bioRxiv preprint doi: https://doi.org/10.1101/2024.01.11.575207; this version posted March 6, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Raw signal segmentation for estimating RNA modifications and structures from Nanopore direct RNA sequencing data

Guangzhao Cheng, Aki Vehtari, and Lu Cheng*

Department of Computer Science, Aalto University, Finland

*Correspondence to: lu.cheng.ac@gmail.com

Abstract

Estimating RNA modifications from Nanopore direct RNA sequencing data is an important task for the RNA research community. Current computational methods could not provide satisfactory results due to the inaccurate segmentation of the raw signal. We develop a new method, SegPore, that utilizes a molecular jiggling translocation hypothesis to segment the raw signal. SegPore is a pure white-box model with a superior interpretability, which significantly reduces structured noise in the raw signal. Based on the improved signal segmentation, SegPore+m6Anet has achieved state-of-the-art performance in m6A identification. Additionally, we demonstrate SegPore's interpretable results and decent performances on inosine modification estimation and RNA secondary structure estimation. An interesting discovery in RNA structure estimation is that the end points of the reads take place at the start of stem structures along the reverse transcription direction. Our results indicate SegPore's capability to concurrently estimate multiple modifications at the individual molecule level from the same Nanopore direct RNA sequencing data, as well as shed light on RNA structure estimation from a novel angle.

Introduction

RNA modifications play important roles in different diseases, such as Acute Myeloid Leukemia (Yankova, Aspris and Tzelepis 2021) and Fragile X Syndrome (Prieto, Folci and Martin 2020), as well as biological processes, such as cell differentiation (Bellodi et al. 2013; Lee et al. 2019) and immune response (Quin et al. 2021). To date, researchers have identified over 150

bioRxiv preprint doi: https://doi.org/10.1101/2024.01.11.575207; this version posted March 6, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

different types of RNA modifications (Boccaletto et al. 2022; Chen, Owens and Liu 2023; Zimna et al. 2023; Ohira and Suzuki 2024), highlighting the complexity and diversity of RNA regulation. RNA modifications are pivotal for RNA secondary structures, e.g., modifications in tRNA are essential for accurate and efficient decoding of the mRNA codons into proteins (Agris et al. 2017). N1-methylpseudouridine (m1Ψ) is used to enhance the efficacy of COVID-19 mRNA vaccines (Nance and Meier 2021), which showcases the practical applications of RNA modifications.

To identify RNA modifications, researchers generally resort to immunoprecipitation methods such as MeRIP-Seq (Meyer et al. 2012), miCLIP (Linder et al. 2015) and m6ACE-Seq (Koh, Goh and Goh 2019), where a m6A antibody is used to target m6A in these protocols. Like ChIP-seq, we could infer m6A locations on the transcripts from the generated next-generation sequencing (NGS) data. RNA structure estimation protocols such as SHAPE-MaP (Siegfried et al. 2014) follow a similar idea by attaching chemical adducts to nucleotides in loop regions but not stem structures, where the chemical adducts are treated as modifications. A reactivity score profile obtained from the sequencing data is then used to estimate the RNA structure. However, there are several limitations of these NGS-based RNA modification estimation methods. First, a modification-specific antibody is needed for each modification, while there are more than 150 modifications. Second, these methods could not directly measure RNA modifications, i.e., they infer m6A locations on transcripts from the NGS data. As a result, researchers could not get modification estimations on single molecules. The direct RNA sequencing (DRS) of Oxford Nanopore Technologies (ONT) is a new sequencing technology that could address both challenges.

The direct RNA sequencing measures the current intensity as an RNA molecule translocates through the pore. For ONT pore version R9.4, five nucleotides reside in the pore, which is termed as *5mer* (Jain et al. 2018). The 5mer blocks certain numbers of ions passing through the pore, the amount of which is measured as the current intensity in a unit of time. During the

bioRxiv preprint doi: https://doi.org/10.1101/2024.01.11.575207; this version posted March 6, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

translocation process, each 5mer stays in the pore for a short period, and the current measurement varies around a baseline level. Due to differences in size, shape, order, and chemical properties of the nucleotides of a 5mer, the current level varies across different 5mers. The basic idea of Nanopore sequencing is to infer the original nucleotides from the currents of different 5mers. As a reference, ONT has provided the mean and standard deviation (std) of the current intensity for each 5mer of RNA based on their training data. Since nucleotides with a larger size could block more ions, e.g., Adenine and Guanine, their current signal levels are lower than Cytosine and Uridine.

The raw signal of DRS could be utilized to estimate RNA modifications and RNA structures. DRS directly records the raw current signals for each nucleotide on the RNA molecule. For a normal nucleotide and its modification, e.g., Adenosine and m6A, different current signals are generated due to their chemical property difference. In theory, we could infer each nucleotide and its modification state from the raw signals of a RNA molecule. Given the capability of estimating RNA modifications, researchers could estimate RNA structures from DRS data (Stephenson et al. 2022). However, these tasks are challenging from the computational perspective.

The central computational problem of DRS is the segmentation of raw current signal. The raw current signal is split into segments and assigned to corresponding 5mers, which determines the RNA sequence, as well as the modifications. Current mainstream methods for segmenting raw signals are represented by Nanopolish (Loman, Quick and Simpson 2015; Simpson et al. 2017) and Tombo (Stoiber et al. 2017). Nanopolish employs a Hidden Markow Model (HMM) for the raw signal segmentation, while Tombo segments the raw signal by scanning for large signal shifts. Due to the lack of the RNA translocation process modelling, these methods could not provide accurate segmentation results. Consequently, the derived raw signal segments contain unwanted noises that deteriorate the performance of downstream tasks such as RNA modification and structure estimation.

bioRxiv preprint doi: https://doi.org/10.1101/2024.01.11.575207; this version posted March 6, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

The second challenge is that an RNA modification may only induce a minor change in the current signal of the corresponding 5mer. This means we can only distinguish 5mers with significant changes in their current signal, i.e., those with a high signal-to-noise ratio. This problem would be mitigated if we could reduce the noise level in the current signal. In other words, downstream tasks such as estimation of RNA modifications and structures, may achieve a better performance given the de-noised raw signal segments. Therefore, it is necessary to model the physical DRS process to gain a better understanding of the noises, which is missing in current RNA modification estimation methods.

Moreover, current RNA modification methods are hindered by limitations in data availability and interpretability. Present RNA modification detection methods can be classified into two groups: comparison-mode group and single-mode group (Zhong et al. 2023). The comparison-mode group, such as DiffErr (Parker et al. 2020), DRUMMER (Price et al. 2020), xPore (Pratanwanich et al. 2021) and Nanocompore (Leger et al. 2021), compares modified and unmodified samples, which requires both modification-abundant and modification-sparse datasets for the model training. The single-mode group, such as m6Anet (Hendra et al. 2022), employs deep neural networks (DNN) to classify modified and unmodified nucleotides by feeding various features (e.g., mean, std, dwell time) of the raw signal segments provided by Nanopolish or Tombo. The training data are generated from modification-abundant samples, where the signal segment features are used as input and estimated modifications given by NGS-base methods are used as the labels. Note that current methods are modification-specific, i.e. one method only handles one type of modification. As a result, the amount of training data increases linearly as the number of RNA modification types to be studied. Additionally, we may not be able to generate the training data due to technical reasons, e.g. removing a certain type of RNA modification from the samples. The RNA modification estimation task would be much simpler if we could take the mean of a raw signal segment and compare it against a reference table to determine its modification state. This brings out the interpretability problem of current DNN-based methods, which utilize the input features in a

bioRxiv preprint doi: https://doi.org/10.1101/2024.01.11.575207; this version posted March 6, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

complicated way for the modification estimation. It is impossible to gain straightforward interpretations why an input raw signal segment corresponds to its estimated modification state.

To address the above challenges, we have developed SegPore, which is capable of diminishing noise using a novel signal segmentation algorithm and offers the flexibility to estimate various modifications by a customized 5mer parameter table. Due to the change of the segmentation algorithm, we have re-implemented almost all components of the traditional RNA modification estimation workflow. SegPore consists of (1) a hierarchical hidden Markov model (HHMM) for segmenting the raw current signal of DRS, (2) alignment algorithms for aligning raw current signal segments with reference sequence, and (3) a Gaussian mixture model (GMM) for RNA modification estimation. SegPore provides interpretable segmentation results, which facilitates various downstream analyses. We demonstrate its capabilities in the estimation of RNA modifications (m6A and inosine) and RNA secondary structure.

Results

RNA translocation hypothesis

It is challenging to segment the raw current signal due to its complexity. To solve this problem, we need to know how exactly the RNA molecule moves through the pore. In traditional basecalling algorithms such as Guppy and Albacore, we implicitly assume that the RNA molecule is translocated through the pore by the motor protein in a monotonic fashion, i.e., the RNA is pulled through the pore unidirectionally. In the DNN training process of Guppy and Albacore, we try to align the current signal with the reference RNA sequence. The alignment is unidirectional, which is the source of the implicit monotonic translocating assumption.

bioRxiv preprint doi: https://doi.org/10.1101/2024.01.11.575207; this version posted March 6, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

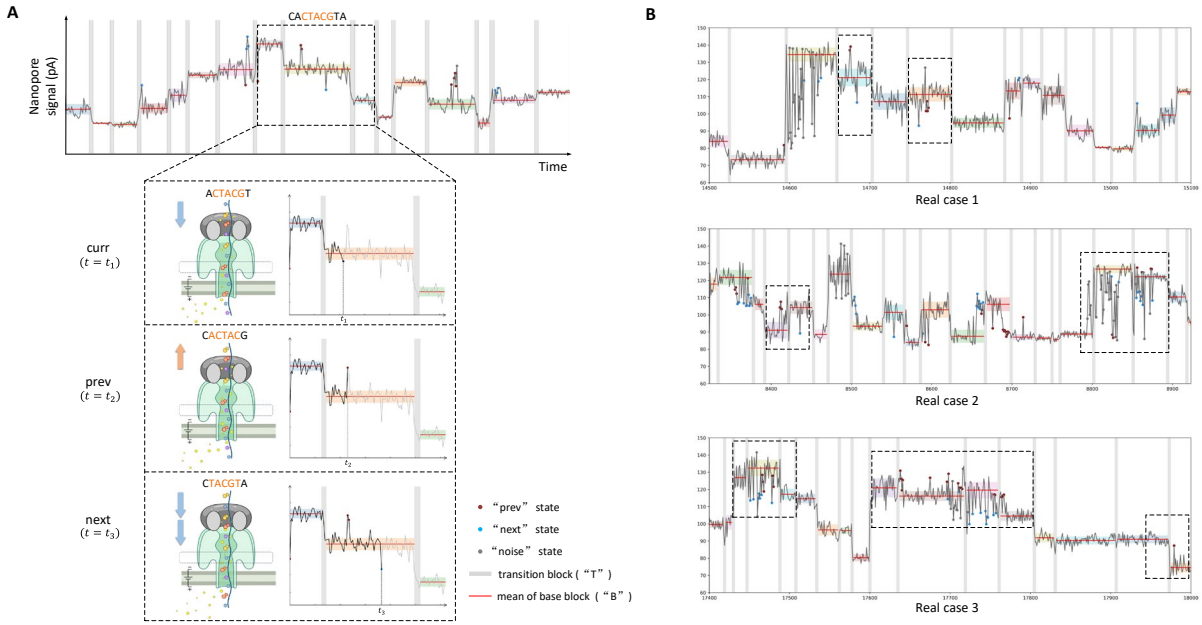

Figure 1. RNA translocation hypothesis. (*A*) Jiggling RNA translocation hypothesis. Top panel shows the raw current signal of Nanopore direct RNA sequencing. The gray areas are SegPore estimated transition blocks. Here we focus on three neighboring 5mers and consider the center 5mer (CTACG) as the current 5mer. The RNA molecule might be translocated forward or backward for a short period during the translocation process of the current 5mer. If the RNA molecule is pulled backward, the previous 5mer ("prev" state) is placed in the pore and the current signal is similar to the previous 5mer's baseline (mean and std highlighted by red lines and shades). If the RNA is pushed forward, the current signal is similar to the next 5mer's baseline ("prev" state). (*B*) Example raw current signals that supports this jiggling hypothesis, with dash rectangles highlighting relevant base blocks. Measurements assigned to the previous and next 5mer are highlighted as red and blue points. It is obvious to observe that red points are close to the previous 5mer's baseline and blue points are close to the next 5mer's baseline. The raw current signals were extracted from mESC WT samples of the training data in the m6A benchmark experiment.

However, the raw current data suggests that the motor protein translocates RNA back and forth. Fig. 1B shows several example fragments of DRS raw current signal (Zhong et al. 2023), each of which roughly corresponds to three neighboring 5mers. The highlighted spikes have similar current intensities, either as the previous 5mer or the next 5mer. Similar patterns are widely observed across the whole data. This suggests that the RNA molecule may move forward and backward while passing through the pore. This observation can also be supported by previous reports (Caldwell and Spies 2017; Craig et al. 2017), in which the helicase (the motor protein) translocates the DNA strand through the nanopore in a back-and-forth manner.

Based on the reported kinetic model (Craig et al. 2017), we hypothesize that the RNA is translocated through the pore in a jiggling manner. On average, the motor protein sequentially translocates 5mers on the RNA strand forward, and each 5mer resides in the pore for a short

bioRxiv preprint doi: https://doi.org/10.1101/2024.01.11.575207; this version posted March 6, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

time. During the short period of a single 5mer, the motor protein may swiftly drive the RNA molecule forward and backward by 0.5~1 nucleotide in the translocation process of the current 5mer (Fig. 1A), which makes the measured current intensity occasionally similar to the previous or the next 5mer. When the motor protein does not move the RNA molecule, we assume that the 5mer inside the pore oscillates thermodynamically, generating current intensities around its baseline. Additionally, we assume there is a sharp change in the current intensity between two consecutive 5mers, which serves as their boundary.

We also assume that the raw current signal of a read can be segmented into a series of alternating base and transition blocks (Fig. 1). In the ideal case, a base block corresponds to the base state where the 5mer resides in the pore and jiggles between neighboring 5mers, i.e., the current 5mer can transiently jump to the previous or the next 5mer. A transition block corresponds to the transition state between two consecutive base states where one 5mer translocates to the next 5mer in the pore. The current signal should be relatively flat in the base blocks, while a sharp change is expected in the transition blocks. In practice, the number of transition blocks is generally overestimated since we are not able to distinguish the transitions within a base block and between base blocks. As a result, multiple base blocks may correspond to only one 5mer, which is determined by the SegPore alignment algorithms.

SegPore Workflow

The SegPore workflow (Fig. 2) contains six steps: (1) preprocess fast5 files to pair the raw current signal segments with reference RNA sequence fragments, (2) segment each raw current signal using the hierarchical hidden Markov model (HHMM) into base and transition blocks, (3) align the derived base blocks with the paired RNA sequence, (4) estimate the modification state for each 5mer of the RNA sequence, (5) fit a two-component Gaussian Mixture Model (GMM) for each unique 5mer across different RNA reference sequences, and (6) use results of step 5 to update relevant parameters. Step 3~6 will iterate several times until

bioRxiv preprint doi: https://doi.org/10.1101/2024.01.11.575207; this version posted March 6, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

171 the estimated parameters stabilize. Detailed descriptions are provided in Methods and
172 Supplementary Note 1-3.

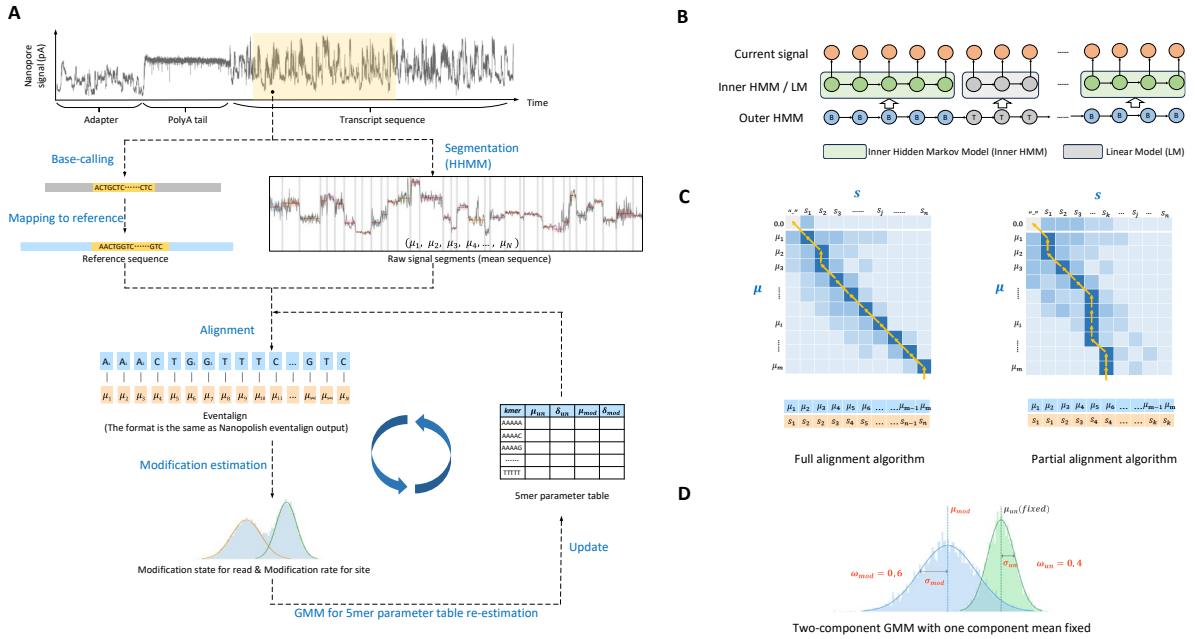

Figure 2. SegPore workflow. (*A*) The general workflow. First, basecalling and mapping are performed using Guppy and Minimap2 such that a raw current signal fragment is paired with a reference sequence fragment. Meanwhile, the raw current signal of a read is split into segments by HHMM and an estimated mean (μ_i) is derived for each segment. Then, the current signal segments (σ_i) are aligned with the 5mer list of the corresponding reference sequence fragment using the full/partial alignment algorithm, given a 5mer parameter table. Here we use A_j to denote A at jth position on the reference. Next, all aligned to the same 5mer at different genomic locations are pooled together and a two-component GMM is fitted to re-estimate the 5mer parameters. One GMM component models the unmodified state and the other models the modified state, while the hidden variable of the GMM specifies the modification state of the 5mer on each read. The parameter estimation process is iterated several times on the training data to gain a stable estimation of the 5mer parameter table. The final 5mer parameter table is used for estimating the modification states on the test data. (*B*) Hierarchical hidden Markov model. The outer HMM partitions current signal into alternating base blocks and transition blocks. An inner HMM approximates the emission probability of a base block by considering neighboring 5mers. A linear model approximates the emission probability of a transition block. (*C*) Full/partial alignment algorithms. Each row is an estimated mean of a base block given by HHMM. Each column is a 5mer of the reference sequence. One 5mer can be aligned to multiple means. (*D*) Gaussian mixture model (GMM) for estimating modification states. The green component codes the unmodified state of a given 5mer. The blue component codes the modified state of the given 5mer. Each component has three parameters: mean (μ), std (σ) and weight (ω).

173 The final outcomes of SegPore are the “eventalign” and modification state estimation. The
174 SegPore eventalign is similar to the output of Nanopolish “eventalign” command, which is the
175 pairing between raw current signal segments and 5mers of the corresponding RNA reference
176 sequences. For selected 5mers, SegPore outputs the modification rate for each site and the
177 modification state of the given site on each read. The key element of SegPore is the 5mer
178 parameter table, which specifies the mean and std of each 5mer in an unmodified or modified
179 state (Fig. 2A). Since the peaks (modified and unmodified state) are separable for a limited

bioRxiv preprint doi: https://doi.org/10.1101/2024.01.11.575207; this version posted March 6, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

number of 5mers, we could only obtain the modification parameters for these 5mers in 5mer parameter table, i.e. SegPore could not provide the modification state estimation for other 5mers.

m6A identification

We demonstrate SegPore's performance on raw data segmentation and m6A identification using independent public datasets as training and test data. It is well known that there are 18 DRACH (where D denotes A, G or U, and H denotes A, C or U) motifs for m6A (Linder et al. 2015). In this study, we concentrate on estimating the m6A modification on the DRACH motif.

We first estimate the 5mer parameter table of m6A modification from the training data using public Nanopore direct RNA sequencing (DRS) data from three wild type samples of HEK293T cells (Pratanwanich et al. 2021). To get the 5mer parameter table for m6A modification, we concatenate fast5 files of all samples and run the full SegPore workflow. We iterate the parameter estimation process five times to gain a stabilized 5mer parameter table, which contains modification parameters for ten 5mers. Their modification state distribution significantly differs from the unmodified state distribution, with moderate support of read counts and genomics locations.

Next, we perform segmentation and m6A identification on test data, which is the DRS data from wild type mouse embryonic stem cells (mESCs) in a benchmark study (Zhong et al. 2023). Zhong et al. benchmarked a set of tools with different input requirements. Considering the similarity of the methods and input requirements, we select single-mode tools as baselines: Tombo, MINES (Lorenz et al. 2020), Nanom6A (Gao et al. 2021), m6Anet, and Epinano (Liu et al. 2019). We process the test data using standard SegPore workflow with 5mer parameter table estimated from the training data. As a result, we get SegPore eventalign for all 5mers, modification rates for selected sites (differentiable motifs), modification states of these sites

bioRxiv preprint doi: https://doi.org/10.1101/2024.01.11.575207; this version posted March 6, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

206 on each read. To demonstrate the performance of SegPore eventalign on downstream tasks,
207 we feed the SegPore eventalign results to m6Anet, which is termed as SegPore+m6Anet.

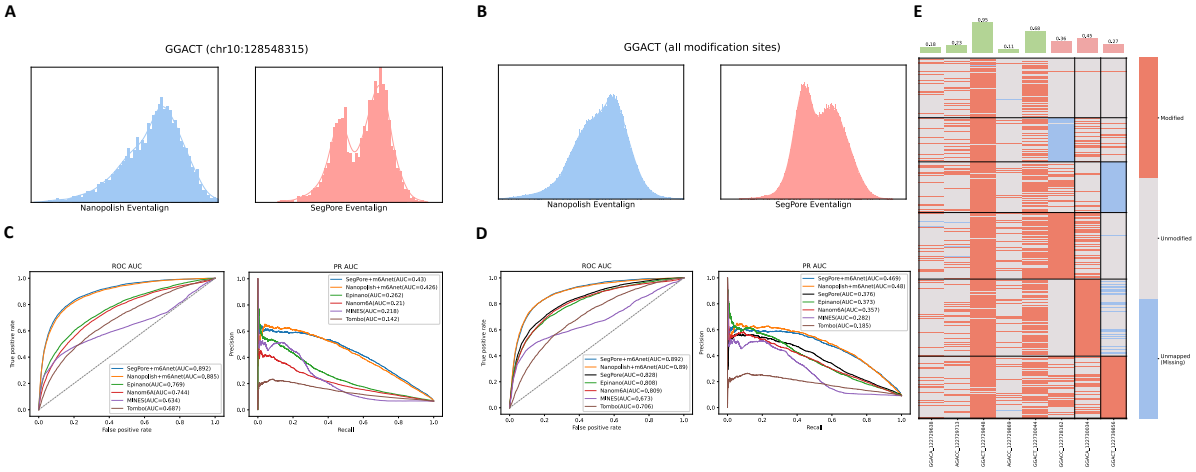

Figure 3. m6A identification. (*A*) Histogram of current signal mapped to an example modified genomic location (chr10:128548315, GGACT) across all reads in training data given by Nanopolish (left) and SegPore (right). (*B*) Histogram of current signal mapped to GGACT at all annotated m6A genomic locations of the training data given by Nanopolish (left) and SegPore (right). (*C*) Benchmark results on all DRACH motifs. (*D*) Benchmark results on six selected motifs. (*E*) Modification rate of selected genomic locations (upper panel) and modification states of all reads mapped to an example gene (ENSMUSG00000003153, lower panel). The black borders in the heatmap highlight the bi-clustering results.

208 SegPore demonstrates the improved segmentation results compared with Nanopolish. Fig.
209 3A shows the eventalign results of Nanopolish and SegPore at an example genomic location
210 with m6A modifications. The histogram shows the distribution of the raw signal segment mean
211 of all reads mapped to the genomic location. Compared with Nanopolish eventalign results,
212 the bimodal distribution is more obvious in SegPore results. We next pooled all reads mapped
213 to modification genomic locations (based on the ground truth) for the classical m6A motif
214 “GGACT”. Fig. 3B illustrates the current signal distribution over all reads mapped to these
215 locations, where prominent peaks are observed in SegPore but not in Nanopolish. The results
216 suggest that SegPore is able to remove a certain degree of noise from the raw current signal,
217 which makes the modification distribution clearer.
218
219 SegPore exhibits decent performance on m6A identification on the test data. Given the ground
220 truth miCLIP2 (Kortel et al. 2021) data, we calculate the area under the curve (AUC) of the
221 receiver operating characteristic (ROC) curve and precision-recall (PR) curve for each

bioRxiv preprint doi: https://doi.org/10.1101/2024.01.11.575207; this version posted March 6, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

method. Fig. 3C shows the benchmark results on all DRACH motifs, where SegPore+m6Anet shows the best performance (ROC AUC=89.2%, PR AUC=43.0%). Next, we demonstrate SegPore's modification estimation performance on selected 5mers. We selected six m6A motifs ("GGACT", "GGACA", "GGACC", "AGACA", "AGACC", "AGACT") based on the following two criteria (1) modified and unmodified peaks are significantly different and (2) the 5mer is both abundant in the training and test data (Supplementary Fig. 1). As shown in Fig. 3D, SegPore's ROC AUC is 82.8%, where the best is 89.2% (SegPore+m6Anet). SegPore's PR AUC is 37.6%, where the best is 48.0% (Nanopolish+m6Anet). The results suggest the decent performance of SegPore in m6A identification.

SegPore naturally identifies m6A modifications at the single molecule level. Fig. 3E demonstrates high modification genomic locations (modification rate > 0.1, also supported by ground truth) of an example gene ENSMUSG00000003153, where rows are reads and columns are genomic locations. Biclustering was performed on the heatmap to illustrate the modification patterns, which resulted in 6 clusters of the reads and 3 clusters of the genomic locations. The heatmap suggests that modifications at the 6th, 7th, and 8th genomic locations are specific to cluster 4, 5, 6 of the reads, which correspond to different modification mechanisms. It is also obvious that the majority of reads at the 3rd and 5th genomic locations are modified across other clusters.

Inosine identification

Thanks to the flexibility of the 5mer parameter table, SegPore can be easily extended to estimate other RNA modifications using the same computational framework. We next test the performance of SegPore on inosine identification, which is another RNA modification of adenine (Slotkin and Nishikura 2013; Nishikura 2016; Eisenberg and Levanon 2018; Nguyen et al. 2022). The *in virto* transcription (IVT) data and H9 human embryonic stem cells (hESCs) data in a public dataset (Nguyen et al. 2022) were used for this task. The 5mer parameter table of inosine was obtained from IVT data, and the H9 data was for testing.

bioRxiv preprint doi: https://doi.org/10.1101/2024.01.11.575207; this version posted March 6, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

First, we explain the training process. The IVT data contained eight pure-G samples and eight pure-I samples, which were merged into one pure-G and one pure-I sample, respectively. The pure-G sample was an unmodified sample where A, C, G, U nucleotides were used in the synthesis of a designed RNA molecule. The designed sequence contains 81 different XXGXX 5mers, where X is A, C or U/T. The pure-I sample was a modified sample where A, C, I (inosine), U were used for synthesizing the same RNA molecule. The “G” nucleotides in the RNA sequence could either be replaced by A or I nucleotides, which provided data to infer distributions of 5mers in the form of XXA/IXX. We run standard SegPore workflow on the pure-G sample and pure-I sample to get a stabilized 5mer parameter table, which contains modification states for 61 selected 5mers (Supplementary Fig. 2A). Note that the mean of XXGXX and XXAXX are fixed, whose values are specified by the 5mer parameter table of ONT.

The derived 5mer parameter table nicely depicts the selected 5mer distributions. Fig. 4A shows three example patterns of 5mer distributions, which are the densities derived from raw signal segments aligned to the selected 5mers on the reference sequence. We obtain the XXGXX distributions from the pure-G sample and XXA/IXX distributions from the pure-I sample. For the pure-G sample, we expect only one peak for XXGXX, while we expect two peaks for XXAXX and XXIXX in the pure-I sample. There are three general patterns: (1) G is replaced by A or I in the pure-I sample, but we cannot differentiate the peaks of XXAXX and XXIXX (top panel) (2) G is replaced by A or I, and the peaks of A and I can be clearly identified (center panel) (3) G is only replaced by I and the peak of I is significantly different to the ONT reference A peak (bottom panel). The selected 5mers belong to the second and third patterns (Supplementary Fig. 2A). Supplementary Fig. 2B shows that the mean of XXIXX is generally between that of XXAXX and XXGXX, which agrees with the findings in the original publication (Nguyen et al. 2022).

bioRxiv preprint doi: https://doi.org/10.1101/2024.01.11.575207; this version posted March 6, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

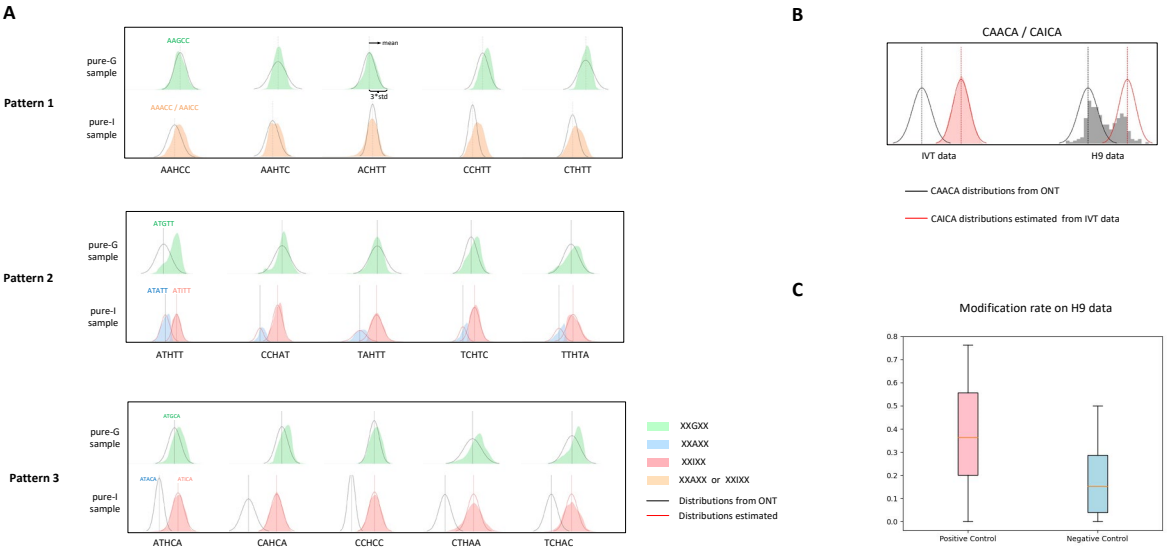

Figure 4. Inosine identification. (*A*) Current signal distributions of example 5mers in IVT data, which are classified into three patterns. The shaded area is the smoothed density of current signal of a given 5mer across different reads and different genomic locations. The fitted Gaussian density curves are given by estimated 5mer parameters. For each pattern, the top panel shows the distributions for 5mers like XXGXX (X is A, C, T/U) in the pure-G sample, while the bottom panel shows distributions of the same 5mers with the center nucleotide replaced by A or I. These examples represent three patterns of 5mers: (1) 5mers that show no difference between A and I (2) 5mers that show significant difference between A and I and (3) 5mers that show significant difference between A and I, but the component weight of A is low. (*B*) Current signal distribution of example 5mers CAACA or CAICA pooled from different reads and different genomic locations in IVT data (training data) and H9 data (test data). (*C*) Modification rate distributions in positive control (inosine modification sites) and negative control (unmodified sites). Modification states are given by the ground truth.

We test the performance of SegPore on H9 DRS data, which consists of ten wild type samples. We used the 5mer table estimated from the training data and ran the standard SegPore workflow on H9 data, which provides the modification rate of each genomic location that matches the selected 5mers. Due to data availability, we only find the ground truth of locations on Chromosome 3 and 11 from the original publication (Nguyen et al. 2022). Out of these locations, 37 inosine modification sites match our selected 5mers, which are used as the positive control, and 241 unmodified sites match our selected 5mers, which are used as the negative control. Fig. 4b shows the current signal distribution of one selected 5mer CAACA by pooling all relevant sites on positive control, which exhibits two clear peaks for CAACA and CAICA. This suggests a high consistency of 5mer distributions between the IVT data and the H9 data. Fig. 4C shows the modification rate distribution in the positive and negative controls. It can be seen that the modification rates of positive control are higher than that of negative control on average (t-test, p-value=2.22e-11).

bioRxiv preprint doi: https://doi.org/10.1101/2024.01.11.575207; this version posted March 6, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

RNA structure estimation

RNA structure can be probed by sequencing based approaches such as SHAPE-MaP and nanoSHAPE. Their idea is to treat the RNA molecules with chemical adducts, which are only attached to single stranded parts but not double stranded parts of RNA molecules. The chemically treated RNA is then subjected to short-read sequencing (SHAPE-Map) or Nanopore sequencing (nanoSHAPE). In short-read sequencing, a nucleotide attached by an adduct turns out to be a mutation, while it is recognized as a modification in Nanopore sequencing. The mutation rate or modification rate is then transformed to a reactivity score that is utilized by the RNAfold software (Lorenz et al. 2011) to generate RNA structure.

SegPore can be used to identify the modifications and predict RNA structure from Nanopore sequencing data. As an illustration, we analyze the public Nanopore sequencing data of pri-miR-17~92 RNA in the nanoSHAPE (Stephenson et al. 2022), for which the ground truth structure (based on SHAPE-MaP) is available. In this dataset, we choose Nanopore sequencing data of two samples: one *in vitro* transcribed (IVT) sample with no modifications and one adduct treated sample with modifications. Following standard SegPore workflow, we first segment raw current signals of reads in both samples. Since all reads are from the same reference sequence, we directly align the derived signal segments with 5mers of the pri-miR-17~92 reference sequence. Then, for each location (5mer) of the reference sequence, we fit a two-component Gaussian mixture model to all signal segments aligned to that location in both samples, with one component representing the unmodified state and the other representing the modified state. Next, we estimate the modification rate and reactivity score for each location of the reference sequence based on the fitted GMM parameters. In the end, pri-miR-17~92 RNA sequence and the reactivity scores are fed into RNAfold software (Lorenz et al. 2011) to generate the final RNA structure.

bioRxiv preprint doi: https://doi.org/10.1101/2024.01.11.575207; this version posted March 6, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

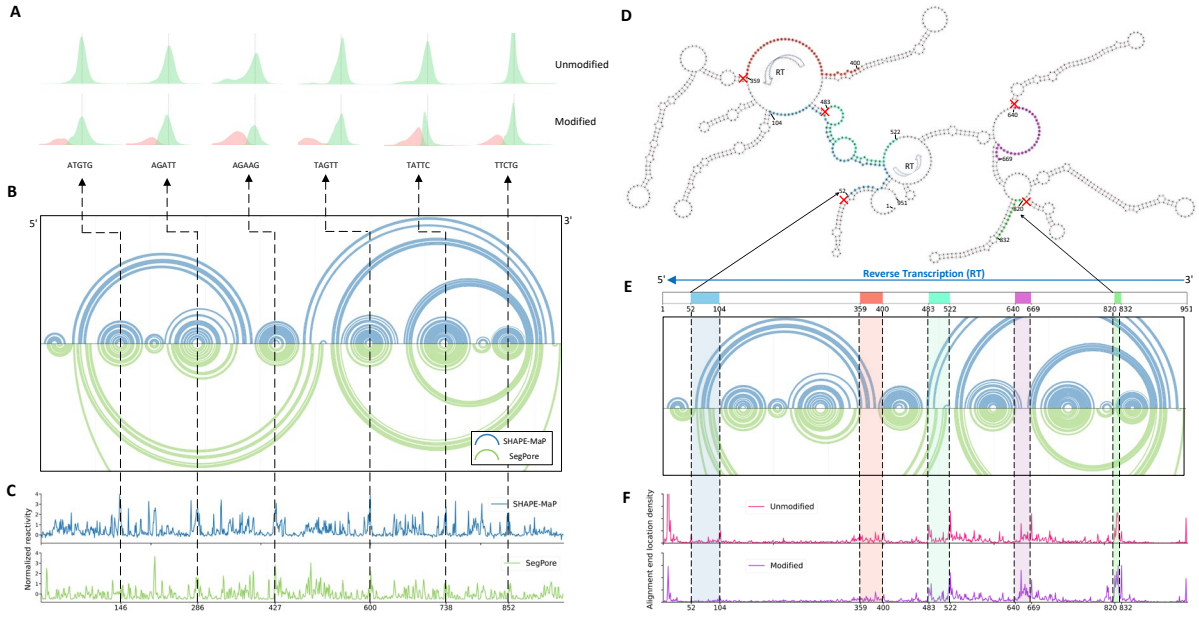

Figure 5. RNA (pri-miR-17~92) structure estimation using SegPore. (*A*) Six example sites chosen from the loop part of six conserved hairpin structures, with the corresponding 5mers displayed at the bottom. The top panel is from the unmodified sample (in vitro transcription) and the lower panel is from the modified sample (adduct treated), whose modified component is highlighted in red. (*B*) Arc diagrams of RNA structure based on SHAPE-Map (blue) and SegPore (green). X-axis is the 5'-3' location on the reference sequence. (*C*) Reactivity scores of SHAPE-MaP and SegPore. A high reactivity score is associated with a high modification rate in the modified sample. (*D*) SegPore predicted RNA structure, with different colors denoting peak regions in the alignment end location distribution. The crossings indicate the stem structures near potential stop locations during reverse transcription. The numbers indicate the 5'-3' locations of the nucleotides. The reverse transcription is from 3' to 5', i.e. from location 951 to 1. (*E*) The same arc diagrams as (*B*) with peak regions highlighted. (*F*) Alignment end location distribution for both the unmodified and modified samples. X-axis indicates 5'-3' locations on the reference sequence. Y-axis is the probability density.

The RNA structure given by SegPore nicely agrees with the ground truth (SHAPE-MaP), as shown in Fig. 5B. The conserved six hairpin structures, each marked by a dashed line in the center, are almost identical between SegPore and SHAPE-MaP predictions. Six example sites in the loop part from each hairpin structure are selected to illustrate the modifications in Fig. 5A, where a large component of modification (highlighted in red) is observed in the modified sample. This result is consistent with the expected outcome of the SHAPE protocol as the loops are single stranded, where adducts are attached and recognized as modifications. The reactivity scores of SHAPE-MaP and SegPore are similar (Fig. 5C), and peaks are located at similar sites. It can be seen that peak regions in Fig. 5C generally correspond to single stranded regions with no base pairing in Fig. 5B.

bioRxiv preprint doi: https://doi.org/10.1101/2024.01.11.575207; this version posted March 6, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

After aligning the raw signal with the reference sequence, large amounts of reads in both unmodified (86.3%) and modified samples (95.7%) do not cover the full-length reference sequence. We are curious about the cause and examine the distribution of the alignment end locations on the reference sequence. Fig. 5F shows that the distributions are very similar between the unmodified and modified samples, which precludes the adduct modifications from being the cause. Comparing the peak regions in Fig. 5F with the RNA structure in Fig. 5D and Fig. 5E, we find that the peak regions are generally located right before stem structures, i.e., a stretch of paired bases. Note that the reverse transcription starts from 3' to 5', it is obvious that the peak regions are right next to the downstream stem structures, the start locations of which are highlighted by crossing signs in Fig. 5D.

The above finding suggests that peaks are associated with reverse transcription in the Nanopore sequencing library preparation. Here, we hypothesize that the reverse transcriptase enzyme loses its momentum when it hits stem structures highlighted by crossings in Fig. 5D, as it takes lots of energy to unwind the stem structures. When the reverse transcriptase enzyme stalls on the RNA molecule, we hypothesize that the RNA molecule fractures near the stalled locations due to the thermodynamic movements of the reverse transcriptase enzyme. Therefore, a partial read corresponds to a fractured RNA molecule in the reverse transcription.

Discussion

The key computational problem in DRS is how to segment the raw current signal. We developed a segmentation algorithm that utilized the jiggling property in the physical process of DRS and demonstrated that the segmentation led to cleaner current signals in m6A, inosine identification, as well as RNA structure estimation. We have shown that m6Anet achieved better performances based on SegPore's improved segmentation. We believe that the de-noised current signals will be beneficial for other downstream tasks. However, some open questions remain to be addressed in the future. In SegPore, we assume a drastic change

bioRxiv preprint doi: https://doi.org/10.1101/2024.01.11.575207; this version posted March 6, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

between two consecutive 5mers, which may hold for 5mers with large difference in their current baselines but may not hold for those with small difference. Another question is the physical meaning of derived base blocks. In the ideal case, one base block corresponds to one 5mer, while multiple base blocks are aligned with one 5mer in the real case. One guess is that the HHMM may partition the current signal of one 5mer into several base blocks, during which the 5mer may oscillate between different sub-states, i.e., each sub-state of the same 5mer has different baselines.

Different from DNN-based methods, SegPore offers great interpretability to the estimation of RNA modifications, which makes it applicable to different modifications. SegPore codes current intensity levels for different 5mers in a parameter table, where unmodified and modified 5mers are modeled using two Gaussian distributions. Given 5mer parameter tables of different RNA modifications, e.g., m6A, inosine, pseudouridine, etc., we can estimate various modifications from the same DRS data. Since we directly model the current level of 5mers with RNA modifications, SegPore naturally provides the modification states on the single RNA molecule level. This capability will be particularly powerful for studying RNA modifications in various disease contexts. However, the amount of 5mers with significant changes in their modification states is relatively small. We may need larger training data to improve the accuracy and expand 5mers to 7mers or 9mers to consider more context information, on which significant baseline change might be observed on more kmers.

In RNA structure experiment, we also found there was an association between the end points of reads and the stem structures of the RNA molecule. We hypothesize that the RNA molecule fractures are near the stem structures. If stronger stem structures are associated with a larger proportion of partial reads, we may use this information to probe the binding energy of the stems and incorporate it in the RNA structure estimation, which is a unique angle provided by DRS. SegPore differs from the original structure estimation method NanoSHAPE in the following aspects (1) we use SegPore's segmentation results instead of Nanopolish's, (2) we

bioRxiv preprint doi: https://doi.org/10.1101/2024.01.11.575207; this version posted March 6, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

379 estimate RNA modifications using GMM while NanoSHAPE uses outlier detection, and (3) the
380 reactivity score profiles are calculated differently. Due to the improved segmentation, SegPore
381 increased the mapping rate of full-length RNA sequences. In our experiment, SegPore
382 identified 21% of reads to be full-length, while Nanopolish only identified 5%.
383
384 Computation speed is another concern in handling the fast5 files. We implemented a GPU-
385 accelerated inference algorithm in SegPore, which has a 10~20 fold speedup compared with
386 the CPU-based implementation. We believe that the GPU-implementation will unlock the full
387 potential of SegPore for a wider range of downstream tasks and larger datasets.

388 Methods

389 5mer parameter table

390 We download the kmer models “r9.4_180mv_70bps_5mer_RNA” from GitHub repository of
391 ONT (https://github.com/nanoporetech/kmer_models). The columns “level_mean” and
392 “level_stdv” were used as the mean and std for unmodified 5mers in a parameter table. We
393 denoted it as the 5mer parameter table of ONT, which was used for initialization in SegPore.

394 SegPore workflow

395 Preprocessing

396 We first perform basecalling of the input fast5 file using Guppy. Then we map the basecalled
397 sequence to the reference sequences using Minimap2 (Li 2018). After that, we obtain the
398 paired raw current signal segments and corresponding fragments of the reference sequence
399 using Nanopolish eventalign. Meanwhile, we obtain the raw current signal segments
400 corresponding to the polyA tail. Finally, we standardize the raw current signal of each read
401 based on its polyA tail, such that the mean and std of the polyA tail are the same across
402 different reads.

bioRxiv preprint doi: https://doi.org/10.1101/2024.01.11.575207; this version posted March 6, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Signal Segmentation via hierarchical Hidden Markov model

The RNA translocation hypothesis naturally leads to a hierarchical hidden Markov model (HHMM) for segmenting the raw current signal. As shown in Fig. 2B, our HHMM has two layers. The outer HMM divides the raw current signal into alternating base and transition blocks, as indicated by hidden states "B" and "T" in Fig. 2B. The inner HMM models a single base block and a linear model is used for each transition block. The inner HMM has four hidden states: "prev", "next", "curr", "noise". The "prev", "next", "curr" states refer to previous, next and current 5mer in the pore, while "noise" refers to random noise. Each raw current measurement is then emitted from one of these hidden states. The linear model with a large absolute slope is used to model the sharp changes in the transition block.

Given the raw current signal $\boldsymbol{y}$ of a read, we denote the hidden states of the outer hidden HMM by $\boldsymbol{g}$. $\boldsymbol{y}$ and $\boldsymbol{g}$ are divided into $2K+1$ blocks $\boldsymbol{c}$, where $\boldsymbol{y}^{(k)}$, $\boldsymbol{g}^{(k)}$ correspond to k th block and $\boldsymbol{c}=(c_1,c_2,...,c_k,...,c_{2K+1}), c_k \in \{"B","T"\}$. Note that blocks with odd index are base blocks, i.e. $k=1,3,5,...,2K+1$, whereas blocks with even index are transition blocks. The likelihood of HHMM is given by

$$p(\boldsymbol{y},\boldsymbol{g}) = p(\boldsymbol{y}\mid\boldsymbol{g})p(\boldsymbol{g}) \tag{1}$$

$$= p(\boldsymbol{y}\mid\boldsymbol{g})\{\pi^{outer}_{g_1}\prod_{i=2}^{N}T^{outer}_{g_{i-1}g_i}\} \tag{2}$$

$$= \{\prod_{i=1}^{N}p(y_i\mid g_i)\}\{\pi^{outer}_{g_1}\prod_{i=2}^{N}T^{outer}_{g_{i-1}g_i}\} \tag{3}$$

$$= \{\prod_{k=0}^{K}p(\boldsymbol{y}^{(2k+1)}\mid c_{2k+1}="B")\}\{\prod_{k=1}^{K}p(\boldsymbol{y}^{(2k)}\mid c_{2k}="T")\}\{\pi^{outer}_{g_1}\prod_{i=2}^{N}T^{outer}_{g_{i-1}g_i}\} \tag{4}$$

where $T^{outer}_{g_{i-1}g_i}$ is the transition matrix of the outer HMM and $\pi^{outer}_{g_1}$ is the probability for the first hidden state. It is obviously seen that the left side of the Eq. 2 are emission probabilities, and the right side are the transition probabilities. It is not possible to directly compute the emission probabilities of the outer HMM (Eq. 3) since there exist dependencies for the current signal

bioRxiv preprint doi: https://doi.org/10.1101/2024.01.11.575207; this version posted March 6, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

measurements within a base or transition block. Therefore, we use the inner HMM and linear model (Eq. 4) to handle the dependencies and approximate emission probabilities.

The inner HMM models the transitions between the hidden states “prev”, “next”, “curr”, and “noise”. For “prev”, “next”, “curr” states, we use the Gaussian distribution as their emission distribution. Uniform distribution is used as the emission distribution for the “noise” state. Given these specifications, we can calculate the marginal likelihood of the inner HMM using the Forward-Backward algorithm, which is used to approximate the emission probabilities of the base blocks in the outer HMM. Similarly, we can approximate the emission probabilities of the transition blocks using the likelihood of the standard linear model.

For any given $\boldsymbol{g}$, we can calculate the joint likelihood (Eq. 1). Therefore, we enumerate different configurations of $\boldsymbol{g}$ and choose the one with the highest likelihood. Detailed model description is provided in Supplementary Note 1.

The parameter inference is challenging given the massive data size of fast5 files, which are generally on the level of terabytes (TB). We developed a GPU-based inference algorithm to accelerate the inference process. A detailed description of the GPU-accelerated inference algorithm is provided in Supplementary Note 2.

In the end, we segment the raw current signal of a read into alternating base and transition blocks, where one or multiple base blocks may correspond to only one 5mer. For each base block, it has a mean and std parameter (Gaussian distribution). The mean values of the base blocks are used as input for the downstream alignment tasks.

Alignment of raw signal segment with reference sequence

After segmenting the raw current signal of a read into base and transition blocks using HHMM, we align the means of base blocks with the 5mer list of the reference sequence.

bioRxiv preprint doi: https://doi.org/10.1101/2024.01.11.575207; this version posted March 6, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

The alignment has three different matching cases. The first case is that one base block matches with one 5mer, which means the base block follows the Gaussian distribution of the given 5mer. Note that the 5mer might have two states: unmodified and modified. The corresponding Gaussian parameters can be found in the 5mer parameter table. The second case is the one base block matches with an indel "-", which means there is an inserted nucleotide in the read. The third case is that an indel (0.0) matches with a 5mer, which means there is a deleted nucleotide in the read.

Our score function in the alignment models the matching cases as follows. For the first case, we calculate the probability of the base block mean sampled from the unmodified 5mer Gaussian distribution, as well as the modified 5mer Gaussian distribution. The larger probability is used as the match score in this case. For the second or third case, we treat it as noise and use a fixed uniform distribution to calculate the match score.

In our alignment, another significant difference with the classical global alignment algorithm is that one or multiple base blocks could be aligned with one 5mer. Given the base block means $\boldsymbol{\mu} = (\mu_1, \mu_2, ..., \mu_i, ..., \mu_m)$ and 5mer list $\boldsymbol{s} = (s_1, s_2, ..., s_j, ..., s_n)$, we define $(m+1) \times (n+1)$ the score matrix as $\boldsymbol{M}$ (Fig. 2C). The first row and column of $\boldsymbol{M}$ are reserved for indels "0.0" and "-", which represent indels of the base block and 5mer, respectively. We denote the score function by f. The recursion formula of the dynamic programming alignment algorithm is given by

$$\boldsymbol{M}(i,j) = max\begin{cases} \boldsymbol{M}(i-1,j-1) + f(\mu_i, s_j) \\ \boldsymbol{M}(i,j-1) + f(0.0, s_j) \\ \boldsymbol{M}(i-1,j) + f(\mu_i, \text{"}-\text{"}) \\ \boldsymbol{M}(i-1,j) + f(\mu_i, s_j) \end{cases}. \quad (5)$$

It can be seen that we can still align μ_i with s_j after we have aligned μ_{i-1} with s_j, which fulfills the special consideration that one or multiple base blocks might be aligned with one 5mer.

bioRxiv preprint doi: https://doi.org/10.1101/2024.01.11.575207; this version posted March 6, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

There are two different types of alignment algorithms given the score matrix $\boldsymbol{M}$: the full alignment algorithm and the partial alignment algorithm (Fig. 2C). The full alignment algorithm tries to align the full list of base block means with the full list of 5mer list, which is similar to the classical global alignment algorithm. This is implemented by tracing back from the $(m+1, n+1)$ position of the score matrix. The partial alignment algorithm tries to align the full list of the base blocks with the first part of the 5mer list. It differs from the full alignment algorithm in two aspects: (1) no indel is allowed in both the base block means and the 5mer list, and (2) trace back the maximum value of the last row of the score matrix $\boldsymbol{M}$. A detailed description of the full and partial alignment algorithm is provided in Supplementary Note 1.

The output of the alignment algorithm is the eventalign, which contains the pairing between the base blocks and 5mer list of reference sequences for each read (Fig. 2C).

Modification estimation

After obtaining the eventalign results, we estimate the modification state of each motif using the 5mer parameter table. If the 5mer is in the modification state, the probability of the base block mean under the modified 5mer Gaussian distribution should be higher than that calculated using unmodified 5mer parameters, and vice versa. Since one 5mer may be aligned with multiple base blocks, we merge all the aligned base blocks and take the weighted mean as the single base block mean aligned with the given 5mer, which provides us the modification state of each site of a read. We then pool all reads mapped to the same genomic location on the reference sequence to get the modification rate of the genomic location, which is the proportion of reads in modification state. Detailed description of the modification state estimation is provided in Supplementary Note 1.

GMM for 5mer parameter table re-estimation

To gain better alignment results and more accurate modification estimation, we use GMM to fine-tune the 5mer parameter table iteratively. As shown in Fig. 2A, the row of the 5mer

bioRxiv preprint doi: https://doi.org/10.1101/2024.01.11.575207; this version posted March 6, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

parameter table are 5mers and the columns are the mean and std of the unmodified and modified states. We denote a 5mer by s and its relevant parameters by $\mu_{s,un}$, $\delta_{s,un}$, $\mu_{s,mod}$, $\delta_{s,mod}$.

Given the alignment results of all reads, we extract all base block means that are aligned to the same 5mer s on different reads and across different genomic locations with high modification rates. Next, we fit a two-component GMM to the collected base blocks corresponding to 5mer s, with the first component mean fixed to $\mu_{s,un}$. From the GMM, we have the updated $\delta_{s,un}$, $\omega_{s,un}$, $\mu_{s,mod}$, $\delta_{s,mod}$, and $\omega_{s,mod}$, where $\omega_{s,un}$, $\omega_{s,mod}$ are the weights for unmodified and modified components. Then, we manually update the 5mer parameter table based on some heuristics such that the modified 5mer distribution is significantly different from that of unmodified 5mer distribution. Detailed description of the GMM re-estimation process is provided in Supplementary Note 1.

The re-estimation of 5mer parameter table is only performed on the training data. We initialize the 5mer parameter table using the 5mer parameter table provided by ONT. Every time after we gain the updated 5mer parameter table, we run the SegPore workflow from the alignment again. The process is repeated 3 to 5 times in general, after which the 5mer parameter table is stabilized. After that, the 5mer parameter table is fixed. For testing data, we estimate the RNA modification states using the fixed 5mer parameter table.

m6A benchmark

The HEK293T wild type (WT) samples were downloaded from ENA database under accession number PRJEB40872, while the HCT116 samples were downloaded from ENA PRJEB44348. The reference sequence (Homo_sapiens.GRCh38.cdna.ncrna_wtChrIs_modified.fa) were downloaded from https://doi.org/10.5281/zenodo.4587661. The ground truth data were obtained from Supplementary Data 1 of Pratanwanich, P.N. et al. (Pratanwanich et al. 2021). Fast5 files of the test dataset (mESC WT samples, mESCs_Mettl3_WT_fast5.tar.gz) were

bioRxiv preprint doi: https://doi.org/10.1101/2024.01.11.575207; this version posted March 6, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

downloaded from NCBI Sequence Read Archive (SRA) database under accession number SRP166020.

During training, the 5mer parameter table was initialized using ONT. Standard SegPore workflow is performed on the training data (HEK293T WT samples), where the full alignment algorithm is used. The 5mer parameter table estimation was iterated five times. During the training, reads were first mapped to cDNA, then converted to genomic locations on the reference genome using Ensembl GTF file (GRCh38, v9), after that the same 5mer at different genomic locations were pooled together. We select 5mers with significant modifications if its read coverage is greater than 1,500 and the distance between two components means in the GMM is greater than 5. The modification parameters were specified for ten significant 5mers, as illustrated in Supplementary Fig. 1A.

With the estimated 5mer parameter table from the training data, we ran SegPore workflow on the test data. Transcript sequences of GENCODE release version M18 were used as the reference sequence for mapping, where the GTF file (gencode.vM18.chr_patch_hapl_scaff.annotation.gtf) downloaded from GENCODE was used to convert transcript locations to genomic locations. Note that we do not estimate 5mer parameter table for test data, and the modification states for each read are estimated only once. Due to the difference between human and mouse, only six out of ten selected 5mers have m6A annotations in the ground truth of the test data (Supplementary Fig. 1C). For a genomic location to be considered as a m6A modification site, we require that it must correspond to one of the six common 5mers and the read coverage must be greater than 20. For SegPore, the modification rate of each genomic location was used to derive the ROC and PR curves in the benchmark study.

In the SegPore+m6Anet analysis, we fine-tuned the m6Anet using the SegPore eventalign results to demonstrate its improved performance on m6A identification. Based on the pre-

bioRxiv preprint doi: https://doi.org/10.1101/2024.01.11.575207; this version posted March 6, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

trained m6Anet network (https://github.com/GoekeLab/m6anet, model version: HCT116_RNA002), we fine-tuned it using SegPore eventalign results of HCT116 samples. SegPore eventalign provides the pairing between each genomic location and its corresponding raw signal segment, from which it generates the normalized mean μ_i, std σ_i, dwell time l_i. For genomic location i, m6Anet extracts a feature vector $x_i = \{\mu_{i-1}, \sigma_{i-1}, l_{i-1}, \mu_i, \sigma_i, l_i, \mu_{i+1}, \sigma_{i+1}, l_{i+1}\}$ to be used as the input of m6Anet. Feature vectors of 80% genomic locations were used as the training set and the rest 20% were used as the validation set. We run 100 epochs to fine-tuning m6Anet and selected the model that performs the best on the validation set.

The ground truth, performances of other methods (Tombo v1.5.1, Nanom6A v2.0, m6Anet v1.0, and Epinano v1.2.0) of mESCs were obtained through personal communications with Prof. Luo Guanzheng, who is the corresponding author of the referenced benchmark study (Zhong et al. 2023).

Inosine identification experiments

Raw nanopore sequencing data of *in virto* transcription data (training data) and H9 human embryonic stem cells (test data) were downloaded from NCBI Sequence Read Archive under accession number SRP363295. The training data consists of eight pure-G samples (Accessions: SRX18177003, SRX14536452, SRX14536451, SRX14536450, SRX14536449, SRX14536448, SRX14536447, and SRX14536446), eight pure-I samples (Accessions: SRX18176999, SRX14535372, SRX14535371, SRX14535370, SRX14535369, SRX14535368, SRX14535367, and SRX14535366), and three reference sequences for synthetic RNA can be download from Supplementary Table 5 of the original publication (Nguyen et al. 2022). The test data consists of ten H9 WT embryo samples (Accessions: SRX14436756, SRX14436755, SRX14436754, SRX14436753, SRX14436752, SRX14448128, SRX14448125, SRX14448129, SRX14604385, and SRX14604391). The

bioRxiv preprint doi: https://doi.org/10.1101/2024.01.11.575207; this version posted March 6, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

reference sequence for the test data is the human GRCh37 cDNA sequences downloaded from the Ensembl database.

Standard SegPore workflow was performed on the training data, where four iterations were used to derive the 5mer parameter table. Considering that there is only one transcript per sample in the IVT data, the partial alignment algorithm is used in the training. The designed reference sequences of the IVT data contain 81 different 5mers in the form of XXGXX, where X is A, C or U/T. We manually classify the 81 5mers into three patterns (Supplementary Fig. 2a), which contain 20, 23, 38 5mers, respectively. For Pattern 2, a two-component GMM is fitted to the data with the unmodified component mean fixed. For Pattern 3, a single Gaussian distribution is fitted to the data with the constraint that the std should not exceed 3.0. The final 5mer parameter table after training contains modification parameters for 61 5mers (Pattern 2 and 3).

The derived 5mer parameter table was then used to identify inosine from the test data (H9 data). Ensembl GTF file (GRCh37, v87) was used to convert transcript locations to genomic locations. For a genomic location to be considered an inosine modification site, its read coverage must be larger than 10. Due to the availability of data, we can only download the ground truth on Chromosome 3 and 11 (test_H9_regression_allannotatedsites_0to1.RData) for the test data from the code repository of the original publication (Nguyen et al. 2022), which contains the modification rates of 4,934 genomic locations. From these genomic locations, we selected sites with a modification rate greater than 0.1 as the positive control, and sites with a modification rate equal to 0 as the negative control. As a result, 1,007 sites were retained from the ground truth, consisting of 129 positive sites (37 unique 5mers) and 878 negative sites. Out of the 37 unique 5mers in the positive control of the ground truth, there are 18 overlapping 5mers with the 61 selected 5mers (Pattern 2 and 3) in the training. The sites that match the 18 overlapping 5mers were kept, after that we retained sites with read coverage larger than

bioRxiv preprint doi: https://doi.org/10.1101/2024.01.11.575207; this version posted March 6, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

10, resulting in 37 positive sites and 241 negative sites. SegPore estimated modification rates of these sites were extracted to generate Fig. 4C.

RNA structure estimation

The raw DRS data of pri-miR-17~92 is downloaded from NCBI under accession number PRJNA634693, which contains an unmodified sample (*in vitro* transcription) and several modified RNA samples treated using 5, 20, 50, 75, 100, 150, or 200 mM AcIm. The IVT sample (unmodified RNA) and the 150mM AcIm sample (modified RNA) were used in this experiment.

Standard SegPore workflow is performed on both samples and partial alignment algorithm is used. From the alignment results, we find a large proportion of the reads (~90%, Supplementary Note 3) do not cover the whole reference sequence. Therefore, we analysis the alignment end position distribution. For getting more accurate reactivity score in RNA structure estimation, we fine-tuned the parameters for unmodified distribution and modified distribution on each site of the reference sequence. Finally, the normalized reactivity score is fed into the RNAfold (v2.4.13) (Vienna) web server to predict the RNA structure and R-chie (Tsybulskyi, Mounir and Meyer 2020) is used for displaying base pairing (arc). More details are provided in Supplementary Note 3.

Data Access

The data utilized in this study are obtained from publicly available repositories. Details regarding the accession number and data processing can be found in Methods. The source code is hosted on GitHub (https://github.com/guangzhaocs/SegPore).

Acknowledgements

We would like to thank Prof. Zhijie Tan from Wuhan University for a useful discussion about the molecule dynamics of Nanopore sequencing, Dr. Dan Zhang from Sichuan University for helpful tutorials about traditional Nanopore analysis workflows. We would like to thank

bioRxiv preprint doi: https://doi.org/10.1101/2024.01.11.575207; this version posted March 6, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

637 Research Council of Finland grants (NO. 335858, 358086) to GC and LC. GC and LC
638 acknowledge the computational resources provided by the Aalto Science-IT project. We also
639 thank Prof. Luo Guanzheng for sharing the m6A benchmark baseline results.

640 Author Contributions

641 GC developed the methods and performed the analyses. AV provided advice about statistical
642 modelling and manuscript writing. LC conceptualized and supervised the project. GC and LC
643 co-implemented the SegPore workflow, co-wrote the manuscript. All authors read and
644 approved the manuscript.

645 Competing Interest Statement

646 The authors declare no competing interests.

647 Supplementary Information

648 **Supplementary Figure 1**: m6A identification kmer motif statistics.
649 **Supplementary Figure 2**: Inosine identification kmer patterns.
650 **Supplementary Note 1**: SegPore workflow.
651 **Supplementary Note 2**: GPU-accelerated Hierarchical Hidden Markov Model parameter
652 inference.
653 **Supplementary Note 3**: SegPore for RNA structure estimation.

654 References

655 Agris PF, Narendran A, Sarachan K, Vare VYP, Eruysal E. 2017. The Importance of Being
656 Modified: The Role of RNA Modifications in Translational Fidelity. *Enzymes* **41**: 1-50.
657 Bellodi C, McMahon M, Contreras A, Juliano D, Kopmar N, Nakamura T, Maltby D,
658 Burlingame A, Savage SA, Shimamura A, Ruggero D. 2013. H/ACA small RNA
659 dysfunctions in disease reveal key roles for noncoding RNA modifications in
660 hematopoietic stem cell differentiation. *Cell Rep* **3**: 1493-1502.

bioRxiv preprint doi: https://doi.org/10.1101/2024.01.11.575207; this version posted March 6, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

661 Boccaletto P, Stefaniak F, Ray A, Cappannini A, Mukherjee S, Purta E, Kurkowska M,
662 Shirvanizadeh N, Destefanis E, Groza P et al. 2022. MODOMICS: a database of
663 RNA modification pathways. 2021 update. *Nucleic Acids Res* **50**: D231-D235.
664 Caldwell CC, Spies M. 2017. Helicase SPRNTing through the nanopore. *Proc Natl Acad Sci*
665 *U S A* **114**: 11809-11811.
666 Chen AY, Owens MC, Liu KF. 2023. Coordination of RNA modifications in the brain and
667 beyond. *Mol Psychiatry* **28**: 2737-2749.
668 Craig JM, Laszlo AH, Brinkerhoff H, Derrington IM, Noakes MT, Nova IC, Tickman BI,
669 Doering K, de Leeuw NF, Gundlach JH. 2017. Revealing dynamics of helicase
670 translocation on single-stranded DNA using high-resolution nanopore tweezers. *Proc*
671 *Natl Acad Sci U S A* **114**: 11932-11937.
672 Eisenberg E, Levanon EY. 2018. A-to-I RNA editing - immune protector and transcriptome
673 diversifier. *Nat Rev Genet* **19**: 473-490.
674 Gao Y, Liu X, Wu B, Wang H, Xi F, Kohnen MV, Reddy ASN, Gu L. 2021. Quantitative
675 profiling of N(6)-methyladenosine at single-base resolution in stem-differentiating
676 xylem of Populus trichocarpa using Nanopore direct RNA sequencing. *Genome Biol*
677 **22**: 22.
678 Jain M, Koren S, Miga KH, Quick J, Rand AC, Sasani TA, Tyson JR, Beggs AD, Dilthey AT,
679 Fiddes IT et al. 2018. Nanopore sequencing and assembly of a human genome with
680 ultra-long reads. *Nat Biotechnol* **36**: 338-345.
681 Koh CWQ, Goh YT, Goh WSS. 2019. Atlas of quantitative single-base-resolution N(6)-
682 methyl-adenine methylomes. *Nat Commun* **10**: 5636.
683 Kortel N, Ruckle C, Zhou Y, Busch A, Hoch-Kraft P, Sutandy FXR, Haase J, Pradhan M,
684 Musheev M, Ostareck D et al. 2021. Deep and accurate detection of m6A RNA
685 modifications using miCLIP2 and m6Aboost machine learning. *Nucleic Acids Res* **49**:
686 e92.

bioRxiv preprint doi: https://doi.org/10.1101/2024.01.11.575207; this version posted March 6, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

687 Lee H, Bao S, Qian Y, Geula S, Leslie J, Zhang C, Hanna JH, Ding L. 2019. Stage-specific
688 requirement for Mettl3-dependent m(6)A mRNA methylation during haematopoietic
689 stem cell differentiation. *Nat Cell Biol* **21**: 700-709.
690 Leger A, Amaral PP, Pandolfini L, Capitanchik C, Capraro F, Miano V, Migliori V, Toolan-
691 Kerr P, Sideri T, Enright AJ et al. 2021. RNA modifications detection by comparative
692 Nanopore direct RNA sequencing. *Nat Commun* **12**: 7198.
693 Li H. 2018. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**: 3094-
694 3100.
695 Linder B, Grozhik AV, Olarerin-George AO, Meydan C, Mason CE, Jaffrey SR. 2015. Single-
696 nucleotide-resolution mapping of m6A and m6Am throughout the transcriptome. *Nat*
697 *Methods* **12**: 767-772.
698 Liu H, Begik O, Lucas MC, Ramirez JM, Mason CE, Wiener D, Schwartz S, Mattick JS,
699 Smith MA, Novoa EM. 2019. Accurate detection of m(6)A RNA modifications in
700 native RNA sequences. *Nat Commun* **10**: 4079.
701 Loman NJ, Quick J, Simpson JT. 2015. A complete bacterial genome assembled de novo
702 using only nanopore sequencing data. *Nat Methods* **12**: 733-735.
703 Lorenz DA, Sathe S, Einstein JM, Yeo GW. 2020. Direct RNA sequencing enables m(6)A
704 detection in endogenous transcript isoforms at base-specific resolution. *RNA* **26**: 19-
705 28.
706 Lorenz R, Bernhart SH, Honer Zu Siederdissen C, Tafer H, Flamm C, Stadler PF, Hofacker
707 IL. 2011. ViennaRNA Package 2.0. *Algorithms Mol Biol* **6**: 26.
708 Meyer KD, Saletore Y, Zumbo P, Elemento O, Mason CE, Jaffrey SR. 2012. Comprehensive
709 analysis of mRNA methylation reveals enrichment in 3' UTRs and near stop codons.
710 *Cell* **149**: 1635-1646.
711 Nance KD, Meier JL. 2021. Modifications in an Emergency: The Role of N1-
712 Methylpseudouridine in COVID-19 Vaccines. *ACS Cent Sci* **7**: 748-756.

bioRxiv preprint doi: https://doi.org/10.1101/2024.01.11.575207; this version posted March 6, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

713 Nguyen TA, Heng JWJ, Kaewsapsak P, Kok EPL, Stanojevic D, Liu H, Cardilla A, Praditya
714 A, Yi Z, Lin M et al. 2022. Direct identification of A-to-I editing sites with nanopore
715 native RNA sequencing. *Nat Methods* **19**: 833-844.
716 Nishikura K. 2016. A-to-I editing of coding and non-coding RNAs by ADARs. *Nat Rev Mol*
717 *Cell Biol* **17**: 83-96.
718 Ohira T, Suzuki T. 2024. Transfer RNA modifications and cellular thermotolerance. *Mol Cell*
719 **84**: 94-106.
720 Parker MT, Knop K, Sherwood AV, Schurch NJ, Mackinnon K, Gould PD, Hall AJ, Barton
721 GJ, Simpson GG. 2020. Nanopore direct RNA sequencing maps the complexity of
722 Arabidopsis mRNA processing and m(6)A modification. *Elife* **9**.
723 Pratanwanich PN, Yao F, Chen Y, Koh CWQ, Wan YK, Hendra C, Poon P, Goh YT, Yap
724 PML, Chooi JY et al. 2021. Identification of differential RNA modifications from
725 nanopore direct RNA sequencing with xPore. *Nat Biotechnol* **39**: 1394-1402.
726 Price AM, Hayer KE, McIntyre ABR, Gokhale NS, Abebe JS, Della Fera AN, Mason CE,
727 Horner SM, Wilson AC, Depledge DP, Weitzman MD. 2020. Direct RNA sequencing
728 reveals m(6)A modifications on adenovirus RNA are necessary for efficient splicing.
729 *Nat Commun* **11**: 6016.
730 Prieto M, Folci A, Martin S. 2020. Post-translational modifications of the Fragile X Mental
731 Retardation Protein in neuronal function and dysfunction. *Mol Psychiatry* **25**: 1688-
732 1703.
733 Quin J, Sedmik J, Vukic D, Khan A, Keegan LP, O'Connell MA. 2021. ADAR RNA
734 Modifications, the Epitranscriptome and Innate Immunity. *Trends Biochem Sci* **46**:
735 758-771.
736 Siegfried NA, Busan S, Rice GM, Nelson JA, Weeks KM. 2014. RNA motif discovery by
737 SHAPE and mutational profiling (SHAPE-MaP). *Nat Methods* **11**: 959-965.
738 Simpson JT, Workman RE, Zuzarte PC, David M, Dursi LJ, Timp W. 2017. Detecting DNA
739 cytosine methylation using nanopore sequencing. *Nat Methods* **14**: 407-410.

bioRxiv preprint doi: https://doi.org/10.1101/2024.01.11.575207; this version posted March 6, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

740 Slotkin W, Nishikura K. 2013. Adenosine-to-inosine RNA editing and human disease.
741 *Genome Med* **5**: 105.
742 Stephenson W, Razaghi R, Busan S, Weeks KM, Timp W, Smibert P. 2022. Direct detection
743 of RNA modifications and structure using single-molecule nanopore sequencing. *Cell*
744 *Genom* **2**.
745 Stoiber M, Quick J, Egan R, Eun Lee J, Celniker S, Neely RK, Loman N, Pennacchio LA,
746 Brown J. 2017. De novo Identification of DNA Modifications Enabled by Genome-
747 Guided Nanopore Signal Processing. *biorxiv* doi:10.1101/094672.
748 Tsybulskyi V, Mounir M, Meyer IM. 2020. R-chie: a web server and R package for visualizing
749 cis and trans RNA–RNA, RNA–DNA and DNA–DNA interactions. *Nucleic Acids*
750 *Research* **48**: e105-e105.
751 Yankova E, Aspris D, Tzelepis K. 2021. The N6-methyladenosine RNA modification in acute
752 myeloid leukemia. *Curr Opin Hematol* **28**: 80-85.
753 Zhong ZD, Xie YY, Chen HX, Lan YL, Liu XH, Ji JY, Wu F, Jin L, Chen J, Mak DW et al.
754 2023. Systematic comparison of tools used for m(6)A mapping from nanopore direct
755 RNA sequencing. *Nat Commun* **14**: 1906.
756 Zimna M, Dolata J, Szweykowska-Kulinska Z, Jarmolowski A. 2023. The expanding role of
757 RNA modifications in plant RNA polymerase II transcripts: highlights and
758 perspectives. *J Exp Bot* **74**: 3975-3986.
759