Localized, Non-parametric Detection of RNA Structural Modification using Nanopore Basecalling.

J. White Bear*
McGill University
Montréal, Québec, CA
jwb@cs.mcgill.ca

Grégoire de Bisschop
Institut de recherche clinique de Montréal
Montréal, Québec, CA
gregoire.de.bisschop@ircm.qc.ca

Éric Lécuyer
Institut de recherche clinique de Montréal
Montréal, Québec, CA
eric.Lecuyer@ircm.qc.ca

Jérôme Waldispühl
McGill University
Montréal, Québec, CA
jerome.waldispuhl@mcgill.ca

ABSTRACT

Recently, much work has been done on chemical probing strategies with nanopore sequencing to identify RNA modifications at the single nucleotide level. Here, we examine the use of Oxford Nanopore's Guppy basecalling to identify structural modifications using localized, non-parametric peak detection. In a novel experiment, we evaluate whether detection of structural modifications is possible using the Guppy's basecalling error and determine the accuracy of our approach for selected RNA control sequences. Next, we use statistical analysis to determine the dominant structural bindings in a set of averaged read errors. Finally, we compare our approach to average reactivity determined by orthogonal experiments from SHAPE-CE and alternative approaches. We show that localized, non-parametric peak detection demonstrates improved accuracy and coverage of structural modifications in selected control RNA and that our method is agnostic to underlying changes in the distribution. Our approach allows for a more generalizeable methodology for detecting structural modification with nanopore sequencing and the subsequent generated probabilities can be used to refine further downstream analysis.

CCS CONCEPTS

• **Applied computing** → **Life and medical sciences**; **Bioinformatics**;

KEYWORDS

Oxford Nanopore, RNA, sequencing, structure prediction, modification detection, structure probing, basecalling

ACM Reference Format:
J. White Bear, Grégoire de Bisschop, Éric Lécuyer, and Jérôme Waldispühl. 2023. Localized, Non-parametric Detection of RNA Structural Modification using Nanopore Basecalling.. In *14th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics (BCB '23), September 3–6, 2023, Houston, TX, USA.* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3584371.3612989

BCB '23, September 3–6, 2023, Houston, TX, USA

ACM ISBN 979-8-4007-0126-9/23/09…$15.00
https://doi.org/10.1145/3584371.3612989

1 INTRODUCTION

RNA is a central player of virtually all cellular processes. As such, it is both a promising target and therapeutic agent for numerous diseases as exemplified by the recent development of mRNA vaccines. The function of RNA molecules closely depends on its structure, which can regulate RNA-protein [12] and RNA-RNA interactions [7]. However, RNA computational structure determination is a complex challenge that benefits from experimental data. Over the years, numerous tools have been developed to probe RNA secondary structure [24]. Popular approaches rely on a small chemical probe that covalently modifies either the nucleobase (DMS, CMCT) or the ribose (SHAPE reagents) of flexible positions. Modified positions are then detected by reverse transcription and sequencing. Depending on the method, the presence of an adduct will either stop reverse transcription or increase the error rate at a given position (See [24]).

Early work on Oxford Nanopore sequencing focused on producing optimal basecalling algorithms using deep learning methods. Guppy, a basecalling algorithm based on a Recurrent Neural Network (RNN), is used to determine the nucleotide base (A,C,G,U) as it passes through the nanopores [25]. Each sequenced RNA, basecalled with Guppy, is a 'read' that can be aligned and mapped to the reference sequence using sequence aligners like Minimap [16]. The sequenced reads may differ from the reference sequence and have errors including insertion, deletion, or mismatch of a nucleotide. Basecalling algorithms attempt to minimize spurious sequencing errors and allow for determination of sequence variations in the reads [21, 25] The understanding these variation have taken on new importance as they can be indicative of various RNA modifications [4, 14, 17, 22, 25].

2 PRIOR WORK

Several tools have been introduced that pursue specific endogenous or experimental modifications, and perform analysis of basecalling errors to determine RNA modification. Often, these tools apply specifically to methylated bases or well known motifs such as DRACH and RRACH which can isolate methylated bases using a distance metric within the motif, specifically focusing on errors at 'A' bases for m^6A modifications, and employ the use of synthetic

sequences [4, 14, 17, 22]. Given these specificities, it is still a challenging task requiring significant number of reads or read depth. Yet, prior methods have been bench-marked with accuracies averaging well above 70% at the lower range (Table 1). However the benchmarks are not consistent across all methods and were not tested on structure probe modifications [27].

For this comparison, we selected Epinano's Epinano_Differr method, which had the highest overall accuracy to test because its methods could be applied to structural modifications. The Epinano-SVM method and other methods developed for basecalling analysis use pretrained models specifically designed for m^6A and could not be applied to basecalling analysis for a novel structural modification (Table 1). However, we calculated the differential between unmodified and modified RNA basecall errors for mismatch, insertion, deletion, and quality features and ranked them to determine which feature was most indicative of a modification by measuring accuracy according to Epinano's protocol [18]. In our controls, the features most indicative of a modification varied by sequence and were not consistent enough to use as a determining metric in our tests (Supplemental 3). We also examined the sum of errors and corresponding z-score outliers using linear regression sum of errors method (LRSUM) specified for basecalling features in Epinano_Differr [18] and observed accuracies similar to statistical outlier method (Supplemental Table 4).

These results may be partially due to the fact that we are identifying structural modification rather than m^6A modifications. Also, the use of synthetic RNA or methylated bases at specific distances, particularly greater than the distances that induces pore constriction on adjacent nucleotides ($\leq$ 5 nucleotides), can significantly modulate the upstream and downstream effects observed in several experiments. These modulated distributions display uniform errors in annotated positions and can more easily be characterized using standard statistical approaches such as contingency tables, and statistical tests. However, they do not always generalize well to structural modification or capturing significant change in noisier data, e.g. adjacent nucleotide modifications which is still a challenging problem [4, 14, 17, 22].

Here, we introduce a localized, non-parametric peak detection method (LNPPD) to address these challenges and create a method that works reasonably well with structural modifications. We examine the use of LNPPD on differential basecall error analysis on five RNA sequences in determining structural modifications using a small chemical probe, 1-acetylimidazole (AcIm), to determine whether basecall error analysis has wider applicability to detection of structural modifications and move towards a more generalizeable approach that could be applied to any RNA sequence modification.

3 ACIM CHEMICAL MODIFICATION

Numerous reagents have been used to probe RNA structure. AcIm, 2-methylnicotinic acid imidazolide (NAI) and Diethyl pyrocarbonate (DEPC) have recently been successfully used with Nanopore sequencing [5, 9, 23]. We used AcIm, which generates the smallest adduct upon reaction with 2' hydroxyl base of the ribose of RNA [23]. It can be used to probe all four nucleotides as it reacts with the ribose, as opposed to DEPC which only reacts with adenines. We apply this probe to *in vitro* transcribed RNA samples and sequence them with Oxford Nanopore direct RNA sequencing. For validation, we perform the same experiments with SHAPE-CE (Selective Hydroxyl-Acylation analyzed by Primer Extension and Capillary Electrophoresis) and compare the detection of modified bases of both experiments [15, 26].

4 DETECTION OF MODIFIED BASES

For detection of modified bases, we examine the Nanopore sequence data from fast5 files of both unmodified RNA sequences and sequences probed with AcIm. The sequences are basecalled with Guppy and aligned with minimap2 [16, 25]. We, then, extract relevant statistical features of the basecalled data and create reactivity profiles using the Guppy error rates indicating mismatch, deletion, or insertion at a given base. These rates are calculated for both umodified and modified RNA. Unmodified error rates are subtracted from the error rates of RNA modified with AcIm rates to remove background modifications that occur even in the presence of no modification, Methods A.2.2. We examined basecall reactivity profiles using three approaches: statistical anomaly detection, LRSUM, and LNPPD.

5 OUTLIER DETECTION OF MODIFIED BASES

We normalize both basecalling reactivity and SHAPE-CE data to unit scale for direct comparison of reactivities. Using the SHAPE-CE threshold for anomalous or highly reactive bases, we observed that anomalies were distributed across RNA sequences with some overlap with SHAPE-CE, but less than 50% of basecall reactivites correlated to SHAPE-CE's indication of structural modification. We compared the basecall reactivity profiles to the SHAPE-CE reactivity profile using both the MannWhitney U test (MWU) and Kolmogorov-Smirnov (KS) test, similarly, observed no significant correlation (Table 2). The KS test is sensitive to differences in the two underlying cumulative distributions, while the MWU test is mostly sensitive to changes in the median of the ranked distribution [19, 20].

We, then, applied standard statistical outlier detection to the basecall reactivity profiles where an outlier is defined as $\mu \pm k\sigma$ Methods A.2.4 and observed varied, sequence dependent correlation for *in vitro* RNA between basecall reactivity profiles and SHAPE-CE reactivity.

Tetrahymena ribozyme exhibited accuracy up to 44% using this method, while HCV IRES, *E. coli* tmRNA, FMN and Lysine riboswitches recorded between 10% and 40% accuracy. Average accuracy across our control sequences using statistical outlier methods including both the SHAPE-CE threshold of .2 Figure 1 and the standard statistical approach remained at or below 72% for all sequences (Methods A.2.3). However, the number of predictions for each sequence was very low and outliers did not reproduce detected modifications in SHAPE-CE data very well (Figure 1) and demonstrating low coverage (Figure 2).

6 LOCALIZED, NON-PARAMETRIC PEAK DETECTION OF MODIFIED BASES

Direct outlier detection of basecall reactivity profiles is more challenging because reactivity profiles are not necessarily Gaussian

Table 1: Comparison of Published Methods

Method	Modification	Benchmark	Method
Diff_err [22]	m^6A	66% (Accuracy)	G-Test
Eligos [14]	m^6A	74-96% (AUROC)	Fisher's Exact & Benjamin-Hochberg
Epinano (LRSUM) [17, 18]	m^6A	87-90% (Accuracy)	Linear regression, z-score outliers (Multiple)
Drummer [4]	m^6A	Comparative overlap w/ Eligos only	Modified G-Test & Odds ratio

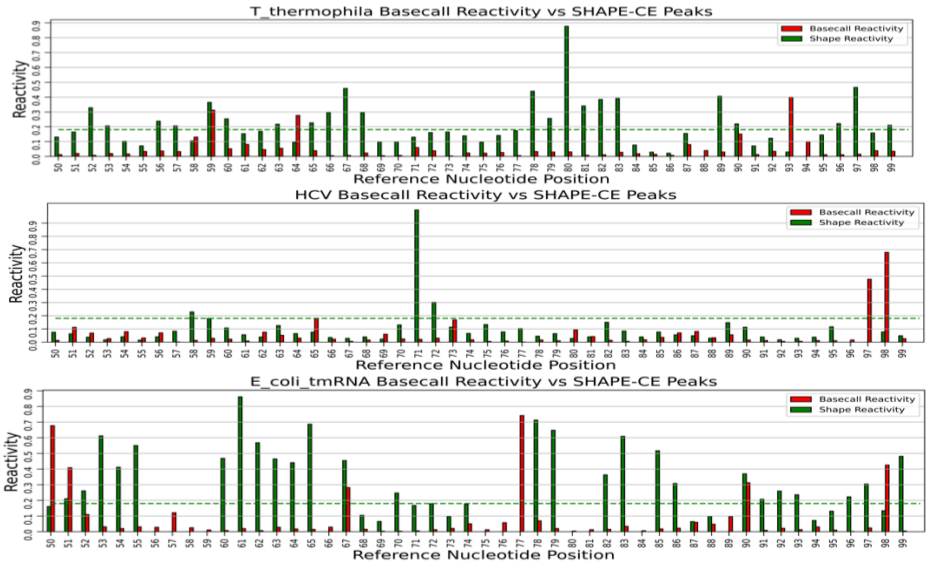

Figure 1: A. Basecall reactivity profiles were plotted against SHAPE-CE reactivity. reactivity profiles were scaled to SHAPE-CE reactivity for direct comparison. The green line indicates the SHAPE-CE reactivity threshold which is approximately .2 when scaled to unit. Both data were scaled to unit vector and the average number of reactive sites, using SHAPE-CE medium and high reactivities as ground truth, detected across our control sequences was 12 and the maximum number was 17. The average distance between the basecall error and SHAPE-CE reactivity was approximately .07. The basecall reactivity profiles were in agreement with SHAPE-CE reactivity, less than the average distance between the two measurements, on average 46 % of all measured sites.

distributions and modified nucleotides exhibit effects in reactivity of up to 2 nucleotides both upstream and downstream of the indicated nucleotide. Moreover, modified nucleotides can show moderate to significant decreases or increases in reactivity that correlated with a modified nucleotide indicated by SHAPE-CE which means the definition of an outlier in this context is highly variable. To account for non-Gaussian and non-parametric behaviour that confounds typical outlier detection, we applied a Gaussian kernel density estimate (KDE) to estimate local relative probabilities (within 3-5 nucleotide) for each base Methods A.2.5.

Next, we applied peak detection to the KDE using a 2 base peak plateau to capture adjacent modifications and compared them to SHAPE-CE reactivities to determine relative probabilities and accuracy of peaks across our control sequences (Methods A.2.6). We, again, applied the MWU and KS tests to the resulting distributions. Both tests are non-parametric tests of the null hypothesis that the distribution underlying two samples are the same. While outlier based methods were not significantly correlated with experimental SHAPE-CE data, p-values rejected the null hypothesis that the distributions are the same. P-values using peak detection indicated similar distributions for all sequences except Tetrahymena ribozyme (Table 2).

6.1 Probabilistic Peak Weights

Finally, we optimized the relative probabilities within the ±2 base ranged of a peak. Our analysis showed a consistent ranking from lowest to highest accuracy for each offset between $[-2, 2]$, where negative indicates upstream and positive indicates downstream, in comparison with SHAPE-CE reactivity. The position 2 bases upstream from a detected peak, had the highest accuracy across control sequences (Methods A.2.3). We used this observation to determine relative probabilities for each offset from $[-2, 2]$ by finding a weight vector, w, that maximized the average accuracy all sequences. Using the control sequences, we calculated the joint probability of a given base being modified at relative position to the detected peak. The w vectors ranked from lowest to highest between $[-2, 2]$ was .35, .25, .25, .1, and .2. We evaluated these relative probabilities at each offset and determined their accuracy.

7 PERFORMANCE

First, we evaluated the distribution correlations using the outlier based method, LRSUM and LNPPD. The outlier based method distributions indicated $p < .05$ meaning the distribution of bases detected *in silico* differed significantly from bases detected by SHAPE-CE using both MWU and KS test. LNPPD demonstrated improved the correlation of *in silico* detected bases to experimentally detected bases compared to both outlier and LRSUM detection method and were significantly correlated ($p > .05$), except for in the case of the Tetrahymena ribozyme where LRSUM was very highly correlated with SHAPE-CE.

Next, we evaluated the average accuracy and coverage of our method against both the outlier detection method and LRSUM [18]. LNPPD demonstrated improved coverage and detection of modified bases indicated by SHAPE-CE compared to both the outlier detection method discussed in this work and LRSUM. Average accuracy of detected modified bases across all sequences increased on average by 12% compared to the outlier method and by an average of 22% compared to LRSUM.

8 DISCUSSION

All methods of analysis on basecall data have several caveats including accommodating changes in underlying methods, imbalanced data sets, the use of averaged reads across many sequences, and inherent noise and variation of sequence structure. Additionally, the basecalling algorithm used by Oxford Nanopore have changed several times [25] and therefore the effectiveness of any method may be limited by these changes. Non-parametric approaches, like LNPPD, may more easily adapt to and accurately capture dynamic effects. Modifications to native or *in vitro* transcribed RNA, may yield imbalanced data sets, unlike synthetic modifications. As a result, different measures of accuracy may be required based on the prediction task. For instance, imbalanced data sets are not well suited to machine learning methods requiring balanced training sets and straight forward anomaly detection may rely too heavily on statistical assumptions. Approaches like LNPPD accommodate a more robust method of anomaly detection. While many works to date are based on averaged reads across many sequences, additional orthogonal experiments with single nucleotide resolution may further improve results. Finally, the noise and variance we see is often sequence dependent as observed in all algorithms the accuracy is related to the structure of the sequence, proximity of modifications and other factors which have yet to be explored. Based on this, we conclude that Guppy basecall errors can be used to determine structural modifications, but due to the inherent noise and variation across sequences there may be a more practical use as an a priori probability. We designed the probabilistic peak weights (Section 6.1) to act as a control for variance in downstream analysis.

9 CONCLUSION

We have presented a novel method for detection of structural modification in Oxford Nanopore's basecall data that does not require the use of synthetic sequences to detect changes. LNPPD does not rely on pre-trained data or synthetic sequences, it is agnostic to the type of change whether it is a m^6A endogenous modification

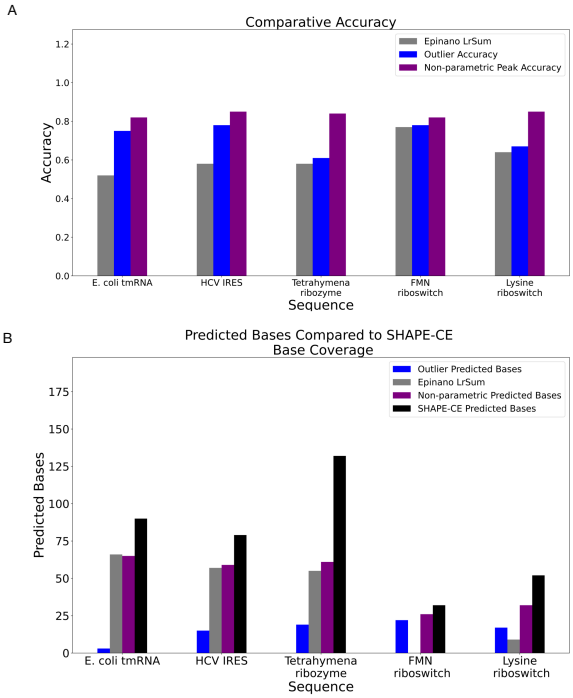

Figure 2: A. Accuracies were measured for each sequence and compared using Epinano's method (gray), outlier detection method (blue) and the non-parametric approach (purple). In all cases non-parametric peak detection outperforms other methods. It specifically outperforms LRSUM by on average 22%. B. Depicts the increased coverage or number of predicted bases for each sequence compared to experimentally predicted bases, SHAPE-CE: LRSUM (gray), outlier detection (blue), non-parametric peak detection (purple), and SHAPE-CE (black). Both LRSUM and LNPPD significantly outperform outlier detection methods. However, the differences in coverage between LRSUM and LNPPD is comparable. Base coverage more than doubled using LNPPD and more closely approached the number of bases detected using SHAPE-CE for FMN riboswitch and Lysine Riboswitch.

or AcIm chemical modification and, therefore, more generalizeable than existing methods. Moreover, our method emphasizes a non-parametric approach that does not rely on the assumption of standard Guassian distributions to detect outliers as modifications. Instead, localized peak detection more accurately distinguishes between persistent non-specific anomalies and actual modified sites. Finally, weighted probabilities allow for fine-tuning of accuracy and coverage over multiple sequences, and optimization for specific modification features. We demonstrated that LNPPD offers improved accuracy and coverage over existing methods and more closely aligns with experimental predictions. This suggests that non-parametric methods, like LNPPD, have wider applicability to diverse sequences and modifications.

In the future, as more work on Oxford Nanopore's basecall and signal analysis becomes available, the estimated probabilities could

Table 2: P-Values of Distribution Correlations

Sequence	Outlier based MWU & KS Test	LRSUM based MWU & KS Test	LNPPD based MWU & KS Test
Tetrahymena ribozyme	2.84e-08, 1.10e-42	0.65, .99	3.75e-09, 4.16e-06
E. coli tmRNA	0.0003, 4.54e-30	.007, .099	0.02, 0.36
HCV IRES	8.05e-11, 1.05e-49	1.82e-05, .001	0.06, 0.07
FMN riboswitch	1.55e-07, 2.79e-19	2.67e-06, .02	0.38, 0.99
Lysine riboswitch	6.39e-17, 4.42e-34	.001, .099	0.01, 0.29

serve as orthogonal information for detection of structural modification along with signal analysis and other sequence characteristics. Future work should focus on expanding this characterization across a large number of RNA sequences and experiments and developing better probabilistic estimates across larger datasets for application to novel RNA.

ACKNOWLEDGMENTS

The authors would like to acknowledge Bruno Sargueil for contribution of RNA sequences and IRCM molecular biology platform for its technical support. Funding: FRQS postdoctoral fellowship to GDB

REFERENCES

[1] [n. d.]. scipy.signal.find_peaks — SciPy v1.11.1 Manual. https://docs.scipy.org/doc/scipy/reference/generated/scipy.signal.find_peaks.html
[2] [n. d.]. scipy.stats.gaussian_kde — SciPy v1.11.1 Manual. https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.gaussian_kde.html
[3] 2023. Pysam. https://github.com/pysam-developers/pysam original-date: 2014-02-05T20:38:10Z.
[4] Jonathan S Abebe, Alexander M Price, Katharina E Hayer, Ian Mohr, Matthew D Weitzman, Angus C Wilson, and Daniel P Depledge. 2022. DRUMMER—rapid detection of RNA modifications through comparative nanopore sequencing. *Bioinformatics* 38, 11 (June 2022), 3113–3115. https://doi.org/10.1093/bioinformatics/btac274
[5] Jong Ghut Ashley Aw, Shaun W. Lim, Jia Xu Wang, Finnlay R. P. Lambert, Wen Ting Tan, Yang Shen, Yu Zhang, Pornchai Kaewsapsak, Chenhao Li, Sarah B. Ng, Leah A. Vardy, Meng How Tan, Niranjan Nagarajan, and Yue Wan. 2021. Determination of isoform-specific RNA structure with nanopore long reads. *Nature Biotechnology* 39, 3 (March 2021), 336–346. https://doi.org/10.1038/s41587-020-0712-z Number: 3 Publisher: Nature Publishing Group.
[6] RCSB Protein Data Bank. [n. d.]. RCSB PDB: Homepage. https://www.rcsb.org/
[7] Jean-Denis Beaudoin, Eva Maria Novoa, Charles E. Vejnar, Valeria Yartseva, Carter M. Takacs, Manolis Kellis, and Antonio J. Giraldez. 2018. Analyses of mRNA structure dynamics identify embryonic gene regulatory programs. *Nature structural & molecular biology* 25, 8 (Aug. 2018), 677–686. https://doi.org/10.1038/s41594-018-0091-z
[8] Katherine E. Berry, Shruti Waghray, Stefanie A. Mortimer, Yun Bai, and Jennifer A. Doudna. 2011. Crystal Structure of the HCV IRES Central Domain Reveals Strategy for Start-Codon Positioning. *Structure* 19, 10 (Oct. 2011), 1456–1466. https://doi.org/10.1016/j.str.2011.08.002 Publisher: Elsevier.
[9] Teshome Tilahun Bizuayehu, Kornel Labun, Martin Jakubec, Kirill Jefimov, Adnan Muhammad Niazi, and Eivind Valen. 2022. Long-read single-molecule RNA structure sequencing using nanopore. *Nucleic Acids Research* 50, 20 (Nov. 2022), e120. https://doi.org/10.1093/nar/gkac775
[10] Petr Danecek, James K Bonfield, Jennifer Liddle, John Marshall, Valeriu Ohan, Martin O Pollard, Andrew Whitwham, Thomas Keane, Shane A McCarthy, Robert M Davies, and Heng Li. 2021. Twelve years of SAMtools and BCFtools. *GigaScience* 10, 2 (02 2021), giab008. https://doi.org/10.1093/gigascience/giab008 arXiv:https://academic.oup.com/gigascience/article-pdf/10/2/giab008/36332246/giab008.pdf
[11] Grégoire De Bisschop and Bruno Sargueil. 2021. RNA Footprinting Using Small Chemical Reagents. In *RNA Scaffolds: Methods and Protocols*, Luc Ponchon (Ed.). Springer US, New York, NY, 13–23. https://doi.org/10.1007/978-1-0716-1499-0_2
[12] Daniel Dominguez, Peter Freese, Maria S. Alexis, Amanda Su, Myles Hochman, Tsultrim Palden, Cassandra Bazile, Nicole J. Lambert, Eric L. Van Nostrand, Gabriel A. Pratt, Gene W. Yeo, Brenton R. Graveley, and Christopher B. Burge. 2018. Sequence, Structure, and Context Preferences of Human RNA Binding Proteins. *Molecular Cell* 70, 5 (June 2018), 854–867.e9. https://doi.org/10.1016/j.molcel.2018.05.001
[13] Jie Fu, Yaser Hashem, Iwona Wower, Jianlin Lei, Hstau Y Liao, Christian Zwieb, Jacek Wower, and Joachim Frank. 2010. Visualizing the transfer-messenger RNA as the ribosome resumes translation. *The EMBO Journal* 29, 22 (Nov. 2010), 3819–3825. https://doi.org/10.1038/emboj.2010.255 Publisher: John Wiley & Sons, Ltd.
[14] Piroon Jenjaroenpun, Thidathip Wongsurawat, Taylor D Wadley, Trudy M Wassenaar, Jun Liu, Qing Dai, Visanu Wanchai, Nisreen S Akel, Azemat Jamshidi-Parsian, Aime T Franco, Gunnar Boysen, Michael L Jennings, David W Ussery, Chuan He, and Intawat Nookaew. 2021. Decoding the epitranscriptional landscape from native RNA sequences. *Nucleic Acids Research* 49, 2 (Jan. 2021), e7. https://doi.org/10.1093/nar/gkaa620
[15] F. Karabiber, J. L. McGinnis, O. V. Favorov, and K. M. Weeks. 2013. QuShape: Rapid, accurate, and best-practices quantification of nucleic acid probing information, resolved by capillary electrophoresis. *RNA* 19, 1 (Jan. 2013), 63–73. https://doi.org/10.1261/rna.036327.112
[16] Heng Li. 2018. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* 34, 18 (Sept. 2018), 3094–3100. https://doi.org/10.1093/bioinformatics/bty191
[17] Huanle Liu, Oguzhan Begik, Morghan C. Lucas, Jose Miguel Ramirez, Christopher E. Mason, David Wiener, Schraga Schwartz, John S. Mattick, Martin A. Smith, and Eva Maria Novoa. 2019. Accurate detection of m6A RNA modifications in native RNA sequences. *Nature Communications* 10, 1 (Sept. 2019), 4079. https://doi.org/10.1038/s41467-019-11713-9 Number: 1 Publisher: Nature Publishing Group.
[18] Huanle Liu, Oguzhan Begik, and Eva Maria Novoa. 2021. EpiNano: Detection of m6A RNA Modifications Using Oxford Nanopore Direct RNA Sequencing. In *RNA Modifications: Methods and Protocols*, Mary McMahon (Ed.). Springer US, New York, NY, 31–52. https://doi.org/10.1007/978-1-0716-1374-0_3
[19] H. B. Mann and D. R. Whitney. 1947. On a Test of Whether one of Two Random Variables is Stochastically Larger than the Other. *The Annals of Mathematical Statistics* 18, 1 (1947), 50 – 60. https://doi.org/10.1214/aoms/1177730491
[20] Frank J. Massey. 1951. The Kolmogorov-Smirnov Test for Goodness of Fit. *J. Amer. Statist. Assoc.* 46, 253 (March 1951), 68–78. https://doi.org/10.1080/01621459.1951.10500769
[21] Don Neumann, Anireddy S. N. Reddy, and Asa Ben-Hur. 2022. RODAN: a fully convolutional architecture for basecalling nanopore RNA sequencing data. *BMC Bioinformatics* 23, 1 (April 2022), 142. https://doi.org/10.1186/s12859-022-04686-y
[22] Matthew T Parker, Katarzyna Knop, Anna V Sherwood, Nicholas J Schurch, Katarzyna Mackinnon, Peter D Gould, Anthony JW Hall, Geoffrey J Barton, and Gordon G Simpson. 2020. Nanopore direct RNA sequencing maps the complexity of Arabidopsis mRNA processing and m6A modification. *eLife* 9 (Jan. 2020), e49658. https://doi.org/10.7554/eLife.49658 Publisher: eLife Sciences Publications, Ltd.
[23] William Stephenson, Roham Razaghi, Steven Busan, Kevin M. Weeks, Winston Timp, and Peter Smibert. 2022. Direct detection of RNA modifications and structure using single-molecule nanopore sequencing. *Cell Genomics* 2, 2 (Feb. 2022), 100097. https://doi.org/10.1016/j.xgen.2022.100097
[24] Xi-Wen Wang, Chu-Xiao Liu, Ling-Ling Chen, and Qiangfeng Cliff Zhang. 2021. RNA structure probing uncovers RNA structure-dependent biological functions. *Nature Chemical Biology* 17, 7 (July 2021), 755–766. https://doi.org/10.1038/s41589-021-00805-7 Number: 7 Publisher: Nature Publishing Group.
[25] Ryan R. Wick, Louise M. Judd, and Kathryn E. Holt. 2019. Performance of neural network basecalling tools for Oxford Nanopore sequencing. *Genome Biology* 20, 1 (June 2019), 129. https://doi.org/10.1186/s13059-019-1727-y
[26] Kevin A Wilkinson, Edward J Merino, and Kevin M Weeks. 2006. Selective 2-hydroxyl acylation analyzed by primer extension (SHAPE): quantitative RNA structure analysis at single nucleotide resolution. *Nature Protocols* 1, 3 (Nov. 2006), 1610–1616. https://doi.org/10.1038/nprot.2006.249
[27] Xichen Zhao, Yuxin Zhang, Daiyun Hang, Jia Meng, and Zhen Wei. 2022. Detecting RNA modification using direct RNA sequencing: A systematic review. *Computational and Structural Biotechnology Journal* 20 (Jan. 2022), 5740–5749.

https://doi.org/10.1016/j.csbj.2022.10.023

10 APPENDICES

A RESEARCH METHODS

A.1 Experimental Methods

A.1.1 ***In vitro*** *transcription.* DNA templates were amplified by PCR using custom oligos in order to add a T7 promoter and a structured cassette 3' of the sequence of interest (as in [26]). RNA were *in vitro* transcribed from PCR amplified templates using T7 RNA polymerase (NEB) for 2 hours at 37°C. DNA templates were digested with DNAse I (NEB) for 15 minutes at 37°C and RNA products were purified using SPRI beads (Beckman). RNA were quantified with a Nanodrop 2000 (Thermo) and their integrity was assessed by gel electrophoresis.

A.1.2 SHAPE-CE. SHAPE-CE was performed as in [11]. 12 pmoles of RNA were folded in 40 µL as for electromobility shift assays. One 18 uL aliquot was added to 2 µL SHAPE probe (0.75 M AcIm or 1M BzCN or DMS) and one 18 µL aliquot was added to 2 µL DMSO as the non-modified control and incubated for 10 minutes. RNA were precipitated with 0.3 M sodium acetate, 1 µg glycogen and 75% ethanol at -20°C. After centrifugation, pellets were washed in ethanol 75%, air dried and resuspended in 10 µL H2O. RNA were reverse transcribed using MMLV reverse transcriptase (Thermo) for 30 minutes at 42°C with 0.3 µM 6-FAM-labeled primer (Applied Biosystems). A sequencing reaction was set in parallel with a NED-labeled primer and 0.5 mM ddTTP (Jena Bioscience). cDNA from probing and sequencing reaction were then mixed and precipitated with 0.3 M sodium acetate, 1 µg glycogen and 75% ethanol, resuspended in HiDi formamide and submitted to capillary electrophoresis (Applied Biosystems). Flurorescence signals were processed using QuSHAPE [15] to determine SHAPE reactivities.

A.1.3 ONT sequencing. RNA were polyadenylated with E. coli polyA polymerase (NEB). Polyadenylation efficiency was assessed by gel electrophoresis and RNA were purified on RNAClean & Concentrator columns (Zymo Research). Polyadenylated RNA were modified with AcIm as for SHAPE-CE, then purified on RNAClean & Concentrator columns. 200 ng of RNA processed for ONT sequencing using SQK-RNA002 kit (ONT) using the provided RTA adapter. Superscript IV (Thermo) was substitued to Supercript III with no notable effect. Ligated, reverse-transcribed RNA were loaded on Flongle flow cells (R9.4.1 chemistry) and run until the number of active pore dropped bellow 5.

A.2 Data Analysis

A.2.1 Basecalling Nanopore Reads. RNA sequences were basecalled with ONT Guppy basecalling software version 5.0.11 [25] with filtering at QC >= 7 during sequencing to ensure at least 20,000 reads were obtained for non-structure probed RNA (control). Subsequently, all RNA were basecalled without filtering and aligned to their respective sequences using Minimap2 version 2.24 [16]. For noncoding reference sequences and structure were obtained from PDB [6] and from existing literature for HCV IRES[8] and E. coli tmRNA [13].

A.2.2 Basecalling Reactivity Profiles and Transformation. Basecalling reactivity profiles were generated for control sequences using Guppy [25] for both unmodified (DMSO) and modified (AcIm) control sequences. Basecalling reactivity profiles were generated using Pysam [3], Samtools [10], and Minimap2 [16] to obtain insertion, deletion, and mismatch errors for each read and each nucleotide.

For all reads, at position, i, in sequence, S, the basecalling error at p_i is the sum of the insertions, I, deletions, D, and mismatches, M at p_i divided by the total number of reads at p_i. The basecalling reactivity profile consists of all $p_i \in S$.

$$p_i = \frac{\sum (I_i + D_i + M_i)}{\|p_i\|} \tag{1}$$

We evaluate the empirical cumulative distribution function (cdf) for each reactivity profile using the standard discrete cdf where $F_{dmso} = \sum_{x_i \leq x} \Pr(x_i)$ and $F_{AcIm} = \sum_{x_i \leq x} \Pr(x_i)$ and difference, Δ_F is calculated as $F_{AcIm} - F_{dmso}$.

A.2.3 Measuring Accuracy. Accuracy is based on a 2 base offset both upstream and downstream are calculated for all controls since nucleotide modifications have been observed exhibiting both upstream and downstream effects. We calculated accuracy based on ±2 nucleotides from an indicated modification. Meaning if a base is identified as having a modification by a method, then it is counted as true if there is a modification indicated by SHAPE-CE within ±2 nucleotides. Accuracy is defined as $\frac{TP+TN}{P+N}$ where TP indicates true positive, TN indicates true negative, p indicates positive, and n indicates negative. Due to imbalanced data sets balanced accuracy may be a better measure of accuracy for some sequences.

A.2.4 Statistical Outliers. Statistical outliers were defined as $\mu \pm k\sigma$, where k represented the factor of standard deviations required to detect outlier. k was varied to account for non-Gaussian distributions present in reactivity profiles. The optimal k was chosen to compute the performance metrics for each sequence and the best k for all sequences was used to define the final metric.

A.2.5 Kernel Density Estimate. We apply scikit-learn's kernel density estimation [2] using a Gaussian kernel and the "scott" bandwith method to $F_{kde} = \text{KDE}(\Delta F)$ (Methods A.2.2) detect local maxima and minima using a 3 base sequential subset of nucleotides for each estimate, we iterate over each $F_{kde}[i : i+2]$. Using all of F_{kde} elicits global maxima and minima that are not ideal as the cdf is a monotonically increasing function to observe local changes. Variation and noise in basecalling reactivity do not necessarily allow for a global threshold.

A.2.6 Peak Detection. We apply scikit-learn's peak detection method to the KDE of F_{kde} (Methods A.2.5) detect local maxima and minima within a maximum width (or plateau)of 2 nucleotides to allow for adjacent nucleotide modification [1]. Initial examination suggests that there may be some further improvement in applying additional transforms such as wavelet to Δ_F before performing peak detection. Our methods applies the non-parametric kernel density estimate (Methods A.2.5).

B SUPPLEMENTAL DATA

B.1 Accuracy by Basecall Column Feature by Sequence

In Table 3, we show that basecall reactivity correlated well with column reactivity, being on average at or above the maximum accuracy for individual columns except in the case of Lysine. We opted to use basecall reactivity (Methods A.2.2) as selecting individual sets of columns as suggested in Epinano [18] yielded little overall advantage in accuracy and the sum in basecall reactivity could aptly capture changes.

B.2 Accuracies for Linear Regression Sum of Errors with Outlier Detection using z-score

In Table 4, we applied Epinano's [18] method of using the reconstruction error between actual errors and errors predicted by linear regression using the basecall columns as features. We used the recommended columns of insertion, mismatch, deletion, and quality for all predictions. Outlier values were tested at $z \geq 1$ and $z \geq 2$, and set at $z \geq 2$ as this produced the best accuracy for this method, however it did reduce coverage and overlap with SHAPE-CE, significantly in some cases. This method was calculated with the same offset values used in peak detection, and performed best with an offset of -2 (Table 3). We observed that varying columns would need to be customized for each sequence and this method did not lend itself well to a fully unsupervised prediction method. As doing feature determination in this way requires a more supervised approach. Therefore, we opted to use all recommended columns for each sequence for a reasonable comparison to our method. We did not apply the optional Bonferroni correction to this method and that may offer some additional improvement in accuracy.

$$\begin{bmatrix} r_2 & r_4 & r_5 & r_1 & r_3 \end{bmatrix} * \begin{bmatrix} w_2 & w_4 & w_5 & w_3 & w_1 \end{bmatrix} = \begin{bmatrix} p_1 & p_2 & p_3 & p_4 & p_5 \end{bmatrix}$$

$$P(s_i = peak \wedge s_i = mod) = \begin{bmatrix} w_1 & w_2 & w_3 & w_4 & w_5 \end{bmatrix}$$

Table 3: Accuracy by Basecall Column Feature by Sequence

Sequence	Mismatch	Insertion	Deletion	Quality	Basecall Reactivity	Most Reactive Column
Tetrahymena ribozyme	.62	.60	.57	0	.61	Mismatch
E. coli tmRNA	.71	.72	.71	0	.75	Insertion
HCV IRES	.71	.76	.79	0	.79	Deletion
FMN riboswitch	.79	.79	.79	0	.78	Equal
riboswitch	.75	.75	.75	0	.68	Equal

Table 4: Accuracies for Linear Regression Sum of Errors with Outlier Detection

Sequence	Accuracy	Number of Predicted Bases	Number of SHAPE-CE Predicted Bases	Coverage
Tetrahymena ribozyme	.52	126	138	66
E. coli tmRNA	.58	123	90	57
HCV IRES	.58	132	79	55
FMN riboswitch	.77	5	55	0
Lysine riboswitch	.64	27	52	9