Oxford NANOPORE Technologies

WHITE PAPER

Nanopore sequencing
New insights into large genomes

OCTOBER 2017

Contents

Introduction

Since *Haemophilus influenzae* became the first organism to have its entire genome sequenced in 1995[1], there are now thousands of completed genomes covering archaea, bacteria and eukarya. However, the term 'completed genome' (or sometimes 'finished genome') is overused, as the majority of sequenced genomes contain numerous gaps[2]. These gaps correspond to repetitive regions and structural variants that are too large to be resolved by the short-read sequencing technologies often used to generate the genomes. This review outlines how researchers are now utilising the advantages of nanopore long-read sequencing to address this challenge, delivering more accurate large genomes for a wide variety of applications.

More accurate large genomes are now achievable with the long reads offered by nanopore sequencing.

1

The importance of accurate whole genomes

For the purposes of this article, we will define a 'large genome' as being over 100 Mb[3] in size. The range and diversity of organisms that are included in this definition are shown in Figure 1.

To unlock the secrets of the genome, it is imperative that more rigorous standards are applied to whole-genome sequencing studies, allowing:

- The completion of existing reference genomes, providing high-quality data for future comparative studies
- Identification of undiscovered genes and their associated function
- Characterisation of uncharted regions such as centromeres and telomeres
- A comprehensive view of all genomic variation, enabling more precise disease association and future improvements in medical practice
- Enhanced animal and crop breeding, enabling the conservation of natural diversity and selection for desirable traits

Availability of long reads coupled with the ability to analyse large amounts of data allows us to study extremely large genomes.

The study of large genomes also enhances our understanding of the genetic mechanisms underlying evolution. Dr Christiaan Henkel of Leiden University in The Netherlands comments: 'Organisms such as amphibians, lung fish and many plants have extremely large genomes and we don't know why as we haven't been able to sequence or study them. However, the good news is that, due to the availability of long reads coupled with the ability to analyse large amounts of data, we can now start to understand this[14].

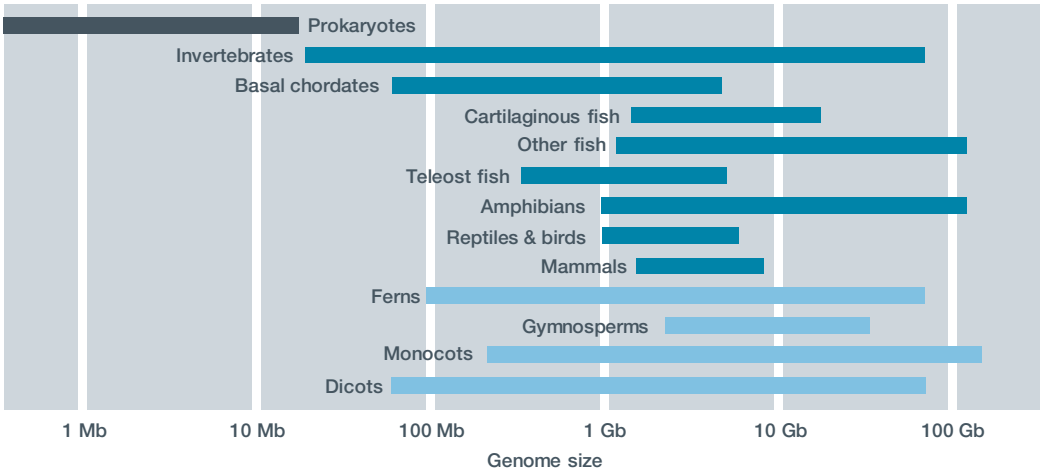


Figure 1
Genome sizes of a range of organisms. Image adapted from Henkel (2017)[4].

2

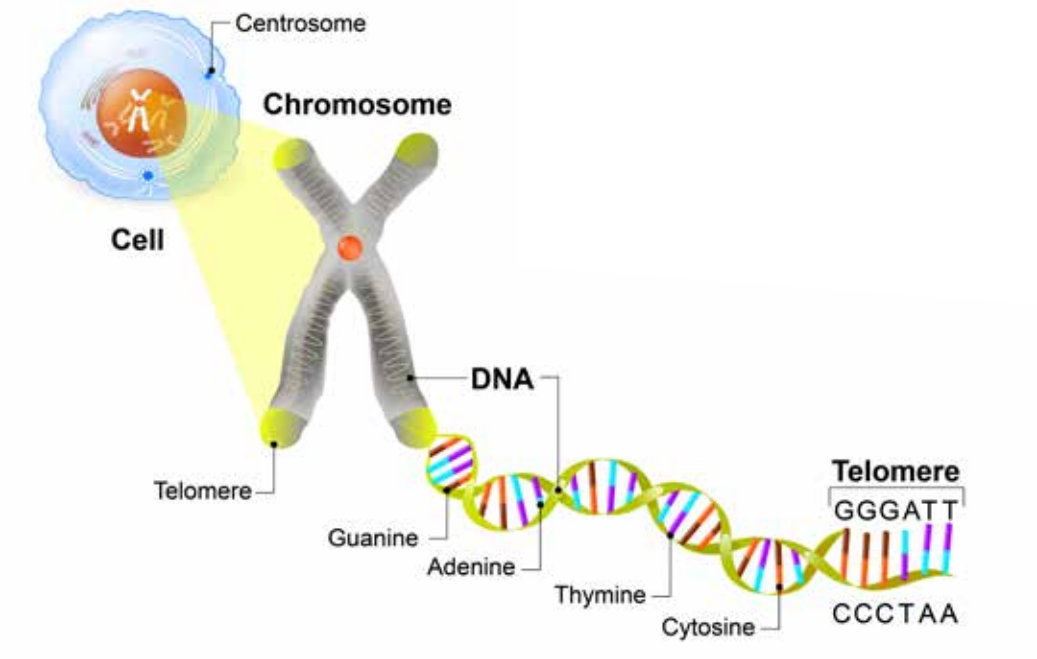
Most current genomes are incomplete

Even some of the most studied genomes are incomplete. For example, it is estimated that approximately 8% of the human genome remains to be sequenced[5]. These 'gaps' in the human genome are predominantly composed of densely packed and repetitive heterochromatic DNA that is found in centromeres and telomeres, plus large structural variation (SV)[6]. Telomeres have long been associated with ageing[7], while SV is associated with many diseases (e.g. autism[8], obesity[9], schizophrenia[10] and cancer[11]), making routine characterisation of these regions important for human health.

The magnitude of the challenge of creating truly complete large genome assemblies is further apparent when considering the significant proportion of the genome that is constituted by repeat regions. In humans, approximately 50% of the genome is composed of repeats[12], while other organisms, such as flowering plants, display an even higher percentage of repetitive DNA of between 55 to 83%[13].

In general, the larger the genome being studied, the more acute the challenge of whole-genome sequencing and assembly.

It is estimated that approximately 8% of the human genome remains to be sequenced.



3

Advantages of nanopore sequencing for studying genomes

Repetitive regions and large structural variation

Traditional short-read sequencing technologies typically generate read lengths of 150-300 bp. In contrast, nanopore-based sequencing processes the DNA fragment that is presented to it. Complete fragments of thousands of kb are routinely processed and ultra-long read lengths approaching 1 Mb have been shown[14]. Clearly, such long reads are more likely to span entire regions of repetitive DNA and structural variation. As a result, nanopore sequencing provides a more complete view of genetic variation, incorporating the phasing of single nucleotide variation (SNV) and characterisation of structural variation and repeat regions.

Long and ultra-long reads which span entire regions provide a complete view of genetic variation, as well as accurate phasing.

Phasing

Another advantage of long and ultra-long reads is that they allow accurate phasing or haplotyping, whereby it is possible to identify which allele or set of alleles are derived from a specific chromosome.

Such information delivers further insights into genome function and, in humans, is becoming increasingly important in the advancement of precision medicine[15]. Researchers can also use haplotyping information to shed further light on gene pool diversity, which is particularly important when studying endangered species[16].

Time to result

For particular whole-genome sequencing applications, such as characterising a rare disease or cancer samples, time to result can be of paramount importance. Unlike traditional sequencing technologies, which typically deliver data in bulk at the end of a run, nanopore sequencing allows real-time data streaming for immediate analysis. This means that, rather than sequence a whole genome, a nanopore sequencing run can be stopped as soon as sufficient data has been generated to answer the biological question – saving valuable time. A range of streamlined library preparation kits are available, which take as little as 10 minutes to perform. Furthermore, utilising nanopore sequencing, researchers have been able to perform the genome assembly step required to generate a highly contiguous plant genome in just 1 hour on a laptop computer (see Case study 4).

Cost-effective analysis

Traditional sequencing platforms typically require large capital investments (>$100k)[17], adaptations to infrastructure and calibration by trained engineers — requiring significant funds and implementation time.

In stark contrast, researchers interested in using long reads for genome assembly can do so through consumable-only purchases which start at $1000 for the MinION™ Starter Pack.

The increased throughput of the benchtop GridION™ X5 and PromethION™ offers users the option to cost-effectively scale their research to meet the demands of sequencing large or extremely large genomes, with no capital expenditure, and at a comparable cost-per-base to traditional sequencing platforms. In addition, the facility to use flow cells independently allows other projects to be run concurrently — delivering highly efficient, on-demand sequencing.

The potential of direct analysis

In order to capture the true genomic diversity of an organism, DNA methylation should also be considered. Cytosine methylation is now known to be an essential component of gene regulation, which in turn is implicated in many genetic diseases including cancer[18]. Unlike traditional short-read sequencing techniques, nanopore-based sequencing has been shown to simultaneously and directly detect cytosine methylation alongside the DNA sequence, adding a further, more detailed level of characterisation to the genome being investigated[18]. As Dr Matt Loose at the University of Nottingham explains: ‘Methylation is key to our understanding of development and gene regulation. The fact that we can now look at it directly is going to be very important’.[19]

Being able to analyse methylation directly alongside the DNA sequence is important for our future understanding of development and gene regulation.

4 Considerations for long-read, whole-genome sequencing of large genomes

From preparing high-quality DNA to selecting the right analysis strategy, there are a number of factors to consider to ensure optimal whole-genome sequencing results. This section presents the thoughts of a selection of leading researchers who are utilising nanopore sequencing to analyse large genomes.

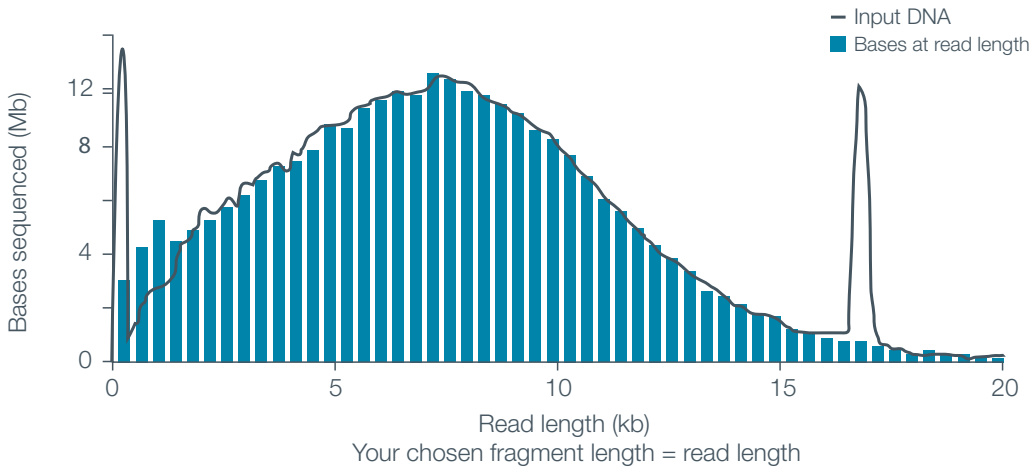
The key to assembly of complex genomes is obtaining reads greater than 20 kb which traverse the main repeat types.

Read length

According to Professor Todd Michael at the J. Craig Venter Institute: 'DNA quality and read length are the most important factors to consider when sequencing large genomes'[20]. Furthermore, he states that: 'While read lengths of 5-10 kb that can be obtained using alternative [non-nanopore] long-read platforms are OK, they are not a game changer in genome assembly. The real key to assembling more complex genomes is obtaining reads greater than 20 kb which traverse the main repeat types of transposons and tandem duplications'[20]. Utilising nanopore sequencing, Prof Michael's team are already generating reads approaching 1 Mb but, looking to the future, he believes that: 'The potential for even longer reads will change the situation from genome assembly to straight genome sequencing'[20]. This sentiment is echoed by Dr Hans Jansen at Future Genomics Technologies who suggests that: 'At the current rate of progress, nanopore-based long-read sequencing has the potential to make the practice of genome assembly obsolete within the next 5 years'[16].

Nanopore sequencing allows the generation of ultra-long reads as the pore processes the fragments (Figure 2) presented to it regardless of their length. This enables researchers to investigate extremely long stretches of repeat elements or large structural variants. The longest DNA fragment sequenced to date using nanopore technology is 882 kb[14]. This was achieved by Josh Quick from the University of Birmingham who observes that: 'The most important factor to generating long reads is sample quality. This is more important than the library prep itself which is relatively simple – especially with the Rapid Sequencing Kit [Oxford Nanopore Technologies] – and that is more important again than the sequencing, which is really just loading the flow cell and pressing go'[21].

Figure 2
The length of the nanopore reads obtained from a library corresponds well to the length of the fragments within that library.



Obtaining intact HMW DNA is critical to the generation of ultra-long reads.

Obtaining intact high molecular-weight (HMW) DNA is critical to the generation of ultra-long reads and researchers are now rediscovering DNA purification techniques that had long been discarded since the advent of modern purification kits. One such technique is phenol:chloroform extraction, which in combination with the streamlined Rapid Sequencing Kit, is currently the favoured method for delivering ultra-long nanopore sequencing reads[16,22,23]. Using this technique, Josh and the University of Birmingham team have been able to routinely sequence DNA fragments of over 800 kb in length[14,22,23]. While much of their optimisation work focused on the bacterium *Escherichia coli*, similarly long read lengths have now also been achieved using human DNA (see Case study 1).

Sample type, DNA quality and purity

Another key consideration for obtaining HMW DNA for whole genome sequencing is sample type. The purification of DNA from different organisms and even from different cell types within the same organism often requires careful optimisation.

In his research, Dr Hans Jansen explores the genomes of many different organisms and, as he acknowledges: ‘obtaining high-quality DNA can be challenging – even within different species of fish, it is never the same body part that gives the best DNA’[16].

Dr Jansen's recommendation to researchers looking to optimise their read lengths is to: 'spend more time upfront figuring out the best DNA isolation method for their specific sample'[16].

It is well known within the field of plant science that obtaining HMW DNA from plant cells can prove more challenging than for other commonly studied organisms[20]. As Professor Michael explains: 'plants have a high content of secondary metabolites such as polysaccharides and polyphenols, which can interfere with DNA extraction and purification'[20]. A common recommendation to combat polysaccharides is to place the plants in the dark for a couple of days prior to DNA extraction. However, as Dr Jansen reports, for the tulip, this step just didn't work[16]. There are chemical techniques to deplete secondary metabolites, such as the use of antioxidants but these examples further serve to highlight the importance of carefully optimising the DNA extraction step to obtain HMW DNA[24].

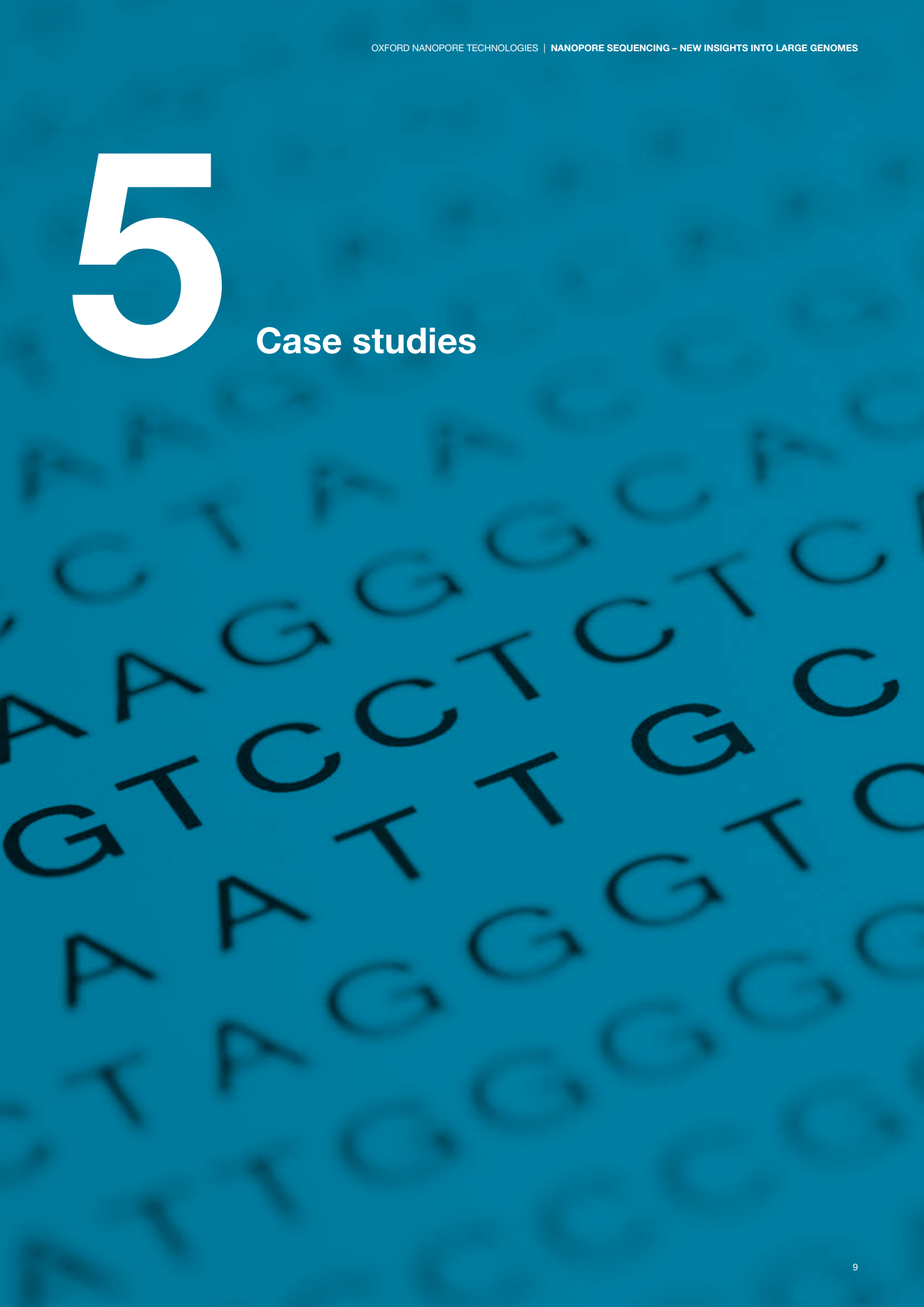
Analysis strategy

According to Professor Michael, the shortened genome assembly time of long-read nanopore sequencing and the facility to perform the analysis in real time on your own laptop is a significant advantage over short-read sequencing technologies[20]. Within just 1 hour his team was able to assemble a highly contiguous *Arabidopsis thaliana* genome (see Case study 4)[36].

Many software tools are now freely available for assembly and analysis of long-read sequencing data. Which tool(s) to use will depend on each individual project; however, there are now a number of research papers that compare the performance of different tools with regard to basecalling accuracy, assembly size, number of contigs and average contig sizes amongst other factors[23,31,32]. Some recently developed tools, such as TULIP from the University of Leiden (see Case study 2), can take advantage of existing short-read genome assemblies to provide a scaffold for genome assembly, streamlining the assembly process and providing significant time savings.

Researchers who are new to large genome sequencing should also carefully consider their data management strategy. Dr Matt Loose comments: 'It is theoretically possible to sequence a single human genome on the PromethION platform at 30x coverage in less than 1 hour, now that's amazing but it does highlight the importance of robust data management plans'[19].

5 Case studies



Case study 1

Improving the human genome assembly

The sequencing of large eukaryotic genomes remains challenging, and even though the well-studied human genome sequence was published over a decade ago, it is still incomplete[2]. Recently a multicentre project involving researchers from the UK, USA and Canada sought to address these issues using long-read nanopore sequencing[23]. The well-characterised human reference genome NA12878 was selected for study allowing accurate benchmarking of the nanopore sequencing data.

An example of the superior genome contiguity was the presence of all the HLA class I genes in a single contig.

Using nanopore sequencing, the team generated 91 Gb of data giving a theoretical 30x coverage of the human genome. The resultant assembly was found to be highly complete and contiguous, comprising 2886 contigs with an NG50* contig size of approximately 3 Mb. The team identified 899 structural differences when compared to the existing NA12878 reference assembly, potentially representing previously unmapped stretches of DNA[23].

An example of the superior genome contiguity provided by nanopore sequencing was evidenced when examining the highly repetitive — and thereby notoriously difficult to assemble — HLA class I region. Not only were all of the HLA class I genes present in a single contig of approximately 3 Mb, but it was also possible to identify haplotype phasing.

The ultra-long read sequencing method employed demonstrated the validity of this approach for improving assembly contiguity, delivering an additional 5x genome coverage with a read length N50† of 99.7 kb and contig NG50 of 6.4 Mb, compared with 10.6 kb and 3 Mb respectively for the standard long-read protocol (Figure 3). The longest full-length mapped read obtained was 882 kb.

Methods	Long read	Ultra-long read
DNA purification	QIAamp DNA Mini Kit	Phenol:chloroform
Library preparation	Ligation Sequencing Kit 1D	Rapid Sequencing Kit
Basecalling	Cloud based (now local)	Albacore (0.8.4)
Assembly	Canu	Canu
Polishing	Nanopolish	N/A

* The NG50 value represents the longest contig such that contigs of this length or greater sum to at least half the haploid genome size.
† The N50 value represents the contiguity of a genome assembly, the higher the value, the more contiguous the assembly.

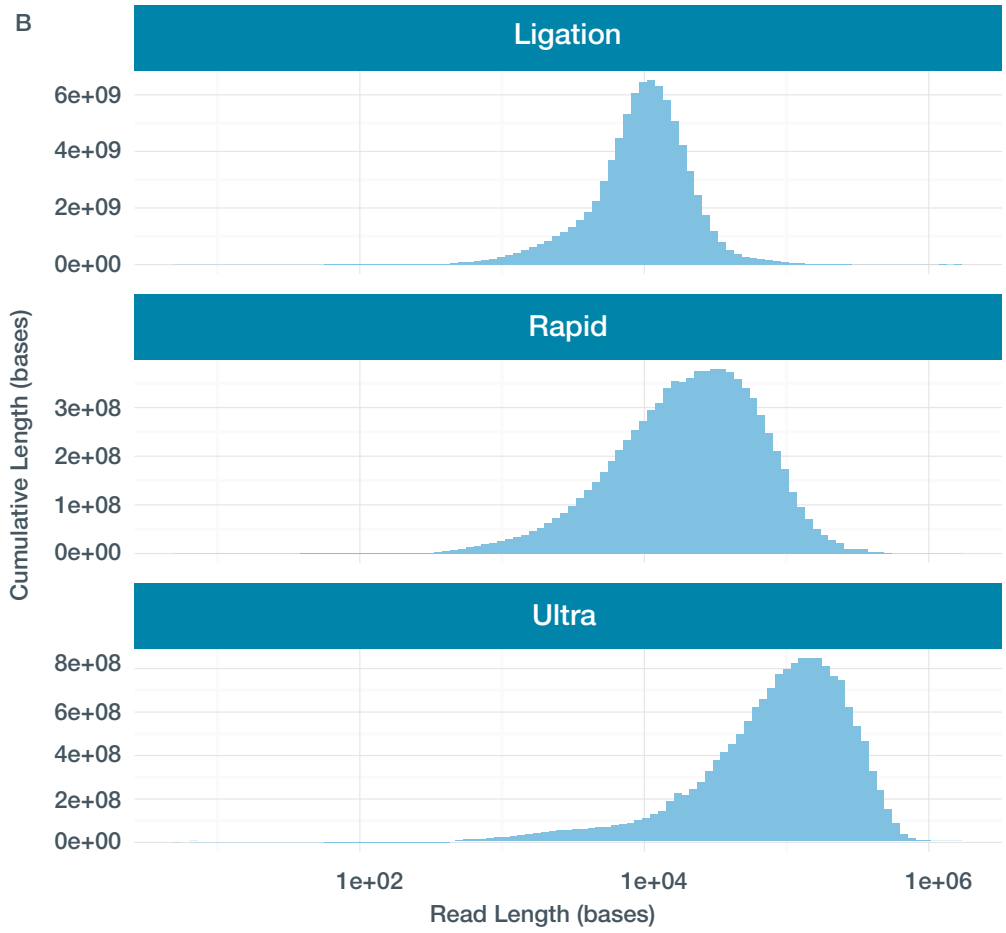


Figure 3
Yield by read length for ligation, rapid and rapid ultra-long library preparations. The longest reads were achieved via DNA extraction direct from cells (rather than purchased DNA) using a modified phenol:cholorofom DNA extraction method with the Oxford Nanopore Rapid Sequencing Kit protocol (Ultra). Image adapted from Jain (2017)[23].

Unlike short-read technology, nanopore sequencing also allowed the detection of cytosine methylation (5-methylcytosine), with the results showing high concordance with previously published data obtained using alternative analysis techniques.

Summarising this research Dr Matt Loose, concludes: 'We've demonstrated the impact of ultra-long reads on assembly contiguity and their facility to resolve areas of the genome that have proven intractable to short-read sequencing, including telomeres, centromeres and highly variable regions such as MHC'[19].

Case study 2

Understanding the European eel genome

The European eel (*Anguilla anguilla*) is an iconic, yet critically endangered, fish species that remains resistant to efficient farming in aquaculture[26]. Through overfishing, parasites and barriers to migration, it is estimated that the number of eels reaching Europe from their breeding grounds in the Sargasso Sea — a 4,000-mile (6,500km) journey — has declined by over 90%[27].

Using TULIP, the 860 Mb European eel genome was assembled in 3 hours using a modest desktop computer.

To support breeding and conservation efforts, researchers, led by ZF-screens B.V. and the University of Leiden in The Netherlands, set out to improve the existing, highly fragmented reference genome using long-read nanopore sequencing[28]. According to Dr Hans Jansen who led the project: 'Long reads are required to generate high-quality genome assemblies from which you can truly understand gene synteny, allowing much more accurate lineage analysis of related species – this just can't be done using short reads'.[16]

Methods	
DNA purification	Genomic-tip 100/G (QIAGEN)
Library preparation	Ligation Sequencing Kit 1D Rapid Sequencing Kit
Basecalling	Cloud-based (now local)
Correction	Racon
Alignment	BWA-MEM
Assembly	TULIP
Polishing	Pilon

The assembly of large genomes can be a time-consuming and computationally intensive process. To combat this, the team created a new assembly strategy named TULIP (The Uncorrected Long-read Integration Process), which instead of performing an all-versus-all alignment, aligns reads to a set of non-repetitive 'seed' sequences that are representative of the genome (Figure 4). In this instance, these seed sequences were derived from the short-read genome assembly; however, the researchers state that this could also be achieved using the ends of long reads.

Impressively, using TULIP, the 860 Mb European eel genome could be assembled in 3 hours on a modest desktop computer. Including subsequent sequence correction, the whole process — from sample to result — could, in theory, be achieved in just 2 days. Summarising, 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'[28].

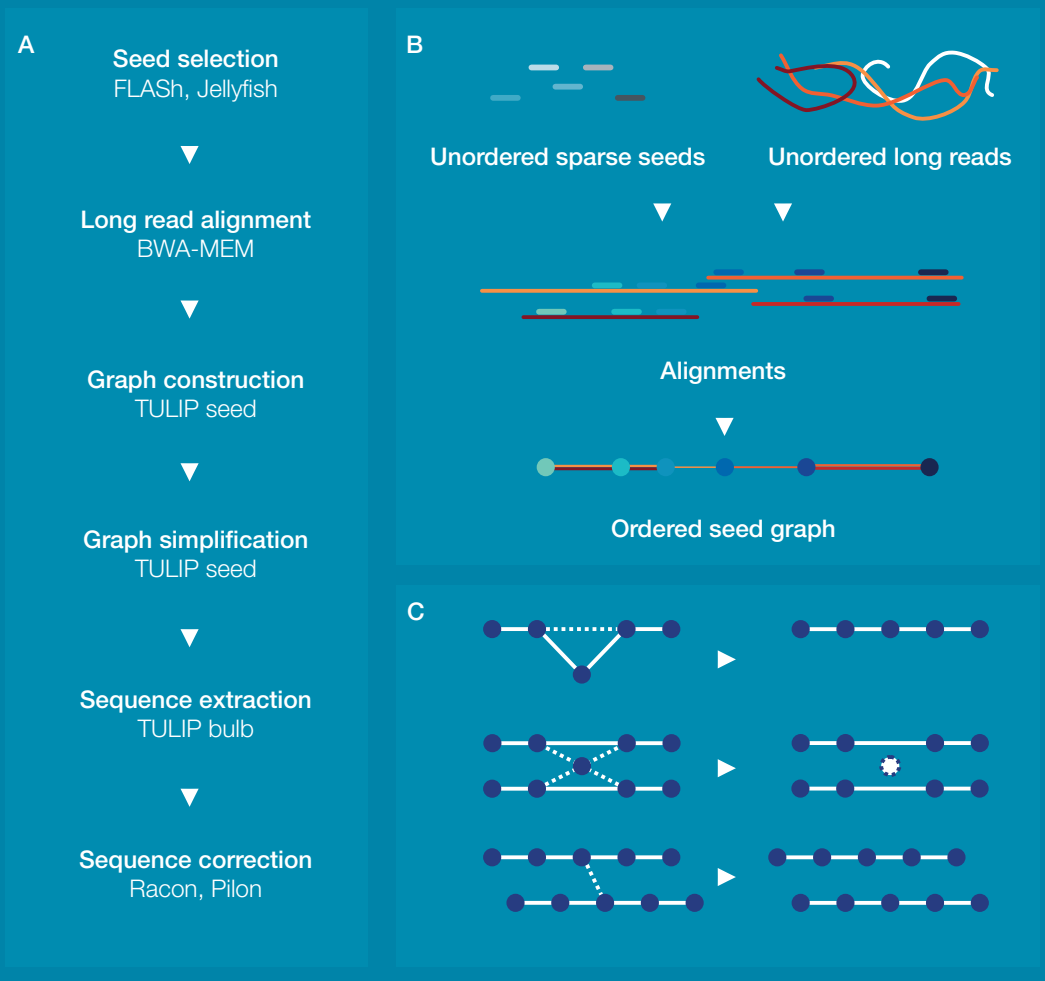


Figure 4
The TULIP assembly strategy. (a) Stages in TULIP. (b) Graph construction based on long read alignments to short seeds. (c) The initial seed graph based on alignments contains ambiguities, caused by missed alignments, repetitive seed sequences and spurious alignments. These are removed during the initial layout process, resulting in linear scaffolds. Image adapted from Jansen, H.J. *et al* (2017)[28].

Case study 3
The problem with plants

The assembly of plant genomes is particularly challenging due to a number of factors including, extreme genome sizes (e.g. 22 Gb Loblolly pine and 20 Gb Norway spruce) and a high level of repetitiveness, which is further exacerbated through the polyploid nature of many plant species[29]. Some plant genomes also contain very large gene families that can comprise up to 700 members with significant homology[30].

According to Professor Björn Usadel of RWTH Aachen University in Germany, these factors make the accurate analysis of plant genomes particularly challenging, potentially even more so than the assembly of vertebrate genomes (Figure 5)[31].

Nanopore sequencing means that researchers can now sequence and assemble a genome on their own bench.

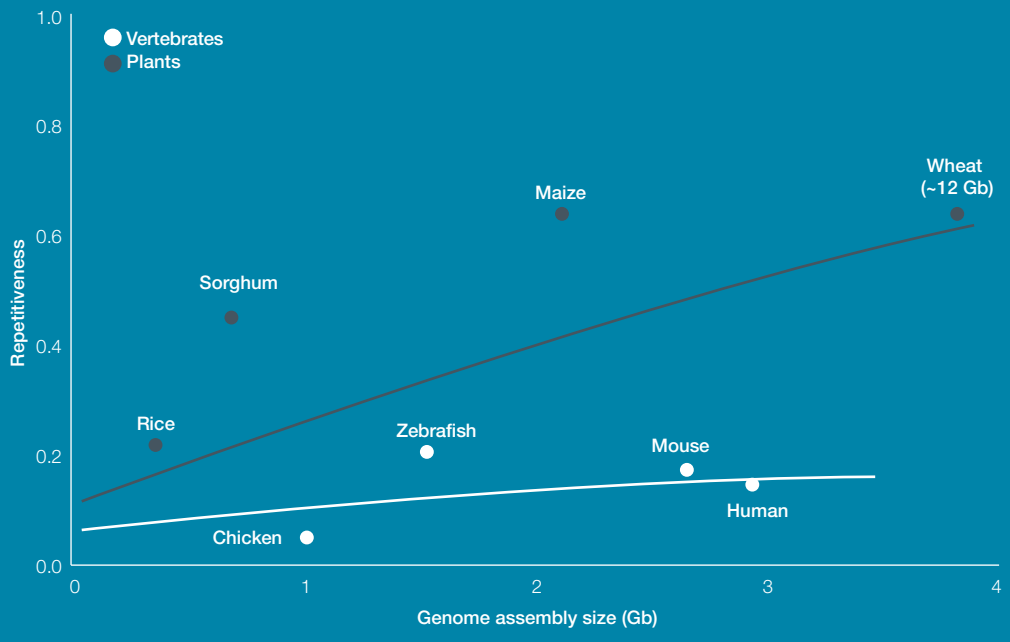


Figure 5
Plants can exhibit extremely large genomes with a high proportion of repetitive DNA, making accurate genome assembly particularly challenging when using traditional short-read sequencing technology. Image adapted from Jiao and Schneeberger (2017)[29].

Professor Usadel and his team are applying the advantages of long-read nanopore sequencing to study the genome of the wild tomato species *Solanum pennellii*. This species of tomato has evolved to live in arid habitats; however, its fruit is inedible[31]. Improved knowledge of the genome of this species could allow the breeding of varieties of the domesticated tomato with enhanced drought resistance.

The team were able to routinely achieve reads over 12 kb in length, with the longest read being 153 kb. In total, 111 Gb of data was generated, which corresponds to approximately 100x coverage of the 1-1.2 Gb genome[32].

In conclusion Professor Usadel comments: 'We got a very good, competitive plant genome within a few months, which we can now use to develop better plants. Furthermore, the project will act as a springboard for our next plant genome sequencing projects. It [nanopore sequencing] means that small labs can now sequence and assemble a genome'[31].

Methods

DNA purification	Nuclei enrichment
Library preparation	Ligation Sequencing Kit 1D with BluePippin DNA size selection
Basecalling	Cloud-based (now local)
Correction	Canu
Assembly	SMARTdenovo
Polishing	Racon Pilon

Case study 4

Advancing the Arabidopsis genome

Due to its short generation time, small size, large number of offspring and relatively small genome size (~150 Mb), *Arabidopsis thaliana* has long been a popular model organism in plant biology[34], and in 2000, it became the first plant to genome to be sequenced[35]. However, in common with most current reference genomes, the short-read technology used to construct the genome precludes the analysis of large structural variation and repetitive regions, such as transposons[36]. To address this challenge, an international team of researchers led by Professor Todd Michael at the J. Craig Venter Institute utilised nanopore long-read sequencing to create an additional, highly contiguous reference genome for *A. thaliana*[36].

In just 4 days, using a single nanopore flowcell and at a cost of just $1000, the researchers were able to achieve a more contiguous genome assembly than provided by the existing 'gold standard' TAIR1035 assembly. The initial assembly step took just 1 hour using a standard laptop. Furthermore, base quality was deemed to be on a par with the current gold-standard reference assembly.

Methods

DNA purification	Phenol:chloroform followed by Zymo Genomic DNA Clean and Concentrator-10 column
Library preparation	Ligation Sequencing Kit 1D
Basecalling	Albacore (0.8.4)
Assembly	Minimap/miniasm
Correction	Racon
Polish	Pilon

Routine generation of 200 – 800 kb reads enabled the identification of nested transposable elements and repeat regions which were previously inaccessible.

The team routinely generated reads between 200 kb and 800 kb in length using nanopore sequencing, enabling them to identify nested transposable elements and repeat regions that, until now, had been inaccessible to DNA sequencing technologies (Figure 6). Prof Michael explains: 'We found that *de novo* assembly with long reads revealed micro-variation that we did not see with short-read technologies. Even the small Arabidopsis genome is riddled with transposable element fragments that have dragged micro-duplications across the genome'.

In addition to improving the genome of this key model organism, Prof Michael states that this study highlights that: 'researchers no longer have to send out their samples to core labs or service providers, they can now do it [whole-genome sequencing] on their own bench, and within a week have an answer to their question. This is something that people can do right now in their own labs – they should do it'[20].

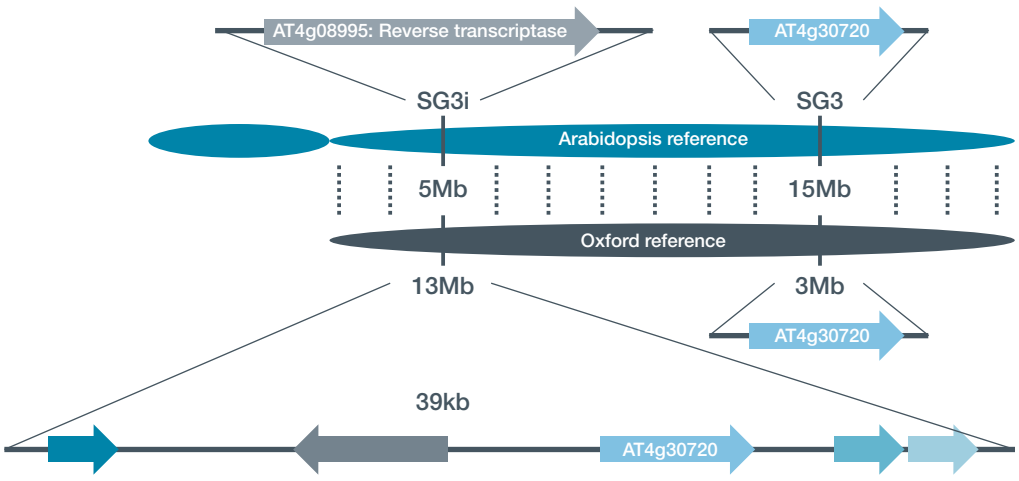


Figure 6
Long-read nanopore-based sequencing resolved a highly repetitive region of nested transposable element fragments. An additional copy of the transposable element (TE) AT4g30720 within a 39 kb expansion (comprising numerous TE fragments) was identified that could not be resolved by short-read sequencing approaches. Image adapted from Michael *et al* (2017)[36].

6

Summary

Nanopore-based sequencing, where the read length is only limited by the length of the DNA fragment itself, addresses many of the challenges researchers face when trying to sequence and assemble large genomes. Using long-read nanopore sequencing, researchers are now able to accurately analyse and map repetitive regions and large structural variations that have proved intractable to traditional short-read sequencing technology. These advances mean that organisms that had previously been deemed too complex or too expensive to warrant whole-genome sequencing can now be studied by researchers in their own laboratories.

As stated by Dr Christiaan Henkel: 'We can finally start looking at… massive genomes which we have previously had to ignore because we couldn't sequence them'[14].

7

About Oxford Nanopore Technologies

Oxford Nanopore Technologies introduced the world's first and only nanopore DNA sequencer, the MinION — a portable, real-time, long-read, low-cost device. Through the utilisation of long reads, which allow the analysis of regions of repetitive DNA and large structural variation, nanopore sequencing technology allows the delivery of more complete genomes than are possible using traditional short-read sequencing technology[4,23,31,36]. A range of platforms are available, suitable for different sized genomes and higher coverage and throughput requirements. The original nanopore platform, the MinION can provide 10s of Gb of data per run, while the GridION X5 (Figure 7) and PromethION provide 100s and 1000s of Gb of data respectively.

These compact benchtop systems utilise the same nanopore technology as the MinION, offering up to 5 and 48 flow cells. 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. Both the GridION X5 and PromethION are available with no capital cost — only consumables need to be purchased — delivering scalable, cost-effective analysis.

For the latest information about sequencing large genomes using long-read nanopore sequencing, visit www.nanoporetech.com/applications.

Figure 7
The GridION X5 is a high-throughput, benchtop system with integrated compute module.

8 References

1 Fleischmann, R.D. *et al* (1995) Whole-genome random sequencing and assembly of *Haemophilus influenzae*. Science 269(5223): 496-512

2 Chaisson, M.J.P. *et al* (2015) Genetic variation and the *de novo* assembly of human genomes. Nat Rev Genet 16(11):62740

3 Schatz, M.C., Delcher, A.L, and Salzberg, S.L. (2010) Assembly of large genomes using second-generation sequencing. Genome Res. 20(9): 1165–1173. doi: 10.1101/gr.101360.109

4 Henkel, C. (2017) Lightweight sequencing of massive genomes. Presentation. Available at: https://nanoporetech.com/index.php/talk/lightweight-sequencing-massive-genomes [Accessed 20 July 2017]

5 Miga, K.H., Eisenhart, C. and Kent, W.J. (2015) Utilizing mapping targets of sequences underrepresented in the reference assembly to reduce false positive alignments. Nucleic Acids Research 43(20): e133-e133

6 Altemose, N. *et al* (2014) Genomic characterization of large heterochromatic gaps in the human genome assembly. PLOS Computational Biology 10(5): e1003628. https://doi.org/10.1371/journal.pcbi.1003628

7 Harley C.B, Futcher A.B., and Greider C.W. (1990). Telomeres shorten during ageing of human fibroblasts. Nature 345: 458–460

8 Sebat, J. *et al* (2007) Strong association of *de novo* copy number mutations with autism. Science 316: 445 – 449

9 Bochukova, E.G *et al* (2010) Large, rare chromosomal deletions associated with severe early-onset obesity. Nature 463: 666–670

10 Stefansson, H. *et al* (2008) Large recurrent microdeletions associated with schizophrenia. Nature 455: 232–236

11 Diskin, S.J. *et al* (2009) Copy number variation at 1q21.1 associated with neuroblastoma. Nature 459(7249): 987–991

12 Treangen, T.J. and Salzberg, S.L. (2011) Repetitive DNA and next-generation sequencing: computational challenges and solutions. Nat Rev Genet. 13(1):36-46. doi: 10.1038/nrg3117

13 Macas, J. *et al* (2015) In depth characterization of repetitive DNA in 23 plant genomes reveals sources of genome size variation in the legume tribe *Fabeae*. PLoS ONE 10(11): e0143424. doi:10.1371/journal.pone.0143424

14 Loman, N. (2017) 950kb, yeah! [Twitter] 08 March. Available from: https://twitter.com/pathogenomenick/status/839604187100934144 [Accessed: 20 July 2017]

15 Ammar, R. *et al* (2015) Long read nanopore sequencing for detection of HLA and CYP2D6 variants and haplotypes. F1000Research 4, 17. http://doi.org/10.12688/f1000research.6037.1

16 Jansen, H. (2017). Personal communication with Oxford Nanopore Technologies on 07 September 2017.

17 Norris, A.L. *et al* (2016) Nanopore sequencing detects structural variants in cancer. Cancer Biol Ther 17(3): 246-253

18 Simpson, J.T. *et al* (2017) Detecting DNA cytosine methylation using nanopore sequencing. Nature Methods 14: 407–410 doi:10.1038/nmeth.4184

19 Loose, M. (2017) Personal communication with Oxford Nanopore Technologies on 14 September 2017.

20 Michael, T.P. (2017). Personal communication with Oxford Nanopore Technologies on 29 August 2017.

21 Quick, J. (2017) Thar she blows! Ultra-long read methods for nanopore sequencing. Presentation. Available at: https://nanoporetech.com/resource-centre/publications/josh-quick-thar-she-blows-ultra-long-read-methods-nanopore-sequencing [Accessed 5 May, 2017]

22 Loman Labs (2017) Thar she blows! Ultra-long read method for nanopore sequencing [online] Available at: http://lab.loman.net/2017/03/09/ultrareads-for-nanopore/ [Accessed 3 Mar 2017]

23 Jain, M. *et al* (2017) Nanopore sequencing and assembly of a human genome with ultra-long reads. bioRxiv doi.org/10.1101/128835

24 Lutz, K.A. *et al* (2011) Isolation and analysis of high quality nuclear DNA with reduced organellar DNA for plant genome sequencing and resequencing. BMC Biotechnol. 11:54. doi: 10.1186/1472-6750-11-54.

25 Oxford Nanopore Technologies (2017) Genome assembly and the advantages of long read sequencing technology. Available at: https://nanoporetech.com/resource-centre/white-papers [Accessed 05 October 2017]

26 Minegishi, Y. *et al* (2012) Genomics in eels — towards aquaculture and biology. Marine Biotechnology 14(5): 583-590

27 Jacoby, D. & Gollock, M. (2014) *Anguilla Anguilla* [online] International Union for Conservation of Nature. Available at: http://www.iucnredlist.org/details/60344/0 [Accessed 04 August 2017]

28 Jansen, H.J. *et al* (2017) Rapid *de novo* assembly of the European eel genome from nanopore sequencing reads. Sci Rep. 7(1):7213-7226 doi: 10.1038/s41598-017-07650-6

29 Jiao, W.B and Schneeberger, K. (2017) The impact of third generation genomic technologies on plant genome assembly. Current Opinion in Plant Biology 36: 64-70

30 Schatz, M.C., Witkowski, J., and McCombie, W.R. (2012) Current challenges in *de novo* plant genome sequencing and assembly. Genome Biology 13:243 DOI: 10.1186/gb-2012-13-4-243

31 Usadel, B. (2017) Complex tomato genomes: Easy with nanopores. Presentation. Available at: https://nanoporetech.com/index.php/talk/complex-tomato-genomes-easy-nanopores [Accessed 20 July 2017]

32 Schmidt, M.H.W. *et al* (2017) Reconstructing the gigabase plant genome of *Solanum pennellii* using nanopore sequencing. bioRxiv doi.org/10.1101/129148

33 Bowler, C. *et al* (2004) Chromatin techniques for plant cells. Plant J. 39:776–789. doi: 10.1111/j.1365-313X.2004.02169.x.

34 Arabidopsis Genome Initiative (2000) Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. Nature 14;408(6814):796-815

35 TAIR (2017) Genome assembly [online] Available at: http://www.arabidopsis.org/portals/genAnnotation/gene_structural_annotation/agicomplete.jsp [Accessed 20 July 2017]

36 Michael, T.P. *et al* (2017) High contiguity *Arabidopsis thaliana* genome assembly with a single nanopore flow cell. bioRxiv doi.org/10.1101/149997

Oxford Nanopore Technologies

phone +44 (0)845 034 7900

email info@nanoporetech.com

twitter @nanopore

www.nanoporetech.com

LG_W1010_v1_revA_12Oct2017