Oxford NANOPORE Technologies

WHOLE-GENOME SEQUENCING

Closing the gap in plant genomes

WHITE PAPER

January 2020

nanoporetech.com/publications

nanoporetech.com

Contents

Introduction

According to the latest estimates, approximately 400,000 plant species are currently known to science; however, 21% of them are at risk of extinction due to factors such as climate change, habitat loss, and disease[1]. Combatting these challenges is a key focus of plant research efforts, particularly in the case of crop production, where we are reliant on a small number of species. It has been reported that just 20 plant species occupy 90% of arable land, highlighting our vulnerability to the potential impacts of climate change and disease[2]. In addition, the growing population, which is anticipated to increase by approximately 30% to 9.8 billion by 2050[3], is placing increasing pressure on crop yields.

Of the 400,000 known plant species, just 400 (0.1%) have had their genomes sequenced[4,5,6].

Bearing in mind the vital importance of plants for food, fuel, medicine, and clothing, it is no surprise that significant emphasis is being placed on plant genetic characterisation, which can allow the preservation of genetic diversity and the identification of beneficial traits for breeding purposes.

With the importance of studying plant genomes being widely acknowledged, the question arises: why have so few been sequenced? To date just 600 plant lines (~400 species) have had their genomes sequenced[4,5,6], which compares rather unfavourably to the >200,000 microbial strains that have so far been sequenced[5]. As we will see, the answer to this question lies in the size and complexity of plant genomes, which make them extremely difficult to sequence using traditional technologies.

This review outlines how researchers are now addressing the challenges of producing high-quality, highly contiguous plant genome assemblies through the use of nanopore sequencing technology — enabling new opportunities in plant conservation and breeding.

The advantages of nanopore sequencing for studying plant genomes

Simplifying large genomes

The size of plant genomes is extremely diverse, ranging from 61 Mb (i.e. *Genlisea tuberosa*)[7] up to 152 Gb (i.e. *Paris japonica*)[8], with the latter being almost 50 times larger than the human genome. They also exhibit a wide variety of ploidy (chromosome copies), from diploid (two copies: e.g. *Arabidopsis thaliana*) through to decaploid (ten copies: e.g. *Fragaria iturupensis*). Such large genomes are intrinsically difficult to sequence using traditional short-read sequencing technologies, which typically generate read lengths of 150–300 nucleotides.

In contrast, nanopore-based sequencing processes the entire length of the DNA fragments that are presented to it. Complete fragments of thousands of kilobases are routinely processed and ultra-long read lengths in excess of 4 Mb have been shown[9]. The long sequencing reads generated by nanopore technology simplify the genome assembly process (**Figure 1**).

Most eukaryotic genome sequencing projects are limited to producing a single haploid consensus sequence which combines the data from each chromosome copy. This is because the nature of traditional short-read sequencing precludes the cost-effective linking of genes on the same chromosome. However, the long sequencing reads delivered by nanopore technology now make haplotyping much more feasible — even for polyploid species. Long-range haplotyping information could reveal new insights into plant evolution and domestication, and may drive future improvements in crop production[11,12].

> ***'Oxford Nanopore enables a reference quality genome in a week for a small plant genome; better than ~20 years of hand curation'[10]***

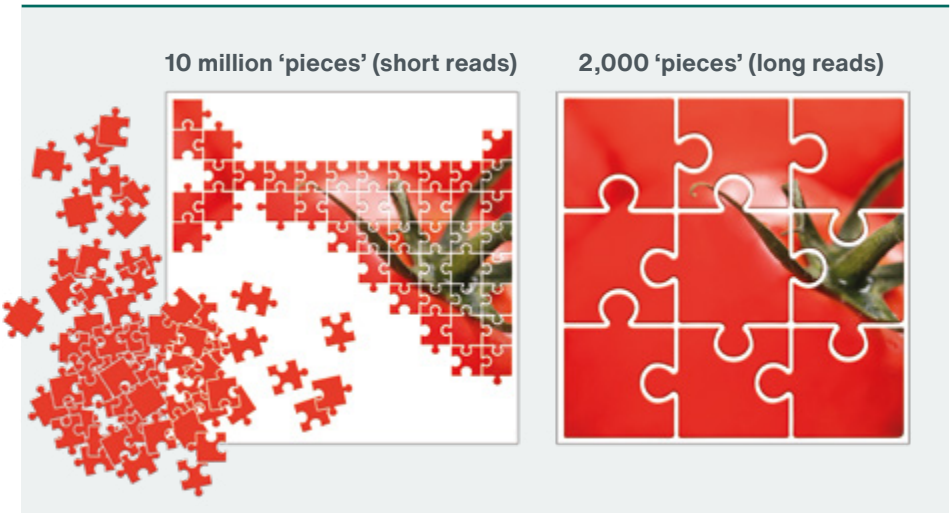


Figure 1
Like a jigsaw puzzle with large pieces, long-read DNA is much easier to assemble than short-read DNA. The tomato genome is approximately 1 Gb in length, which equates to 10 million short reads of 100 bp or 2,000 long reads of 500 kb. The high proportion (60%) of repetitive DNA further complicates assembly when using short sequencing reads.

Closing the gap — resolving repetitive regions and structural variation

Repetitive DNA sequence is the largest component of most eukaryotic genomes, and nowhere is this more apparent than in plants[13], with some species (e.g. maize) exhibiting over 80% repetitive DNA[14] (**Figure 2**). Repetitive sequences can be divided into three main classes: transposable elements (TEs), tandem repeats, and high copy number genes — with TEs comprising the largest component[15].

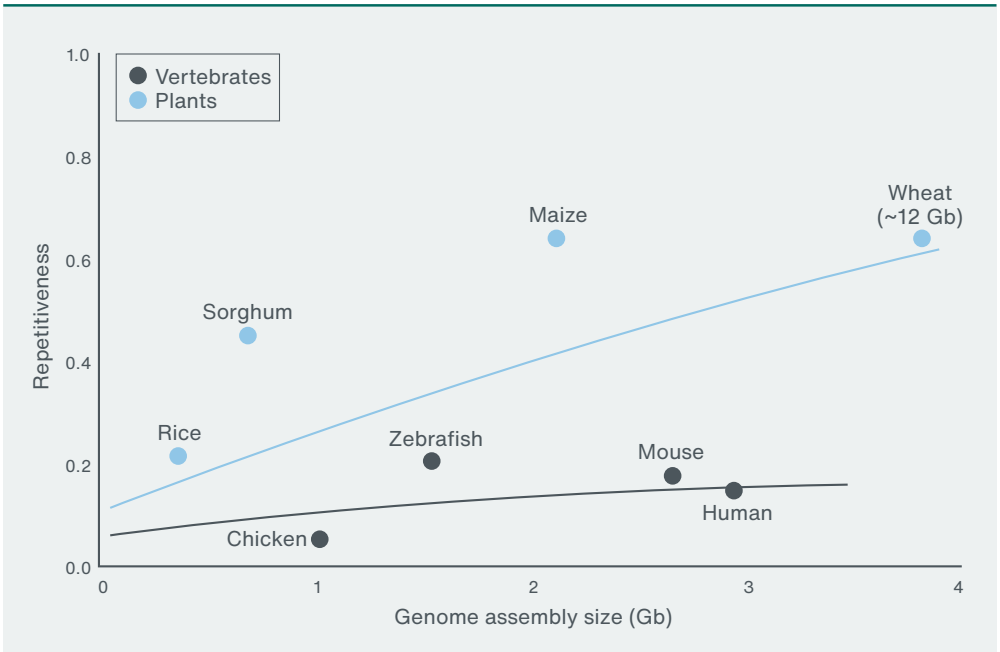
These DNA repeats can both duplicate and insert into new chromosomal locations, and, as a result, play a major role in genome evolution[16]. Therefore, the facility to accurately locate and characterise such transposable elements could shed new light on plant evolution and adaptation, with significant utility for genetic manipulation.

Repetitive DNA poses a major challenge to genome assembly when using short-read sequencing approaches, where the sequencing read does not span the complete repetitive region. A similar situation is observed for large regions of structural variation, which cannot be easily or cost-effectively resolved using short-read sequencing technologies alone. As a consequence, most existing genome assemblies, which have been created using short reads, exhibit numerous gaps corresponding to repetitive regions and structural variants.

Conversely, the long reads delivered by nanopore sequencing are able to span these regions, delivering more contiguous genome assemblies[18,19]. As a result, researchers are now using nanopore technology for both *de novo* sequencing of reference genomes and, in some cases, correction of existing genome assemblies that had been created using short-read approaches[19].

Using a single MinION™ Flow Cell, a team from the J. Craig Venter Institute created a highly contiguous genome assembly for the model organism *Arabidopsis thaliana* within just four days at a cost of $1,000[19]. Furthermore, the assembly displayed higher contiguity than the existing ‘gold standard’ reference genome, while base accuracy was deemed to be on a par. According to the lead researcher, Professor Todd Michael: ‘[the long-read nanopore assembly] *revealed microvariation that we did not see with short-read technologies*’[19].

Figure 2
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[17].



On-demand, scalable sequencing — from lab to field

The sharing of germplasm (living tissue from which new plants can be grown) across international borders has been central to the breeding of enhanced plant lines; however, increasingly complex regulatory requirements for transporting such material are threatening to slow down progress in this area[20]. The largest impact of this may be felt in resource-limited environments, which have reduced access to the capital-intensive genomic technologies required to enhance their plant breeding or conservation efforts.

Unlike traditional sequencing platforms that require large capital investments (>$50k–1M)[21], significant infrastructure and calibration by trained engineers, the MinION Starter Pack from Oxford Nanopore costs just $1,000 (including two flow cells and sequencing reagents) and is powered by the USB port on a laptop or the MinIT™ accessory. The uniquely transportable and affordable MinION and MinION Mk1C (that includes an integrated touchscreen and powerful integrated compute — for real-time data analysis) provide any researcher with access to real-time, long-read capable sequencing technology — in both laboratory and field settings (**Figure 3**). As stated by Professor Björn Usadel of RWTH Aachen University: '[the MinION] *means that small labs can now sequence and assemble a genome*'[18]. A sentiment that is echoed by Professor Todd Michael of the J. Craig Venter Institute, who commented: '*Researchers no longer have to send out their samples to core labs or service providers, they can do it now at their own bench, and within a week have an answer to their question — they should do it*'[22].

GridION™ Mk1, PromethION™ 24, and PromethION 48 offer respectively 5, 150, and 300 times the yield of the MinION, providing users with the facility to cost-effectively scale their research to meet the demands of sequencing extremely large or multiple large genomes — for example, the characterisation of seed collections and transformed plant lines (**Figure 3**). These devices are available with no capital expenditure and deliver 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.

Oxford Nanopore also offers Flongle™, a flow cell adapter for MinION and GridION devices, designed to provide even more cost-effective analyses of smaller, more frequently performed tests and experiments (**Figure 3**). Potential applications include rapid, low-cost optimisation and quality checking of sequencing performance prior to running a large genome sequencing project.

Figure 3
Oxford Nanopore sequencing platforms (from left to right): Flongle, a flow cell adapter for MinION and GridION; the portable MinION and the latest addition, MinION Mk1C; GridION, with capacity for five Flongle or MinION Flow Cells; and the high-throughput PromethION (P24 or P48) platform, with the capacity for up to 24 (P24) or 48 (P48) high-yield flow cells.

The problem with plants

It is notoriously difficult to obtain pure, high-quality, high molecular weight (HMW) DNA from plant cells due to the high levels of secondary metabolites such as polysaccharides and polyphenols. The presence of such contaminants can have a large impact on downstream analysis applications and is one of the most common reasons for low sequencing yields. As Dr Alexander Wittenberg, a scientist at KeyGene in the Netherlands, points out: '*There is no single protocol that covers all plant species, tissues, or conditions*'[23].

As a result, experienced plant researchers typically recommend spending time carefully optimising the sample preparation methodology, not just for each species, but also for each tissue type from a given species. In agreement with this practice, Dr. Anthony Bolger from RWTH Aachen University in Germany suggests that researchers '*tune one parameter per run by a small amount…don't jump around, don't add 50% more material, just tweak them*'[24].

The sequencing of multiple test samples with different preparation parameters could rapidly increase experimental costs. A solution to this potential challenge comes in the form of Flongle (see previous section).

RNA sequencing, transcriptomics, and gene annotation

RNA sequencing is a powerful tool for gene discovery and transcript expression analysis. As previously discussed, the majority of plant genomes are yet to be sequenced and those that have been assembled are often incomplete, potentially missing many important genes. RNA sequencing provides an effective method for identifying all of the genes in a given organism, enabling gene expression analysis and supporting genome annotation.

Traditional RNA-Seq methodologies typically utilise short sequencing reads of 50—100 nucleotides, which must be assembled using complex computational techniques; however, a study by Steijger *et al.*[25] revealed that automated transcript assembly methods fail to identify all constituent exons in over half of the transcripts analysed. Furthermore, of those transcripts with all exons identified, over half were incorrectly assembled[25].

With nanopore sequencing, read length is equal to RNA (or DNA) fragment length, allowing the unambiguous analysis of full-length transcripts — simplifying gene annotation and enabling accurate characterisation and quantification of gene expression at the isoform level (**Figure 4**). Using nanopore sequencing, full-length transcripts in excess of 20 kb in length have been generated[26].

In addition to offering PCR and PCR-free full-length cDNA sequencing kits, Oxford Nanopore also provides the Direct RNA Sequencing Kit, which allows sequencing of native RNA. This unique technique eliminates two potential sources of bias, namely PCR and reverse transcription, and, importantly, also allows the retention and subsequent direct identification of RNA base modifications.

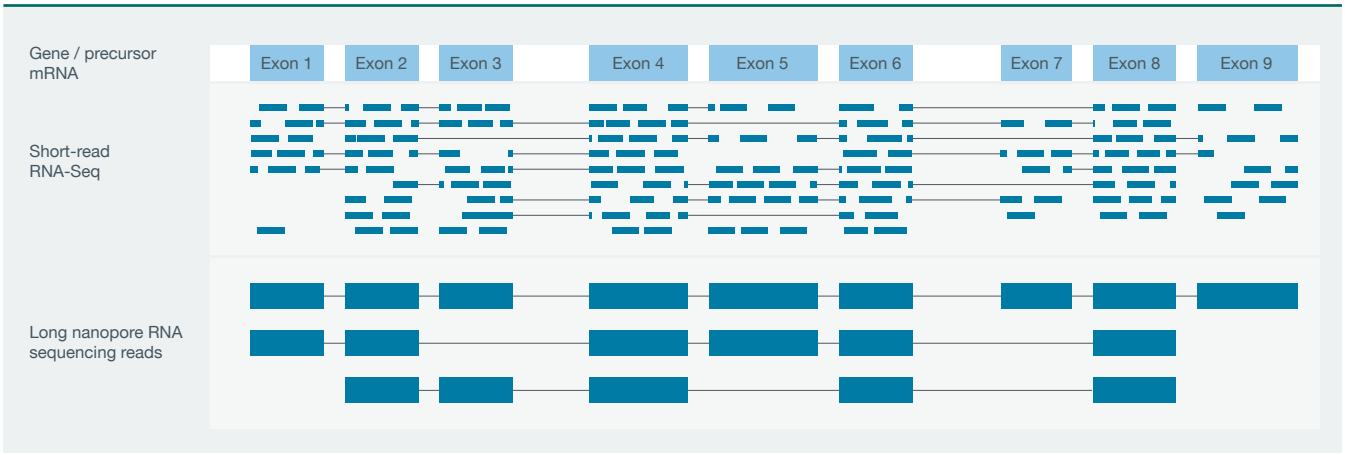


Figure 4
Alternative splicing can give rise to numerous mRNA isoforms per gene, which in turn can alter protein composition and function. The short reads generated by traditional RNA sequencing techniques lose positional information, making the correct assembly of alternative mRNA isoforms challenging. Long nanopore reads can span full-length transcripts, simplifying their identification.

The potential of direct analysis

The action of DNA methylation in plants was first noticed by Barbara McClintock in the 1950s who described a novel mobile element that could switch between active and inactive forms, in what she termed 'changes of phase'[27]. It is now known that DNA methylation in plants occurs in the contexts of CG, CHG, and CHH (where H can be A, C, or T)[28]. In eukaryotes, over 200 different RNA modifications have been reported. Whilst the majority of research into RNA modifications has been undertaken in animals, increasing focus is now being given to plants, with recent studies exploring the role of modifications such as N6-methyladenosine (m6A), 5-methylcytosine (m^5C), and uridylation in plant development[29,30].

Methylated bases play a key role in gene expression, and, as a result, are of increasing importance in agricultural research and plant breeding initiatives. For example, differential methylation has been shown in response to stressors such as salinity, water deficit, and pesticides[30]. Unlike traditional sequencing techniques, nanopore-based sequencing allows simultaneous and direct detection of DNA or RNA methylation and other base analogues alongside the nucleotide sequence, adding a further, more detailed level of genomic characterisation[31,32].

Nanopore sequencing allows simultaneous and direct detection of DNA or RNA methylation and other base analogues alongside the nucleotide sequence.

Case studies

CASE STUDY 1

Revealing the complexities of genetically modified plant genomes

Following their successful sequencing and assembly of a highly contiguous and accurate *Arabidopsis thaliana* genome in just 4 days using a single MinION Flow Cell[19], the team at the J. Craig Venter Institute, together with researchers from the Salk Institute, turned their attention to using long nanopore sequencing reads to characterise transformed, genetically modified lines[10,34].

Nanopore sequencing allowed the resolution of T-DNA structures up to 36 kb in length[34].

One of the most common methods for introducing new genetic material into plant cells is through the use of the bacterium *Agrobacterium tumefaciens*. This plant pathogen randomly inserts DNA contained within its plasmid to double-strand breaks within the host genome. For both scientific and regulatory reasons, it is important to characterise the insertion sites and copy number of this transfer DNA (T-DNA); however, traditional analysis techniques, such as Southern blotting, can be laborious and lack resolution.

To address this challenge, the research team utilised both optical mapping and long nanopore sequencing reads to examine the genome of two transformed and one reference line of *A. thaliana* — with each line being fully sequenced on single MinION Flow Cells. Highly contiguous genomes were assembled with complete chromosome arms being contained within just one or two contigs. Genome analysis allowed the team to resolve T-DNA structures up to 36 kb in length and revealed large-scale T-DNA associated translocations and exchange of chromosome arm ends.

Moreover, sequence contigs for the two transgenic lines (SAIL_232 and SALK_059379) captured up to 39 kb of assembled T-DNA insertion sequence, providing sufficient information to better understand the complexity of such *Agrobacterium*-mediated transgene insertions (e.g. rearrangements, insertions, deletions, etc.) and the effect of these insertions on proximal genes.

This study provides new insights into the structural impact of engineering plant genomes and demonstrates the utility of state-of-the-art, long-range sequencing technologies to rapidly identify unanticipated genomic changes. The team now plans to utilise the nanopore sequencing data to identify the methylation status of their transformed genomes, with a view to the potential replacement of separate bisulfite sequencing-based methylation analysis. Commenting on this research, Professor Todd Michael remarked: '*It has been known that [T-DNA] insertions have variable length and that there are many of them but, up until Oxford Nanopore technology with long reads, it was really impossible see what those insertions looked like*'[10].

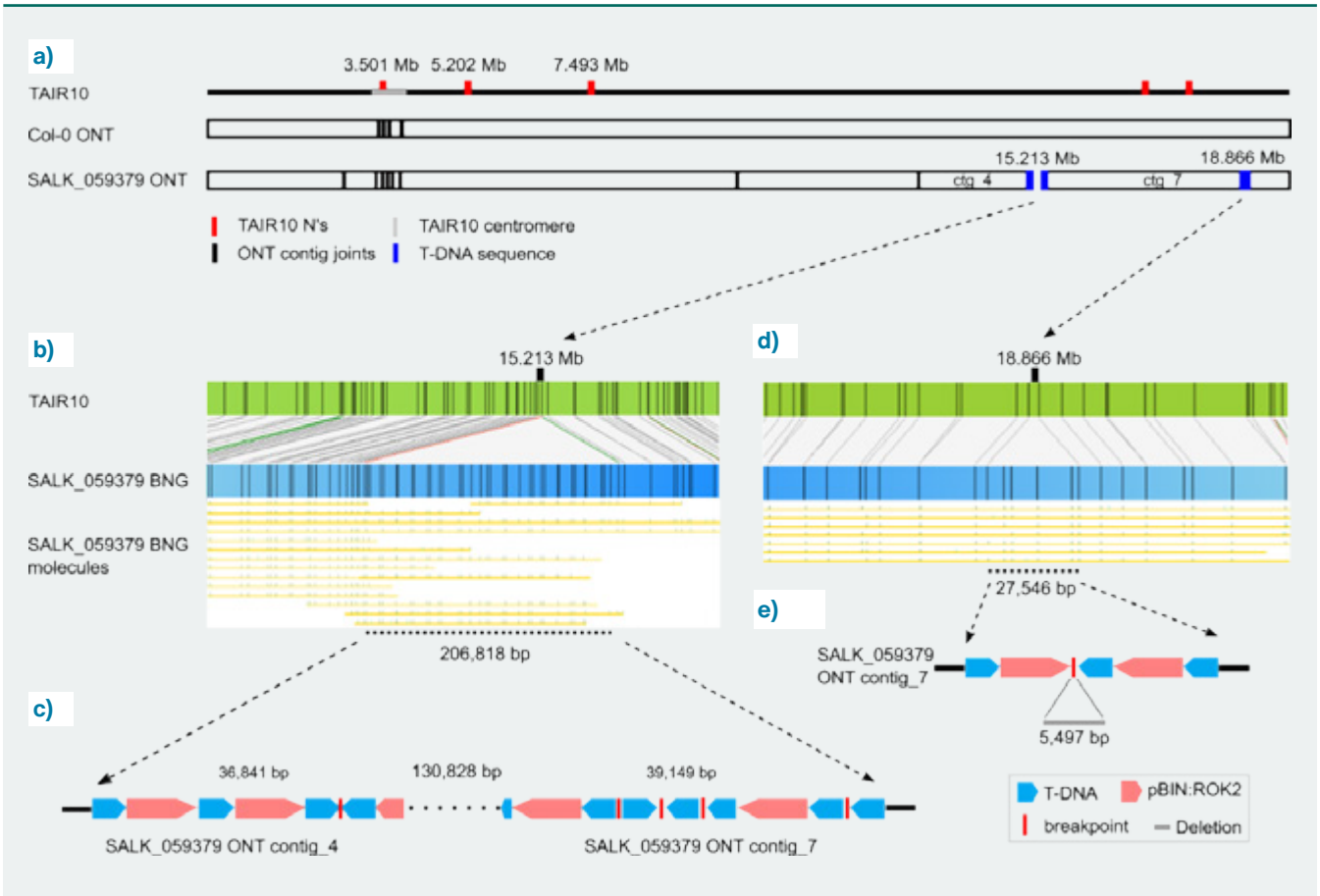


Figure 5
Nanopore sequencing using the MinION resolves assembly errors in the short-read *Arabidopsis thaliana* reference genome and the T-DNA insertion sites in the transformed lines. (a) Chromosome 2 of: TAIR10 wild-type reference line, the nanopore-sequenced Col-0 wild-type line, and SALK_059379 transformed line. Both chromosome arms are present in single Col-0 contigs and misassembles in the TAIR10 reference are corrected. Blue boxes in the SALK line represent T-DNA insertions, which were apparent in the (b,d) optical maps. (c,e) The nanopore sequencing data allowed complete resolution of the 27,546 bp insertion and partial resolution of the 206,818 bp insertion. Figure courtesy of Professor Todd Michael, J. Craig Venter Institute, US.

Recently, researchers at University of Nebraska-Lincoln combined a target enrichment strategy with long nanopore sequencing reads to cost-effectively characterise the insertion sites of the 1,166 bp maize *Dissociation (Ds)* transposable element in transgenic soybean lines[36]. *Ds* element DNA was captured and enriched in 51 soybean lines using a biotinylated oligo probe, prior to barcoding and running all samples on a single MinION Flow Cell. The team reported complete concordance with previously obtained results for each of these well-characterised plant lines and insertion sites. Importantly, the total sequencing cost for all 51 samples was $1,360 ($27/sample) and all results were generated within just one week. Furthermore, the on-target coverage, which for one cell line was as high as 3,555x, could be further reduced. This, coupled with the potential to increase enrichment efficiency, suggests that more samples could be analysed in a single sequencing run, enabling further cost savings. In addition, the methodology also allowed accurate mapping of multiple transgene insertions within individual soybean lines. According to the researchers: '*These results demonstrate that this nanopore-based sequencing method is rapid, convenient, reliable, cost efficient, and high throughput*'[36].

CASE STUDY 2

Sequencing and assembling mega-genomes of mega-trees

Native to the east coast of the United States, the giant sequoia (*Sequoiadendron giganteum*) and coast redwood (*Sequoia sempervirens*) are two of the largest living organisms on the planet. Over the past century, 95% of ancient redwoods, which can live to over 2,000 years, have been lost, lowering the pool of genetic diversity and leaving these species endangered.

In order to support conservation and breeding efforts, researchers at the University of Californa, Davis and Johns Hopkins University initiated an ambitious project to sequence the massive genomes of these massive organisms[37]. At 8.2 Gb for sequoia and 26.5 Gb for redwood, the genomes of these organisms are, respectively, 2.6 and 8.3 times larger than that of humans. To tackle this challenge, the team deployed a 'hybrid' genome assembly strategy, utilising both short-read sequencing technology and long nanopore sequencing reads. As stated by Professor Steven Salzberg, one of the project leaders, with lengths in excess of 10 kb the nanopore sequencing reads are able to span nearly all common repeats, simplifying the assembly process[37].

The team deployed the MaSuRCA hybrid assembler, an open source tool developed by Aleksey Zimin, a senior scientist in Professor Salzberg's lab. Briefly, this uses a k-mer lookup to extend short sequencing reads base by base, at both the 5' and 3' ends (as long as the extension is unique), to form much longer 'super-reads'. The combination of super-reads and long nanopore sequencing reads then enable the generation of even larger 'mega-reads'.

The sequoia sample was sequenced to a depth of 135x using short-read technology and 22x using nanopore sequencing on the MinION. Assembly using the short-read data alone generated 2,507,175 contigs; however, addition of the long-read nanopore data delivered a 30-fold reduction in contig number (**Table 1**).

	Sequence in contigs	Contig N50 (bp)	Number of contigs
Short read only	7.91 Gb	12,036	2,507,175
Short read + nanopore long reads	8.12 Gb	359,531	49,676

Table 1
Addition of long nanopore sequencing reads provided a 50-fold reduction in the number of contigs, and a 20-fold increase in contig sizes, for the sequoia genome assembly. Data courtesy of Professor Steven Salzberg, Johns Hopkins University, US.

To further enhance assembly contiguity, the team collaborated with Dovetail Genomics to use Hi-C chromosome conformation in conjunction with Dovetail's HiRise assembly algorithm, a technique that they had previously successfully used to generate a chromosome-level assembly of the walnut (*Juglans regia L*) genome[38]. Comparing the walnut and sequoia nanopore sequencing reads, Professor Salzberg commented that the more recent sequoia reads were significantly longer, reflecting the rapid development of nanopore technology and optimisation of their sequencing workflow.

Assembly using the HiRise algorithm generated 11 'enormous' chromosome-size scaffolds ranging from 443 Mb to 985 Gb in size. Describing such large scaffolds as 'spectacular' and 'transformative', Professor Salzberg noted that these are the largest scaffolds ever assembled for any genome.

The 26.5 Gb hexaploid (six copies of each chromosome) coast redwood genome provided the team with an even sterner assembly challenge. In total, 3.2 trillion bases of short-read data and 582 billion bases of nanopore sequencing data were generated, representing 122x and 21x genome coverage respectively. Confirming the scale of the task, subsequent genome assembly took 6 months (or approximately 700,000 CPU hours post error correction). Hi-C scaffolding is currently ongoing; however, the initial hybrid assembly strategy delivered a N50 contig size of 110 kb and a longest contig of 2.4 Mb. Professor Salzberg suggests that, using the final assembly, it may be possible to segment the redwood genome into its three sub-genomes, shedding new light on the evolutionary history of this iconic species[37].

The Redwood Genome Project is led by Professor David Neale at UC Davis and Steven Salzberg at Johns Hopkins University, and is funded by the non-profit conservation group, Save The Redwoods League.

CASE STUDY 3

The importance of structural variation in crop breeding

Brassica napus (oilseed rape) is a major oil crop worldwide, with widespread application in cooking, biofuel, and animal feed. The 1.2 Gb *B. napus* genome is allotetraploid, with one set of chromosomes from *B. oleracea* (e.g. cabbage; sub-genome C) and another from *B. rapa* (e.g. turnip; subgenome A) revealing the organism's evolutionary history.

The genome of *B. napus* displays extensive gene- and chromosome-level structural variation (SV), which underlies important phenotypic traits, such as flowering time, disease resistance, and seed quality. Precise resolution of these SVs could support improvement of these economically important crops.

Short-read sequencing technology has been utilised to describe many SVs; however, due to the genome's tetraploidy and the propensity of short sequencing reads to map to more than one location, resolution of these aberrations to the sub-genome level is extremely challenging.

In order to more accurately resolve SVs at the sub-genome level in *B. napus*, researchers at the Justus Liebig University utilised nanopore sequencing reads, which, due to their long lengths, are far less likely to multimap[39].

The team sequenced four diverse *B. napus* lines taken from sites around the world, including North America (N99, spring flowering type), China (PAK85912, semi-winter flowering type), and Europe (Express 617 and R53, winter flowering types). To ensure accurate delineation of large SVs, the researchers implemented a size selection step using the Circulomics Short Read Eliminator XL Kit, which is designed to deplete DNA fragments less than 40 kb in length. The most recent runs, which also utilised a nuclease wash to maximise pore availability and enhance sequencing yield, delivered over 30 Gb of data on a single MinION Flow Cell, with a read N50 of approximately 40 kb.

The resulting sequencing data were aligned to a reference using NGMLR[40] prior to SV calling using the Sniffles[41] algorithm.

> ***'[We] identified insertions, which were almost impossible to detect with small read length technologies'[39]***

The team observed that the majority of SVs across all plant lines ranged from 100–1,000 bp in length, with lead researcher Harmeet Singh Chawla commenting that such SVs would be 'almost impossible' to detect using short-read sequencing technology (**Figure 6a**)[39]. It was also evident that larger SVs were detected in spring flowering genotypes (N99 and PAK85912) when compared to winter flowering genotypes (R53 and Express 617) (**Figure 6b**)

Interestingly, between 5-8% of genes were found to contain SVs, with lower SV diversity observed in the C-subgenome compared to the A-subgenome. According to Harmeet, this is likely to reflect the breeding history of the crop, as many traits have been artificially bred in the cabbage (C-subgenome) where the turnip (A-subgenome) remains relatively unaltered.

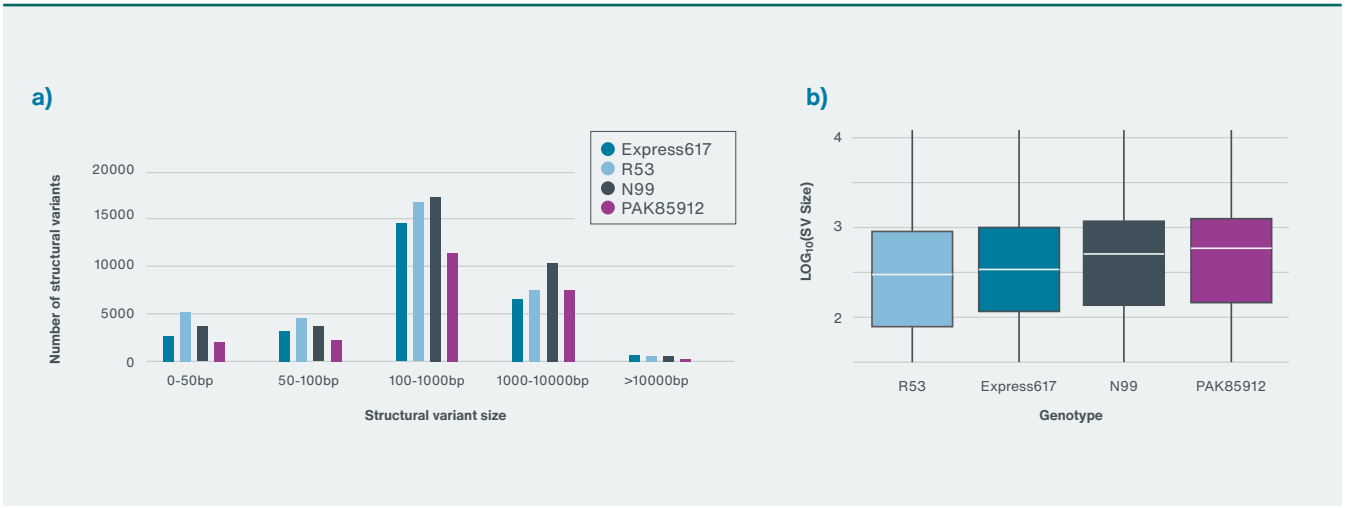


Figure 6
The majority of SVs detected across all *B. napus* lines were between 100 bp and 1,000 bp in length (a). Overall, the spring flowering lines N99 and PAK85912 contained larger SVs (b). Figure courtesy of Harmeet Singh Chawla, Justus Liebig University, Germany[39].

Examining genes known to be involved in geographical adaptation, the team observed a number of SVs, including a 90 bp insertion in *BnVIN3* — a gene associated with flowering time. Interestingly, this insertion was found in only one of the two winter flowering lines, Express 617.

SVs associated with disease resistance were also identified, including a 725 bp deletion in the 4-Coumarate:CoA ligase gene of the R53 line (**Figure 7**). According to Harmeet, this variant explains 20% of the *Verticillium* (a major fungal pathogen of *B. napus*) resistance phenotype[39].

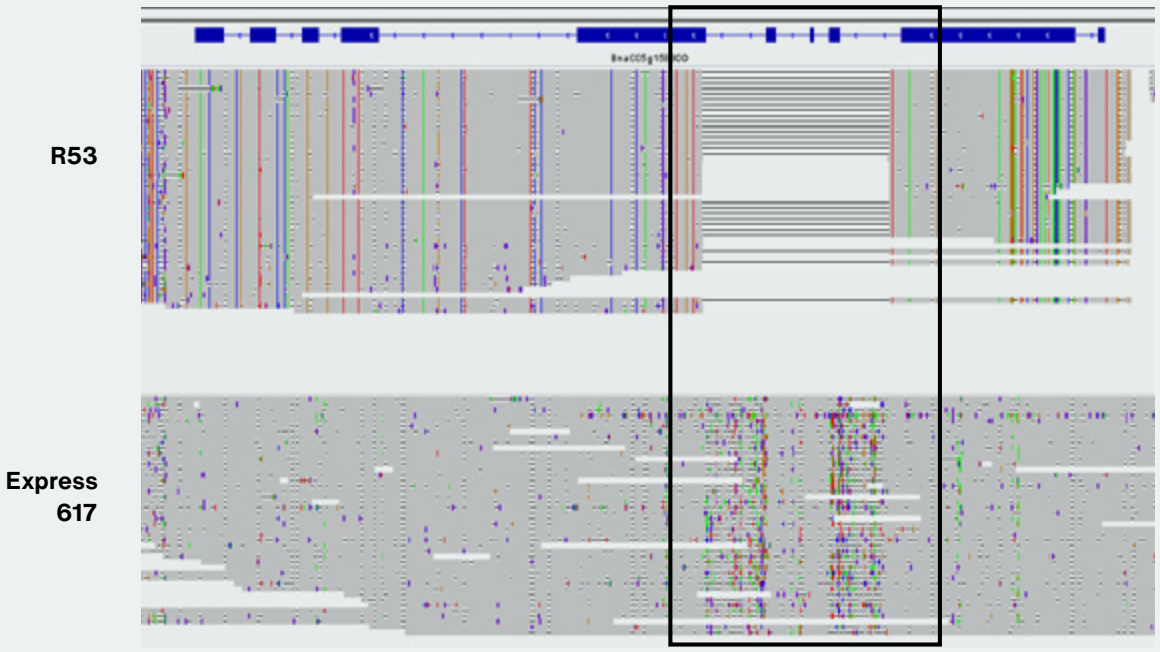


Figure 7
Long nanopore sequencing reads enabled the identification of a 725 bp deletion in the 4-Coumarate:CoA ligase gene, which was observed in the R53 but not the Express 617 winter flowering line. Figure courtesy of Harmeet Singh Chawla, Justus Liebig University, Germany[39].

'There is much more than SNPs to explain the observable phenotype in oilseed rape'[39]

Researchers at Johns Hopkins University are also utilising nanopore sequencing to support comprehensive SV analysis of important crops[42]. Using the high-throughput PromethION platform, 100 diverse tomato genomes were sequenced in just 100 days, identifying between 25,000–45,000 SVs per genome, including an 83 kb tandem duplication that increases fruit yield.

For more information on this project and the use of nanopore sequencing to accurately and rapidly characterise the SV landscape, download the Structural variation white paper at www.nanoporetech.com/resource-centre/white-papers.

CASE STUDY 4

Enhancing gene annotation and gene expression analysis using full-length RNA sequencing

Duckweed is the fastest growing plant on Earth, and this attribute, combined with its small genome, minimal gene set, aquatic lifestyle, and transformation system are behind its recent resurgence as a model research organism. Furthermore, some species of duckweed such as *Wolffia arrhizal*, known as khai-nam (or 'eggs of the water') in Thailand, are edible, offering the potential for a new generation of fast-growing, nutritious, and sustainable crops.

Comprised of five genera (*Spirodela, Landoltia, Lemna, Wolffiella, and Wolffia*), duckweed genome sizes span an order of magnitude — from 150 Mb (*Spirodela*) to 1,881 Mb (*Wolffia*).

Possessing the largest body, smallest genome, and fewest genes, researchers at the J. Craig Venter Institute (JCVI) sought to further characterise the genome of *Spirodela polyrhiza* and study gene expression using full-length nanopore cDNA sequencing reads[43]. Initial genomic analysis using long nanopore sequencing reads and optical mapping allowed the generation of a highly contiguous genome assembly with chromosome-arm level resolution[44]. The assembly also allowed the identification of errors in previous reference genomes created using short-read sequencing technology.

Daily light-dark cycles and the internal circadian clock drive most plant gene processes to specific times of the day, and, as such, time-of-day (TOD) sampling can provide greater insight into different gene expression networks. Utilising the long-read capability of nanopore technology, the JCVI team performed full-length cDNA (FL-cDNA) transcriptome sequencing of *Spirodela polyrhiza* samples taken every four hours over two days. The resultant expression data was found to be highly concordant to that obtained using short-read sequencing technology; however, as the nanopore reads were full

> ***'We demonstrate here the capability of the low-cost, long-read sequencing technologies such as the Oxford Nanopore platform to provide genome-wide, sequence-based validation of sequence assemblies at high resolution as well as to identify likely regions of mis-assembly in a genome draft'[44]***

length, far fewer of them were required to identify cycling genes (i.e. genes whose expression levels change dependent on time of day)[43].

The FL-cDNA reads were also found to provide more accurate gene models than those obtained using traditional short-read technology (**Figure 8**).

Critically, the long nanopore reads also enabled the identification of many more alternative gene transcripts, enabling more powerful, comprehensive analyses (**Figure 9**).

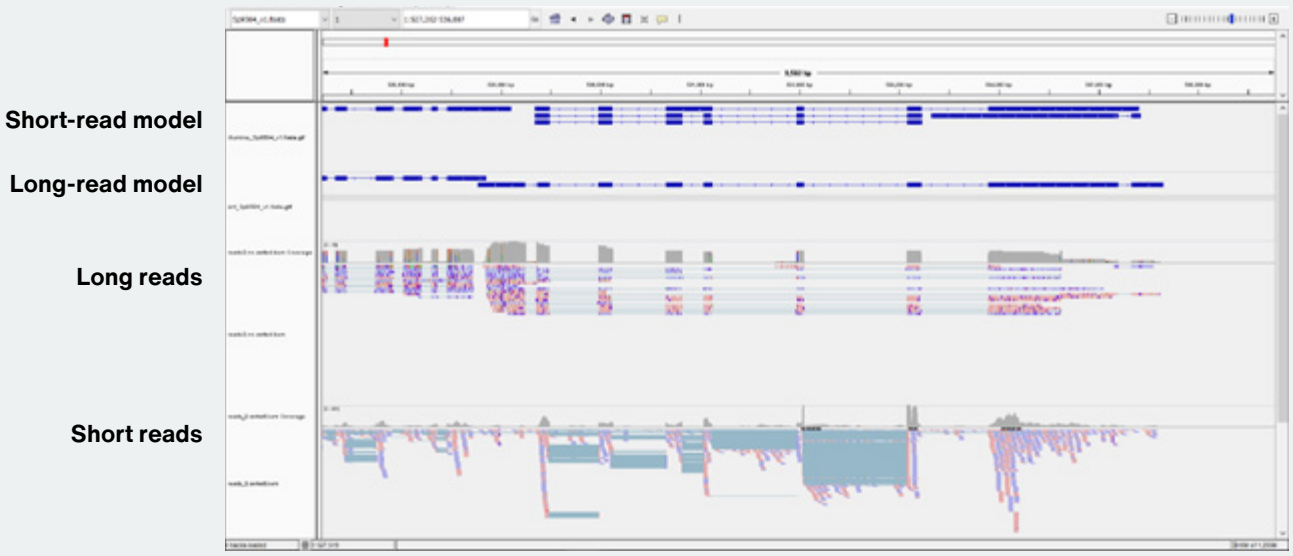
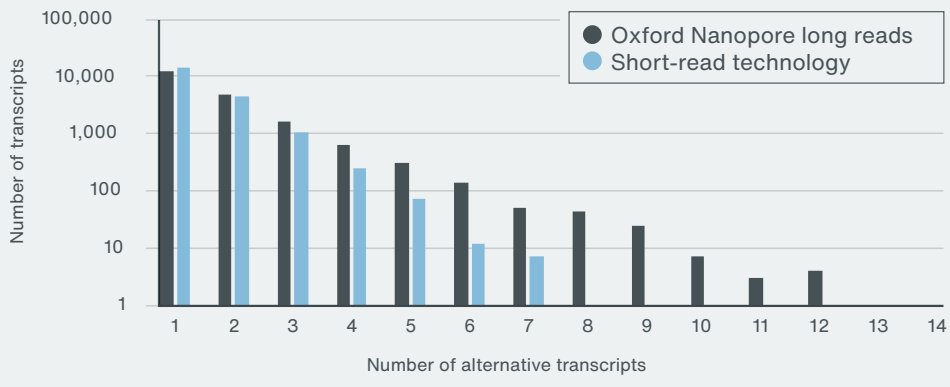


Figure 8
Full-length cDNA sequencing using nanopore technology allowed the creation of more accurate gene models than provided by short-read sequencing technology. Figure kindly provided by Professor Todd Michael, JCVI, US[43].

Figure 9
Full-length nanopore cDNA sequencing reads enabled the identification of more transcript isoforms than a traditional short-read RNA-Seq methodology. Figure kindly provided by Professor Todd Michael, JCVI, US[43]



Using FL-cDNA reads to analyse the expression of LHY, a key circadian clock gene, the team were able to correlate alternative isoform expression with different timepoints. In addition, a number of novel transcripts were identified that the team plan to further validate.

Furthermore, the FL-cDNA allowed expression analysis of paralogous genes (caused by tandem repeats and whole genome duplications), revealing that these genes often have distinctly different expression levels and cycling. Lead researcher Professor Todd Michael described this as a 'game-changer' in plant gene expression analysis, allowing greater insights into polyploidy and recent genome duplications[43]. Highlighting the accessibility of nanopore technology, all of the library preparation and sequencing work was undertaken by an undergraduate student.

Summary

In the year 2000, *Arabidopsis thaliana* became the first plant to have its genome sequenced and assembled. This landmark achievement took 10 years to complete and cost approximately $100 million. Now this feat can be achieved within a few days using a single MinION Flow Cell at the cost of $1,000[19].

With the facility to generate long and ultra-long sequencing reads, nanopore technology offers many advantages over traditional short-read approaches, allowing researchers to assemble and explore previously intractable plant genomes and transcriptomes — providing new and important insights into plant evolution and breeding strategies. These advances are essential in order to cost-effectively address some of the world's most pressing challenges, such as reliably feeding a rapidly growing population while facing the agricultural threats of climate change, pest resistance, and reliance on a limited number of genetically homogeneous crop species.

A significant requirement for plant research is the capability to characterise and access the genetic variation held within large plant repositories, such as seed banks and transgenic lines. This need is now being met by the flexible, high-throughput PromethION platform, which can deliver up to 14 Tb of data in 72 hours* — providing researchers with the facility to completely and rapidly characterise large numbers of plant genomes.

* Theoretical max output when system is run for 72 hours at 420 bases / second. Outputs may vary according to library type, run conditions, etc.

About Oxford Nanopore Technologies

Oxford Nanopore Technologies introduced the world's first nanopore DNA sequencer, the MinION — a portable, real-time, long-read, low-cost device — followed by the larger GridION Mk1 and PromethION 24 and 48 devices. The latest addition to the range, MinION Mk1C, combines the portability and power of the MinION with high-performance compute and an integrated touchscreen, providing a complete, go-anywhere solution for nanopore sequencing.

For smaller-scale analyses, such as optimising sample parameters prior to a large genome sequencing experiments, Flongle — a flow cell adapter for the MinION and GridION platforms — provides a highly affordable option.

Through the utilisation of long reads, which allow the analysis of regions of repetitive DNA, large structural variation, and full-length RNA isoform characterisation, nanopore sequencing technology allows the delivery of more complete genomic and transcriptomic analyses than is possible using traditional short-read sequencing technology[19].

A range of platforms are available, suitable for different sized genomes and higher coverage and throughput requirements (**Table 2**).

	Flongle	MinION & MinION Mk1C	GridION Mk1	PromethION 24	PromethION 48
Read length	Fragment length = read length. Longest read now >4 Mb[9]				
Run time	1 min – 16 hrs	1 min – 72 hrs	1 min – 72 hrs	1 min – 72 hrs	1 min – 72 hrs
Number of flow cells per device	1	1	5	24	48
DNA sequencing yield per flow cell*	Up to 2.8 Gb	Up to 50 Gb	Up to 50 Gb	Up to 290 Gb	Up to 290 Gb
DNA sequencing yield per device*	Up to 2.8 Gb	Up to 50 Gb	Up to 250 Gb	Up to 7 Tb	Up to 14 Tb
Multiplexing	1 – 96 samples	1 – >2,000 samples	1 – >2,000 samples	1 – >2,000 samples	1 – >2,000 samples

* Theoretical max output when system is run for 72 hours (or 16 hours for Flongle) at 420 bases / second. Outputs may vary according to library type, run conditions, etc.

Table 2
A range of nanopore sequencing devices is available to suit all applications.
Data correct at time of print. Visit www.nanoporetech.com for the latest information.

For the latest information about sequencing plant genomes using nanopore technology, visit www.nanoporetech.com/applications.

References

1. Royal Botanic Gardens Kew. 2016. State of the world's plants. Available at: https://stateoftheworldsplants.com/2016/report/sotwp_2016.pdf [Accessed: 10 December 2019]

2. Finkers, R. Know your onion — The impact of long reads on large genomes. Presentation. Available at: https://nanoporetech.com/resource-centre/know-your-onion-impact-long-reads-large-genomes [Accessed: 10 December 2019]

3. United Nations. 2017. World population projected to reach 9.8 billion in 2050, and 11.2 billion in 2100. Available at: https://www.un.org/development/desa/en/news/population/world-population-prospects-2017.html [Accessed: 10 December 2019]

4. NCBI National Center for Biotechnology Information. Genomes information by organism. Available at: https://www.ncbi.nlm.nih.gov/genome/browse#!/overview/ [Accessed: 10 December 2019]

5. Royal Botanic Gardens Kew. 2017. State of the world's plants. Available online: https://stateoftheworldsplants.com/2017/report/SOTWP_2017.pdf

6. Michael, T.P. (2019) Personal communication with Oxford Nanopore Technologies on 10 October 2019.

7. Fleischmann, A. et al. Evolution of genome size and chromosome number in the carnivorous plant genus *Genlisea* (*Lentibulariaceae*), with a new estimate of the minimum genome size in angiosperms. Annals of Botany. 114 (8): 1651–1663 (2014).

8. Pellicer, J., Fay, M.F., and Leitch, I. J. The largest eukaryotic genome of them all? Botanical Journal of the Linnean Society. 164 (1) (2010).

9. Oxford Nanopore Technologies. Ultra-Long DNA Sequencing Kit. Available at: https://store.nanoporetech.com/ultra-long-dna-sequencing-kit.html [Accessed: 23 August 2021]

10. Michael, T. The complex architecture of plant T-DNA transgene insertions. Presentation. Available at: https://nanoporetech.com/resource-centre/complex-architecture-plant-t-dna-transgene-insertions [Accessed: 10 December 2019]

11. Yang, J. et al. Haplotype-resolved sweet potato genome traces back its hexaploidization history. Nature Plants. 3, 696–703 (2017).

12. Minio, A., Lin, J., Gaut, B.S. and Cantu, D. How single molecule real-time sequencing and haplotype phasing have enabled reference-grade diploid genome assembly of wine grapes. Front Plant Sci. 8:826 (2017).

13. Feschotte, C., Jiang, N. and Wessler, S.R. Plant transposable elements: where genetics meets genomics. Nat Rev Genet 3(5):329-41 (2002).

14. Schnable, P.S. et al. The B73 maize genome: complexity, diversity, and dynamics. Science. 326:1112–1115 (2009).

15. Lee, S.-I and Kim, N.-S. Transposable elements and genome size variations in plants. Genomics Inform. 12(3): 87–97 (2014).

16. Debladis, E., Llauro, C., Carpentier, M., Mirouze, M. and Panaud, O. Detection of active transposable elements in *Arabidopsis thaliana* using Oxford Nanopore Sequencing technology. BMC Genomics. 18:537 (2017).

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

18. Usadel, B. (2017) Complex tomato genomes: Easy with nanopores. Presentation. Available at: https://nanoporetech.com/index.php/talk/complex-tomato-genomes-easy-nanopores [Accessed: 10 December 2019]

19. Michael, T.P. et al. High contiguity *Arabidopsis thaliana* genome assembly with a single nanopore flow cell. Nat. Commun. 9(1):541 (2018).

20. International Food Policy Research Institute. International germplasm exchanges are getting more complicated, and here's why that matters to global food security. 2017. Available at: http://www.ifpri.org/blog/international-germplasm-exchanges-are-getting-more-complicated-and-heres-why-matters-global [Accessed: 10 December 2019]

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

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

23. Wittenberg, A. PromethION sequencing of complex plant genomes. Presentation. Available at: https://nanoporetech.com/resource-centre/talk/promethion-sequencing-complex-plant-genomes. [Accessed: 10 December 2019]

24. Bolger, A. LOGAN: Lossless graph-based analysis of NGS data sets. Presentation. Available at: https://nanoporetech.com/resource-centre/logan-lossless-graph-based-analysis-ngs-datasets [Accessed: 10 December 2019]

25. Steijger, T. et al. Assessment of transcript reconstruction methods for RNA-seq. Nat. Methods 10, 1177–1184 (2013).

26. Timp, W. & Jain, M. Direct RNA cDNA sequencing of the human transcriptome. Presentation. Available at: https://nanoporetech.com/resource-centre/direct-rna-cdna-sequencing-human-transcriptome [Accessed: 22 October 2019]

27. Ravindran, S. Barbara McClintock and the discovery of jumping genes. Proc Natl Acad Sci U S A. 109(50):20198-9 (2012).

28. He, H.-J., Chen, T. and Zhu, J.-K. Regulation and function of DNA methylation in plants and animals. Cell Res. 21(3): 442–465 (2011).

29. Parker, M.T. et al. Nanopore direct RNA sequencing maps an Arabidopsis N6 methyladenosine epitranscriptome. bioRxiv 706002 (2019).

30. Shen, L., Liang, Z., Wong, C.E., and Yu, H. Messenger RNA modifications in plants. Trends Plant Sci. 24(4):328-341 (2019).

31. Lanciano, S. and Mirouze, M. DNA methylation in rice and relevance for breeding. Epigenomes 1(2), 10 (2017).

32. Simpson, J.T. et al. Detecting DNA cytosine methylation using nanopore sequencing. Nature Methods. 14, 407–410 (2017).

33. Rand, A.C. et al. Mapping DNA methylation with high-throughput nanopore sequencing. Nat Methods. 14(4):411-413 (2017).

34. Jupe, F. et al. The complex architecture and epigenomic impact of plant T-DNA insertions. PLoS Genet. 15(1): e1007819 (2019).

35. Fraley, R.T. et al. Expression of bacterial genes in plant cells. Proc. Natl. Acad. Sci. U.S.A. 80 (15): 4803–07 (1983).

36. Li, S. et al. Mapping of transgenic alleles in soybean using a nanopore-based sequencing strategy. J. Exp. Bot. 7;70(15):3825-3833 (2019).

37. Salzberg, S. Sequencing and assembling the mega-genomes of mega-trees: the giant sequoia and coast redwood genomes. Presentation. Available at: https://nanoporetech.com/resource-centre/sequencing-and-assembling-mega-genomes-mega-trees-giant-sequoia-and-coast-redwood-0 [Accessed: 15 December 2019]

38. Marrano, A. et al. High-quality chromosome-scale assembly of the walnut (Juglans regia L) reference genome. bioRxiv 809798 (2019).

39. Chawla, H.S. Long reads reveal small-scale genome structural variations in allotetraploid canola. Presentation. Available at: https://nanoporetech.com/resource-centre [Accessed: 15 December 2019]

40. GitHub. NGMLR. Available at: https://github.com/philres/ngmlr [Accessed: 15 December 2019]

41. GitHub. Sniffles. Available at: https://github.com/fritzsedlazeck/Sniffles [Accessed: 15 December 2019]

42. Schatz, M. 100 genomes in 100 days: The structural variant landscape in tomato genomes. Presentation. Available at: https://nanoporetech.com/resource-centre/michael-schatz-100-genomes-100-days-structural-variant-landscape-tomato-genomes [Accessed: 15 December 2019]

43. Michael, T. Full-length cDNA sequencing coupled to time-of-day sampling enables enhanced gene prediction in the fastest growing plant on Earth. Presentation. Available at: https://nanoporetech.com/resource-centre [Accessed: 15 December 2019]

44. Hoang, P.N.T. et al. Generating a high-confidence reference genome map of the Greater Duckweed by integration of cytogenomic, optical mapping, and Oxford Nanopore technologies. Plant J. 96(3):670-684 (2018).

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

WP_1052(EN)_V2_24Aug2021