Trends in Genetics

CellPress

Review

Beyond sequencing: machine learning algorithms extract biology hidden in Nanopore signal data

Yuk Kei Wan,[1,2] Christopher Hendra,[3,1] Ploy N. Pratanwanich,[1,4,5] and Jonathan Göke[1,6,*]

Nanopore sequencing provides signal data corresponding to the nucleotide motifs sequenced. Through machine learning-based methods, these signals are translated into long-read sequences that overcome the read size limit of short-read sequencing. However, analyzing the raw nanopore signal data provides many more opportunities beyond just sequencing genomes and transcriptomes: algorithms that use machine learning approaches to extract biological information from these signals allow the detection of DNA and RNA modifications, the estimation of poly(A) tail length, and the prediction of RNA secondary structures. In this review, we discuss how developments in machine learning methodologies contributed to more accurate basecalling and lower error rates, and how these methods enable new biological discoveries. We argue that direct nanopore sequencing of DNA and RNA provides a new dimensionality for genomics experiments and highlight challenges and future directions for computational approaches to extract the additional information provided by nanopore signal data.

Highlights

Nanopore sequencing accuracy has increased to 98.3% as new-generation base callers replace early generation hidden Markov model basecalling algorithms with neural network algorithms.

Machine learning methods can classify sequences in real-time, allowing targeted sequencing with nanopore's ReadUntil feature.

Machine learning and statistical testing tools can detect DNA modifications by analyzing ion current signals from nanopore direct DNA sequencing.

Nanopore direct RNA sequencing profiles RNAs with their modification retained, which influences the ion current signals emitted from the nanopore.

Machine learning and statistical testing tools analyze ion current signals from direct RNA sequencing, enabling RNA modification detection, RNA secondary structure prediction, and poly(A) tail length estimation.

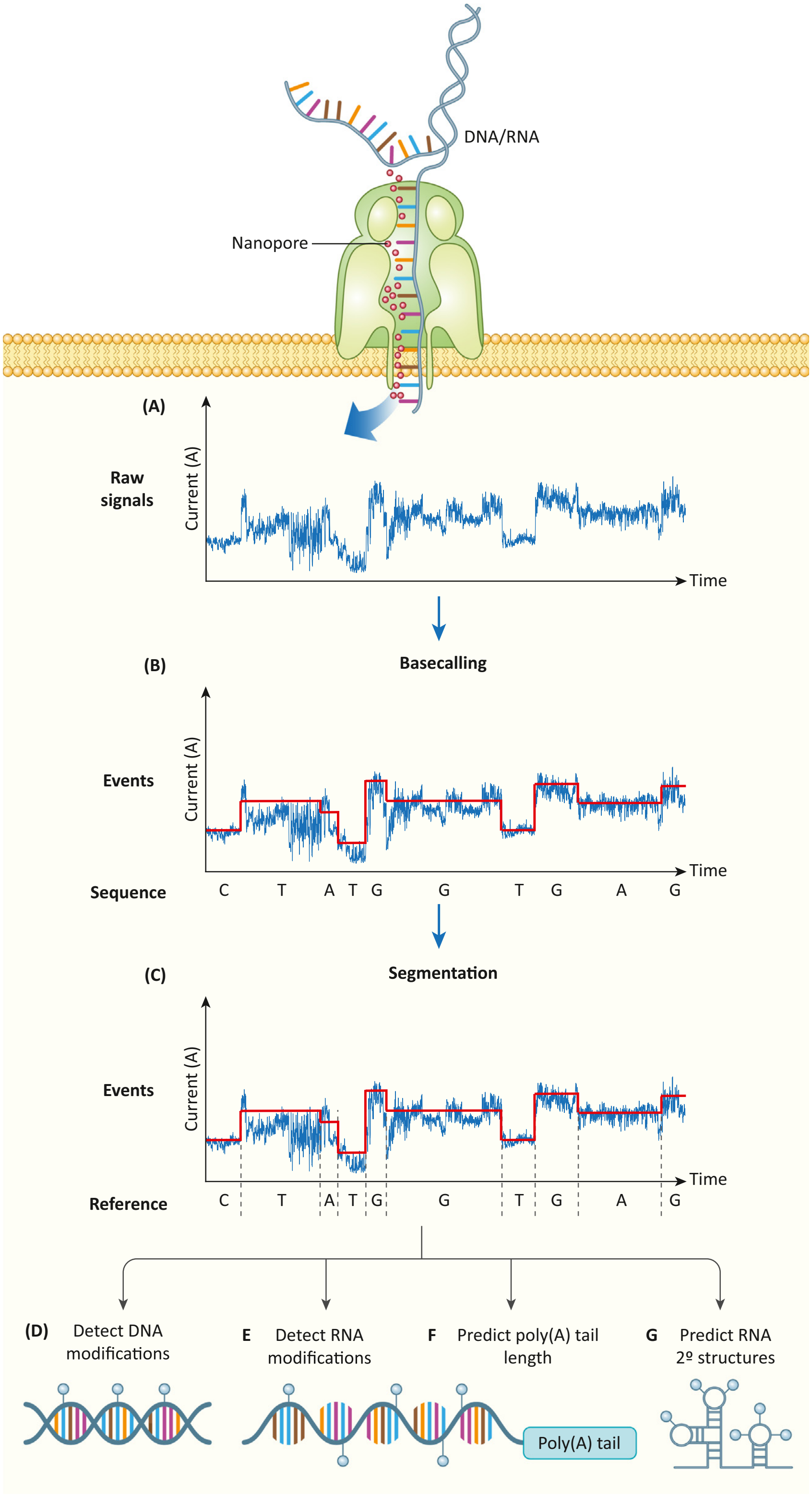
Nanopore sequencing – more than just sequences

High-throughput short-read sequencing has played a pivotal role in broadening our understanding of biology. Short-read sequencing technologies have advanced the understanding of genetic diversity [1,2], provided insights into transcriptomes and cell profiles in healthy populations [3,4], and helped deciphering disease biology [5–8]. On top of the nucleotide sequences are epigenetic modifications that influence gene expression [9] and **epitranscriptomic** (see Glossary) modifications that impact RNA processing, stability, and translation efficiency [10]. By coupling high-throughput sequencing with wet lab techniques, approaches such as MeRIP (methylated RNA immunoprecipitation)-seq [11], miCLIP (m6A individual-nucleotide-resolution cross-linking and immunoprecipitation)-seq [12], and bisulfite sequencing [13] allow the profiling of DNA and RNA modifications [14]. Although short-read sequencing on DNA and RNA are easily scalable strategies, the profiling of epigenetic and epitranscriptomic modifications involves highly specialized protocols.

Oxford Nanopore Technologies (ONT) provides a sequencing method (nanopore sequencing) that allows the profiling of genome and epigenome, or transcriptome and epitranscriptome with a single assay [15,16]. Nanopore sequencing generates long reads as each DNA or RNA molecule directly translocates through a nanopore. As the nucleic acids move through the nanopores in different nucleotide combinations, the changes in electrical current are measured (Figure 1A). This measured signal not only enables the determination of sequence bases, but also the detection of DNA and RNA modifications, and the prediction of poly(A) tail length and RNA secondary structures through computational methodologies developed for these purposes.

[1] Laboratory of Computational Transcriptomics, Genome Institute of Singapore, Singapore 138672
[2] Yong Loo Lin School of Medicine, National University of Singapore, Singapore
[3] Institute of Data Science, National University of Singapore, Singapore
[4] Department of Mathematics and Computer Science, Faculty of Science, Chulalongkorn University, Bangkok, Thailand
[5] Chula Intelligent and Complex Systems Research Unit, Chulalongkorn University, Bangkok, Thailand
[6] National Cancer Centre Singapore, Singapore

https://doi.org/10.1016/j.tig.2021.09.001



Trends in Genetics

(See figure legend at the bottom of the next page.)

*Correspondence:
gokej@gis.a-star.edu.sg (J. Göke).

Because of the complex nature of the nanopore raw ion current signal, computational methods that use approaches from machine learning have been key to extracting the additional layers of information. In this review, we will provide an overview of computational approaches that facilitate the analysis of nanopore signal data with a GitHub page of the tools described [i]. We will highlight the different machine learning concepts that advanced basecalling, illustrate how they are applied for targeted sequencing, and introduce supervised and unsupervised approaches for identifying DNA and RNA modifications, RNA secondary structure prediction, and poly(A) tail length estimation. Finally, we provide an outlook into the future directions that should further enable the discovery of complex biological information from nanopore signal analysis with computational methods.

Associating signal with sequence

Basecalling: from signals to nucleotides

Basecalling is the process that translates raw ion current signal data from nanopore sequencing to a sequence of bases (Figure 1B). The signal data correspond to the measured ion current changes from one nucleotide sequence of five (RNA) or six (DNA) bases (k-mer) to another during the translocation of a nucleic acid molecule through the pore. The noisy nature of the ion current signal makes determining the associating k-mers based solely on the signal data difficult as many of the k-mers share similar ranges of ion current signal values, which is especially true with the presence of homopolymers [17]. Early generation basecallers employ an error-prone and time-consuming segmentation process, which divides raw data series into k-mer-corresponding signal segments and translates these signal segments into k-mers [18]. These basecallers generate reads with an accuracy of 85% or lower [ii]. Since then, improvements in basecallers have been a major driver to increase nanopore sequencing accuracy, achieving over 98.3% of correctly identified bases[iii].

Hidden Markov model-based basecallers

The first basecallers including ONT's cloud-based Metrichor[iv] and the open-source software Nanocall [19], an offline alternative of Metrichor, utilize the hidden Markov model (HMM) for decoding the signal data. Assuming a nucleic acid moves through a pore one nucleotide at a time, these HMM-based basecallers treat the ion current signals as a chain of observable events while the k-mers as states within the HMM [20]. As the first nucleotides of each state overlap with the last nucleotides of the previous state, joint probabilities of a sequence of nucleotides can be calculated, and the path with the maximum total joint probability represents the final predicted sequence [20] (Figure 2A). To improve the accuracy of the predicted sequence, the basecalling algorithm PoreSeq introduces artificial mutations to the sequence and replaces short regions of the original best sequences with the same regions of the mutated sequence having a higher probability [21].

Recurrent neural network-based basecallers

As HMM basecallers predict sequences based on the short-range dependencies of one k-mer to its next, they may overlook the long-range dependencies in nanopore sequencing. Furthermore, using a nucleotide sequence model that inaccurately describes the expected current values of the k-mers can cause basecalling biases with HMM basecallers [22]. To overcome these constraints, ONT's Albacore (prior version 2.0.1)[v] and nanonet[vi], and the open-source software DeepNano [22] and BasecRAWller [23] use a **recurrent neural network (RNN)** framework for basecalling. A unidirectional RNN takes in information from the ion current input vector and the previous

Figure 1. Workflow of nanopore sequencing signal analysis and its applications. (A) Ion current signal over time for a nucleic acid sequenced with nanopore. (B) The basecalling process translates the ion current signal to sequences. (C) The segmentation process aligns the raw ion current signals to a reference sequence. (D–G) The applications of nanopore ion current signal analysis include the detection of (D) DNA or (E) RNA modifications, (F) the estimation of poly(A) tail length, and (G) the prediction of RNA secondary structure.

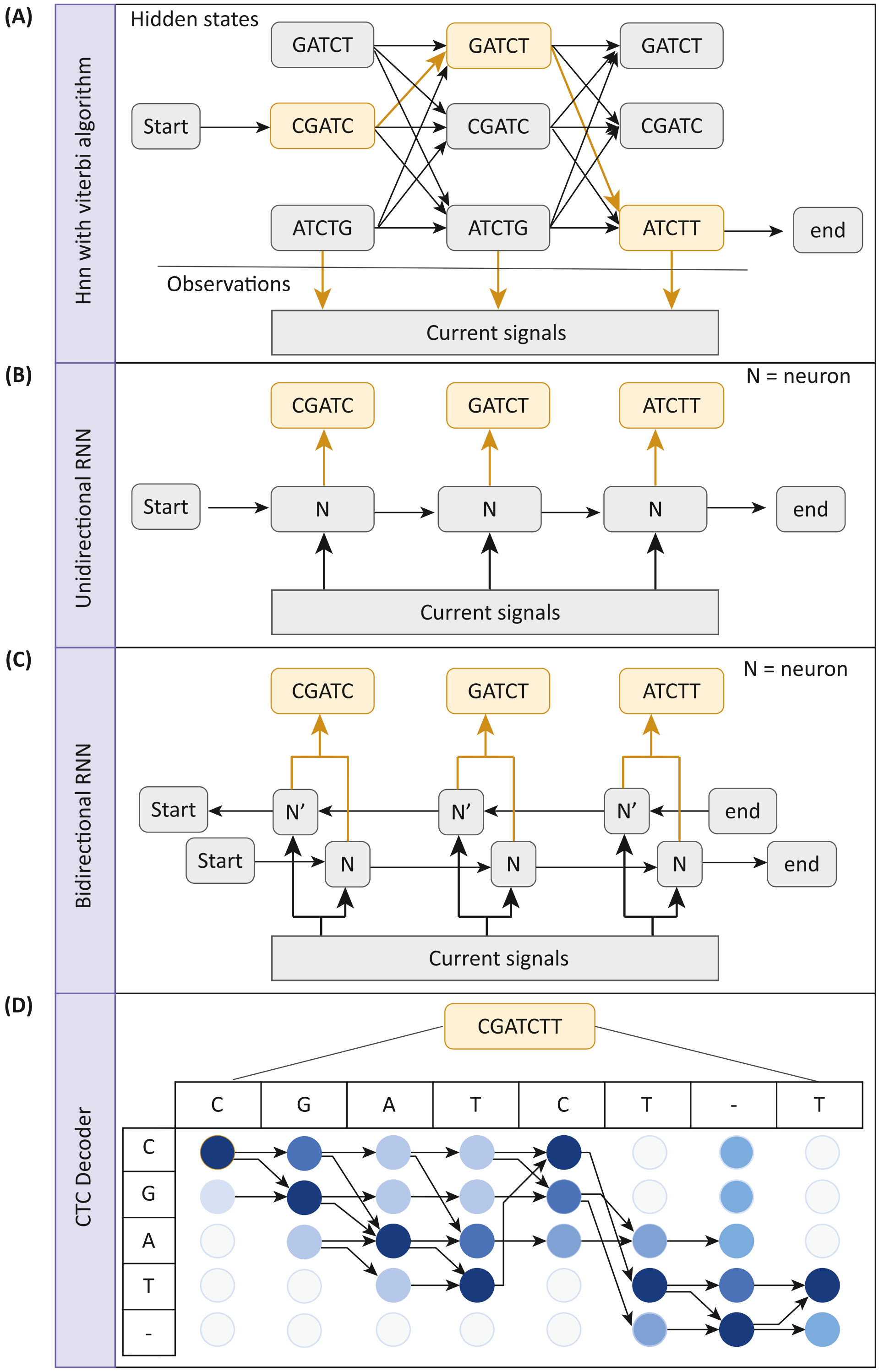
Glossary

Epitranscriptomics: the study of RNA modifications and their associated biological functions.

Multiple instance learning: a supervised learning approach for problems with weak knowledge about the labels of the training samples, where a model is trained with labels assigned to bags containing multiple instances.

Recurrent neural network (RNN): a type of artificial neural network that utilizes a data's sequential patterns from either the previous state (unidirectional) or both the previous and future states (bidirectional) to predict the next likely output.

Temporal convolutional network: a type of CNN that allows parallel processing and uses causal convolutions, where the output at each time point is based on elements from previous time points and the final output sequence has the same length as the input sequence.



Trends in Genetics

(See figure legend at the bottom of the next page.)

hidden state to calculate the current hidden state and the associated probability distribution of bases [22] (Figure 2B). Albacore, nanonet, and DeepNano use a bidirectional RNN, which incorporates information from previous and future states of the ion current input vector to improve prediction accuracy [22] (Figure 2C). Still, bidirectional RNNs are time consuming; therefore, BasecRAWller, with the aim to achieve real-time basecalling, uses two unidirectional RNNs to both segment and basecall the sequence in a streaming fashion, resulting in overall faster run time [23].

Segmentation-free basecallers

These early basecallers depend on the segmentation process to define boundaries of segments based on a sharp change in signals (Figure 1). Segmentation can be error prone due to the varying translocation speed and the noisy signal [18]. To address this, segmentation-free basecallers have been developed, such as ONT's Albacore version 2.0.1[v] and the open-source software Chiron [18]. To eliminate the segmentation step, Chiron combines a convolutional neural network (CNN) for extracting signal features and an RNN for predicting nucleotide probability. Then, it implements a connectionist temporal classification (CTC) decoder to select the base with the highest probability at each position (Figure 2D) and does many-to-one mapping to finalize the complete sequence [18].

Basecallers using convolutional networks and CTC

Although Chiron's segmentation-free approach outperforms the segmentation-dependent methods, the RNN framework's reliance on results from previous time points results in long running time. To speed up basecalling, Causalcall allows parallel processing by inputting segmented ion current measurements as a matrix into a **temporal convolutional network**, which models the ion current signal and calculates the nucleotide base occurrence probability at each time point. It uses a CTC decoder to output the base sequence with the highest probability for each fixed-size signal input and overlaps the base sequences to finalize the complete sequence [24]. The combination of a CNN and a CTC decoder is also used in ONT's research basecaller Bonito[vii], which has achieved an unprecedentedly high basecalling accuracy of 98.3%, making the accuracy of nanopore sequencing comparable to that of next-generation sequencing[iii].

Real-time mapping – selecting which reads to sequence

Nanopore sequencing has a unique ReadUntil feature that can eject reads in real time, and thereby free up the pore for sequencing specific reads of interest. To determine whether a read is a target in real time, ReadUntil requires rapid read classification based on as few nucleotides as possible from the reads. The ReadUntil feature can increase the sequencing depth for specific genomic regions, which enables targeted sequencing for applications such as sequencing-based diagnosis or novel microbial genome discovery from metagenomic samples [25,26].

Approaches that enable utilizing the ReadUntil feature includes Readfish [25], UNCALLED [26], and SquiggleNet [27]. The Readfish pipeline translates raw signals to nucleotide sequences in real time with guppy, aligns sequences to the reference, and then decides whether to eject the reads from the pores [25]. Similar success in real-time mapping is seen with the UNCALLED

Figure 2. HMM- and neural network-based signal analysis. (A) HMM-based basecallers traverse the hidden states in the HMM with the Viterbi algorithm. (B and C) RNN-based basecallers use either a unidirectional RNN or a bidirectional RNN framework, where the former takes in information from the previous state only, whereas the latter takes in information from both the previous and future states. (D) The CTC decoder is used in segmentation-free basecallers to output the most likely nucleotide at each position and uses an arbitrary letter (noted as '-' here) to indicate the same nucleotide at the next position, which will not get mixed up with duplicated nucleotide entries. Abbreviations: CTC, connectionist temporal classification; HMM, hidden Markov model; RNN, recurrent neural network.

algorithm [26]. UNCALLED first converts signals into events (k-mers) with an HMM and then searches through the reference genome for matches that are consistent with the event-matched k-mers. After clustering consistent reads and reference coordinates, UNCALLED filters out false positives and reports the best-supported location [26]. Using a neural-network framework, SquiggleNet uses a CNN and makes classification using a model that was learned on the reference training data [27]. These approaches have allowed the ReadUntil feature to classify target sequences in real-time.

The application of these methods can be used effectively for targeted sequencing of microbial genomes and human cancer genes [25,26], leading to an enrichment of the sequence of interest without the requirement for additional experiments.

Segmentation: aligning raw signals to reference genomic bases

Along with the basecalled sequences, downstream analyses of direct DNA and RNA sequencing also require reference sequence-aligned raw signals as inputs (Figure 1C). Segmentation describes the process that performs this raw signal-to-reference sequence alignment. Two methods for performing segmentation are tombo's resquiggle[viii] (previously nanoraw [28]) and nanopolish's eventalign [29]. Tombo's resquiggle first identifies event boundaries based on large shifts of signal level as these are associated with a change in the nucleotide that occupies the nanopore. Tombo then assigns these signal events to their corresponding reference sequences using a dynamic time warping algorithm. Nanopolish's eventalign assigns ion current signals to the reference sequence using an adaptive banded alignment that identifies the most likely sequence associated with the signal for each read. By aligning raw signals to a reference sequence, both tombo's resquiggle and nanopolish's eventalign allow the extraction of biological information from direct DNA and RNA sequencing with the nanopore signal data for downstream analyses.

Analyzing signals from direct DNA sequencing

Analyzing reference genome-aligned signals from direct DNA sequencing enables the extraction of biological information such as DNA modifications and/or chromatin accessibility (Figure 1D). The most common DNA modifications include N4-cytosine (4mC), 5-methylcytosine (5mC), 5-hydroxymethylcytosine (5hmC), and N^6-methyladenine (6mA) [28–37][x], which are known to regulate transcription and alter biological processes, with some of them also likely to have clinical relevance [35]. The detection of the GpC modification can also allow the profiling of chromatin accessibility [33]. Through analyzing direct DNA sequencing signals with supervised machine learning methods or statistical methods that do not rely on training data, computational tools can discover novel DNA modifications and infer chromatin accessibility in a high-throughput manner (Table 1).

Supervised learning methods for detecting specific DNA modifications and chromatin accessibility

To detect specific DNA modifications, supervised learning-based modification detection tools can be trained using data sets of experimentally validated modification sites (labeled data). Labels for modified cytosines, including the GpC modifications used for inferring chromatin accessibility [33], can be obtained through bisulfite sequencing [16,31–33] while m^{6}A labels can be obtained from artificially methylated nucleotides with methyltransferases [16] or orthogonal PacBio sequencing of naturally existing modifications [34,36].

Supervised learning methods for DNA modification identification include nanopolish, signalAlign, mCaller, DeepSignals, and tombo's detect_modifications module's alternative model mode [28][x] (Table 1). Nanopolish uses an HMM for detecting 5mC and has since expanded the model to detect endogenous CpG methylation and exogenous GpC methylation, where the detected GpC

Table 1. Overview of computational methods used for analyzing direct DNA and RNA sequencing data (a comprehensive list of nanopore analysis tools is available online[i])

Application	Tool	Data	Type/modification analysis	Refs
Infer chromatin accessibility	nanoNOMe (nanopolish extension)	DNA	CpG, GpC methylation	[33]
Detect DNA modification	nanopolish call-methylation	DNA	5mC	[31]
	SignalAlign	DNA	5mC,5hmC,6mA	[16]
	mCaller	DNA	6mA	[34]
	DeepSignals	DNA	6mA	[32]
	NanoMod	DNA	*De novo* DNA modification detection	[30]
Detect RNA modification	tombo detect_modifications	DNA/RNA	Alternate base detection	[28]
	MINES	RNA	m^{6A}	[42]
	EpiNano	RNA	m^{6A}	[43]
	Nanom6A	RNA	m^{6A}	[44]
	m6anet	RNA	m^{6A}	[45]
	nano-ID	RNA	5-EU	[46]
	nanoRMS	RNA	Ψ, Nm, and comparative RNA modification detection	[47]
	Yanocomp	RNA	Comparative RNA modification detection	[48]
	DiffErr	RNA	Comparative RNA modification detection	[ix]
	ELIGOS	RNA	Comparative RNA modification detection	[49]
	nanoDoc	RNA	Comparative RNA modification detection	[50]
	nanocompore	RNA	Comparative RNA modification detection	[51]
	DRUMMER	RNA	Comparative RNA modification detection	[52]
	xPore	RNA	Differential RNA modification rate analysis	[53]
Predict RNA 2° structure	nanoSHAPE	RNA	RNA structure (Nm, 2′-O-acetyl)	[54]
	PORE-cupine	RNA	RNA structure (NAI-N3)	[55]
Estimate poly (A) tail length	nanopolish polya	RNA	PolyA tails	[56]
	tailfindr	RNA	PolyA tails	[56,57]

methylation serves as a label for inferring chromatin accessibility in nanopolish's nanoNOMe extension [31,33]. Also utilizing an HMM, signalAlign expands modification detection to 5mC, 5hmC, and 6mA [16]. Both mCaller and DeepSignal use a neural network framework for modification detection, where mCaller detects 6mA with a neural network binary classifier [34] while DeepSignal constructs features from signals and sequences and uses a deep learning neural network classifier for methylation detection [32]. Using a statistical testing framework, tombo (alternative model mode) detects modifications based on the signal difference between a sample and the expected signal values of specific modified nucleotides, namely, 5mC and 6mA, provided by tombo [28][x].

Computational analysis of nanopore signal data allowed epigenetic profiling of *Drosophila* genome [37], human tandem repeat regions [38], human transposons [39], and the complete telomere-to-telomere assembly of the human X chromosome [40], extending the study of DNA modifications into regions otherwise not accessible by short-read sequencing.

Comparative and unsupervised learning methods for detecting DNA modifications

Comparative and unsupervised learning-based DNA modification detection tools can identify modified genomic positions by comparing naturally and/or artificially methylated DNAs with their unmethylated counterparts. These tools include tombo's detect_modifications module's *de novo* and sample compare modes [28][x] and NanoMod [30]. Tombo (*de novo* and sample compare modes) detects modifications by statistically testing the signal difference between a sample and the expected unmodified distribution in *de novo* mode or between samples from two groups in sample compare mode [28][x]. NanoMod takes *P* values of neighboring positions into account as a modified base may affect signals emitted by its neighbors [30]. Despite not being able to specify the modification detected, these statistical testing tools can provide insights when supervised methods cannot be applied [28,30][x]. NanoMod has been applied to detect various artificial thymine modifications even in the absence of training data [41].

Analyzing signal data from direct RNA sequencing

Direct RNA sequencing sequences RNAs with their modifications retained, which influence the ion current signals emitted from the nanopores. Analyzing these direct RNA sequencing signals with machine learning- and statistical testing-based tools allows rapid detection of naturally existing RNA modifications (Figure 1E), known to affect human diseases such as cancer [58], and artificial RNA modifications from chemical probing for RNA secondary structure prediction (Figure 1F), providing insights into RNA structure-influenced gene regulation [55]. Furthermore, as nanopore RNA sequencing overcomes the next-generation sequencing's short read length limitation (Figure 1G), poly(A) tail length can now be estimated at the isoform level, allowing a better understanding of how poly(A) tails affect gene expression [56,57] (Table 1).

Detecting naturally existing RNA modifications

To date, more than 168 post-transcriptional RNA modifications have been identified, and most of these identified modifications are on tRNAs and rRNAs [59,60]. The development of high-throughput assays to profiling RNA modifications has increased the number of modifications identified on mRNAs and noncoding RNAs, with some being disease related [58]. Utilizing direct RNA sequencing, RNA modification detection no longer requires laborious and bias-prone wet lab assays and has reached a high level of efficiency and accuracy.

Supervised learning methods for detecting specific RNA modifications

Starting with the most abundant mRNA modification, current supervised learning-based RNA modification detection tools have trained their models with data sets containing m^{6A} labels. These m^{6A}-labeled data sets can be obtained through wet lab assays such as m6ACE-seq and CLIP-seq [42,61] or artificially methylating adenosines with methyltransferases [43]. Such supervised methods include MINES [42], EpiNano [43], Nanom6A [44], and m6anet [45]. As m^{6A} is preferentially found in DRACH/RRACH motifs [12], MINES uses this aspect by modeling m^{6A} sites using a random forest classifier for each of the DRACH motifs using CLIP-seq-identified m^{6A} sites as positive samples [42]. Similarly, EpiNano and Nanom6A limit their analyses to RRACH motifs to increase specificity [43,44]. EpiNano trains a support vector machine that predicts candidate m^{6A} sites from basecalling errors that are presumably caused by the presence of m^{6A} [43] while Nanom6A trains an eXtreme Gradient Boosting (XGBoost) model with the raw signal features [44]. Predicting the probability of a site being modified with m^{6A}, m6anet employs a **multiple instance learning** framework and takes the entire differentially labeled reads into account [45].

Direct RNA-seq methods allow the discovery of specific RNA modifications beyond m^{6A}, including 5-methylcytosine, 5-ethynyluridine (5-EU), pseudouridine (Ψ), and 2′ O-methyl (Nm)[x] [46,47].

Tombo (alternative model mode) extended its 5-methylcytosine detection functionality, originally for DNA, to RNA[x]. nano-ID trained a neural network for 5-EU detection [46]. nanoRMS single mode uses basecalling errors to predicts the stoichiometries of Ψ and Nm [47].

Tombo's detect_modifications module's alternative model mode has contributed to the investigation of the coronavirus modification landscape [62–64] while EpiNano has successfully profiled the landscape of m^{6A} in polyadenylated nuclear RNAs [65] and unpolyadenylated plant circular RNAs [66]. Furthermore, the basecalling error features used in EpiNano and nanoRMS can be applied in detecting pseudouridine and inosine [47,67], illustrating some of the applications of machine learning-based tools in specific RNA modification profiling and their potential to expand the scope of RNA modifications that can be detected.

Comparative and unsupervised learning methods for detecting RNA modifications

Comparative and unsupervised learning-based tools can detect multiple kinds of modifications by comparing unmodified and modified samples. These tools include tombo (*de novo* and sample compare modes) [28][x], DRUMMER [52], nanocompore [51], xPore [53], Yanocomp [48], DiffErr[ix], nanoDoc [50], ELIGOS [49], and nanoRMS paired mode [47]. Tombo (*de novo* and sample compare modes) can perform on both DNA and RNA inputs[xii]. Specifically developed for RNA modification detection, DRUMMER, nanocompore, xPore, Yanocomp, DiffErr[ix], and nanoDoc model the distributions of the unmodified and modified samples and statistically test whether the ion current signal distributions of the two samples differ significantly [48,50,51,53,68]. Instead of comparing the differences in ion current distribution, ELIGOS and nanoRMS paired mode compare the error profiles of ONT signals between modified and unmodified samples [47,49]. Other than the ion current signal features, the dwell time parameter, the period of the nucleic acid in the pore, is used in nanocompore, nanoRMS, and nanoSHAPE, where the latter aims to detect 2′-O-methylation (Nm), which is prevalent in the 5′ cap region of mRNA [47,51,54]. Unlike the other methods, xPore can analyze differential RNA modifications from samples without an unmodified control. xPore infers the modification rate in each sample, thereby providing an estimate of the modification stoichiometry and enabling the quantitative comparison of RNA modifications from nanopore signal data [53]. Although most of these tools cannot specify the modification types, they can detect modified positions spanning various kinds of modifications. These tools have contributed to RNA modification discovery in multiple contexts. They include transcriptome-wide m^{6A} identification in human cell lines and clinical samples [53], coronavirus modification landscape investigation [64], and pseudouridine detection at known sites [47,51].

Detecting artificial RNA modifications for RNA secondary structures prediction

RNA folds into secondary and tertiary structures, serving as a mechanism for gene regulation [69]. RNA structure profiling has been done by adding artificial RNA modifications to the secondary structures with chemical reagents. Mutational profiling of the mutations induced by the chemically added modifications during next-generation sequencing library preparation can then provide insights into RNA structures [70,71]. To make this approach compatible with nanopore direct RNA sequencing, these reagents can be adopted to insert computationally detectable RNA modifications [54,55]. Structural probing for the nanopore platform include nanoSHAPE [54] and PORE-cupine [55]. nanoSHAPE's structural probing reagent introduces a smaller 2′-O-acetyl adduct, which can be detected through statistical testing of ion current signals and dwell time between modified and unmodified control samples [54]. PORE-cupine structural probes RNAs with a SHAPE-like reagent, where the modifications are detected by a one-class support vector machine trained using unmodified samples [55]. nanoSHAPE and

PORE-cupine are consistent with the SHAPE-MaP predictions. However, by combining structural probing with long reads these approaches facilitate the analysis of individual isoforms [55], demonstrating the ability of using direct RNA sequencing for high-throughput high-resolution RNA secondary structure profiling.

Estimating poly(A) tail length

During RNA processing, a poly(A) tail is added to an mRNA and influences the mature mRNA's nuclear export, stability, and translation efficiency [72,73]. Short-read transcriptome-wide poly(A) tail length estimation imposes a size limit and PCR biases [74]. While long reads overcome the size limitation, they are particularly error prone in homopolymer regions that make poly(A) tail length estimation challenging [75]. Hence, tools including nanopolish's polya [56] and tailfindr [57] utilize the translocation rate of a poly(A) tail for estimating its length [56,57]. Nanopolish's polya estimates poly (A) tail lengths with the estimated translocation rates while signals are being segmented [56]. With an alternative approach, tailfindr refines the boundaries of the poly(A) tail defined, based on the ONT adaptor location, and normalizes the boundaries with a read-specific nucleotide translocation rate [57]. Poly(A) tail lengths estimated by either nanopolish's polya or tailfindr are consistent with the expected poly(A) tail lengths from the control data sets used for method validation [56,57], and the measurement of isoform-level poly(A) tail length distribution has been applied to infer noncanonical poly(A) polymerase regulation [76].

Concluding remarks

By analyzing nanopore sequencing signals with machine learning algorithms, computational methods can reveal biological information such as poly(A) tail length and DNA or RNA modifications[i]. Furthermore, advances in the computational methods for basecalling nanopore reads have achieved comparable accuracy to short-read sequencing, reaching 98.3%.

While further improvements in the accuracy of computational methods are expected to be achieved, the increase in throughput and broader adaptation open new challenges as well (see Outstanding questions). Supervised learning algorithms rely on accurate training data, yet the influence of the training data on their performance has not yet been comprehensively evaluated. The species, nucleotide composition, or the process that generates training data can influence the accuracy, and systematic benchmark data sets will be essential to fully evaluate this aspect [77]. Besides improvements in accuracy, improvements in data handling will become central as the raw data are multifold larger than those obtained from short read data. Methods that improve compression of fast5 files and more space-efficient alternative file types for storing raw nanopore data are currently being developed [78,79][xi], and graphics processing unit acceleration is used routinely [80]. However, further improvements to reduce file sizes, standardizing file formats, and compute and memory-efficient algorithms will greatly reduce the barrier for larger-scale applications and adaptation.

Improvements in accuracy and more efficient data handling and processing, combined with the availability of larger data sets and systematic benchmarking studies, will facilitate the broad use of nanopore signal data analysis to extract the many diverse features of nucleic acids beyond their sequence. With additional studies that highlight the biological insights that can be obtained, such computational methods will be one of the key factors to make nanopore sequencing ion current signal data analysis a routine task in genomics.

Declaration of interests

No interests are declared.

Outstanding questions

Basecalling can be combined with modification detection. Can basecalling accuracy be further improved while expanding the panel of bases to be classified?

Training data have a major influence on machine learning tools for nanopore signal analysis. What are the best training data sets (synthetic, organism-specific, etc.), and how will they influence or improve the performance of these tools?

Can high compute and storage requirements of nanopore signal analysis methods be reduced to enable broad adaptation for routine genomics analysis of signal-level data?

What are the advantages and disadvantages of the different tools? Systematic benchmark studies and comprehensively annotated benchmark data sets can help to develop evaluation metrics and compare different methods.

Resources

[i] https://github.com/GoekeLab/awesome-nanopore

[ii] http://nanoporetech.com/about-us/news/new-research-algorithms-yield-accuracy-gains-nanopore-sequencing

[iii] http://nanoporetech.com/accuracy

[iv] https://metrichor.com

[v] https://community.nanoporetech.com

[vi] https://github.com/ProgramFiles/nanonet

[vii] https://github.com/nanoporetech/bonito

[viii] https://nanoporetech.github.io/tombo/resquiggle.html

[ix] https://github.com/bartongroup/differr_nanopore_DRS

[x] https://nanoporetech.github.io/tombo/

[xi] https://github.com/nanoporetech/vbz_compression

References

1. 1000 Genomes Project Consortium *et al.* (2015) A global reference for human genetic variation. *Nature* 526, 68–74
2. Wu, D. *et al.* (2019) Large-scale whole-genome sequencing of three diverse Asian populations in Singapore. *Cell* 179, 736–749.e15
3. GTEx Consortium (2013) The Genotype-Tissue Expression (GTEx) project. *Nat. Genet.* 45, 580–585
4. Regev, A. *et al.* (2017) The Human Cell Atlas. *eLife* 6, e27041
5. Weinstein, J.N. *et al.* (2013) The Cancer Genome Atlas Pan-Cancer analysis project. *Nat. Genet.* 45, 1113–1120
6. PCAWG Transcriptome Core Group *et al.* (2020) Genomic basis for RNA alterations in cancer. *Nature* 578, 129–136
7. ICGC/TCGA Pan-Cancer Analysis of Whole Genomes Consortium (2020) Pan-cancer analysis of whole genomes. *Nature* 578, 82–93
8. Hoadley, K.A. *et al.* (2018) Cell-of-origin patterns dominate the molecular classification of 10,000 tumors from 33 types of cancer. *Cell* 173, 291–304.e6
9. Allis, C.D. and Jenuwein, T. (2016) The molecular hallmarks of epigenetic control. *Nat. Rev. Genet.* 17, 487–500
10. Roundtree, I.A. *et al.* (2017) Dynamic RNA modifications in gene expression regulation. *Cell* 169, 1187–1200
11. Meyer, K.D. *et al.* (2012) Comprehensive analysis of mRNA methylation reveals enrichment in 3′ UTRs and near stop codons. *Cell* 149, 1635–1646
12. Linder, B. *et al.* (2015) Single-nucleotide-resolution mapping of m^{6}A and m^{6}Am throughout the transcriptome. *Nat. Methods* 12, 767–772
13. Frommer, M. *et al.* (1992) A genomic sequencing protocol that yields a positive display of 5-methylcytosine residues in individual DNA strands. *Proc. Natl. Acad. Sci. U. S. A.* 89, 1827–1831
14. Novoa, E.M. *et al.* (2017) Charting the unknown epitranscriptome. *Nat. Rev. Mol. Cell Biol.* 18, 339–340
15. Garalde, D.R. *et al.* (2018) Highly parallel direct RNA sequencing on an array of nanopores. *Nat. Methods* 15, 201–206
16. Rand, A.C. *et al.* (2017) Mapping DNA methylation with high-throughput nanopore sequencing. *Nat. Methods* 14, 411–413
17. Branton, D. *et al.* (2008) The potential and challenges of nanopore sequencing. *Nat. Biotechnol.* 26, 1146–1153
18. Teng, H. *et al.* (2018) Chiron: translating nanopore raw signal directly into nucleotide sequence using deep learning. *Gigascience* 7, giy037
19. David, M. *et al.* (2017) Nanocall: an open source basecaller for Oxford Nanopore sequencing data. *Bioinformatics* 33, 49–55
20. Timp, W. *et al.* (2012) DNA base-calling from a nanopore using a Viterbi algorithm. *Biophys. J.* 102, L37–L39
21. Szalay, T. and Golovchenko, J.A. (2015) *De novo* sequencing and variant calling with nanopores using PoreSeq. *Nat. Biotechnol.* 33, 1087–1091
22. Boža, V. *et al.* (2017) DeepNano: deep recurrent neural networks for base calling in MinION nanopore reads. *PLoS One* 12, e0178751
23. Stoiber, M. and Brown, J. (2017) BasecRAWller: streaming nanopore basecalling directly from raw signal. *bioRxiv* Published online May 1, 2017. https://doi.org/10.1101/133058
24. Zeng, J. *et al.* (2019) Causalcall: nanopore basecalling using a temporal convolutional network. *Front. Genet.* 10, 1332
25. Payne, A. *et al.* (2021) Readfish enables targeted nanopore sequencing of gigabase-sized genomes. *Nat. Biotechnol.* 39, 442–450
26. Kovaka, S. *et al.* (2021) Targeted nanopore sequencing by real-time mapping of raw electrical signal with UNCALLED. *Nat. Biotechnol.* 9, 431–441
27. Bao, Y. *et al.* (2021) Real-time, direct classification of nanopore signals with SquiggleNet. *bioRxiv* Published online January 20, 2021. https://doi.org/10.1101/2021.01.15.426907
28. Stoiber, M. *et al.* (2017) *De novo* identification of DNA modifications enabled by genome-guided nanopore signal processing. *bioRxiv* Published online December 15, 2016. https://doi.org/10.1101/094672
29. Loman, N.J. *et al.* (2015) A complete bacterial genome assembled *de novo* using only nanopore sequencing data. *Nat. Methods* 12, 733–735
30. Liu, Q. *et al.* (2019) NanoMod: a computational tool to detect DNA modifications using Nanopore long-read sequencing data. *BMC Genomics* 20, 78
31. Simpson, J.T. *et al.* (2017) Detecting DNA cytosine methylation using nanopore sequencing. *Nat. Methods* 14, 407–410
32. Ni, P. *et al.* (2019) DeepSignal: detecting DNA methylation state from nanopore sequencing reads using deep-learning. *Bioinformatics* 35, 4586–4595
33. Lee, I. *et al.* (2020) Simultaneous profiling of chromatin accessibility and methylation on human cell lines with nanopore sequencing. *Nat. Methods* 17, 1191–1199
34. McIntyre, A.B.R. *et al.* (2019) Single-molecule sequencing detection of N^6-methyladenine in microbial reference materials. *Nat. Commun.* 10, 579
35. Jin, Z. and Liu, Y. (2018) DNA methylation in human diseases. *Genes Diseases* 5, 1–8
36. Flusberg, B.A. *et al.* (2010) Direct detection of DNA methylation during single-molecule, real-time sequencing. *Nat. Methods* 7, 461–465
37. Shah, K. *et al.* (2019) Adenine methylation in *Drosophila* is associated with the tissue-specific expression of developmental and regulatory genes. *G3* 9, 1893–1900
38. Giesselmann, P. *et al.* (2019) Analysis of short tandem repeat expansions and their methylation state with nanopore sequencing. *Nat. Biotechnol.* 37, 1478–1481
39. Ewing, A.D. *et al.* (2020) Nanopore sequencing enables comprehensive transposable element epigenomic profiling. *Mol. Cell* 80, 915–928.e5
40. Miga, K.H. *et al.* (2020) Telomere-to-telomere assembly of a complete human X chromosome. *Nature* 585, 79–84
41. Georgieva, D. *et al.* (2020) Detection of base analogs incorporated during DNA replication by nanopore sequencing. *Nucleic Acids Res.* 48, e88
42. Lorenz, D.A. *et al.* (2020) Direct RNA sequencing enables mA detection in endogenous transcript isoforms at base-specific resolution. *RNA* 26, 19–28

43. Liu, H. *et al.* (2019) Accurate detection of mA RNA modifications in native RNA sequences. *Nat. Commun.* 10, 4079
44. Gao, Y. *et al.* (2021) Quantitative profiling of N-methyladenosine at single-base resolution in stem-differentiating xylem of *Populus trichocarpa* using nanopore direct RNA sequencing. *Genome Biol.* 22, 22
45. Hendra, C. *et al.* (2021) Detection of m6A from direct RNA sequencing using a Multiple Instance Learning framework. *bioRxiv* Published online September 22, 2021. https://doi.org/10.1101/2021.09.20.461055
46. Maier, K.C. *et al.* (2020) Native molecule sequencing by nano-ID reveals synthesis and stability of RNA isoforms. *Genome Res.* 30, 1332–1344
47. Begik, O. *et al.* (2021) Quantitative profiling of pseudouridylation dynamics in native RNAs with nanopore sequencing. *Nat. Biotechnol.* Published online May 13, 2021. https://doi.org/10.1038/s41587-021-00915-6
48. Parker, M.T. *et al.* (2021) Yanocomp: robust prediction of m^{6}A modifications in individual nanopore direct RNA reads. *bioRxiv* Published online June 16, 2021. https://doi.org/10.1101/2021.06.15.448494
49. Jenjaroenpun, P. *et al.* (2021) Decoding the epitranscriptional landscape from native RNA sequences. *Nucleic Acids Res.* 49, e7
50. Ueda, H. (2020) nanoDoc: RNA modification detection using nanopore raw reads with Deep One-Class Classification. *bioRxiv* Published online September 13, 2020. https://doi.org/10.1101/2020.09.13.295089
51. Leger, A. *et al.* (2019) RNA modifications detection by comparative nanopore direct RNA sequencing. *bioRxiv* Published online November 15, 2019. https://doi.org/10.1101/843136
52. Price, A.M. *et al.* (2020) Direct RNA sequencing reveals mA modifications on adenovirus RNA are necessary for efficient splicing. *Nat. Commun.* 11, 6016
53. Pratanwanich, P.N. *et al.* (2021) Identification of differential RNA modifications from nanopore direct RNA sequencing with xPore. *Nat. Biotechnol.* Published online July 1, 2021. https://doi.org/10.1038/s41587-021-00949-w
54. Stephenson, W. *et al.* (2020) Direct detection of RNA modifications and structure using single molecule nanopore sequencing. *bioRxiv* Published online June 01, 2020. https://doi.org/10.1101/2020.05.31.126763
55. Aw, J.G.A. *et al.* (2021) Determination of isoform-specific RNA structure with nanopore long reads. *Nat. Biotechnol.* 39, 336–346
56. Workman, R.E. *et al.* (2019) Nanopore native RNA sequencing of a human poly(A) transcriptome. *Nat. Methods* 16, 1297–1305
57. Krause, M. *et al.* (2019) tailfindr: alignment-free poly(A) length measurement for Oxford Nanopore RNA and DNA sequencing. *RNA* 25, 1229–1241
58. Barbieri, I. and Kouzarides, T. (2020) Role of RNA modifications in cancer. *Nat. Rev. Cancer* 20, 303–322
59. Boccaletto, P. and Bagiński, B. (2021) MODOMICS: an operational guide to the use of the RNA modification pathways database. *Methods Mol. Biol.* 2284, 481–505
60. Wetzel, C. and Limbach, P.A. (2016) Mass spectrometry of modified RNAs: recent developments. *Analyst* 141, 16–23
61. Koh, C.W.Q. *et al.* (2019) Atlas of quantitative single-base-resolution N^6-methyl-adenine methylomes. *Nat. Commun.* 10, 5636
62. Viehweger, A. *et al.* (2019) Direct RNA nanopore sequencing of full-length coronavirus genomes provides novel insights into structural variants and enables modification analysis. *Genome Res.* 29, 1545–1554
63. Kim, D. *et al.* (2020) The architecture of SARS-CoV-2 transcriptome. *Cell* 181, 914–921.e10
64. Miladi, M. *et al.* (2020) The landscape of SARS-CoV-2 RNA modifications. *bioRxiv* Published online July 18, 2020. https://doi.org/10.1101/2020.07.18.204362
65. Martin, S.E. *et al.* (2021) The m^{6}A landscape of polyadenylated nuclear (PAN) RNA and its related methylome in the context of KSHV replication. *RNA* 27, 1102–1125
66. Wang, Y. *et al.* (2020) Profiling of circular RNA N^6-methyladenosine in moso bamboo (*Phyllostachys edulis*) using nanopore-based direct RNA sequencing. *J. Integr. Plant Biol.* 62, 1823–1838
67. Ramasamy, S. *et al.* (2020) Chemical probe-based nanopore sequencing to selectively assess the RNA modifications. *bioRxiv* Published online May 21, 2020. https://doi.org/10.1101/2020.05.19.105338
68. Ding, H. *et al.* (2020) Gaussian mixture model-based unsupervised nucleotide modification number detection using nanopore-sequencing readouts. *Bioinformatics* 36, 4928–4934
69. Wan, Y. *et al.* (2011) Understanding the transcriptome through RNA structure. *Nat. Rev. Genet.* 12, 641–655
70. Siegfried, N.A. *et al.* (2014) RNA motif discovery by SHAPE and mutational profiling (SHAPE-MaP). *Nat. Methods* 11, 959–965
71. Yang, S.L. *et al.* (2021) Comprehensive mapping of SARS-CoV-2 interactions in vivo reveals functional virus-host interactions. *Nat. Commun.* 12, 5113
72. Fuke, H. and Ohno, M. (2008) Role of poly (A) tail as an identity element for mRNA nuclear export. *Nucleic Acids Res.* 36, 1037–1049
73. Eckmann, C.R. *et al.* (2011) Control of poly(A) tail length. *Wiley Interdiscip. Rev. RNA* 2, 348–361
74. Nilsen, T.W. (2015) Measuring the length of poly(A) tails. *Cold Spring Harb Protoc* 2015, 413–418
75. Rang, F.J. *et al.* (2018) From squiggle to basepair: computational approaches for improving nanopore sequencing read accuracy. *Genome Biol.* 19, 90
76. Bilska, A. *et al.* (2020) Immunoglobulin expression and the humoral immune response is regulated by the non-canonical poly(A) polymerase TENT5C. *Nat. Commun.* 11, 2032
77. Chen, Y. *et al.* (2021) A systematic benchmark of nanopore long read RNA sequencing for transcript level analysis in human cell lines. *bioRxi* Published online April 22, 2021. https://doi.org/10.1101/2021.04.21.440736
78. Chandak, S. *et al.* (2020) Impact of lossy compression of nanopore raw signal data on basecalling and consensus accuracy. *Bioinformatics* 36, 5313–5321
79. Gamaarachchi, H. *et al*. SLOW5: a new file format enables massive acceleration of nanopore sequencing data analysis. *bioRxiv* Published online June 30, 2021. https://doi.org/10.1101/2021.06.29.450255
80. Gamaarachchi, H. *et al.* (2020) GPU accelerated adaptive banded event alignment for rapid comparative nanopore signal analysis. *BMC Bioinforma.* 21, 1–13