Oxford NANOPORE Technologies

RNA SEQUENCING

The value of full-length transcripts without bias

WHITE PAPER

October 2019

Contents

Introduction

RNA sequencing (referred to as RNA-Seq with traditional sequencing technologies) has led to unprecedented advances in all fields of biology and medicine. It has been an invaluable tool for the study of human genetics and the pathology associated with disease. Transcript isoform expression and usage, for example, is a prominent source of variation between healthy and diseased tissues in a number of medical conditions, including cancer[1,2]. RNA sequencing is also instrumental in identifying fusion transcripts present in a growing number of disorders[3].

Transcript isoform expression and usage is a prominent source of variation between healthy and diseased tissues.

Sequencing of cDNA has also significantly aided viral pathogen characterisation and timely detection[4,5,6], drastically improving time-to-result compared to the gold-standard viral isolation in cell culture. RNA sequencing of viral-infected samples improves sensitivity[7], specificity, and is responsive to rapid viral evolution, overcoming the issues of the alternative methods which rely on ELISA and RT-PCR. The numerous applications of RNA sequencing are not, however, restricted to the human health research field. The approach has also been utilised in agricultural settings, for example, to research the drought-induced stress response in plants[8].

Lastly, RNA sequencing use in developmental biology has helped elucidate transcriptional program changes associated with various developmental events[9].

Traditional sequencing technologies have, undoubtedly, made comprehensive transcriptome analysis possible and have led to numerous important developments in science. Nevertheless, there are some important limitations in the field that need to be addressed. Here, we will focus on the facility of nanopore technology to tackle challenges in the areas of full-length transcript identification, isoform characterisation and quantification, and viral detection.

Transcriptome analysis challenges

Obtaining and analysing full-length transcripts

High-throughput RNA sequencing approaches often encounter difficulties with the generation of full-length transcripts, mostly due to premature termination of reverse transcription[10]. When using an oligo-dT primer, this could be reflected in an over-representation of 3' transcript ends at the expense of complete sequences[11]. A mitigation to this problem is presented by the reverse transcriptase 'template-switching' method used for full-length transcript enrichment[12]. However, due to the fragmentation required by most traditional sequencing methods, accurate assembly of complete transcripts is still difficult, especially where a read maps to more than one location (e.g. highly conserved stretches of sequence) (**Figure 1**).

With nanopore technology, entire fragments are sequenced, regardless of their length. Hence, a full-length cDNA can be sequenced as a single fragment, reducing significantly multiple-locus alignment issues later on (**Figure 1**). The technology allows for the template-switching method to reach its maximum potential, as demonstrated in a study by Oikonomopoulos *et al.*[13], where 90.8% of their nanopore reads were of full-length transcripts. Importantly, the use of the template-switching protocol is not a prerequisite for obtaining full-length transcripts with nanopore sequencing, as shown in an investigation by Hargreaves and Mulley[14] where 88% of the reads represented full-length transcripts. Long reads provided by nanopore sequencing devices range from 1–100s kb, making the identification of full-length transcripts and *de novo* transcriptome assembly easier than using traditional sequencing platforms.

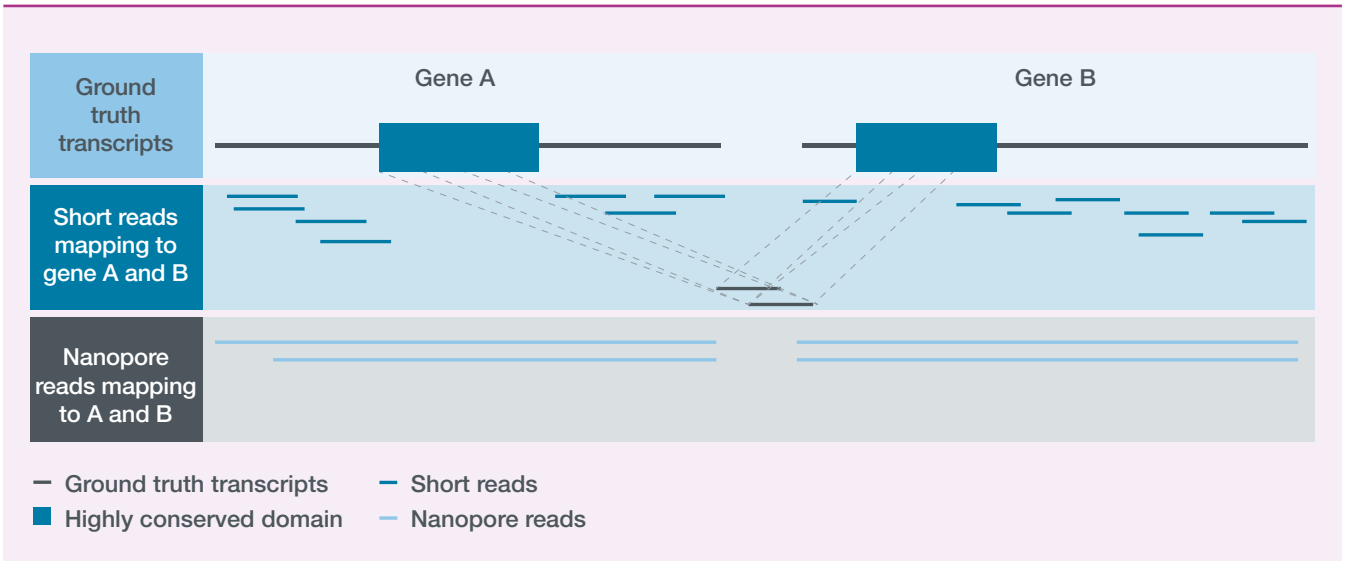


Figure 1
Assembly of full-length transcripts with short reads and long nanopore sequencing reads.

Accurate identification of multiple isoforms per gene

The process of alternative splicing, whereby a precursor mRNA molecule can give rise to multiple mature mRNA molecules that differ in their exon content, allows many proteins to be encoded by a single gene[15]. In order to correctly identify these unique mRNA isoforms, it is necessary to sequence several consecutive exons, which, due to their length, is beyond the capabilities of short-read sequencing technologies (**Figure 2**).

Long nanopore reads, generating complete transcript isoform sequences, circumvent issues with assembly.

As a result, isoforms are reconstructed using computational techniques; however, a study by Steijger *et al.*[16] 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. These complications are further compounded where reads from highly similar transcripts, such as those of paralogous genes, are under investigation. Rare isoforms could also remain altogether undetected[17].

Long nanopore reads, generating complete transcript isoform sequences, circumvent these issues with assembly. Their advantage in splice variant identification has been repeatedly demonstrated[14,15,18,19]. Of note, nanopore sequencing allowed the identification of over 7,000 full-length isoforms of *Dscam1, 'the most complicated alternatively spliced gene known in nature'*[15].

Reliable isoform resolution by nanopore technology makes the *de novo* assembly of complex alternatively spliced transcriptomes possible.

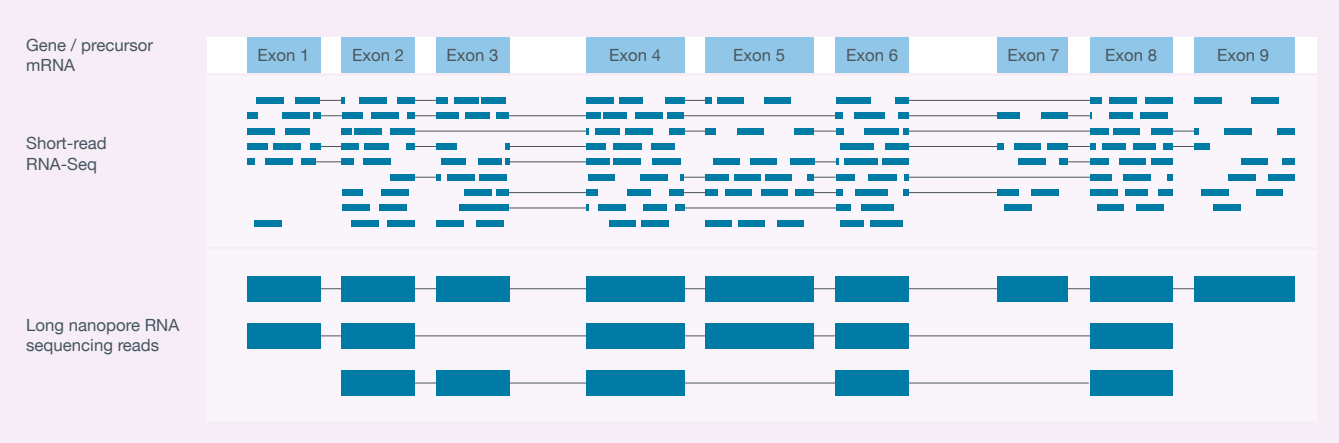


Figure 2
Alternative splicing can give rise to numerous mRNA isoforms per gene, which in turn can alter protein composition and function. The short reads generated by traditional RNA sequencing techniques lose positional information, making the correct assembly of alternative mRNA isoforms challenging. Long nanopore reads can span full-length transcripts simplifying their identification.

Problematic isoform identification and quantification

As a consequence of not being able to accurately identify isoforms, traditional sequencing platforms have also encountered difficulties with the quantification of isoform levels. Nanopore cDNA sequencing was shown to perform comparably to other well-established technologies in single-cell gene expression analysis, as indicated by a strong correlation between short-read and nanopore expression data[20]. Additionally, the nanopore platform is capable of generating data for reliable isoform quantification as a stand-alone tool[20] or as part of a hybrid sequencing strategy[18].

Direct RNA sequencing — unbiased, full-length transcript and base modification analysis

Until recently, sequence-based analysis of RNA required the conversion of RNA to complementary DNA (cDNA), a process that can introduce bias through reverse transcription or amplification. Research shows that PCR amplified libraries tend to have reduced complexity when compared to the total mRNA pool[21]. Furthermore, not all transcripts amplify with the same efficiency, causing drop-out of some RNA species and excessive amplification of other species[21]. These issues can be exacerbated by the use of traditional short-read sequencing technologies which are known to exhibit GC bias, where sequences with low or high levels of GC content are underrepresented.

An additional limitation of cDNA-based studies is that the process of PCR amplification loses all information on modified bases. Such base modifications are known to have a role in modulating the activity and stability of RNA and are therefore of increasing interest to researchers. Nanopore sequencing overcomes all of these challenges through the facility for direct RNA sequencing — delivering unbiased, full-length, strand-specific RNA sequences (**Figure 3**)[22].

With nanopore sequencing, the read length is only limited by the size of the RNA (or DNA) fragment presented to the pore, making it possible to sequence extremely long, full-length transcripts. The longest transcript processed by direct RNA sequencing currently stands at over 20 kb in length[23].

> ***'Direct RNA methylation analysis will radically change how people study RNA in the next decade'[24]***

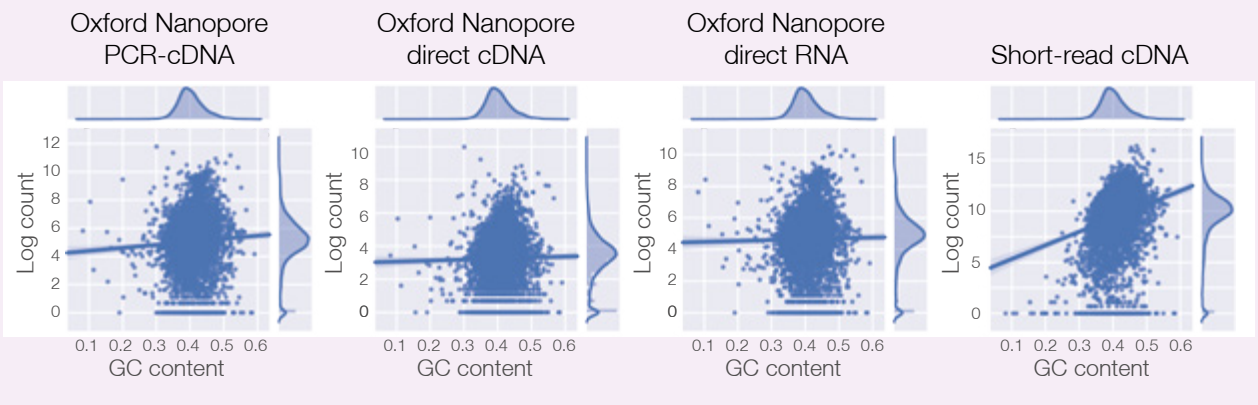


Figure 3
Sequencing workflows that incorporate amplification are vulnerable to sequence-specific biases. Yeast transcriptome libraries were prepared using three nanopore sequencing techniques (PCR-cDNA, direct cDNA and direct RNA) and a typical short-read cDNA technique. In all cases, GC bias in the nanopore data sets was lower than in the short-read data set[22].

Direct RNA sequencing also allows the detection of epigenetic modifications alongside sequence data. To date, the RNA modifications detected using nanopore sequencing technology include, but are not limited to, pseudouridine[23], N6-methyladenosine (m6A)[21,25] and 5-methylcytosine (5mC)[21,26]. In the opinion of Dr. Miten Jain of the University of California, Santa Cruz: *'The facility of nanopore technology for direct RNA methylation analysis will radically change how people study RNA in the next decade'*[24].

A further benefit of direct RNA sequencing is the ability to accurately measure poly-A tail length. In eukaryotes, messenger RNA (mRNA) is augmented with a series of adenosine bases at the 3' end known as the poly-A tail. These tails can vary in size, with the largest being over 250 nucleotides in length and therefore beyond the typical analysis capabilities of short-read sequencing technologies[23,27]. Research suggests that poly-A tail length is an important factor in post-transcriptional regulation and further study may provide new insights into gene expression and disease[23,27]. Using nanopore sequencing, researchers have identified dynamic poly-A tail length distributions across the development stages of the model organism *Caenorhabditis elegans*[28]. Different poly-A tail lengths for alternative human transcript isoforms has also been described[29].

***'An advantage of profiling poly(A) tail lengths with direct RNA-seq versus short-read sequencing is that poly(A) tail lengths are directly coupled to information about the splice isoforms and 3'UTR isoforms of the associated read'*[28]**

Direct RNA sequencing is already being used by clinical researchers in order to develop new methods for viral pathogen detection, drastically reducing time to result. In the case of influenza virus, researchers from Centers for Disease Control and Prevention, USA, have developed a direct RNA sequencing protocol that reduces pathogen detection and characterisation time from days to just a few hours[30].

RNA sequencing depth guidelines for transcriptome analysis

Experiments aiming to compare transcriptional profiles between two samples may require 'modest' sequencing depth, while discovery of novel transcripts and quantification of isoforms usually requires greater sequencing depth (ENCODE Guidelines and Best Practices for RNA-Seq: Revised December 2016[31,32]). **Table 1** provides general guidance on the amount of data required for different RNA sequencing experiments.

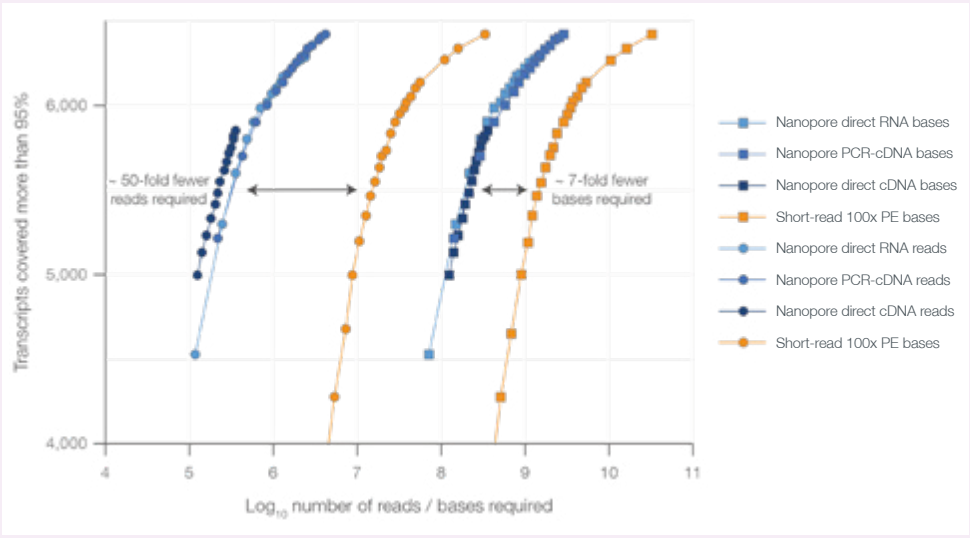
In a single-cell sequencing study, Byrne *et al.* demonstrated that obtaining 3–4 times fewer nanopore reads produced highly concordant results.

These recommendations are based on short-read sequencing technology, where *de novo* assembly and isoform detection are particularly difficult. Much smaller numbers of long nanopore reads are likely to be needed to achieve similar results. For example, in a single-cell sequencing study, Byrne *et al.*[20] demonstrated that obtaining 3–4 times fewer nanopore reads than short reads per cell was enough to produce highly concordant

Table 1
Guidelines for the number of reads required for different types of RNA sequencing experiments. The information is based on a mammalian genome with an approximate size of 3 Gb using short-read sequencing. The values are informed by Liu *et al.*[33,34], and ENCODE Guidelines and Best Practices for RNA-Seq[31,32].

Detection or application	Recommended reads per sample (millions)
Differential expression profiling	10–25 M paired-end reads (~1.5 Gb–3.75 Gb)
Alternative splicing	50–100 M paired-end reads (~7.5 Gb–15 Gb)
Isoform differential expression	>100 M paired-end reads (>15 Gb)
De novo assembly	>100 M paired-end reads (>15 Gb)

Figure 4
Nanopore RNA sequencing techniques, which can deliver long, full-length reads, require less data to cover the same number of transcripts when compared to a typical short-read technology[22].



expression profiles between the two different platforms. Furthermore, for >70% of the cells, an expression level quantification saturation was reached[20]. Data comparing nanopore PCR-cDNA, PCR-free direct cDNA, and direct RNA sequencing with a typical short-read cDNA technology revealed that the nanopore techniques required 50-fold fewer reads and 7-fold fewer bases to cover the same number of transcripts (**Figure 4**)[22].

Oxford Nanopore offers scalable solutions for transcriptomic research with Flongle™, MinION™, GridION™, and PromethION™ 24, and PromethION 48, offering up to 2.8 Gb, 50 Gb, 250 Gb, 7 Tb and 14 Tb of data, respectively.*

A need for more rapid identification and characterisation of RNA viruses

Sequencing has been extremely useful in the detection of viral pathogens in clinical research samples. It allows monitoring of rapid viral mutations, unlike other available methods. However, most sequencing platforms are difficult to transport and are not suitable for field use. This means that samples need to be shipped to laboratories with the relevant infrastructure, resulting in potential delays that can be detrimental for scenarios demanding a fast response, such as viral outbreaks. Furthermore, reliable detection sometimes requires a culturing step if the viral titre is low.

Nanopore sequencing is capable of dramatically improving the speed of viral detection, making it possible to achieve sample-to-result times of approximately three hours[35].

Oxford Nanopore offers the hand-held, portable MinION and MinION Mk1C devices, which are suitable for use in both the field and in the lab[4]. A cDNA-based RNA sequencing protocol used in conjunction with nanopore sequencing is capable of dramatically improving the speed of viral detection, making it possible to achieve sample-to-result times of approximately three hours[35]. Moreover, the technology can identify viruses at titres as low as 10^5 copies/ml of blood directly from a metagenomic sample[5].

* 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.

Data-analysis tools

There are a number of tools and pipelines available for the analysis of both direct RNA and cDNA nanopore sequencing reads. Depending upon the aims of the study, RNA sequencing reads are typically either mapped to the transcriptome or genome. Transcriptome mapping is more convenient for transcript identification and quantification but does not allow the discovery of novel isoforms. Alternatively, mapping to the genome allows more comprehensive analysis of splice variants and alternative isoforms; however, it is computationally more challenging.

In addition to a number of community-developed tools for direct RNA and cDNA analysis, Oxford Nanopore currently offers three best practice pipelines that deliver robust and reproducible results.

Pipeline-transcriptome-de[36] utilises a transcriptome mapping approach to enable differential gene expression (DGE) and differential transcript usage (DTU) analysis with long nanopore reads. The pipeline is based on an approach described by Love *et al*.[37]

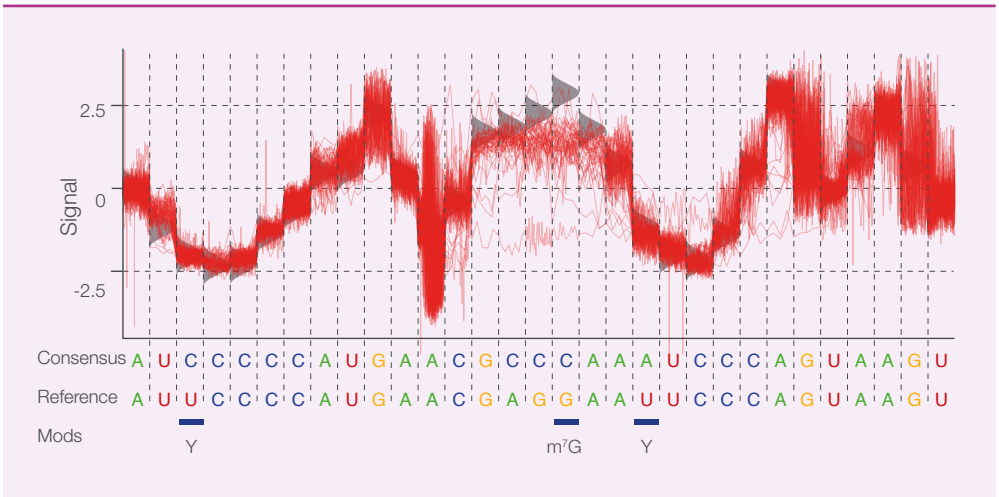
For genome annotation, Oxford Nanopore provides the pipeline-pinfish-analysis workflow, which incorporates pinfish[38] — a collection of tools for transcriptomic data analysis — and the alignment tool minimap2[39], to generate annotations in GFF2 file format, from long cDNA or direct RNA reads and a reference genome.

Lastly, pychopper[40] is a tool that selects for full-length cDNA reads by identifying reads that feature both reverse transcription primer and strand-switching primer sequences. It is recommended that pychopper is run for cDNA reads prior to using pipeline-pinfish-analysis for genome annotation.

Nanopore direct RNA sequencing does not require amplification or strand synthesis meaning that, during sequencing, modified bases pass through the pore, and provide a distinct signature that can be detected in the raw signal. Oxford Nanopore has developed the Tombo[41] analysis tool to enable modified base detection from direct RNA or DNA sequencing (**Figure 5**). Tombo allows researchers to investigate modified bases using a number of different computational methods, providing flexibility with regard to application type and experimental design. The platform can also be trained to identify additional non-standard bases.

Step-by-step tutorials are available for each Oxford Nanopore analysis pipeline in the Nanopore Community at www.nanoporetech.com/knowledge.

Figure 5
The Tombo data analysis package allows identification of a range of modified bases from raw nanopore signal.



CASE STUDY 1

Nanopore sequencing accurately resolves alternatively spliced transcripts

Short-read sequencing technologies struggle with transcripts of genes with multiple isoforms, due to difficulties in measuring exon connectivity where many exons are located further away from each other than the maximum read length[15,42]. Bolisetty *et al.*[15] used nanopore sequencing and the Drosophila model system for a transcriptomic analysis of genes famous for having a large number of isoforms (*Dscam1, Rdl, MRP, Mhc*).

The team detected a total of 7,899 full-length isoforms for these four genes using a gene-specific RT-PCR approach. *Dscam1* is the most abundantly alternatively spliced gene known and it has the potential to generate a total of 38,016 isoforms (115 exons, 95 of which are alternatively spliced, forming 4 clusters)[43]. For the purpose of the study, the authors focused on the region covering exons 3–10 (19,008 potential isoforms). They were able to reliably resolve 7,874 *Dscam1* isoforms with an average identity of >90% for full-length alignments, despite the fact that the exons within each cluster shared 80% nucleotide identity with each other (**Figure 6**).

Bolisetty *et al.*[15] demonstrated that long nanopore reads are particularly suited for the characterisation of multiple-isoform transcripts, even in cases where exons have significant similarity to one another.

A total of 7,874 *Dscam1* isoforms, with an average identity of >90% for full-length alignments, were reliably resolved.

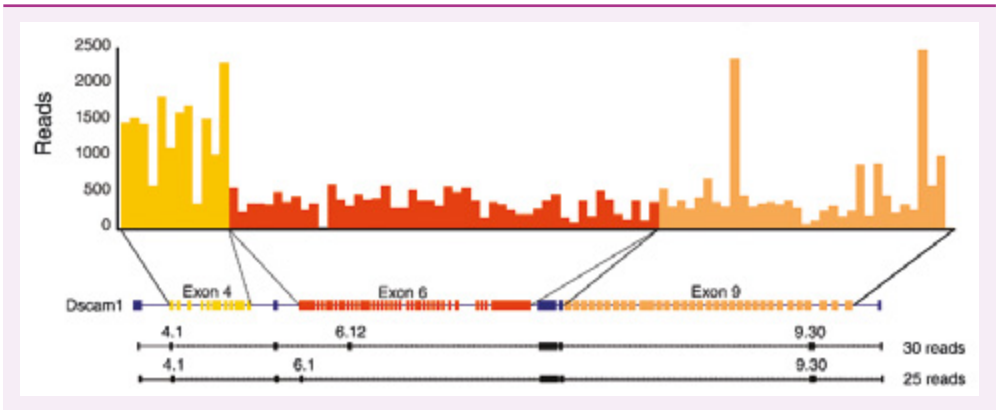


Figure 6
Nanopore sequencing of *Dscam1* isoforms. Exon cluster variants (top) and the isoform connectivity of two highly expressed *Dscam1* isoforms (bottom). Image taken with permission from Professor Brenton Graveley, University of Connecticut Health Center, Farmington, Connecticut, USA; image originally part of a figure from an article[15] published by BioMed Central.

CASE STUDY 2

Transcription dynamics of the developing olive fruit fly embryo

The olive fruit fly (*Bactrocera oleae*) is one of the most important pests of cultivated olive trees, causing an estimated $800 million of damage per year. Understanding the mechanisms of sex determination in such pests may reveal opportunities for genetic control. While it is known that sex determination in *B. oleae* is initiated within the first six hours of embryonic development, transcription dynamics of the developing embryo has not been explored and the sex determining factor remains elusive. To address this, Anthony Bayega and colleagues used the MinION to perform full-length cDNA sequencing of poly-A$^+$ RNA from mixed-sex embryos, collected at hourly intervals over the first six hours of development (**Figure 7**)[44,45].

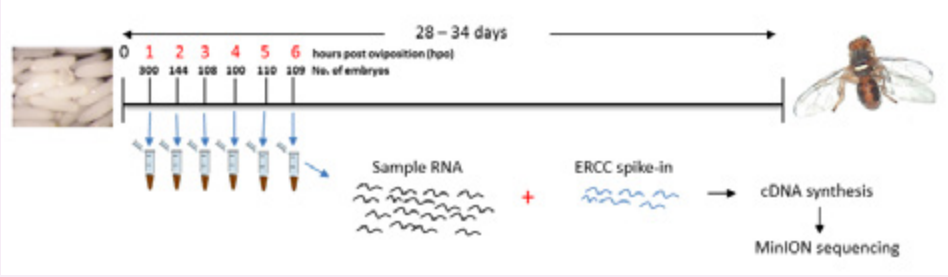
Over 50% of expressed genes were shown to have at least one full-length read. Reads containing a 5' adapter, poly-A$^+$ tail, and 3' adapter were selected to focus on full-length transcripts. Comparing sequences from this approach to those generated using short-read sequencing technology revealed that forty-fold fewer reads were required to detect the same number of genes.

De novo transcriptome assembly identified 3,553 novel genes, 8,330 genes matching the predicted NCBI genes, and a total of 79,810 transcripts. Overall, a four-fold increase in transcriptome diversity compared to the NCBI predicted transcriptome was achieved. Furthermore, 38 genes incorrectly modelled by NCBI were corrected with this dataset.

Adding an RNA standard (ERCC) during the cDNA synthesis stage further allowed absolute transcript numbers to be determined, revealing a 50% reduction in the number of transcripts between the first and second hours of embryo development. At three hours, the number of transcripts increased 143% compared to the 2-hour level. A similar transcript abundance pattern has previously been reported in the closely related Diptera *Drosophila melanogaster*.

Full-length nanopore sequencing reads were also used to explore the isoform complexity of genes involved in sex determination, revealing the prominence of a longer isoform of the *double sex (dsx)* gene transcript during early embryonic development, which shifts to the prevalence of a shorter isoform in adult tissue. Of particular note, the inclusion of exon 4 was only observed in adult female samples. These findings may enable the development of novel genetic control mechanisms to reduce the destructive impact of *B. oleae* and other pests on economically important crops.

Figure 7
Schematic of experimental procedure, from sample collection to sequencing. Adapted from Bayega *et al*[44].



CASE STUDY 3

The utility of full-length transcripts for cancer research

The facility of nanopore sequencing to deliver full-length transcripts and to accurately identify and quantify multiple isoforms has made the technology particularly useful for cancer research. The technology has been applied to the study of many cancer types including, leukaemia[46], breast cancer[47], ovarian cancer[47], and lung cancer[48].

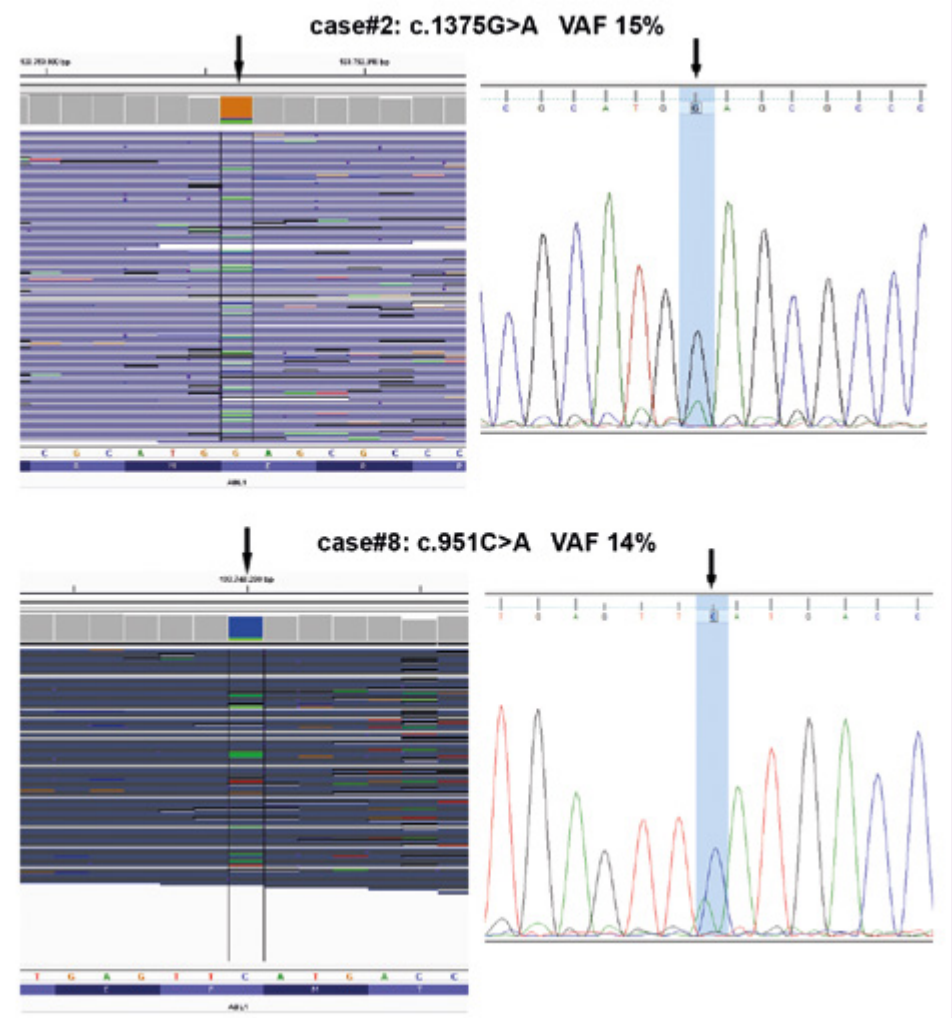
Suzuki *et al.*[48] employed nanopore cDNA sequencing to detect a variety of common mutation types, including single base substitutions, short deletions, exon skipping, and fusions, in a set of genes implicated in lung cancer (*EGFR, KRAS, NRAS, NF1, ALK* and *RET*). Further, the team utilised long nanopore sequencing reads to characterise the phasing of two genomic DNA mutations that are associated with anti-cancer drug efficacy. Following this successful study, the researchers commented that: '*The simple methods of MinION sequencing could possibly enable small/mid-scale research centers and hospitals to conduct research studies by genotyping driver genes and selecting suitable therapeutic approaches*'[48].

At the University of Bari, Italy, researchers are investigating the potential of nanopore sequencing to detect *BCR-ABL1* kinase domain (KD) mutations[46]. The *BCR-ABL1* gene fusion, which codes for a non-regulated tyrosine kinase protein, occurs in 95% of people with chronic myeloid leukaemia (CML)[49]. KD mutations in the *BCR-ABL1* gene contribute to and indicate resistance to first-line therapy with tyrosine kinase inhibitors[46]. The current gold-standard test for variant detection in *BCR-ABL1* is Sanger sequencing; however, with a sensitivity of just 20%, it is unsuitable for identifying low-level variants.

> ***'…MinION is markedly superior to Sanger sequencing in terms of sensitivity, costs and timesaving…'*[46]**

In a 24-sample study, the team assessed the performance of nanopore sequencing to identify KD mutations in a 1.7 kb cDNA amplicon at both 100x and 1,000x sequencing depth. The nanopore data allowed the identification of clinically important mutations in 2 samples which were not initially evident using Sanger sequencing (**Figure 8**). The team also reported that the lower nanopore sequencing depth of 100x was sufficient to identify all mutations. The long reads delivered by nanopore sequencing further enabled the detection of mutations that are in the same clone (*in cis*; compound mutants), something which is particularly challenging for Sanger or short-read sequencing approaches. The researchers summarised that: *'Our data indicates that MinION is markedly superior to Sanger sequencing in terms of sensitivity, costs and timesaving, and has the added advantage of determining the clonal configuration of multiple mutations'*[46].

Figure 8
Nanopore sequencing (left side) allowed the detection of mutations in 2 samples that were not initially evident in the Sanger sequencing data (right side), demonstrating the enhanced sensitivity of nanopore technology. Image courtesy of Dr. Crescenzio Minervini, University of Bari, Italy.



CASE STUDY 4

Novel isoform identification in a neuropsychiatric disease gene

Bipoloar disorder and schizophrenia have a strong genetic component, with genome-wide association studies (GWAS) identifying many risk loci. One such locus, *CACNA1C*, which codes for a voltage-gated calcium channel subunit, is a potential therapeutic target; however, its extreme length (>13 kb), number of exons (>50), and high level of alternative splicing means that *CACNA1C* isoform structure and expression is poorly understood[42].

Researchers from the University of Oxford and The Earlham Institute utilised nanopore sequencing to investigate *CACNA1C* expression in six regions of post-mortem human brain (cerebellum, striatum, and 4 cortical sub-regions)[42]. Long nanopore reads enabled the complete exonic structure of *CACNA1C* (6.5 kb) to be defined, allowing the identification of 38 putative novel exons and 90 isoforms[50]. Out of 40 previously annotated isoforms, only 7 were identified. The team suggests that one of the reasons for this discrepancy is the presence of false positives in the previous data, caused by the inherent complexities of assembling transcript sequences generated using alternative sequencing approaches.

Interestingly, 9 of the top 10 expressed isoforms were novel transcripts (**Figure 9**)[50]. Initial investigation of the 83 novel isoforms identified many that would alter the protein and its function. The team were also able to show that isoform profiles vary between brain regions and that the splicing of *CACNA1C* is more complex than currently appreciated[50].

The lead researcher, Dr. Michael Clark, highlighted that: *'These results are the first step in evaluating* CACNA1C *as a potential therapeutic target and demonstrate the power of long-read nanopore sequencing to elucidate the true nature of expressed genes'*[42].

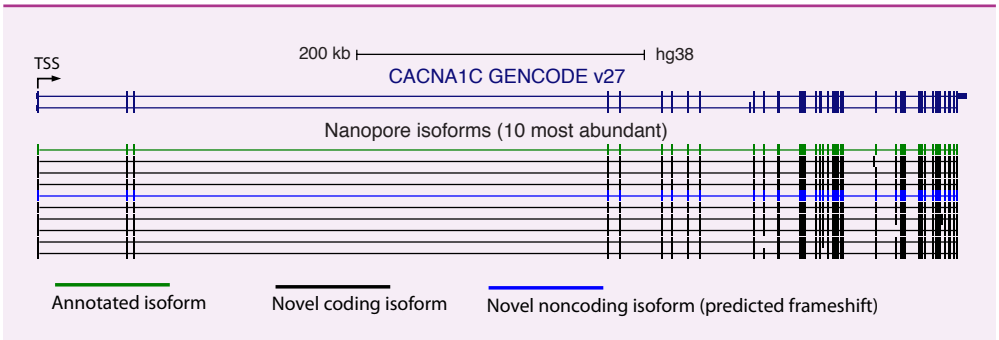


Figure 9
Nine of the top ten most abundant *CACNA1C* isoforms identified in brain tissue were novel, highlighting the benefits of long reads for isoform characterisation. Image courtesy of Dr. Michael Clark, University of Oxford, UK.

CASE STUDY 5

Accurate quantification of full-length cDNA molecules

Oikonomopoulos *et al.*[13] were interested in studying how nanopore technology compares to other sequencing methods used for cDNA-based transcriptomic research. The team used the External RNA Controls Consortium (ERCC) RNA Spike-In mix and a cDNA library prepared from human HEK-293 cells for comparing performance of different sequencing machines. A template-switching cDNA protocol was used for all tested platforms. Most (90.8%) of the nanopore reads were full-length and no transcript length bias was observed.

There was a very strong correlation between nanopore data-based estimates of transcript abundance and the known concentration of the ERCC mix (Pearson r=0.98). There was also a strong correlation between short-read data generated transcript abundance estimates and nanopore data-based estimates (Pearson r=0.75) for the HEK-293 sample. For the same sample, nanopore data abundance estimates were also in strong agreement with estimates based on another long-read capable sequencing platform data (Pearson r=0.82). These and other results[22] suggest that nanopore technology can be used to reliably quantify full-length cDNA molecules of any size offering the potential to reduce the analysis burden.

CASE STUDY 6

Transcriptional landscape analysis using direct RNA sequencing

The *Saccharomyces cerevisiae* strain CEN.PK113-7D is used extensively in academic and industrial research due to a combination of ease of genetic manipulation and a fast growth rate. In order to provide detailed insight into this organism, researchers at the University of Arkansas first utilised nanopore sequencing to create a complete genome assembly, using long reads to improve upon the existing, fragmented short-read based assembly[51]. The team then performed direct RNA sequencing to determine gene expression patterns under two different growth conditions (diauxic growth). Examining the data revealed distinctive gene expression profiles for yeast replicates grown on glucose versus ethanol (**Figure 10**). As expected, organisms using ethanol as a growth substrate had significant increases in transcripts related to ethanol metabolism and cellular stresses, while up-regulated genes in the glucose group were associated with rapid growth processes and ethanol fermentation.

The team found that approximately 70% of the reads provided by direct RNA sequencing corresponded to full-length transcripts and, in addition, full-length transcripts over 5 kb could be detected. A comparison of direct RNA sequencing using nanopore technology with data obtained using a short-read sequencing technology revealed that, while the total amount of data was approximately half of that achieved using short reads, the mean coverage depth was comparable (**Figure 11**). Furthermore, there was evidence to suggest that the nanopore reads exhibited lower GC-bias than short-read RNA analysis technology[52].

Direct RNA sequencing also allowed the identification of many polyadenylated non-coding RNAs, including rRNA, telomerase RNA, and long non-coding RNA (lncRNA)[51].

Summarising, the team stated: *'We believe that direct RNA sequencing will become a versatile tool for transcriptome analysis in the "complete genome era" of the future'.*

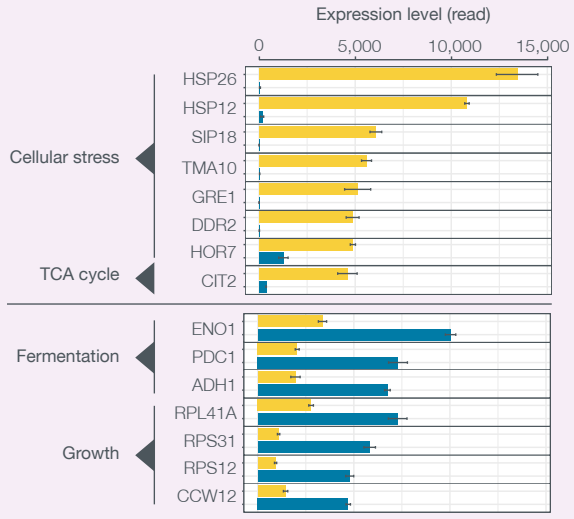


Figure 10
Direct RNA sequencing allowed the identification of distinctive gene expression profiles for yeast replicates grown on glucose (blue) or ethanol (yellow). Image courtesy of Dr. Intawat Nookaew, University of Arkansas, USA.

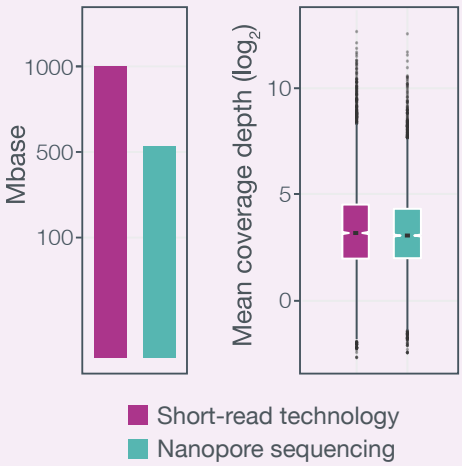


Figure 11
The dynamic range of transcripts generated from 500 Mb of direct RNA sequencing data was similar to that obtained from 1,000 Mb of short-read sequencing technology data as shown by the mean coverage depth box plot. Image courtesy of Dr. Intawat Nookaew, University of Arkansas, USA.

Summary

RNA sequencing with traditional methods has helped answer a variety of biological questions. However, the approach suffers from limitations such as a restricted capability to obtain full-length transcripts, problematic isoform identification and quantification, and various library-preparation-associated biases (PCR, reverse transcription, etc.). Some of these issues are addressed by other commercially available long-read capable sequencing machines, but their use for large-scale experiments is often cost-prohibitive.

Nanopore sequencing produces full-length transcripts suitable for characterisation of isoforms and their abundance.

Nanopore sequencing offers all the transcriptomics analysis advantages bestowed by long reads. The technology produces full-length transcripts suitable for accurate characterisation of isoforms and their abundance. Portability, rapid library-preparation protocols, and real-time analysis pipelines also make nanopore sequencing advantageous in the area of viral pathogen investigation, particularly in a field setting[6]. The advent of direct RNA sequencing further boosts transcriptomic analysis, since biases introduced by reverse transcription and PCR amplification are not applicable to this method. Finally, direct RNA sequencing also enables the identification of multiple RNA modifications[21] from single library preps, without the need for lengthy procedures like bisulfite conversion or antibody precipitation. These features, combined with the scalability offered by nanopore sequencing devices, make the technology, to quote Dr. Marinov from Indiana University Bloomington, *'the most exciting development in transcriptomic research'*[53].

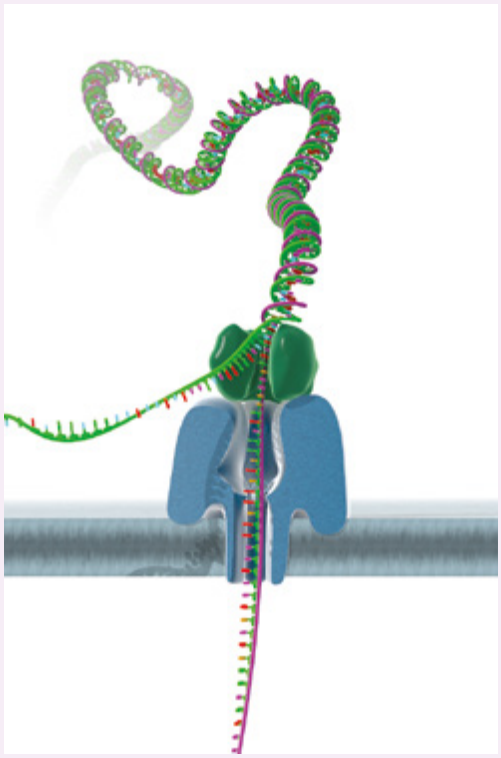


Figure 12
An RNA molecule passing through the nanopore. The optional cDNA first strand provides maximum sequencing efficiency, but does not pass through the pore.

About Oxford Nanopore Technologies

Oxford Nanopore Technologies has revolutionised sequencing by introducing the first nanopore sequencing device, the MinION — a portable, real-time, low-cost device, capable of generating any length of sequencing read.

Nanopore technology is the only available sequencing technology that can sequence RNA directly, removing potential sources of bias and allowing identification of base modifications alongside nucleotide sequence. It is also increasingly being used for the analysis of cDNA, where long read lengths enable sequencing of full-length transcripts (**Table 2**).

MinION Mk1C and Flongle are the most recent additions to the nanopore sequencing portfolio. MinION Mk1C combines the portability and high yields of MinION with a fully integrated touchscreen and high performance compute for on-board analysis. Flongle is a flow cell adapter for MinION and GridION designed to offer the most cost-effective approach for smaller sequencing experiments and assays.

For higher throughput requirements, the GridION Mk1, PromethION 24, and PromethION 48 are available. These compact benchtop systems utilise the same nanopore technology as the MinION, offering up to 5, 24, and 48 flow cells respectively. Each flow cell can be used independently, with the researcher choosing how many are used at any one time, enabling different experiments to be run in parallel.

For the latest information about full-length RNA sequencing without bias, visit www.nanoporetech.com/RNA

Table 2
An overview of the Oxford Nanopore RNA sequencing kits by application.

Applications	PCR-cDNA Sequencing Kit SQK-PCS109	Direct-cDNA Sequencing Kit SQK-DCS109	Direct RNA Sequencing Kit SQK-RNA002
Gene expression	***	***	*
Splice variants and fusion proteins	***	***	**
Viral RNA	***	***	***
Sequencing native RNA allowing RNA modifications to be explored			***

References

1. Beyer, K. et al. Differential expression of alpha-synuclein, parkin, and synphilin-1 isoforms in Lewy body disease. Neurogenetics 9, 163-172 (2008).
2. Gonzalez, E. & McGraw, T. E. The Akt kinases: isoform specificity in metabolism and cancer. Cell Cycle 8, 2502-2508 (2009).
3. Bao, Z.-S. et al. RNA sequencing of 272 gliomas revealed a novel, recurrent PTPRZ1-MET fusion transcript in secondary glioblastomas. Genome Res. 24, 1765-1773 (2014)
4. Quick, J. et al. Real-time, portable genome sequencing for Ebola surveillance. Nature 530, 228-232 (2016).
5. Greninger, A. L. et al. Rapid metagenomic identification of viral pathogens in clinical samples by real-time nanopore sequencing analysis. Genome Med. 7, 99, doi:10.1186/s13073-015-0220-9 (2015).
6. Batovska, J., Lynch, S.E., Rodoni, B.C., Sawbridge, T.I., & Cogan, N.O. Metagenomic arbovirus detection using MinION nanopore sequencing. J. Virol. Methods 249:79-84 (2017).
7. Broadhurst, M. J., Brooks, T. J. & Pollock, N. R. Diagnosis of ebola virus disease: Past, present, and future. Clin. Microbiol. Rev. 29(4):773-93 (2016).
8. Fracasso, A., Trindade, L. M. & Amaducci, S. Drought stress tolerance strategies revealed by RNA sequencing in two sorghum genotypes with contrasting WUE. BMC Plant Biology 16, 115 (2016).
9. Trapnell, C. et al. Transcript assembly and quantification by RNA sequencing reveals unannotated transcripts and isoform switching during cell differentiation. Nat. Biotechnol. 28, 511-515 (2010).
10. Kazuo, M. & Sumio, S. Oligo-capping: a simple method to replace the cap structure of eukaryotic mRNAs with oligoribonucleotides. Gene 138, 171-174 (1994).

11. Finotello, F. et al. Reducing bias in RNA sequencing data: a novel approach to compute counts. BMC Bioinformatics 15, S7 (2014).

12. Zhu, Y., Machleder, E., Chenchik, A., Li, R. & Siebert, P. Reverse transcriptase template switching: A SMART™ approach for full-length cDNA library construction. Biotechniques 30, 892-897 (2001).

13. Oikonomopoulos, S., Wang, Y. C., Djambazian, H., Badescu, D. & Ragoussis, J. Benchmarking of the Oxford Nanopore MinION sequencing for quantitative and qualitative assessment of cDNA populations. Sci. Rep. 6 (2016).

14. Hargreaves, A. D. & Mulley, J. F. Assessing the utility of the Oxford Nanopore MinION for snake venom gland cDNA sequencing. PeerJ 3: e1441 (2015).

15. Bolisetty, M. T., Rajadinakaran, G. & Graveley, B. R. Determining exon connectivity in complex mRNAs by nanopore sequencing. Genome Biol. 16, 204 (2015).

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

17. Martin, J. A. & Wang, Z. Next-generation transcriptome assembly. Nat. Rev. Genet. 12, 671-682 (2011).

18. Weirather, J. L. et al. Comprehensive comparison of Pacific Biosciences and Oxford Nanopore Technologies and their applications to transcriptome analysis. F1000Research 6 (2017).

19. Hardwick, S.A. Targeted, high-resolution RNA sequencing of non-coding genomic regions associated with neuropsychiatric functions. Front. Genet. 10:309 (2019).

20. Byrne, A. et al. Nanopore long-read RNAseq reveals widespread transcriptional variation among the surface receptors of individual B cells. Nat. Commun. 8 (2017).

21. Garalde, D. R. et al. Highly parallel direct RNA sequencing on an array of nanopores. Nat. Methods 15(3):201-206 (2018).

22. Oxford Nanopore Technologies. Quantitative RNA-seq: PCR-cDNA, PCR-free Direct cDNA and Direct RNA sequencing. Poster. Available at: <https://nanoporetech.com/resource-centre/quantitative-rna-seq-pcr-cdna-pcr-free-direct-cdna-and-direct-rna-sequencing > [Accessed: 22 October 2019]

23. 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]

24. Jain, M. Personal communication with Oxford Nanopore Technologies on 27 February 2018.

25. Liu, H. et al. Accurate detection of m6A RNA modifications in native RNA sequences. Nat. Commun. 10(1):4079 (2019).

26. Viehweger, A. Direct RNA nanopore sequencing of full-length coronavirus genomes provides novel insights into structural variants and enables modification analysis. Genome Res. 29(9):1545-1554 (2019).

27. Jalkanen, A.L., Coleman, S.J. & Wilusz, J. Determinants and implications of mRNA poly(A) tail size - Does this protein make my tail look big? Semin. Cell Dev. Biol. 34: 24–32 (2014).

28. Roach, N.P. et al. The full-length transcriptome of C. elegans using direct RNA sequencing. bioRxiv 598763 (2019).

29. Workman, R.E. Nanopore native RNA sequencing of a human poly(A) transcriptome. bioRxiv 459529 (2018).

30. Keller, M. Direct RNA sequencing of Influenza viral RNA using the MinION nanopore sequencer. Presentation. Available at: <https://nanoporetech.com/resource-centre/direct-rna-sequencing-influenza-viral-rna-using-minion-nanopore-sequencer> [Accessed: 22 October 2019]

31. Consortium, E. P. An integrated encyclopedia of DNA elements in the human genome. Nature 489, 57-74 (2012).

32. Sloan, C. A. et al. ENCODE data at the ENCODE portal. Nucleic Acids Research 44, D726-D732 (2016).

33. Liu, Y. et al. Evaluating the impact of sequencing depth on transcriptome profiling in human adipose. PLoS One 8, e66883 (2013).

34. Liu, Y., Zhou, J. & White, K. P. RNA sequencing differential expression studies: more sequence or more replication? Bioinformatics 30, 301-304 (2014).

35. Chiu, C. Personal communication with Oxford Nanopore Technologies on 1 December 2017.

36. GitHub. Pipeline-transcriptome-de. Available at: <https://github.com/nanoporetech/pipeline-transcriptome-de [Accessed: 03 October 2019]

37. Love, M.I., Soneson, C. and Patro, R. Swimming downstream: statistical analysis of differential transcript usage following Salmon quantification. F1000Research 7:952 (2018).

38. GitHub. Pinfish. Available at: <https://github.com/nanoporetech/pinfish> [Accessed: 03 October 2019]

39. GitHub. Minimap2. Available at: <https://github.com/lh3/minimap2> [Accessed: 03 October 2019]

40. GitHub. Pychopper. Available at: < https://github.com/nanoporetech/pychopper> [Accessed: 03 October 2019]

41. GitHub. Tombo. Available at: <https://nanoporetech.github.io/tombo> [Accessed: 22 October 2019]

42. Clark, M. Elucidating the expression and splicing patterns of neuropsychiatric disease genes in human brain. Presentation. Available at: <https://nanoporetech.com/resource-centre/elucidating-expression-and-splicing-patterns-neuropsychiatric-disease-genes-human> [Accessed: 22 October 2019]

43. Schmucker, D. et al. Drosophila Dscam is an axon guidance receptor exhibiting extraordinary molecular diversity. Cell 101(6):671-84 (2000).

44. Bayega, A. et al. Transcriptome landscape of the developing olive fruit fly embryo delineated by Oxford Nanopore long-read RNA-seq. bioRxiv 478172 (2018).

45. Bayega, A. Transcriptome landscape of the developing olive fruit fly embryo delineated by Oxford Nanopore long-read RNA-seq. Presentation. Available at: <https://nanoporetech.com/resource-centre/anthony-bayega-transcriptome-landscape-developing-olive-fruit-fly-embryo-delineated> [Accessed: 21 October 2019]

46. Minervini, C.F. et al. Mutational analysis in BCR-ABL1 positive leukemia by deep sequencing based on nanopore MinION technology. Exp. Mol. Pathol. 103(1):33-37 (2017).

47. de Jong, L.C. et al. Nanopore sequencing of full-length BRCA1 mRNA transcripts reveals co-occurrence of known exon skipping events. Breast Cancer Res. 19:127 (2017).

48. Suzuki, A. et al. Sequencing and phasing cancer mutations in lung cancers using a long-read portable sequencer. DNA Res. 24(6): 585–596 (2017).

49. Moore, F.R., Rempfer, C.B., and Press, R.D. Quantitative BCR-ABL1 RQ-PCR fusion transcript monitoring in chronic myelogenous leukemia. Methods Mol. Biol. 999:1-23 (2013).

50. Clark, M.B. et al. Long-read sequencing reveals the splicing profile of the calcium channel gene CACNA1C in human brain. bioRxiv 260562 (2018).

51. Jenjaroenpun, P. et al. Complete genomic and transcriptional landscape analysis using third-generation sequencing: a case study of *Saccharomyces cerevisiae* CEN.PK113-7D. Nucleic Acids Res. 6(7):e38 (2018).

52. Nookaew, I. Transcriptional landscapes analysis through direct RNA sequencing. Presentation. Available at: <https://nanoporetech.com/resource-centre/transcriptional-landscapes-analysis-through-direct-rna-sequencing> [Accessed: 21 October 2019]

53. Marinov, G. K. On the design and prospects of direct RNA sequencing. Brief. Funct. Genomics 16(6):326-335 (2017).

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

WP_1047(EN)_V2_23Aug2021