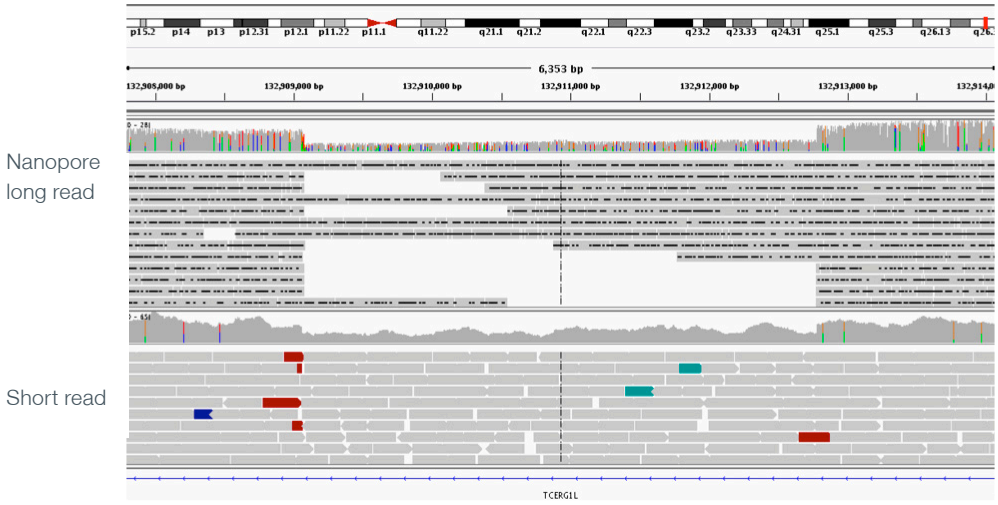
Similarly, Professor Michael Simpson from Genomics plc. together with researchers at the Wellcome Trust Centre for Human Genetics (WTCHG) utilised the MinION to study haplotype phasing in a sample displaying an immune disorder associated with cytopenia and cerebellar ataxia.[12]

Nanopore sequencing identified a large structural variation in a number of regions of a particularly well-characterised human reference genome, which were not evident via short-read sequencing alone.

Using the long reads provided by the MinION, the WTCHG team were able to demonstrate that two previously characterised *de novo* missense variants located 2.2 kb apart in the *SAMD9L* gene were in fact on the same haplotype.

The WTCHG and Genomics plc teams also used the MinION to identify large structural variation in a number of regions of NA12878, a particularly well-characterised human reference genome, which were not evident via short-read sequencing alone - further demonstrating the advantages of long-read sequencing technology (Figure 8).

Figure 8
The long reads generated by nanopore sequencing allowed the identification of a large 7 kb deletion on chromosome 10 of the well-characterised human reference genome, NA12878. This deletion was not evident using short-read sequencing technology. Data courtesy of Michael Simpson, Genomics plc and WTCHG, UK.



Case study 3 Improved genome assembly using long reads

European eel genome

Dr. Hans Jansen at ZF-screens B.V is applying nanopore sequencing to create a more accurate reference genome of the European eel to support breeding, conservation and fundamental research efforts[31]. The team compared long- and short-read approaches and noted several points of difference. One striking finding was that short-read sequencing indicated a genome size of 923 Mb, while nanopore long-read sequencing determined a genome size of 860 Mb.

Upon detailed examination of the data, the team suggest this discrepancy was caused by overestimating the size of the gaps between the contigs in the scaffolding process.

The study also identified another limitation of the short-read sequencing approach: the underrepresentation of high GC content. Impressively, assembly of the long reads using TULIP, a new and efficient assembly tool, took just two days on a desktop computer and the resultant genome was two orders of magnitude less fragmented than the previous short-read based draft (Figure 9).

In summary, the team commented:
'The resulting genome assembly significantly improves on a previous draft based on short reads only, both in terms of contiguity and structural quality.'

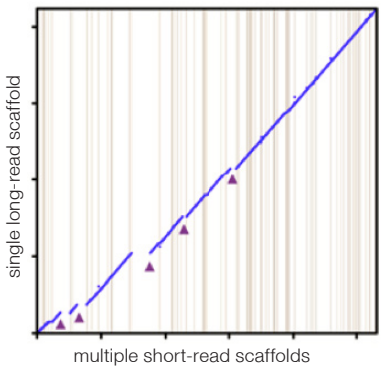


Figure 9
Using TULIP, multiple short-read scaffolds (x-axis) can be aligned against a single long-read scaffold (y-axis). Boundaries of the short-read scaffolds are indicated by the grey vertical lines, highlighting how many short-read scaffolds fit within a single long-read scaffold. Discrepancies between the long- and short-read scaffolds are indicated by the arrowheads. Close examination of the data shows the long-read scaffold to be correct.

S. cerevisiae

To assess the application of long-read sequencing technology for genome assembly, Istace *et al* (2017) used the MinION from Oxford Nanopore Technologies for *de novo* sequencing of 21 *Saccharomyces cerevisiae* isolates.[13] The results showed that assembly contiguity was 14 times higher than when using short-read assemblies, obtaining one or two long contigs for 65% of the chromosomes.

This high contiguity allowed the accurate detection of large structural variations across the 21 studied genomes. Moreover, because of the high completeness of the MinION assemblies, the team were able to produce a complete cartography of transposable element insertions and inspect structural variants that are generally missed using a short-read sequencing strategy (Figure 10).

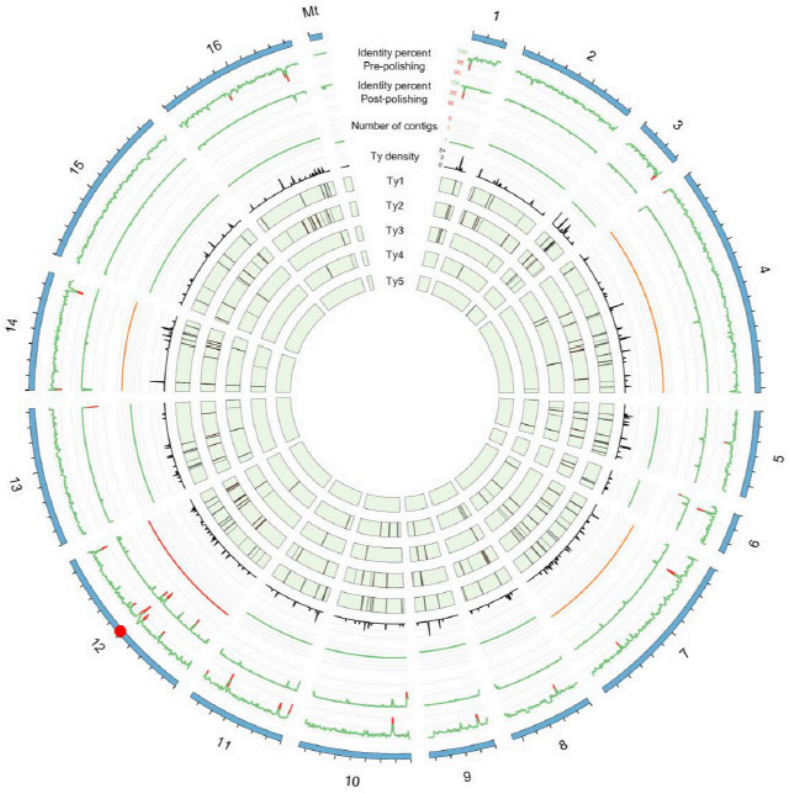


Figure 10
Complete cartography of the Ty transposon family in *Saccharomyces cerevisiae.* Long-read nanopore sequencing allowed the identification of transposable elements and structural variants that are generally missed using a short-read sequencing strategy. The inside tracks show the location of the Ty1, Ty2, Ty3, Ty4, and Ty5 transposons across all strains. Figure courtesy of Benjamin Istace, Genoscope, France.

5

Summary

Long-read sequencing technology offers a real alternative to solve the genome assembly difficulties and improve the completeness of genome assemblies where short-read sequencing has failed[13], particularly the facility to accurately sequence and map repetitive regions and structural variations[9,12,13]. In addition, long-read sequencing offers exciting possibilities for other applications where the accurate identification of structural variation is important, for example in the study of cancer progression[29].

It is clear that the growing usage of long-read sequencing technology and associated improvements in genome assemblies will bring additional and rapid insight into genomics, further refining the relationship of phenotype to genotype.

6

About Oxford Nanopore Technologies

Figure 11
The MinION from Oxford Nanopore is a pocket-sized, portable sequencing device.

Oxford Nanopore Technologies is at the forefront of genomics, having introduced the world's first nanopore DNA sequencer, the MinION - a portable, real time, long-read, low-cost device. Through the utilisation of long reads, the MinION delivers high-quality whole genome sequencing with the facility to span repetitive regions and structural variations, offering significant advantages over traditional short-read sequencing technology. The analysis of larger genomes or higher throughput requirements are met by the GridION X5™ and PromethION™ devices.

These compact benchtop systems utilise the same nanopore technology as the MinION, offering up to 5 and 48 flow cells respectively. Each flow cell can be used independently, with the user choosing how many are used at any one time, enabling different experiments to be run in parallel.

A number of protocols are available for the nanopore sequencing, enabling optimised whole genome analysis for a range of sample types and DNA input amounts. Library preparation can be performed in less than 10 minutes, allowing the generation of 5–10 Gb of data from just 200 ng of starting DNA.

Find out more at:
www.nanoporetech.com

7 References

1. Mak, A.C. *et al* (2016) Genome-Wide Structural Variation Detection by Genome Mapping on Nanochannel Arrays. *Genetics* 202(1):351–62
2. Huang, L. (2013) Short-read alignment with populations of genomes. *Bioinformatics* 29:361–370
3. Norris, A.L. *et al* (2016) Nanopore sequencing detects structural variants in cancer. *Cancer Biol Ther* 17(3):246–53
4. Schatz, M. (2014). *De novo* assembly of complex genomes using single molecule sequencing. Presentation. Available at: http://schatzlab.cshl.edu/presentations/2014-01-14.PAG.Single%20Molecule%20Assembly.pdf [Accessed 2 March 2017]
5. Loman, N. (2017) 950kb, yeah! [Twitter] 08 March. Available from: https://twitter.com/pathogenomenick/status/839604187100934144 [Accessed: 09 March 2017] http://lab.loman.net/2017/03/09/ultrareads-for-nanopore/?rev1
6. Chaisson, M.J.P. *et al* (2015) Genetic variation and the *de novo* assembly of human genomes. *Nat Rev Genet* 16(11):627–40
7. Demharters, S. Available at: https://github.com/demharters/assemblyTutorial [Accessed: 5 January 2017]
8. Eichler, E. (2015) Genome structural variation. Presentation. Available at: http://evomicsorg.wpengine.netdna-cdn.com/wp-content/uploads/2015/01/011815CeskyKrumlowcompgenomicse.pdf [Accessed 05 January 2017]
9. Stankiewicz, P. and Lupski, J.R. (2010) Structural variation in the human genome and its role in disease. *Annu Rev Med* 19:455–77
10. Zarrei, M. *et al* (2015) A copy number variation map of the human genome. *Nat Rev Genet* 16(3):172–83
11. Kloosterman, W. (2016) Characterization of structural variations and chromothripsis in nanopore sequencing data of human genomes. Presentation. Available at: https://nanoporetech.com/events/ncm16#193982637 [Accessed 05 January 2017]
12. Simpson, M. (2016) Sequencing human genomes with nanopore technology. Presentation. Available at: https://nanoporetech.com/events/ncm16#193976224 [Accessed: 05 January 2017]
13. Istace, B. *et al* (2017) *De novo* assembly and population genomic survey of natural yeast isolates with the Oxford Nanopore MinION sequencer. GigaScience http://dx.doi.org/10.1093/gigascience/giw018
14. Jansen, H. (2016) *De novo* genome assembly with MinION long reads. Presentation. Available at: https://nanoporetech.com/events/knowledge-exchange-*de-novo*-assembly-minion-long-reads [Accessed: 05 January 2017]
15. Datema, E. *et al* (2016). The megabase-sized fungal genome of *Rhizoctonia solani* assembled from nanopore reads only. bioRxIV 084772; doi: https://doi.org/10.1101/084772
16. Loman, N.J, Quick, J. and Simpson, J.T. (2015) A complete bacterial genome assembled *de novo* using only nanopore sequencing data. *Nat Methods* 12:733–5
17. Berlin, K. *et al* (2015) Assembling large genomes with single-molecule sequencing and locality-sensitive hashing. *Nat Biotechnol* 33(6):623–30
18. Oxford Nanopore (2016). Hybrid assembly pipeline released (using Canu, Racon and Pilon) [Online]. Available at: https://nanoporetech.com/publications/hybrid-assembly-pipeline-released-using-canu-racon-and-pilon [Accessed: 05 January 2017]
19. Li, H. (2016) Minimap and miniasm: fast mapping and *de novo* assembly for noisy long sequences. *Bioinformatics* 32:2103–10
20. Vaser, R. (2017) Fast and accurate *de novo* genome assembly from long uncorrected reads. *Genome Res*. doi: 10.1101/gr.214270.116. [Epub ahead of print]
21. Sovic, I. *et al* (2016) Evaluation of hybrid and non-hybrid methods for *de novo* assembly of nanopore reads. *Bioinformatics* 32:2582–9
22. Gnerrea, S. *et al* (2011) High-quality draft assemblies of mammalian genomes. *Proc Natl Acad Sci USA* 108(4):1513–8
23. Bankevich, A. *et al* (2012) SPAdes: a new genome assembly algorithm and its applications to singlecell sequencing. *J Comput Biol.* 19(5):455–77
24. Ruan, J. Available at: https://github.com/ruanjue/smartdenovo [Accessed: 5 January 2017].
25. Lin, Y. *et al* (2016) Assembly of long error-prone reads using de Bruijn graphs. *Proc Natl Acad Sci USA* 113(52):E8396-E8405. doi: 10.1073/pnas.1604560113
26. Lu, H., Giordano, F. and Ning, Z. (2016) Oxford Nanopore MinION Sequencing and Genome Assembly. *Genomics Proteomics Bioinformatics* 14(5):265–79
27. Cherukuri, Y. and Janga, S.C. (2016) Benchmarking of *de novo* assembly algorithms for Nanopore data reveals optimal performance of OLC approaches. *BMC Genomics* 22: 17 Suppl 7:507. doi: 10.1186/s12864-016-2895-8
28. Judge, K. *et al* (2016) Comparison of bacterial genome assembly software for MinION data and their applicability to medical microbiology. *Microbial Genomics*. doi: 10.1099/mgen.0.000085
29. Norris, A.L. *et al* (2016) Nanopore sequencing detects structural variants in cancer. *Cancer Biol Ther* 17(3):246–53
30. Wittenberg, A. (2016) A novel approach to elucidate the genomic structure of plants and pathogens. 2016. Presentation. Available at: https://nanoporetech.com/publications/novel-approach-elucidate-genomic-structure-plants-and-pathogens [Accessed: 5 January 2017]
31. Jansen, H.J. *et al* (2017) Rapid *de novo* assembly 1 of the European eel genome from nanopore sequencing reads. bioRxiv doi. Available at: http://dx.doi.org/10.1101/101907

Oxford Nanopore Technologies
phone +44 (0)845 034 7900
email info@nanoporetech.com
twitter @nanopore
www.nanoporetech.com

WA_W1003_v1_revA_24Mar2017