Applications and potentials of nanopore sequencing in the (epi)genome and (epi)transcriptome era

Shangqian Xie,1,8 Amy Wing-Sze Leung,2,8 Zhenxian Zheng,2 Dake Zhang,2 Chuanle Xiao,4,8 Ruibang Luo,2,8 Ming Luo,5,8 and Shoudong Zhang6,7,8

*Correspondence: xiaochuanle@126.com (C.X.); rbluo@cs.hku.hk (R.L.); luoming@scbg.ac.cn (M.L.); shoudongzhang@cuhk.edu.hk (S.Z.)

Received: June 9, 2021; Accepted: August 9, 2021; Published Online: August 11, 2021; https://doi.org/10.1016/j.xinn.2021.100153

© 2021 The Author(s). This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

Graphical abstract

Public summary
- Nanopore-seq can dissect native DNA/RNA molecules from any organisms at unlimited length
- A wide variety of algorithms greatly increase the accuracy of signal decoding in Nanopore-Seq
- Nanopore-Seq significantly facilitates genome assembly and structural variant calling, and can simultaneously detect base modifications
- These advantages ensure its great potentials in future medical and agricultural practices
Applications and potentials of nanopore sequencing in the (epi)genome and (epi)transcriptome era

Shangqian Xie, Amy Wing-Sze Leung, Zhenxian Zheng, Dake Zhang, Chuanle Xiao, Ruibang Luo, Ming Luo, and Shoudong Zhang

1Key Laboratory of Ministry of Education for Genetics and Germplasm Innovation of Tropical Special Trees and Ornamental Plants, College of Forestry, Hainan University, Haikou 570228, China
2Department of Computer Science, The University of Hong Kong, Hong Kong 999077, China
3Beijing Advanced Innovation Centre for Biomedical Engineering, Key Laboratory for Biomechanics and Mechanobiology of Ministry of Education, School of Biological Science and Medical Engineering, Beihang University, Beijing 100083, China
4State Key Laboratory of Ophthalmology, Zhongshan Ophthalmic Centre, Sun Yat-sen University, Guangzhou 510060, China
5Agriculture and Biotechnology Research Center, Guangdong Provincial Key Laboratory of Applied Botany, Center of Economic Botany, Core Botanical Gardens, South China Botanical Garden, Chinese Academy of Sciences, Guangzhou 510650, China
6School of Life Sciences, The Chinese University of Hong Kong, Shatin, Hong Kong 999077, China
7Center for Soybean Research of the State Key Laboratory of Agrobiotechnology, The Chinese University of Hong Kong, Shatin, Hong Kong 999077, China
8These authors contributed equally

Received: June 9, 2021; Accepted: August 9, 2021; Published Online: August 11, 2021; https://doi.org/10.1016/j.xinn.2021.100153
© 2021 The Author(s). This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).
Citation: Xie S., Leung A.W.-S., Zheng Z., et al., (2021). Applications and potentials of nanopore sequencing in the (epi)genome and (epi)transcriptome era. The Innovation 2(4), 100153.

The applications of DNA and RNA sequencing have greatly promoted research in the life sciences and catalyzed biological research into the genomic and post-genomic era. The success of the Human Genome Project (HGP) has promoted the development of sequencing technologies and their large-scale application. The HGP started in 1990 and its draft genome was published in 2001, and finally completed and reported in 2004. It initiated the flourishing period of next-generation sequencing (NGS), in which diverse platforms sprang up, from short-read sequencing to long-read sequencing. In 2004, Pacific Biosciences was founded, focusing on single-molecule real-time sequencing. The following year, Oxford Nanopore was founded. Nanopore sequencing provides single-nucleotide detection and analytical capabilities that are achieved by electrophoretically driving molecules in solution through a nano-scale pore. Nevertheless, the concept behind nanopore sequencing can be traced back to 1989, much earlier than the NGS technologies. However, its first commercial products were not released until 2014. Demanding features of the sequencing methodology hindered its refinement, which slowed down its wide application. Currently, related algorithms and tools have been developed and released, and the use of nanopore sequencing is becoming increasingly frequent. Its complexity, the diversity of associated algorithms and tools, and the requirement for high-purity nucleic acids for library preparation may confuse potential users lacking sufficient background information. In the review, we focus on issues related to nanopore direct DNA sequencing and direct RNA sequencing, so that readers can understand how nanopore sequencing works and how to use the new technology to serve their interests.

A brief history of nanopore sequencing

The development of nanopore sequencing may have also been a byproduct of the HGP, which was first proposed in 1985 but was finally funded in 1990. During this period, many scientists began to imagine effective methods for sequencing human genomic DNA in its entirety. One such scientist was Dr. David Deamer, who first proposed an idea to sequence DNA with membrane electrophoresis in 1989 while vacationing in Oregon. The idea he jotted in his notebook at the time is very similar to the nanopore sequencing technology in current use. At the time, his lab was focusing on lipochemistry. Thus, the need for, and ability of, cells to pass specific molecules through membranes may have influenced his approach to sequencing, maybe the HGP budget authorized in 1988 made him think about the possibility. While Kasiyanowicz, who was working on DNA sequencing with protein-made pores, was invited to collaborate on the idea Deamer jotted in his notebook and published the seminal paper in PNAS in 1996, where they resorted to monitor alterations in ion current blockades to detect the sequence of bases in single molecules of DNA or RNA when they pass through a minute opening—a nanopore—in a lipid bilayer membrane. Afterward, Kasiyanowicz returned to focusing on dissecting DNA sequence through protein-comprised nanopores embedded in a lipid bilayer membrane independent of his collaboration with Deamer on that paper. Meanwhile, George Church was struggling to scale up the sequencing throughput of the HPG, and thus proposed a similar idea to sequence double-stranded DNA via electronically monitoring phage packing motors embedded in a lipid bilayer. Hagan Bayley was first to artificially synthesize hemolysin nanopores in a lipid bilayer.
membrane to allow single-strand DNA or RNA molecules to pass through under certain voltage, the ion current signals produced from bases going through the pores could be observed, thus after he moved to Oxford University, he co-founded the Oxford Nanopore company in 2005. At first, the major challenge for nanopore sequencing was to control the speed of translocation from the cis- to the trans-chambers, which are separated by a lipid bilayer membrane. Finally, in 2010, phi29 DNA polymerase was found to solve this issue by effectively regulating the translocation speed of single-strand DNA.30,31 Meanwhile, the alpha-hemolysin pore was found not to be the best choice for discerning DNA nucleotides.4 Later, MspA, a Mycobacterium smegmatis porin, was found to be more suitable as the protein for nanopore, with a funnel-like geometry that narrowed to an \( \sim 1.2\)-nm-diameter and a \( \leq 0.6\)-nm-long tube, which is optimal for single-strand nucleic acids.32 The coupling of MspA and phi29 DNA polymerase showed significantly increased performance in the sequencing of the phi X174 genome.33 However, the porin used by Oxford Nanopore for commercial flow cells is CsgG,34 since the height (85 Å) of the channel made of CsgG35 even can set two detectors along the channel to monitor the ion current signals. In 2014, some 18 years after Deamer’s initial inspiration, commercial flowcell R7.3 was finally released to the genomics community.4 Figure 1 gives an overview of how nanopore sequencing works.

**Tools and algorithms developed for nanopore sequencing**

Because nanopore sequencing detects alterations in ion current when different bases pass through protein nanometer-scale pores at a specified voltage, the device needs to be sensitive enough to discern signals reflecting different DNA nucleotides, with or without modification. In addition, DNA fragments pass through nanopores at high speed, and the device can only record signals for 5 to 6 bases as a pulse. The deduction of correct DNA bases within each pulse increases the difficulty in data analysis. Subsequently, diverse algorithms have been developed to recognize the sequenced bases; however, their accuracy was still vastly inferior to that achieved by the base-calling method of sequencing by synthesis. Although deep-learning algorithms show better performance, the relatively high error rate requires additional algorithms specific for downstream applications and analysis. Therefore, we first review the tools and algorithms developed specifically for nanopore sequencing.

**Development of base-calling algorithms.** Initially, Hidden Markov Model (HMM) and Viterbi decoding approaches, such as Metrichor and Nanocall, were exclusively adopted for base-calling, albeit unfortunately with low accuracy. In current years, deep neural networks have been widely implemented in base-calling tools, including convolutional neural networks, recurrent neural networks (RNN), and connectionist temporal classification (CTC) decoders, which greatly improved base-calling accuracy to more than 98% for DNA sequencing from less than 80% HMM algorithm called.36

For HMM and Viterbi decoding approaches, Timp et al. (2012)37 demonstrated the method’s potential for usage in the real world, because decoding 3-base pair (bp) resolution nanopore electrical measurements into a DNA sequence using an HMM can reach around 98% accuracy. However, when the flowcell R7.3 was released, each documented ion current signal comprising 5 to 6 bases, nearly two times longer than the 3-bp length, made the algorithm dramatically decrease its accuracy. To find a better resolution for base-calling, Teng et al.38 reported an algorithm called Chiron, which was the first deep-learning model to achieve end-to-end base-calling, and not necessary to segment the ion current signals from a sequenced nucleic acid molecule to detect the corresponding bases. This means it directly translates the raw current signal to a DNA sequence without the error-prone segmentation step, which may avoid the error caused by segmentation, but may increase computing time. To further improve the quality of base-calling, Zeng et al.39 presented Causalcall, an end-to-end temporal convolution-based deep-learning model for accurate and fast nanopore base-calling. It directly identifies base sequences of varying lengths from current measurements in long time series, but still cannot decrease artificial deletion caused by low ion current signals. Zhang et al.40 also proposed a refined U-net model (thus, a UR-net model, an enhanced U-net model for 1D sequence segmentation) called URnano to improve previous end-to-end deep-learning models. Early neural-network base-callers (such as DeepNano41 and BasecRAWller42) relied on a preprocessing step that segmented the current measurements into discrete events that may artificially introduce errors when segmenting the ion current signals. To avoid such flaws, Konishi et al.43 developed an improved base-caller, Halcyon, which utilizes an encoder-decoder model incorporating neural-network techniques to improve base-calling accuracy, but may increase computing time.

Although the Nanopore company developed Guppy44 and Bonito base-callers can make the accuracy of DNA nanopore sequencing reach more than 99%, the third parties still have been attempting to develop more accurate algorithms for the nanopore sequencing community. One example comes from a trial that embedded a chip with a complementary metal-oxide-semiconductor (CMOS) base-caller alongside nanopore sensors to predict the molecule’s base-pair constitution.45 Another test46 is to use an accelerator consisting of a low-power real-time field-programmable gate array (FPGA) integrated circuit for the base-calling task.46 Although these algorithms currently are not better than Nanopore company-developed base-callers, the different strategies used to develop new algorithms may benefit nanopore base-calling in the future.

For nanopore sequencing, not only can a single strand of a genomic region be sequenced, but also both strands in a genomic region can be sequenced via a hairpin adapter. Single-strand sequencing is called 1D sequencing, while double-strand sequencing was previously called 2D sequencing, because of a dispute with the Pacific Biosciences company, now the 2D sequencing is called 1D\(^2\) (1D squared). 1D\(^2\) reads reflect consensus sequence generated from the base-calling of both forward- and reverse-strand DNA connected with a hairpin adapter. Currently, 1D\(^2\) reads cannot be base-called with common open-source algorithms, e.g. Nanocall. Although Nanocall has lower base-calling accuracy (\(\sim 68\)%), it supports offline base-calling for Oxford Nanopore sequencing data.47 The offline base-calling may help users to
The Innovation matches are chosen based on their rareness traditional alignment algorithms to process such long-reads. Unlike seed-and-extend algorithms in mRNA reads for direct RNA sequencing (DRS).49,50 It requires new alignment

98.1%.36 Table 1 shows the tools and algorithms developed for base-calling.

Tools and algorithms developed for alignment. Recent advances in nanopore sequencing technologies promise ultra-long-reads with N50 >100 kb and read lengths up to 882 kb,48 which enables production of full-length mRNA reads for direct RNA sequencing (DRS).49,50 It requires new alignment algorithms to process such long-reads. Unlike seed-and-extend algorithms in traditional alignment—e.g., BLAST and LAST use adaptive seeds, whose matches are chosen based on their rareness—these tools use fixed-length matches, thereby guaranteeing the number of matches of a given sequence length, which achieves fast and sensitive comparison of ultra-long sequences.51 Nevertheless, the runtime of LAST will increase linearly with sequence length. In addition, for those quite short DNA reads, LAST may also take a longer time than those with slightly long-reads for the same initial data. GraphMap is designed to handle high-error Nanopore reads, which progressively refine candidate alignments with a fast graph traversal to align long-reads with speed and high precision (>95%), but may need more memory resources.52 Currently, Minimap2 is a popular tool for long-read alignment. It does split-read alignment (aka chimeric read alignment), employs concave gap cost for long insertions and deletions, and introduces new heuristics to reduce spurious alignments, making the tool 30 times faster than other long-read genomic or cDNA mappers, while obtaining higher accuracy.53

To quickly find unknown DNA fragments in metagenomic sequencing data, ONT devices provide a unique solution for real-time targeted sequencing known as Readuntil. It allows nanopore devices to selectively eject reads from pores in real time via real-time alignment while sequencing (bear in mind that Nanopore reads are typically quite long). By precluding known genomic sequences, could result in the sequencing of purely unknown genomic sequences. Kovaka et al.54 developed the UNCALLED tool, an open-source

| Tool       | Description          | Algorithm                  | Advantages                                | Rate      | Disadvantages                  | Link                                      | Reference (PMID) |
|------------|----------------------|----------------------------|------------------------------------------|-----------|-------------------------------|------------------------------------------|-----------------|
| Chiron     | Basecalling          | deep learning              | no segmentation                          | 2000 (bp/s)| Not suitable for large genomes| https://github.com/haotianteng/chiron    | 29648610        |
| Causcall   | Basecalling          | Temporal Convolutional Network | directly identifies base sequences of varying lengths | 7000 (bp/s)| base deletions                | https://github.com/scubioinformatics/causcall | 32038706        |
| URnano     | Basecalling          | deep neural networks       | model sequential dependencies for a one-dimensional segmentation task | 3600 (bp/s)| segmentation                   | https://github.com/yaozhong/URnano       | 32321433        |
| DeepNano   | Basecalling          | Deep recurrent neural networks | open-source                           | 1250 (bp/s)| Not suitable for large genomes| https://github.com/jeamim/deepnano       | 28582401        |
| BasecRAWller | Basecalling         | unidirectional recurrent neural networks | 1) streaming basecalling, 2) tunable ratio of insertions to deletions, and 3) potential for streaming detection of modified bases | 200 (bp/s)| non-detectable covalently modified bases | https://github.com/rrwick/Basecalling-comparison |                |
| Halcyon    | Basecalling          | Convolutional Neural Network and recurrent neural network | no segmentation and semantic correspondence | 250 (bp/s)| decrease speed                  | https://github.com/relastle/halcyon       | 33165508        |
| CMOS       | nanopore sensors    | complementary metal-oxide-semiconductor | -                                         | -         | -                             | -                                          | 28269559        |
| FPGA       | nanopore sensors    | Field-programmable gate array | -                                         | -         | -                             | -                                          | 31825872        |
| Nanocall   | Basecalling          | Hidden Markov Model         | offline, free and private                | 700 (bp/s)| not currently integrate '2D' read | https://github.com/matedavid/nanocall     | 27614348        |
| Guppy      | Basecalling          | taxon-specific dataset and neural network model | reduction of errors in methylation motifs and no segmentation | 120,000 (bp/s)| a custom model using a larger neural network and/or training data from the same species | https://community.nanoporetech.com        | 31234903        |
| PoreOver   | Basecalling          | CTC-trained neural network and hidden Markov models | compatible with multiple nanopore basecallers | 450 (bp/s)| not currently integrate '2D' read | https://github.com/jordisr/poreover        | 33468205        |
source mapper that probabilistically considers k-mers that could be represented by the signal, and then prunes the candidates based on the reference encoded within a Ferragina-Manzini index. The UNCALLED alignment tool can selectively sequence and analyze targeted genomic data in metagenomic samples, also have structural variants or DNA modification information, which would be lost with conventional targeted sequencing methods. Tools and algorithms developed for genome assembly. Genome assembly is one of the most important research topics in genome biology. Long-read sequencing can significantly improve the assembly accuracy of assembled genomes.55–56 It will also play an important role in solving the assembly of (peri)centromeric genomic regions full of transposable elements (TEs) and centromeric satellites.58–60 For instance, with N50 read length around 100 Kb, ONT DDS super-long-reads can readthrough genomic repeat regions unsolvable in short-read sequencing.48,56–58 Long-reads produced from third-generation sequencing platforms, e.g., Pacific Bioscience and Oxford Nanopore Technologies, are not only long but also have relatively high error rates, which require a genome assembly strategy that differs from previous algorithms.61 However, most of at least 97 assemblers (https://bioinformaticshome.com/tools/wga/wga.html) are only suitable for NGS short-reads. By far, the relatively low base-calling accuracy of ONT DDS still needs improved algorithms and elegant strategies.63–64 Moreover, the error distribution of nanopore sequencing is more complicated than that of PacBio sequencing. The higher-error sub-sequences from ONT DDS reads are widely distributed, and sometimes even reached a 50% error rate in 1,000-bp length reads with earlier base-calling sequences from ONT DDS reads are widely distributed, and sometimes even more complicated than that of PacBio sequencing. The higher-error sub-sequences from ONT DDS reads are widely distributed, and sometimes even reached a 50% error rate in 1,000-bp length reads with earlier base-calling sequences from ONT DDS reads are widely distributed, and sometimes even reached a 50% error rate in 1,000-bp length reads with earlier base-calling sequences from ONT DDS reads are widely distributed, and sometimes even reached a 50% error rate in 1,000-bp length reads with earlier base-calling sequences from ONT DDS reads are widely distributed, and sometimes even reached a 50% error rate in 1,000-bp length reads with earlier base-calling sequences from ONT DDS reads are widely distributed, and sometimes even reached a 50% error rate in 1,000-bp length reads with earlier base-calling sequences from ONT DDS reads are widely distributed, and sometimes even reached a 50% error rate in 1,000-bp length reads with earlier base-calling sequences from ONT DDS reads are widely distributed, and sometimes even reached a 50% error rate in 1,000-bp length reads with earlier base-calling sequences from ONT DDS reads are widely distributed, and sometimes even reached a 50% error rate in 1,000-bp length reads with earlier base-calling sequences from ONT DDS reads are widely distributed, and sometimes even reached a 50% error rate in 1,000-bp length reads with earlier base-calling sequences from ONT DDS reads are widely distributed, and sometimes even reached a 50% error rate in 1,000-bp length reads with earlier base-calling sequences from ONT DDS reads are widely distributed, and sometimes even reached a 50% error rate in 1,000-bp length reads with earlier base-calling sequences from ONT DDS reads are widely distributed, and sometimes even reached a 50% error rate in 1,000-bp length reads with earlier base-calling sequences from ONT DDS reads are widely distributed, and sometimes even reached a 50% error rate in 1,000-bp length reads with earlier base-calling sequences from ONT DDS reads are widely distributed, and sometimes even reached a 50% error rate in 1,000-bp length reads with earlier base-calling sequences from ONT DDS reads are widely distributed, and sometimes even reached a 50% error rate in 1,000-bp length reads with earlier base-calling sequences from ONT DDS reads are widely distributed, and sometimes even reached a 50% error rate in 1,000-bp length reads with earlier base-calling sequences from ONT DDS reads are widely distributed, and sometimes even reached a 50% error rate in 1,000-bp length reads with earlier base-calling sequences from ONT DDS reads are widely distributed, and sometimes even reached a 50% error rate in 1,000-bp length reads with earlier base-calling sequences from ONT DDS reads are widely distributed, and sometimes even reached a 50% error rate in 1,000-bp length reads with earlier base-calling sequences from ONT DDS reads are widely distributed, and sometimes even reached a 50% error rate in 1,000-bp length reads with earlier base-calling sequences from ONT DDS reads are widely distributed, and sometimes even reached a 50% error rate in 1,000-bp length reads with earlier base-calling sequences from ONT DDS reads are widely distributed, and sometimes even reached a 50% error rate in 1,000-bp length reads with earlier base-calling sequences from ONT DDS reads are widely distributed, and sometimes even reached a 50% error rate in 1,000-bp length reads with earlier base-calling sequences from ONT DDS reads are widely distributed, and sometimes even reached a 50% error rate in 1,000-bp length reads with earlier base-calling sequences from ONT DDS reads are widely distributed, and sometimes even reached a 50% error rate in 1,000-bp length reads with earlier base-calling sequences from ONT DDS reads are widely distributed, and sometimes even reached a 50% error rate in 1,000-bp length reads with earlier base-calling sequences from ONT DDS reads are widely distributed, and sometimes even reached a 50% error rate in 1,000-bp length reads with earlier base-calling sequences from ONT DDS reads are widely distributed, and sometimes even reached a 50% error rate in 1,000-bp length reads with earlier base-calling sequences from ONT DDS reads are widely distributed, and sometimes even reached a 50% error rate in 1,000-bp length reads with earlier base-calling sequences from ONT DDS reads are widely distributed, and sometimes even reached a 50% error rate in 1,000-bp length reads with earlier base-calling sequences from ONT DDS reads are widely distributed, and sometimes even reached a 50% error rate in 1,000-bp length reads with earlier base-calling sequences from ONT DDS reads are widely distributed, and sometimes even reached a 50% error rate in 1,000-bp length reads with earlier base-calling sequences from ONT DDS reads are widely distributed, and sometimes even reached a 50% error rate in 1,000-bp length reads with earlier base-calling sequences from ONT DDS reads are widely distributed, and sometimes even reached a 50% error rate in 1,000-bp length reads with earlier base-calling sequences from ONT DDS reads are widely distributed, and sometimes even reached a 50% error rate in 1,000-bp length reads with earlier base-calling sequences from ONT DDS reads are widely distributed, and sometimes even reached a 50% error rate in 1,000-bp length reads with earlier base-calling sequences from ONT DDS reads are widely distributed, and sometimes even reached a 50% error rate in 1,000-bp length reads with earlier base-calling sequences from ONT DDS reads are widely distributed, and sometimes even reached a 50% error rate in 1,000-bp length reads with earlier base-calling sequences from ONT DDS reads are widely distributed, and sometimes even reached a 50% error rate in 1,000-bp length reads with earlier base-calling sequences from ONT DDS reads are widely distributed, and sometimes even reached a 50% error rate in 1,000-bp length reads with earlier base-calling sequences from ONT DDS reads are widely distributed, and sometimes even reached a 50% error rate in 1,000-bp length reads with earlier base-calling sequences from ONT DDS reads are widely distributed, and sometimes even reached a 50% error rate in 1,000-bp length reads with earlier base-calling sequences from ONT DDS reads are widely distributed, and sometimes even reached a 50% error rate in 1,000-bp length reads with earlier base-calling sequences from ONT DDS reads are widely distributed, and sometimes even reached a 50% error rate in 1,000-bp length reads with earlier base-calling sequenc...
detects five types of SVs (deletions, insertions, inversions, duplications, tandem duplications, and translocations) and is more sensitive but less accurate than Sniffles. CuteSV collects the signatures of various types of SVs and employs a clustering-and-refinement method to implement sensitive SV detection. NanoVar was designed to identify SVs with low-depth, long-read whole-genome sequencing data, while SENSv has further lowered the depth requirement to the throughput of a single MinION flow cell using algorithmic advancements. NanoVar and SENSv enabled cost-effective and comprehensive SV studies using ONT DDS, making it a more practical application for disease diagnosis in the future.

Computational solutions for SV identification should integrate more factors than those for base-calling, including split reads (SRs), depth of coverage, and local de novo assembly. Long-read ONT DDS reads, such as those produced by ONT DDS, can directly readthrough the entire region of SVs or have more reads covering the SV boundaries, but short-read sequencing generally does not have high confidence in the discovery of large structural variants due to limited reads from SV boundaries as evidence. Researchers have developed several tools to detect SVs in accordance with ONT DDS long-read traits (Figure 4). NanoSV, developed at the earlier stage, clusters SRs to identify SV breakpoint junctions, and has successful applications in clinical and research scenarios. In addition to SRs, Picky also combines the seed-and-extend process to detect SVs with ONT DDS long-reads. Clairvoyante represents a successful attempt to use a deep-learning algorithm to find SVs, including small SVs, using less computational time. Currently, an analysis pipeline named CAMPHOR can identify deletions (≥100 bp), insertions (≥100 bp), inversions, and intra-chromosomal translocations.

**Tools developed for methylation identification.** Various types of DNA and RNA base modifications play crucial roles in fundamental biological processes. In DNA, 5-methylcytosine (5mC) contributes to maintain genome integrity and stability, and regulate pre-mRNA transcription initiation and processing, and even regulate poly(A) tail length, while N6-methyladenosine (m6A) modifications in mRNA could affect RNA splicing, degradation, and translation. NGS-based solutions, such as whole-genome bisulfite sequencing and methylated RNA immunoprecipitation sequencing, are widely implemented in the latest studies. Nevertheless, biases introduced in experimental procedures, particularly from the random fragmentation and DNA amplification, need to be further solved. ONT DDS and DRS sequence without converting bases and without amplification, so they could theoretically be used to directly detect DNA and RNA modifications when the trained model is sensitive enough (Figure 5).

Several software packages are available to detect DNA and RNA modifications in nanopore sequencing data. Nanopolish includes an HMM-trained model from a 6-mer model of CpG motifs to differentiate between 5mC and unmethylated cytosine in the CpG context of ONT DDS reads of human genomic DNA. In the training data as initial input, mCaller utilizes four machine learning algorithms (neural network, random forest, logistic regression, and naive Bayes classifiers) with 6-mers around positions of interest to improve the accuracy of detection of m6A modifications. Researchers wonder if previous DNA methylation detection models may not fully utilize nanopore electric ion current signals, and Liu et al. trained a bidirectional RNN model with so-called long short-term memory in DeepMod. Its training data were based on bisulfite sequencing confirmed full methylated or unmethylated DNA, and it was evaluated with respect to the genomes of three different species, by comparing the DeepMod results with both naturally occurring and synthetically introduced modifications. Its average precision ranged from ~0.9–0.99 for 5mC and N6-methyladenine (m6A, as opposed to 6mA) detection. In the same period, DeepSignal, another popular DNA modification detection tool, was released and it is by far the best 5mC caller for human CpG modifications according to personal communication. Recently, the DeepSignal has a derived version, DeepSignal-plant, adjusted according to 5mC modifications in plants, which have CHG and CHH methylation in addition to CG methylation. However, these machine learning–based detection algorithms need prior training via extensive datasets. To address this limitation, some other tools examine the statistical difference between signals from native samples and from modification-free controls; for instance, NanoMod relies on the Kolmogorov-Smirnov test to compare raw ion current signals of sequenced samples with two sets of standard ion currents, and then judge whether the related bases are modified or not. NanoMod is perfect for Escherichia coli data analysis, but may not be suitable for methylation detection in eukaryotic cells.

For RNA modification detection, there have been several widely used options. For instance, EpNano achieves an overall accuracy of ~90% in the identification of m6A RNA modification based on multiple trained support vector machines with m6A modified and unmethylated synthetic sequencing as training data. ELIGOS (Epitranscriptional Landscape Inferring from Gichetes of ONT Signals) detects ribonucleotide modification based on the...
error rates of specific bases, which are lower for unmodified RNA than for native RNA. Meanwhile, ELIGOS, taking various types of synthetic modified RNAs as training datasets, can be used for both rRNA and mRNA. Its accuracy reaches over 93% for the known classes of modifications in rRNA from E. coli, yeast, and human cell lines. Recently, nanom6A is released with an accuracy of 97%, higher than other algorithms. Its high accuracy was partially supported from the comparison between its results and those from methylated RNA immunoprecipitation sequencing and m6A-sensitive RNA-endo-ribonuclease–facilitated sequencing. The dysregulation of m6A in mRNAs may lead to tumorigenesis; therefore, using ONT DRS to investigate the relationship between m6A modification and tumorigenesis will be clearer and easier. For instance, the inhibitor of METTL3-MTTL14–STM2457 specifically prevents m6A occurrence and thus can be used to treat acute myeloid leukemia (AML). Therefore, ONT DRS, can directly identify m6A modifications that contribute to AML.

Besides the tools developed for ONT DDS and DRS data analysis, ONT DRS data also contain poly(A) tail length information that can be used to evaluate mRNA stability and translation efficiency. Recently, poly(A) tail length has been confirmed to be positively correlated with DNA methylation around transcription termination sites. Currently, Nanopolish can be used to evaluate poly(A) tail length for each ONT DRS read. Tailfindr is another tool for evaluating poly(A) tail length. It uses an R package to estimate poly(A) tail length for unaligned, base-called data.

ONT DDS and its applications

The first application of ONT DDS was to sequence the model organism E. coli K-12 sub-strain MG1655 in 2014 with the R7.3 flow cell. Since 2017, genome assembly using nanopore sequencing did not get too much attention because of its very high error rate (around 30%). Since 2017, genome assembly using nanopore sequencing has been completed for diverse organisms, e.g.,
yest,\textsuperscript{113,116} fish,\textsuperscript{115,116} plants,\textsuperscript{56,117–119} humans,\textsuperscript{138,139} fungi,\textsuperscript{121} Drosophila,\textsuperscript{122,123} and Caenorhabditis elegans.\textsuperscript{124} Recently, there is a rapid rise of ONT DDS for genome assembly. Over 100 publications in this field were published in 2020, and there have already been over 50 nanopore sequenced genomes reported in the first quarter of this year. The explosion of ONT DDS application is due to its recent progress, including significantly reduced cost and increased ability to provide sequence information at higher accuracy, benefiting from optimized base-calling and long-read alignment algorithms.\textsuperscript{53}

One successful example for ONT DDS application is to profile centromeric DNA sequence and the related epigenetic traits in Arabidopsis.\textsuperscript{125} Centromeric DNA comprises satellite DNA with a 100- to 200-bp repeat sequence. Satellite DNA is highly variable in sequence and length among different species; therefore, it is very challenging to assemble centromeric DNA with traditional sequencing technologies. However, ONT DDS provides such a chance for de novo assembling of Arabidopsis centromeric DNA and to profile its DNA methylation. Although the Arabidopsis thaliana genome was sequenced in 2000, until recently the centromeric DNA reference was assembled with ONT DDS ultra-long-reads.\textsuperscript{126} It clearly shows that there are 66,129 centromeric satellite repeats with an approximately 180-bp sequence in the five chromosomal centromeric regions, and each chromosome possesses largely private satellite variants, higher-order CEN180 repeats are prevalent within centromeres. The investigators also found that ATHILA LTR retrotransposons interrupted centromeric genetic and epigenetic organization by invading the satellite array in centromeres. A clear picture of centromeres in Arabidopsis can be acquired mainly depending on ONT DDS.

Another example for ONT DDS application is to reprofile epigenetic patterns in a complete human genome with ONT DDS long-reads.\textsuperscript{126} In this study, the authors used T2T CHMI3 complete human genome data that were assembled with ONT DDS long-reads to further profile its epigenetic patterns, especially for those newly complete genomic regions (e.g., peri-centromeres, centromeres, acrocentric chromosome arms, sub-telomeres, segmental duplications, tandem repeats), and can distinguish epigenetically heterogeneous and homogeneous regions based on the clustered reads with methylation traits.\textsuperscript{126}

An additional example of ONT DDS application is an adaptive nanopore sequencing for the golden lion tamarin (Leontopithecus rosalia) mitochondrial genome and epigenome.\textsuperscript{127} Adaptive sequencing with ONT DDS can enrich targeted reads and thus can increase the coverage of target DNA. Adaptive sequencing is particularly useful for environmental DNA, e.g., DNA from soil or feces of animals. Using ONT DDS adaptive sequencing, the authors acquired 258x coverage of the mitogenome of the closely related black lion tamarin (Leontopithecus chrysopygus) from feces of golden lion tamarin, which was successfully used to profile the mitogenome of golden lion tamarin. The differences between mitogenomes of the black lion tamarin and golden lion tamarin are 38 SNP, and in the golden lion tamarin mitogenome, some hydromethylation sites were also identified.\textsuperscript{127} For biodiversity preservation research, environmental DNA detection can be performed with ONT DDS adaptive sequencing.

Another application for ONT DDS is nanopore Cas9 targeted sequencing (nCATS), which combines the Cas9 nuclease to cut the targeted DNA from the genome, then uses an adaptor to ligate the enriched targeted DNA for ONT DDS.\textsuperscript{128} The strategy not only allows for the sequencing of genomic regions of interest for SNPs and SVs but can also acquire epigenomic information of enriched targeted genomic regions.\textsuperscript{129} In terms of using ONT DDS for small variant detection, several variant callers have been developed that could be used to provide therapeutic targets for the timely treatment of diseases related to such variation.\textsuperscript{78–80} Clinical diagnosis using both small variants (SNPs and InDels) and SVs can be a precise and effective strategy, especially for therapeutic target identification, and monitoring of hematological malignancies, etc.\textsuperscript{129}

For ONT DDS, DNA quality is crucial for successful sequencing, and current commercial genomic DNA extraction kits are not enough for ONT DDS library preparation because the purity and length of ONT DDS required is higher than the standards of commercial kits. Because the ONT company has some mature protocol, readers can follow the protocol the company provides to obtain qualified genomic samples (https://community.nanoporetech.com/extraction_methods). Currently, a single MinION flow cell can produce up to 50 Gb of data if the library quality is high enough. For beginners, the main problem is how to obtain high-quality genomic DNA for ONT DDS. Currently, Oxford Nanopore provides different protocols for extracting high-quality genomic DNA for different species (https://community.nanoporetech.com/extraction_methods). To download the related protocols, interested parties need to register with the Nanopore community. To obtain high-quality long genomic DNA and avoid contamination with short genomic DNA, researchers can use BluePippin (Sage Science, Beverly, MA) to get the defined sizes of genomic DNA for sequencing,\textsuperscript{130} or use the Short-Read Eliminator Kit (Ciculomics, Baltimore, MD) (SKU SS-100-101-D1). Also note that for different applications, the strategy for ONT DDS library preparation is different. For example, for
assembling an unknown genome, it is better to acquire the longest reads with the Ultra-Long DNA Sequencing Kit (Oxford Nanopore Technologies, Oxford, UK) (SQK-JUL001), so that each sequencing read can cover the largest possible genomic region. However, for those species with reference genomes, if you only want to learn about SVs and/or DNA modifications, it is better to fragment genomic DNA to 8 kb with g-Tube ( Covaris, Woburn, MA) (Cat# S20079), so that you can retrieve as much data as possible for downstream analysis. Readers can find more related information from the following web page: https://community.nanoporetech.com/. Published tools for alignment (genome assembly, structural variants, and methylation detection) can be found in Table 2.

ONT DRS and its applications

Another advantage is its unrestricted read length, which breaks the limit for current commercial reverse transcription kits. ONT DRS has confirmed its ability to sequence super-long RNA molecules,49,50 after it was first demonstrated for native RNA in 2018.131 Using this technology, Zhang et al.52 provided the new splicing patterns detected by sequenced native RNA reads with experimental evidence and corrected the annotation of Arabidopsis At4g17140 splicing sites. The transcripts from At4g17140 are ultra-long RNA molecules (longer than 13,000 nt). This is beyond the ability of common commercial reverse transciptase kits, which can only convert RNA sequences of less than 12,000 nt to complete cDNA molecules. Another great example attesting to the advantages of ONT DRS comes from a novel gene activated in met1-3 (MET1) and ddc2 (DDM1DDR2CM272CM73) quadruple mutants. This gene covers a 17.3-kb genomic region with eight small exons interspersed within the intergenic regions of several annotated genes.92 Very interestingly, as hinted at above, ONT DRS can also be used to correct previously assembled genomes. In Arabidopsis, transcripts from At2g40980 not only contain the previously annotated Arabap11 transcripts (with no additional exons found in the second intron) but also contain around 20 transcripts from a cryptic exon within the second intron. That cryptic exon in the second intron had not been annotated in the Arabidopsis genome, but the original BAC clone and follow-up experimental evidence support the existence of the cryptic exon.50

Unlike traditional NGS, ONT DRS sequences full-length transcripts without converting RNA to cDNA, and it can directly capture alternative splicing events in 15% of human hereditary diseases and cancers.132 Currently, some rare diseases, e.g., myelodysplasia syndromes, were found to be caused by aberrant splicing because of spliceosome mutation in somatic cells.133 To effectively detect aberrant splicing relating to human diseases, a pipeline-FRASER for NGS RNA-seq and ONT DRS analysis was developed.134 Besides spliceosome mutation causing aberrant splicing in human diseases, some splice site mutations also cause alternative splicing and related human diseases.135 Actually, the SpliceDisease database was published in 2012,126 and documents most common splice-related diseases. Using ONT DRS, we can find novel isoforms in a specific disease, which may be omitted by NGS RNA-seq. However, for ONT DRS application for aberrant splicing detection or novel-isoform discovery in a disease, researchers should keep in mind that rG4 structure in some transcripts may impede the sequencing because of the huge structures, while human mRNA, at least, in vitro, easily forms the rG4 structure if no lithium ion (Li+) is present.137 To better detect the related alternative splicing events in human diseases, the purified RNAs should be treated with Li+ before making an ONT DRS library.

With the development of more base modification detection methods, epitranscriptomes of different species have been reported:49,96,103–105,128–142 It has been reported that m6A modifications may have some negative connections with readthrough transcripts, because m6A writer mutant vir-1 has more readthrough transcripts, whereas Smc modifications also have a high correlation with mobile mRNAs because they harbor more Smc modifications compared with total mRNAs.92 Previously, the Kragler lab43 confirmed that in graft junction-mobile methylated mRNAs TRANSLATIONALLY CONTROLLED TUMOR PROTEIN 1 (TCTP1) and HEAT SHOCK COGNATE PROTEIN 70.1 (HSC70.1), the mRNA transport of which is diminished in mutants deficient in Smc mRNA methylation. Furthermore, with ONT DRS, m6A modifications in adenosinyl transcripts have been identified. Modification of m6A particularly affects the splicing of viral transcripts, which was confirmed with the m6A writer mutant.149 Besides m6A and Smc modifications, other kinds of modifications of mRNA also were reported using ONT DRS reads.141,142

For modification detection models, there are two strategies. One is to use differential error calling, e.g., ELIGOS uses the error rate of specific bases between native RNA sequencing data and cDNA sequencing data of the same sample.104,106 A similar strategy was used for m6A detection in the wild-type Arabidopsis ecotype col-0, by employing an m6A writer mutant (vir-1 mutant),49 and also with the normal and m6A writer mutant, for the differential ion current associated with adenovalin transcripts.140 The other strategy adopts trained models. The EpiNano algorithm was based entirely on synthetic fully methylated and unmethylated RNAs.153 Another example using this strategy is nanonm6A,153 which also used the synthetic fully methylated and unmethylated RNAs to train its detection model. Another related approach for modification detection models is to use machine or deep learning. MINES (m6A identification using nanopore sequencing) is an example, which uses a random forest classifier with mCLIP m6A sites within DRACH motifs.56 NanoCompore was also developed based on a machine learning algorithm.156

Any modification in mRNA can be detected if there are perfectly related training data. Oxford Nanopore provides the Megalodon model training software (https://nanoporetech.github.io/megalodon/model_training.html), which uses the Tayaki algorithm (https://github.com/nanoporetech/tayaki) to build up modified base detection models with in vitro or in vivo training data.

ONT DRS can sequence native RNA, and its good application is to profile NAD-capped RNA (Figure 6). Recently, it has been found that mRNAs not only have 7-methylguanosine (m7G) caps, but also have other caps, e.g., NAD-capped mRNA in E. coli,147 yeast,148 humans,130 and Arabidopsis.149–151 Previously, to purify NAD-capped RNA, the NAD caps should be labeled with biotin via click chemistry. Then streptavidin beads would be used to enrich the biotin-labeled RNA molecules. Finally, after elution from streptavidin beads, the RNAs were used in NGS.152 This protocol has been used to prepare RNA from different organisms.152,153,155,156 However, the big problem with the protocol is eluting RNA from the streptavidin beads, which introduces false-positive results because of its huge background. Another strategy for accurately identifying NAD-capped RNAs is to use the click chemistry to introduce an RNA adaptor to the NAD caps, then use biotin-labeled DNA, which is complementary to the synthetic RNA adaptor to enrich the NAD-capped RNA. The enriched NAD-capped RNAs were used for ONT DRS, because the NAD-capped RNA molecules can be identified according to their adaptor sequences.151,154 This strategy can greatly decrease the false-positive rate of NAD-capped RNA molecules, allowing researchers to learn the characteristics of NAD-capped RNA, e.g., whether there are differences in splicing patterns, poly(A) tail length, etc., between NAD-capped RNA and normal RNA molecules. The original protocol for labeling NAD-capped RNA was to use copper ions to catalyze click chemistry. Unfortunately, this splices a lot of RNA, and thus a huge input of RNA was required. To conquer the problem, SPAAC (strain-promoted azide-alkyne cycloaddition) was used to replace copper ions, which also allowed for a great reduction in input RNA. However, during the development of the protocol, it was found that the ADPRC enzyme can catalyze m7G-capped RNA, although with very low efficiency.149 However, because m7G-capped RNA is far more abundant than NAD-capped RNA, it is possible that previously identified NAD-capped RNA from eukaryotic cells may have been contaminated with m7G-capped RNA, especially those from humans and yeast. One way of removing such contamination is to eliminate m7G RNA when performing the ADPRC reaction.149 Another way to identify NAD-capped RNA from organelles or prokaryotic cells. One problem that still needed to be resolved was to add poly(A) tails to the RNA before preparing ONT DRS libraries, as Zhang et al. did.152
| Tool      | Description                  | Algorithm                                      | Advantages                                                                 | Rate                       | Disadvantages                                                                 | Link                                                                 | Reference (PMID) |
|-----------|------------------------------|-----------------------------------------------|---------------------------------------------------------------------------|----------------------------|--------------------------------------------------------------------------------|----------------------------------------------------------------------|------------------|
| LAST      | Alignment                    | adaptive seeds                                | Adaptive seeds are matches that are chosen based on their rareness, instead of using fixed-length matches | -                          | the running time increases linearly with sequence length and short DNA reads   | https://gitlab.com/mcfrith/last                                      | 21209072         |
| Minimap2  | Alignment                    | split-read alignment                          | DNA or long mRNA, higher accuracy, faster, and full length of reads        | -                          | not suitable for chimeric alignments                                          | https://github.com/lh3/minimap2                                    | 29750242         |
| GraphMap  | Alignment                    | candidate alignments and fast graph traversal  | long reads with speed, high sensitivity                                    | -                          | large-memory                                                                  | https://github.com/isovic/graphmap                                  | 27079541         |
| UNCALLED  | Alignment                    | Ferragina-Manzini index                        | mapping during sequencing and the leftmost mapping                        | -                          | not full length                                                                | https://github.com/skovaka/UNCALLED                                 | 33257863         |
| tailfindr | poly(A)                      | measures poly(A) tail length                   |                                                                           | -                          | -                                                                              | https://github.com/adnaniazi/tailfindr                              | 31266821, 33835460 |
| NaS       | Assembly                     | illumina hybrid                                | entirely and with no error                                                 | -                          | Not suitable for large genomes                                                  | https://www.genoscope.cns.fr/externe/nas/                           | 25927464         |
| LQS       | Assembly                     | multiple-alignment corrected                  | corrected by a multiple-alignment and 99.5% nucleotide identity           | -                          | Not suitable for large genomes                                                  | https://github.com/jts/nanopore-paper-analysis                      | 26076426         |
| Canu      | Assembly                     | tf-idf weighted MinHash and graph construction | halves depth-of-coverage requirements, improves assembly continuity and reduces runtime on large genomes | -                          | accuracy depends on signal-level polishing                                      | https://github.com/marbl/canu                                       | 28298431         |
| Miniasm   | Assembly                     | No correction                                  | magnitude faster                                                          | -                          | error rate is as high as raw reads                                             | https://github.com/lh3/miniasm                                      | 27153593         |
| Nanopolish| Variant caller/ Methylation detection | Hidden Markov Model                           | calculate an improved consensus sequence for a draft genome assembly, detect base modifications, call SNPs and indels | -                          | signal-level analysis                                                           | https://github.com/jts/nanopolish                                 | 26076426         |
| Clairvoyante | Variant caller/ SV caller/ Methylation detection | convolutional neural network                 | SV calling, small variants and genotype                                    | -                          | higher sequencing depth                                                         | https://github.com/aquaskyline/Clairvoyante                         | 30824707         |
| Clair     | Variant caller               | Deep neural network                            | faster and complex variants with multiple alternative alleles             | -                          | accuracy depends on pileup data and greater computational demands             | https://github.com/quay/clair                                       |                  |
| NanoSV    | SV caller                    | split- and gapped-aligned reads                | genotyping                                                                | -                          | non-detectable inversion, complex repeat regions and segmental duplications   | https://github.com/mroosmalen/nanosv                                | 29109544         |
| Picky     | SV caller                    | seed-and-extend process and split-read         | micro-insertions and phased SV                                            | -                          | high specificity                                                               | https://github.com/TheJacksonLaboratory/Picky                      | 29713081         |

(Continued on next page)
| Tool       | Description         | Algorithm               | Advantages                                      | Rate     | Disadvantages                                                                                                                                  | Link                                                                 | Reference (PMID) |
|------------|---------------------|-------------------------|-------------------------------------------------|----------|-------------------------------------------------------------------------------------------------------------------------------------------------|----------------------------------------------------------------------|------------------|
| NanoVar    | SV caller           | artificial neural network | low-depth (8X)                                   | -        | the alignment profile of each read requires re-training                                                                                | https://github.com/benoukraflab/nanovar                              | 32127024         |
| SENSV      | SV caller           | Deep neural network     | low-depth                                        | -        | balanced translocation missed                                                                                                                | https://github.com/HKU-BAL/SENSV                                    |                  |
| CAMPHOR    | SV caller           | SV breakpoints          | polymorphic SVs and somatic SVs                  | -        | removed indels in short repeats, the average read length 5 kbps and non-detectable indels < 100 bp                                          | https://github.com/afujimoto/CAMPHOR                                | 33910608         |
| NanoMod    | Methylation detection | signal intensities    | raw signal data and 5mC                          | -        | two pair sample reads                                                                                                                        | https://github.com/WGLab/NanoMod                                   | 30712508         |
| DeepSignal | Methylation detection | deep learning       | 6mA/5mC, lower coverage, and predict methylation states | -        | train DeepSignal to detect more types of base modification                                                                              | https://github.com/bioinformaticsCSU/deepsignal                      | 30994904         |
| mCaller    | Methylation detection | neural network     | 6mA and detect known or confirm suspected methyltransferase target motifs | -        | only bacteria genome                                                                                                                        | https://github.com/al-mcintyre/mCaller_analysis_scripts             | 30718479         |
| DeepMod    | Methylation detection | recurrent neural network | 6mA/5mC, strand-sensitive and has single-base resolution | -        | non-detectable other types of modifications or other different motifs, not suitable for RNA, neighboring bases influence, elided on alignment tool to find correct reference positions of bases | https://github.com/WGLab/DeepMod                                    | 31164644         |
| MINES      | Methylation detection | random forest        | m6A sites within DRACH motifs                    | -        | lost small difference modification sites and not suitable for DNA                                                                         | https://github.com/YeoLab/MINES                                     | 31624092         |
| Nanom6A    | Methylation detection | XGBoost model        | m6A at single-base resolution and quantified abundance of m6A sites | -        | not suitable for DNA                                                                                                                        | https://github.com/gaoyubang/nanom6A                                | 33413586         |
| FLAIR      | Isoform detection   | correct and realign   | assessing 3’ poly(A) tail length, base modifications, and transcript haplotypes | -        | combined short Illumina reads                                                                                                               | https://github.com/BrooksLabUCSC/flair                              | 31740818         |
| TrackCluster | Isoform detection  | read tracks          | read classification, a transcript isoform with numerous exons, stage-specific or cell-specific expression of isoforms | -        | not suitable for large genomes                                                                                                             | https://github.com/Runsheng/trackcluster                            | 32024662         |
The Innovation

Figure 6. Flowchart for ONT DRS of NAD-capped RNAs. The 5' and 3' indicate the NAD-capped RNA direction. The NAD structure was illustrated and connected to the 5' end of the NAD-capped RNA, after ADPRC catalysis, the azide moiety was linked to NAD via replacing nicotinamide, the azide contained NAD-capped RNA then can react with SPAAC, and the synthetic RNA adapter with DBCO at its 3' end can conjugate with azide functionalized NAD-capped RNA. The tagged NAD-capped RNA then can be used for ONT DRS library preparation and further sequenced and analyzed. ADPRC, ADP-ribosyl cyclase; NAD, nicotinamide adenine dinucleotide; SPAAC, strain-promoted azide-alkyne cycloaddition; tagRNA, synthetic adaptor RNA to tag NAD-capped RNA; ONT DRS, Oxford Nanopore Technologies Direct RNA Sequencing.

Another elegant application for ONT DRS is denoting RNA secondary structure. The strategy is to label single-strand RNA bases, because the labeled bases will produce different ion currents that can be used to distinguish the positions of bases (in single- or double-stranded RNA). Stepheenson et al.\textsuperscript{156} used acetylimidazole to exogenously label RNA bases, while Aw et al.\textsuperscript{155} used 2-methylnicotinic acid imidazole azide, dimethyl sulfate, and 1-cyclohexyl-3-(2-morpholinoethyl) carbodiimide metho-p-toluenesulfonate to label RNA bases. But both papers successfully used ONT DRS to detect labeled bases and to decipher RNA structures.

For ONT DRS, to get full-length RNA reads, only RNA primary structure should remain when preparing ONT DRS libraries. Reverse transcription (RT) can destroy the secondary (and higher) structure of most RNA molecules; this is the reason that RT is necessary for preparing ONT DRS libraries, although cDNA is never used for sequencing. However, RNAs may also have G-quadruplex (G4, but rG4 when referring to RNA) secondary structures,\textsuperscript{157,158} which can cause reverse transcriptase stalling.\textsuperscript{137} It has been thought the rG4 structures are globally unfolded in eukaryotic cells,\textsuperscript{159} but RNA can quickly form rG4 structure in vivo in Na+ or K+ solution in vitro,\textsuperscript{159} while Li+ can eliminate rG4 structures.\textsuperscript{137,159} In Arabidopsis, rG4 structures are common in mRNA,\textsuperscript{158} ONT DRS library preparation with and without Li+ treatment has different average RNA molecular lengths.\textsuperscript{50}

ONT DRS reads also can be used to detect poly(A) tail length.\textsuperscript{49,50,115,160,161} Poly(A) tails are homopolymers and proximal to the sequencing adapters, which can both be used to evaluate poly(A) tail length according to the dwell time of the poly(A) bases.\textsuperscript{162} Using Nanopolish to estimate poly(A) tail length, we have found that TE transcripts have longer poly(A) tails compared with other transcripts, while transcripts from housekeeping genes tend to have shorter poly(A) tails.\textsuperscript{92}

Challenges for ONT DDS and DRS

Although ONT DDS and DRS were quickly applied in different research fields and greatly enhanced our understanding of genome profiling, accuracy of base-calling is still a challenge. The problem is becoming more obvious for ONT DRS data because more than 100 base modifications of RNA molecules have been discovered.\textsuperscript{162} To accurately detect all modified bases remains an enormous challenge because any modification may require a trained model for accurate detection, especially the training data used from the completely unmethylated transcriptome and fully methylated transcriptome. Although the completely unmethylated transcriptome and fully methylated transcriptome are relatively easy to get via converting the transcriptome to cDNA, and then using cDNA for \textit{in vitro} transcription with modified bases or normal bases to get these training data, some RNA modifications may make nanopore sequencing impossible, e.g., base glycosylation with oligosaccharide,\textsuperscript{163} which may completely block the nanopore. For such modifications, new strategies may need to be developed.

Currently, ONT DRS is mainly used for mRNA sequencing. For other RNA sequencing, e.g., rRNA or RNAs from prokaryotic cells, the length of RNA normally is longer than 100 nt; therefore, small RNA sequencing, even for tRNA sequencing, still is impossible. To sequence small RNA molecules, especially those less than 30 nt, molecule glue or RNA adapters should be developed to ligate small RNA into long chimeric RNAs for sequencing.

Another challenge may come from ONT DRS library preparation because the RNA used for an ONT DRS library needs to have a poly(A) tail. In the case of noncoding RNA and RNA from prokaryotic cells, a poly(A) tail, the poly(A) tailing step needs to be performed before ONT DRS library preparation.\textsuperscript{168} For ONT DRS, an additional challenge is how to get full-length RNAs, because 15 to 50 nt at the 5' end of mRNAs cannot be detected because of loss of control of the speed mediated by the motor protein.\textsuperscript{51} One way to resolve the problem is to ligate an RNA adapter to the 5' end of the sequenced RNAs, as was done for NAD-capped RNA with the TagSeq method.\textsuperscript{151,154,164}

The relatively higher base error rate, especially in the low-complexity genomic regions, has long prevented genetic testing and microbial detection from utilizing the full power of ONT DDS. However, the miniaturization of equipment and fast turnaround time provide hope that the two applications will become more approachable. With significant efforts made to improve the performance of small variant and structural variant detection in the past 3 years, the impact is likely imminent. Nevertheless, we should also be aware of the limitations that remain. While the sensitivity and accuracy of SNPs detected from ONT DDS have surpassed those from Illumina sequencing (both sensitivity and accuracy have reached over 99.5%),\textsuperscript{65} InDel detection remains...
an unsolved problem (sensitivity around 60% and accuracy around 90%). Most of the incorrect and missing InDels are in low-complexity genomic regions, such as tandem repeats, repetitive elements, and MHC proteins. One should be cautious when drawing conclusions from ONT DDS-detected InDels in these regions, and not make any medical decisions from any ONT DDS-detected InDels unless the InDels can be validated with additional experiments. The per-base cost of ONT DDS is still a few times higher than short-read sequencing, resulting in lower depth coverage per sample by some early adopters of ONT DDS with the same budget as for short-reads. Although some structural variant detection algorithms are designed to cope with the lower depth of coverage, it imposes certain limitations, such as the type of structural variants that can be reliably detected. As ONT DDS is still in its early development and there is no one-size-fits-all solution for variant calling, one should obtain a deeper understanding of the algorithms and their limitations before using them, especially when patients are involved.

FUTURE PERSPECTIVE
ONT nanopore not only represents a new sequencing technology, but also greatly alters our strategies for answering some basic biological questions, e.g., for identification of NAD-capped RNA, and detection of the secondary structure of mRNA. Based on the merits of ONT DDS and DRS, they may also be used as biomarkers for correctly identifying species with close relatives (e.g., for identification of authentic medicinal herbs), and distinguishing different tissues from the same species (e.g., transcripts from roots and leaves, respectively, may have different mRNA modification patterns in their transcripts or for some genomic regions they may have different DNA methylation patterns). In addition, because of their low cost but high output, they can be used for profiling more (epi)genomes (epi)transcriptomes of different species. They may be especially suitable for research on the effect of the environment on the regulation of gene expression. We have discussed the limitations of using ONT DDS for small variant and structural variant calling; however, the most significant limitation remains its high base-calling error rate. It is noteworthy that the ONT DDS error rate has improved remarkably in recent years through better chemistries and better base-calling algorithms. ONT made the new R10.3 chemistry available externally early last year for public testing. With a longer barrel and dual reader head of testing data, the new model reduced the base error rate from ~8% to ~5%. The primary base-caller “Guppy,” made by ONT, has implemented a flip-flop model along with a few deep-learning techniques with support and feedback from a large community of ONT users. With the same set of testing data, the new model reduced the base error rate from ~9% to ~7%. The most recently released ONT base-caller, "Bonito" (https://github.com/nanoporetech/bonito), has shown promising results that further reduced the base error rate to ~5%. Focusing on genomic regions of interest to increase sequencing coverage has also shown its potential in reducing errors for using ONT DDS in the clinical context. While the community keeps developing new variant calling methods to make the most out of the ONT data, we foresee that the advancements in sequencing chemistry, base-calling, and protocol will result in a decisive moment for ONT DDS/DRS to take over more applications that are currently dominated by short-read sequencing.

REFERENCES
1. Lander, E.S., Linton, L.M., Birren, B., et al. (2001). Initial sequencing and analysis of the human genome. Nature 409, 860–921.
2. International Human Genome Sequencing. C. (2004). Finishing the euchromatic sequence of the human genome. Nature 431, 931–945.
3. Barba, M., Czosek, H., and Hadidi, A. (2014). Historical perspective, development and applications of next-generation sequencing in plant virology. Viruses 6, 106–136.
4. Deamer, D., Akeson, M., and Branton, D. (2016). Three decades of nanopore sequencing. Nat. Biotechnol. 34, 518–524.
5. Dunne, K.A., Chaudhuri, R.R., Rossiter, A.E., et al. (2017). Sequencing a piece of history: complete genome sequence of the original Escherichia coli strain. Microb. Genom. 2, mgz00106.
6. Branton, D., Deamer, D.W., Marziali, A., et al. (2008). The potential and challenges of nanopore sequencing. Nat. Nanotechnol. 3, 1146–1153.
7. Kasianowicz, J.J., Brandin, E., Branton, D., and Deamer, D.W. (1994). Characterization of individual polynucleotide molecules using a membrane channel. Proc. Natl. Acad. Sci. U S A 91, 13770–13773.
8. Bezrukov, S.M., and Kasianowicz, J.J. (1997). The charge state of an ion channel controls neutral polymer entry into its pore. Eur. Biophys. J. 26, 471–476.
9. Braha, O., Walker, B., Cheley, S., et al. (1997). Designed protein pores as components for biosensors. Chem. Biol. 4, 597–505.
10. Butler, T.Z., Gundlach, J.H., and Troll, M. (2007). Iionic current blockades from DNA and RNA molecules in the alpha-hemolysin nanopore. Biophys. J. 93, 3229–3240.
11. Kasianowicz, J.J., Burden, D.L., Han, L.C., et al. (1999). Genetically engineered metal ion binding sites on the outside of a Channel’s transmembrane beta-barrel. Biophys. J. 75, 837–845.
12. Henriksen, S.E., Misakian, M., Robertson, B., et al. (2000). Driven DNA transport into an asymmetric nanotube-scale pore. Phys. Rev. Lett. 85, 3057–3060.
13. Howorka, S., Cheley, S., and Bayley, H. (2001). Sequence-specific detection of individual DNA strands using engineered nanopores. Nat. Biotechnol. 19, 636–639.
14. Kasianowicz, J.J., Henriksen, S.E., Weetall, H.H., et al. (2001). Simultaneous multi-analyte detection with a nanometer-scale pore. Anal. Chem. 73, 2268–2272.
15. Halverson, K.M., Panchal, R.G., Nguyen, T.L., et al. (2005). Anthrax biosensor, protective antigen ion channel asymmetric blockade. J. Biol. Chem. 280, 34056–34062.
16. Merzyk, P.G., Capistrano, M.F., Valeva, A., et al. (2005). Conductance and ion selectivity of a mesoscopic protein nanopore probed with cysteine scanning mutagenesis. Biophys. J. 89, 3059–3070.
17. Hromada, L.P., Nablo, B.J., Kasianowicz, J.J., et al. (2008). Single molecule measurements within individual membrane-bound ion channels using a polymer-based bilayer lipid membrane chip. Lab. Chip 8, 602–608.
18. Kasianowicz, J.J., Robertson, J.W., Chan, E.R., et al. (2008). Nanoscopic porous sensors. Annu. Rev. Anal. Chem. (Palo Alto Calif.) 1, 737–766.
19. Nablo, B.J., Halverson, K.M., Robertson, J.W., et al. (2008). Sizing the Bacillus anthra-
20. Reiner, J.E., Balijepalli, A., Robertson, J.W., et al. (2012). The effects of diffusion on an exonuclease/nanopore-based DNA sequencing engine. J. Chem. Phys. 137, 214903.
21. Walker, B., Krishnaswasty, M., Zorn, L., et al. (1992). Assembly of the oligomeric mem-
22. Walker, B., Krishnaswasty, M., and Bayley, H. (1993). Functional complementation of staphylococcal alpha-hemolysin fragments. Overlaps, nicks, and gaps in the glycine-rich loop. J. Biol. Chem. 268, 5281–5292.
23. Walker, B., and Bayley, H. (1995). Restoration of pore-forming activity in staphylococcal alpha-hemolysin by targeted covalent modification. Protein Eng. 8, 491–495.
24. Walker, B., and Bayley, H. (1995). Key residues for membrane binding, oligomerization, and pore forming activity of staphylococcal alpha-hemolysin identified by cysteine scanning mutagenesis and targeted chemical modification. J. Biol. Chem. 270, 23065–23071.
25. Song, L., Hobaugh, M.R., Shustak, C., et al. (1996). Structure of staphylococcal alpha-hemolysin, a heptameric transmembrane pore. Science 274, 1859–1866.
26. Cheley, S., Malghani, M.S., Song, L., et al. (1997). Spontaneous oligomerization of a staphylococcal alpha-hemolysin conformationally constrained by removal of resi-
27. Reiner, J.E., Balijepalli, A., Robertson, J.W., et al. (2012). The effects of diffusion on an exonuclease/nanopore-based DNA sequencing engine. J. Chem. Phys. 137, 214903.
28. Stoddart, D., Heron, A.-J., Klingelhofer, J., et al. (2010). Nucleic base recognition in ssDNA at the central constriction of the alpha-hemolysin pore. Nano Lett. 10, 3633–3637.
29. Lieberman, K.R, Chef, G.M., Doody, M.J., et al. (2010). Processive replication of single DNA molecules in a nanopore catalyzed by phi29 DNA polymerase. J. Am. Chem. Soc. 132, 17961–17972.
30. Chef, G.M., Lieberman, K.R, Rashid, H., et al. (2012). Automated forward and reverse ratcheting of DNA in a nanopore at 5 A precision. Nat. Biotechnol. 30, 344–348.
31. Niederweis, M., Ehr, S., Heinz, C., et al. (1999). Cloning of the mspA gene encoding a porin from Mycobacterium smegmatis. Mol. Microbiol. 33, 933–945.
32. Laszlo, A.H., Derrington, I.M., Ross, B.C., et al. (2014). Decoding long nanopore sequencing reads of natural DNA. Nat. Biotechnol. 32, 829–833.
33. Deamer, D., Akeson, M., and Branton, D. (2016). Author response to John Kasianowicz and Sergey Bezrukov. Nat. Biotechnol. 34, 482.
34. Goyal, P., Kraeva, V.Y., Van Gerven, N., et al. (2014). Structural and mechanistic in-
35. Kasianowicz and Sergey Bezrukov. Nat. Biotechnol. 34, 482.
36. Silvestre-Ryan, J., and Holmes, I. (2021). Pair consensus decoding improves accuracy of neural network basecallers for nanopore sequencing. Genome Biol. 22, 38.

37. Temp, W., Corner, J., and Aksimentiev, A. (2012). DNA base-calling from a nanopore using a Viterbi algorithm. Biophys. J. 102, L37–L39.

38. Teng, H., Cao, M.D., Hall, M.B., et al. (2018). Chiron: translating nanopore raw signal directly into nucleotide sequence using deep learning. Gigascience 7, giy037.

39. Zeng, J., Cai, H., Peng, H., et al. (2019). Causcall: nanopore basecalling using a temporal convolutional network. Front. Genet. 10, 1332.

40. Zhang, Y.Z., Akdemir, A., Tremmel, G., et al. (2020). Nanopore basecalling from a perspective of instance segmentation. BMC Bioinformatics 21, 136.

41. Boza, V., Brenova, B., and Vinar, T. (2017). DeepNano: deep recurrent neural networks for base calling in MinION nanopore reads. PLoS One 12, e0178751.

42. Stoiber, M., and Brown, J. (2017). BasecRAWller: Streaming nanopore basecalling directly from raw signal. BioRxiv. https://doi.org/10.1101/130358.

43. Timp, W., Comer, J., and Aksimentiev, A. (2012). DNA base-calling from a nanopore using a Viterbi algorithm on pileup data for germline variant calling. Nat. Methods 9, 1053–1055.

44. Li, H. (2018). Minimap2: pairwise alignment for nucleotide sequences. Nucleic Acids Res. 46, 7700–7711.

45. Vaser, R., and Sikić, M. (2020). Raven: a de novo genome assembler for long reads. BioRxiv. https://doi.org/10.1101/2021.04.20.440583.

46. Keller, A., Yamanaka, T., Tameguchi, H., et al. (2021). SMARTenovo: a de novo assembler using long noisy reads. Gigabyte 1, 2021. https://doi.org/10.4457/gigabyte.15.

47. Balasubramanian, S., Wang, Y., et al. (2018). Compendium of open source basecallers for nanopore DNA sequencing. Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. 2016, 5745–5748.

48. Wu, Z., Hammad, K., Ghafar-Zadeh, E., et al. (2020). FPGA-accelerated 3rd generation PeakCalling: an open source basecaller exploiting an encoder-decoder model with monotonic attention. Bioinformatics. https://doi.org/10.1093/bioinformatics/btaa953.

49. Liu, H., Wu, S., Li, A., et al. (2021). MECAT: fast mapping, error correction, and variant calling of nanopore reads with GraphMap. Nat. Commun. 12, 4935.
154. Zhang, H., Zhong, H., Wang, X., et al. (2021). Use of NAD tagSeq II to identify growth phase-dependent alterations in E. coli RNA NAD(+) capping. Proc. Natl. Acad. Sci. U S A 118, e2026183118.

155. Aw, J.G.A., Lim, S.W., Wang, J.X., et al. (2021). Determination of isoform-specific RNA structure with nanopore long reads. Nat. Biotechnol. 39, 336–346.

156. Stephenson, W., Razaghi, R., Busan, S., et al. (2020). Direct detection of RNA modifications and structure using single molecule nanopore sequencing. BioRxiv. https://doi.org/10.1101/2020.05.31.126763.

157. Yang, S.Y., Lejault, P., Chevrier, S., et al. (2018). Transcriptome-wide identification of transient RNA G-quadruplexes in human cells. Nat. Commun. 9, 4730.

158. Mullen, M.A., Olson, K.J., Dallaire, P., et al. (2010). RNA G-Quadruplexes in the model plant species Arabidopsis thaliana: prevalence and possible functional roles. Nucleic Acids Res. 38, 8149–8163.

159. Guo, J.U., and Bartel, D.P. (2016). RNA G-quadruplexes are globally unfolded in eukaryotic cells and depleted in bacteria. Science 353, aaf5371.

160. Workman, R.E., Tang, A.D., Tang, P.S., et al. (2019). Nanopore native RNA sequencing of a human poly(A) transcriptome. Nat. Methods 16, 1297–1305.

161. Roach, N.P., Sadowski, N., Alessi, A.F., et al. (2020). The full-length transcriptome of C. elegans using direct RNA sequencing. Genome Res. 30, 299–312.

162. Shen, L., Liang, Z., Wong, C.E., et al. (2019). Messenger RNA modifications in plants. Trends Plant Sci. 24, 328–341.

163. Flynn, R., Pedram, K., Malaker, S., et al. (2021). Small RNAs are modified with N-glycans and displayed on the surface of living cells. Cell. https://doi.org/10.1016/j.cell.2021.04.023.

164. Shao, X., Zhang, H., Yang, Z., et al. (2020). NAD tagSeq for transcriptome-wide identification and characterization of NAD(+) -capped RNAs. Nat. Protoc. 15, 2813–2836.

165. Karst, S.M., Ziels, R.M., Kirkegaard, R.H., et al. (2021). High-accuracy long-read ampli-con sequences using unique molecular identifiers with Nanopore or PacBio sequencing. Nat. Methods 18, 165–169.

ACKNOWLEDGMENTS

This work is supported by the Key-Areas Research and Development Program of Guangdong Province (2020B020220004), the Youth Innovation Promotion Association, Chinese Academy of Sciences (2017399), the Science and Technology Program of Guangzhou (202002030097), the Hong Kong Research Grants Council Area of Excellence Scheme (AoE/M-403/16), the ECS (27204518), and TRS of the HK SAR government (T21-705/20-N).

AUTHOR CONTRIBUTIONS

M.L. and S.Z. conceived and coordinated the review. S.X., W.L., C.X., R.L., M.L., and S.Z. wrote the manuscript. S.X., Z.Z., S.Z., and D.Z. drew the figures. All authors were involved in the preparation of the final manuscript.

DECLARATION OF INTERESTS

The authors declare no competing interests.