

Deep learning and direct sequencing of labeled RNA captures transcriptome dynamics

Vlastimil Martinek ^{1,2,3,†}, Jessica Martin ^{1,4,†}, Cedric Belair ¹, Matthew J. Payea ¹, Sulochan Malla ¹, Panagiotis Alexiou ^{5,‡} and Manolis Maragkakis ^{1,*,‡}

¹Laboratory of Genetics and Genomics, National Institute on Aging, Intramural Research Program, National Institutes of Health, Baltimore, MD 21224, USA

²Central European Institute of Technology, Masaryk University, 625 00 Brno, Czech Republic

³National Centre for Biomolecular Research, Faculty of Science, Masaryk University, 625 00 Brno, Czech Republic

⁴Center for Alzheimer's and Related Dementias, National Institute on Aging and National Institute of Neurological Disorders and Stroke, National Institutes of Health, Bethesda, MD, USA

⁵Centre for Molecular Medicine & Biobanking, University of Malta, MSD 2080 Msida, Malta

*To whom correspondence should be addressed. Tel: +1 410 454 8429; Email: maragkakis@nih.gov

†The first two authors should be regarded as Joint First Authors.

‡The last two authors should be regarded as Joint Last Authors.

Abstract

In eukaryotes, genes produce a variety of distinct RNA isoforms, each with potentially unique protein products, coding potential or regulatory signals such as poly(A) tail and nucleotide modifications. Assessing the kinetics of RNA isoform metabolism, such as transcription and decay rates, is essential for unraveling gene regulation. However, it is currently impeded by lack of methods that can differentiate between individual isoforms. Here, we introduce RNAkinet, a deep convolutional and recurrent neural network, to detect nascent RNA molecules following metabolic labeling with the nucleoside analog 5-ethynyl uridine and long-read, direct RNA sequencing with nanopores. RNAkinet processes electrical signals from nanopore sequencing directly and distinguishes nascent from pre-existing RNA molecules. Our results show that RNAkinet prediction performance generalizes in various cell types and organisms and can be used to quantify RNA isoform half-lives. RNAkinet is expected to enable the identification of the kinetic parameters of RNA isoforms and to facilitate studies of RNA metabolism and the regulatory elements that influence it.

Introduction

The life cycle of RNA from transcription to decay is tightly regulated in cells. Unraveling of the kinetic parameters of RNA metabolism, such as transcription and decay rate, is critical for understanding gene regulation and cellular response to environmental cues. Existing methods for studying RNA dynamics in mammalian cells are based on short-read RNA sequencing (RNA-Seq) and involve the exposure of cells to nucleoside analogs such as 4-thiouridine (4sU) or 5'-bromouridine (BrU) that get incorporated during transcription into newly synthesized RNAs (1–8). Subsequent steps involve either the chemical isolation of metabolically labeled RNAs before sequencing or the bioinformatic inference of labeling through PCR-generated mutations after sequencing (9,10). These methods have substantially simplified the methodology, but all share common limitations. First, due to the use of short-read sequencing, established methods cannot make high-confidence assignments of individual RNA isoforms, a substantial component of RNA regulation with relevance in biology and medicine (11). Second, they do not collect information on individual RNA molecules but rather involve cDNA synthesis and amplification steps that erase information encoded in RNA post-transcriptionally. Consequently, existing methods are limited in their ability to associate RNA kinetics with *cis*-acting transcriptional regulators such as transcription start site and poly(A) site selection, RNA modifi-

cations, and poly(A) tail length, known regulators of RNA metabolism (12).

Nanopore long-read sequencing with the ability to directly detect nucleotides based on electrical current intensity as RNA molecules pass through nanopores has emerged as a promising avenue to address these limitations (13,14). Notably, besides canonical nucleotides, tools have been developed to also detect naturally occurring modified nucleotides such as pseudouridine and N6-methyladenosine. Some of these tools are comparative and require control samples to detect shifts in signal-based features that correlate with the presence of modifications (15–18) while others use machine learning on training data to detect specific modifications in single samples (19–23). The ability to detect modified RNA nucleotides should also enable the design of a new set of studies that involve the direct detection of metabolically incorporated modified nucleotides in RNA molecules. Detection of such molecules following metabolic labeling would allow the *in-silico* separation of newly synthesized RNAs and thus the quantification of RNA kinetic parameters along with transcriptional and post-transcriptional features of RNAs at single-molecule resolution (24).

In this work, we present RNAkinet, a computationally efficient, convolutional, and recurrent neural network (NN) that identifies individual 5EU-modified RNA molecules following direct RNA-Seq (dRNA-Seq). RNAkinet predicts the modification status of RNA molecules directly from the raw

nanopore signal without using basecalling or reference sequence alignment. We show that RNAkinet shows improved generalization to sequences from unique experimental settings, cell types, and species compared to existing methods (24). Tested on independent biological datasets, we find that RNAkinet accurately quantifies RNA kinetic parameters, from single time point experiments. Being able to interrogate whole RNA molecules, we anticipate RNAkinet will allow future applications on the combined study of RNA metabolism of single isoforms and *cis*-acting RNA regulatory elements, such as alternative splicing, poly(A) tail length and post-transcriptional RNA modifications.

Materials and methods

Cell culture and metabolic labeling

NIH3T3 cells (ATCC CRL-1658) were cultured at 37°C, 5% CO₂, 90% humidity in Dulbecco's Modified Eagle's Medium (ThermoFisher) supplemented with 10% bovine calf serum (GeminiBio), 1% MEM-nonessential amino acids (Invitrogen), 2 mM L-glutamine. Cells were passed weekly by gentle dissociation with trypsin-EDTA 0.25%. HeLa cells (ATCC CCL-2) and HEK-293T (ATCC CRL-3216) were cultured at 37°C, 5% CO₂, 90% humidity in Dulbecco's Modified Eagle's Medium (ThermoFisher) supplemented with 10% fetal bovine serum (GeminiBio), 1% MEM-nonessential amino acids (Invitrogen), 2 mM L-glutamine. iPSC-derived i3Neurons were cultured and differentiated as previously described (25). For metabolic labeling, cells were cultured in media containing 400 or 500 μM 5-ethynyl-uridine (ThermoFisher) for 2–24 h as indicated. Total RNA was extracted using TRIzol reagent (Invitrogen) according to manufacturer's instructions followed by DNase I treatment (MilliporeSigma). RNA concentration and integrity were determined using a Nanodrop ND-1000 (ThermoFisher) and Qubit RNA IQ assay (ThermoFisher), respectively. Library preparation for direct RNA sequencing was performed as previously described (26) with modifications. Briefly, poly(A) RNAs were purified from 50 μg of total RNA using Oligo d(T)₂₅ Magnetic Beads (New England Biolabs). 500 ng of poly(A) mRNA was used for library preparation using SQK-RNA002 sequencing kit (Oxford Nanopore Technologies). The final library was quantified using Qubit dsDNA High Sensitivity assay kit (ThermoFisher) and sequenced on a MinION device using FLO-MIN106 flow cells (Oxford Nanopore Technologies).

Data preparation for training, validation, and testing

In order to prevent data leakage and overoptimistic results that could emerge from using *k*-fold cross-validation with random shuffling, the reads were split into training, validation, and testing sets based on the chromosome they originated from. An alternative, by randomly splitting the data into 70% for training, 20% for testing, and 10% for validation was also tested (Supplementary Note). The reads were basecalled with Guppy 6.4.8, aligned to the Ensembl human genome (GRCh38) and transcriptome with Minimap2 (27), and then separated into splits based on their mapped chromosome. All reads that did not map to any chromosome and secondary reads were discarded. Chromosome 1 was used for testing, chromosome 20 was used for validation (utilized for early stopping), and the rest were used for training. The first

5000 raw signal values from each read were cropped to avoid sequencing artifacts and any reads shorter than 5000 raw signal values or longer than 400 000 raw signal values were discarded. The filtered reads were then normalized by median absolute deviation before being passed through the neural network.

Neural network design and training

Due to the temporal nature of nanopore signals, the neural network architecture was chosen to contain a combination of convolutional layers and recurrent layers. The purpose of the convolutional layers was to transform the input signal into local feature vectors corresponding to areas of the input signal. Correspondingly, the purpose of the recurrent layers was to aggregate these feature vectors along the entire read. Notably, the convolution was performed along the entire input signal and thus produced a vector of varying size that could be consumed by the subsequent recurrent block to produce a final fixed length vector, regardless of the read length (Figure 1B).

Specifically, the convolutional section consisted of five blocks each containing an 1D-convolutional, a ReLU activation, and a max pooling layer. The number of filters in the convolutional layer doubled with each subsequent convolutional block, starting with 8 filters in the first block and ending with 128 filters in the last block (Figure 1B). To transform the CNN-extracted local feature vectors into vectors that contain information computed from the whole read, bidirectional recurrent layers (GRU) were placed on top of the convolutional layers. The GRU layer processed the CNN-extracted features along the entire read length, outputting 32 GRU hidden states. All hidden states were then pooled along the read length with max and average pooling. These pooled vectors were then concatenated with the last hidden states of the recurrent layers, similar to (28). This process was performed for the forward and reverse read directions, and the output of both directions was concatenated in a fixed-size vector of size 192 that was independent of the input signal length (Figure 1B). This vector was then fed into a small dense feed-forward network with 30 hidden neurons followed by ReLU and 1 output neuron followed by sigmoid to output the final SEU modification score (Figure 1B).

The network was trained with data from the training split described in the data preparation section and was optimized to perform a binary classification of raw nanopore signals into either modified or unmodified categories. The input for the network was a one-dimensional array of the normalized nanopore electrical current signal produced by the nanopore sequencer for each read in a FAST5 file. These signals were normalized and filtered as described in the data preparation section.

The property of the convolutional and recurrent blocks to consume a signal regardless of its length and to produce a fixed-size vector, allowed the network to accept input signals of varying lengths without requiring padding. For training the network on sequences of variable length, a batch size of 1 was used. Additionally, to simulate minibatch learning, gradients were accumulated over 64 sequences. The network was trained with a learning rate of 0.001 and weight decay of 0.01 using the AdamW optimizer. The model was trained for 1000 epochs with early stopping on the validation set AUC-ROC metric with threshold 0 and patience 50 evaluation steps. The first 1000 learning steps were used as warm

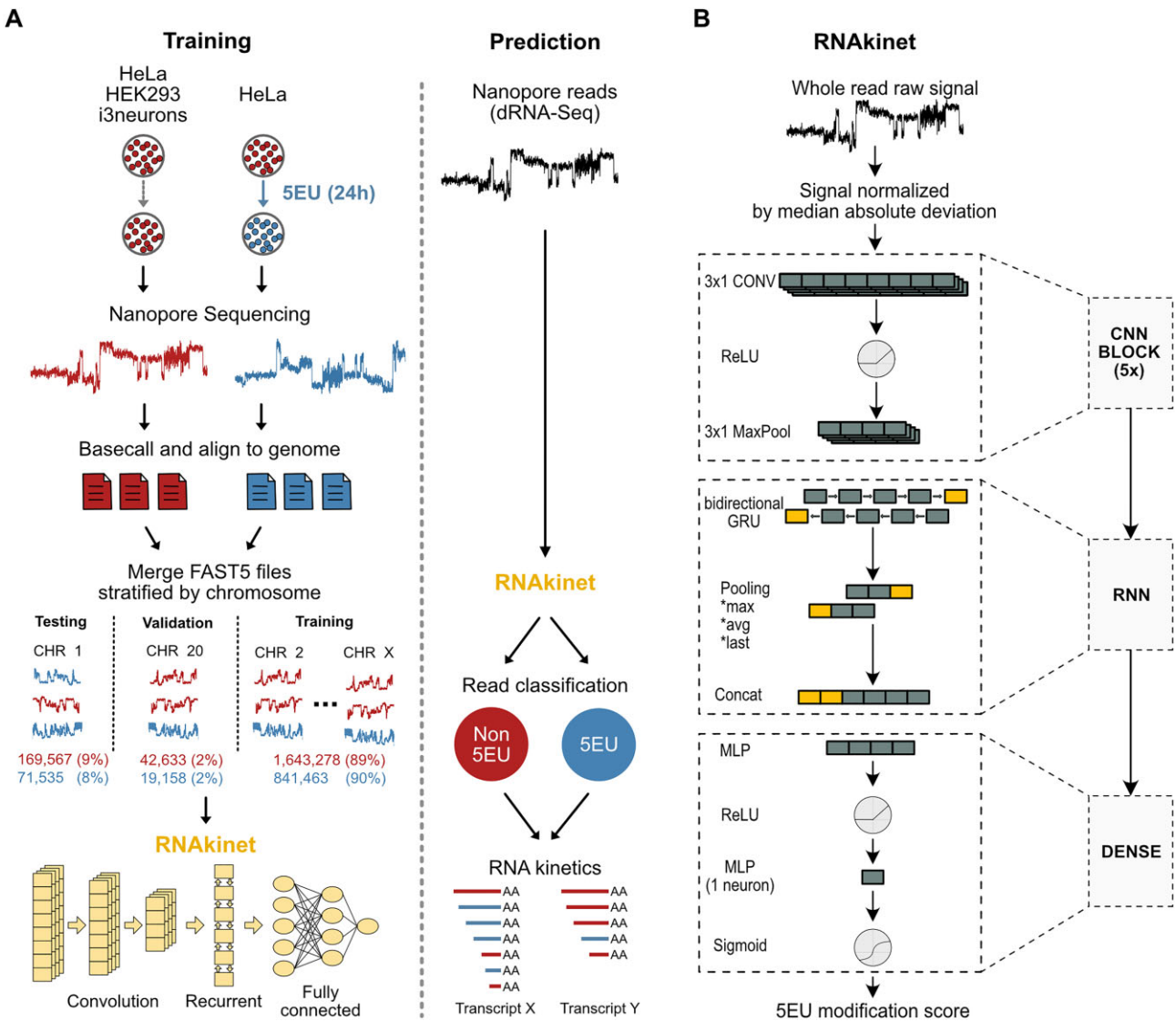


Figure 1. Data preparation and neural network design. **(A)** Schematic of training and prediction workflow. Reads from chromosomes 1 and 20 were retained for testing and validation only. During prediction RNAkinet uses only the raw signal. **(B)** Schematic representation of the architecture of RNAkinet.

up where the learning rate linearly increased from $\frac{1}{3} * 10^{-4}$ to the final learning rate of 0.001. For network training, a single A100 Nvidia GPU was utilized. The implementation and training of the network were done using the PyTorch (29) and PyTorch Lightning frameworks. Snakemake (30) workflows were developed for the entire process of data splitting, model creation, and evaluation to be reproducible, scalable to large computational clusters, and adaptable to be used on newly sequenced datasets.

Calculation of RNA isoform half life

With time denoted as t , the number of modified reads of a given RNA isoform at time $t = 0$, when metabolic labelling starts, is $M_0 = 0$. Similarly, the number of non-modified reads at $t = 0$ is $N_0 = N$, where N is the total normalized number of reads. The cells are metabolically labeled for labeling time $t = t_L$ resulting in M_{t_L} modified reads and N_{t_L} non-modified reads at the end of labeling. Under an exponential decay model, the number of non-modified reads N_t during metabolic labeling

is given by the equation:

$$N_t = N_0 e^{-\lambda t}, \quad t < t_L$$

where λ is an exponential decay constant. Defining h as the half-life time when N_h would be equal to $N_0/2$, the equation becomes

$$N_t = N_0 2^{-\frac{t}{h}}, \quad t < t_L$$

Solving for h , we get that for $t < t_L$ the half-life h is given by the equation:

$$h = -t \frac{\ln(2)}{\ln\left(\frac{N_t}{N_0}\right)}$$

Under a steady state assumption, at the end of metabolic labeling at time t_L we have that

$$N = N_0 = N_{t_L} + M_{t_L}. \text{ Therefore } h \text{ is given by:}$$

$$h = -t_L \frac{\ln 2}{\ln\left(\frac{N_{t_L}}{N_{t_L} + M_{t_L}}\right)} = -t_L \frac{\ln 2}{\ln\left(1 - \frac{M_{t_L}}{N_{t_L} + M_{t_L}}\right)}$$

Results

Data preparation and neural network design

We aimed to develop a prediction tool, RNAkinet, that would (a) accurately distinguish 5EU-labeled RNA from nanopore direct RNA sequencing, (b) have scalable time and space requirements for execution, (c) be future-proof and depend solely on the sequencer raw electrical signal and avoid dependencies on basecalling or alignment software, (d) allow robust quantification of the kinetics of RNA and, (e) be able to generalize to unseen datasets.

To obtain a 5EU classification dataset, we prepared data following labeling of HeLa cells with 5EU for 24 h as previously described (24). Given that the half-life of the majority of RNAs in cells is in the order of a few hours, it was expected that the majority of RNA molecules would incorporate a 5EU within a 24 h window (1,31,32). We defined this as our positive dataset while RNA from HeLa cells cultured without addition of 5EU served as our negative dataset (Supplementary Table S1). Moreover, to enhance domain representation and build a more accurate representation of the experimental background and inherent nuisances in the data, we enriched our list of negative samples with independent dRNA-Seq experiments performed separately by different operators on distinct cell lines including HeLa, HEK293 cells and induced pluripotent stem cell-derived neurons (Supplementary Tables S1 and S2).

To prevent data leakage during training, testing, and validation of RNAkinet, we followed a rigorous approach for data preparation. We first performed basecalling and alignment to the human genome and subsequently separated the raw reads by chromosome. We subsequently isolated reads on chromosome 1 and 20, accounting for 8–9% and 2% of the total reads, to be used for testing and validation, respectively (Figure 1A). This approach ensured that the network never interacted with sequences that existed in these two chromosomes during training, thus removing sequence level confounds from performance evaluation. We showed that this leave-a-chromosome-out approach improved model selection, compared to random split of the data into training, testing, and validation sets (Supplementary Note). Overall, our training data preparation aimed to reflect diversity among distinct experimental settings and to enhance model generalizability.

Following direct RNA sequencing on a MinION device, we extracted the raw electrical signal of each read and encoded it in a one-dimensional array of varying size that was dependent on the read length. Varying input length can be challenging for NNs and has been typically addressed by padding to the length of the longest input. Given that in a typical dRNA-Seq experiment read lengths vary dramatically from a few bases to several kilobases in length (Supplementary Figure S1A), most sequences would need to be heavily padded. Processing of such signals would be computationally expensive, but more importantly, would change the inherent semantics of the continuous time series values of the electrical signal. For example, unlike pad tokens in Natural Language Processing networks (33), padding with zeros or another real value would interfere with the physical meaning of the electrical signal generated from the sequencer. To avoid this, other tools have extracted fixed length windows of the input signal (34,35). However, this is not ideal for detecting 5EU modifications, since the 5EU labeling efficiency per nucleotide is low (2–3%) (24) raising the possibility of a labeled sequence not containing a modification in a given window. To address this, we designed the

network to accept reads of any length without padding the input. For the same reasons and since the exact site of 5EU incorporation is unknown, we decided to feed the entire signal to the network without segmentation. Specifically, the raw electrical signal of a read was extracted into an array and normalized by median absolute deviation. This 1-D array of normalized current values was then used as the only input to the network.

We designed RNAkinet as a NN that utilizes both convolutional (36) and recurrent layers, aiming to integrate long- and short-range interactions between the electrical signal in a read to predict the probability of a read containing a 5EU modification (Figure 1A, B). The network design was intended to predict 5EU modification probability for signals of any length. This was enabled by using convolutional layers, that share weights across the read, and pooling operations over the GRU hidden states, which allowed feature reduction into a fixed size vector representing the whole read (Figure 1B, see Materials and methods). Specifically, the network consisted of five convolutional blocks followed by a bidirectional GRU layer. The hidden states of the GRU layer were pooled with average and max pooling, and the result concatenated with the last hidden state into a fixed-size vector, independent of read length. This vector was then fed into a dense neural network with 1 hidden layer that outputted the final 5EU modification probability score. To support inference and training on varying input sizes, we passed data to the network in minibatches of size 1 thus avoiding the limitation of varying lengths not being supported in a minibatch. Based on whether a read comes from a positive or a negative set, we assigned it a binary label corresponding to either modified or unmodified, respectively and used it as prediction target for the neural network.

Our training process involved randomly sampling from our positive and negative sets uniformly to ensure a balanced ratio for training. Since we used multiple datasets as negatives, those were sampled proportionally based on the number of reads they contained. Due to the large size of our dataset, we adopted a multi-read FAST5 file format for quick loading, and we only loaded files when required to reduce the memory footprint. The same approach was used for inference, which ensured the network used the minimum computational resources. Finally, we limited the input of the model to raw nanopore signal, not utilizing any additional information about the sequence, such as basecalling, alignment, or other metadata. This makes RNAkinet adaptable for future applications since it is independent of external basecaller or alignment software.

RNAkinet accurately classifies RNA molecules labeled with 5EU

To test the performance of RNAkinet, we calculated the area under the receiver operating characteristic curve (AUC-ROC) for the testing dataset which consists of reads in chromosome 1, never seen by the NN (Figure 2A). RNAkinet scored an AUC-ROC of 0.87. When using nano-ID, a dense neural network previously developed (24) for detecting 5EU-labeled molecules, our results showed that it did not accurately distinguish labeled from unlabeled sequences, scoring an AUC-ROC of 0.54. Calculation of the balanced accuracy (BA) as the average of sensitivity and specificity and again showed that RNAkinet outperformed nano-ID (Figure 2B). In real experiments, the ratio of labeled over unlabeled reads depends on the experimental design and is thus not possible to

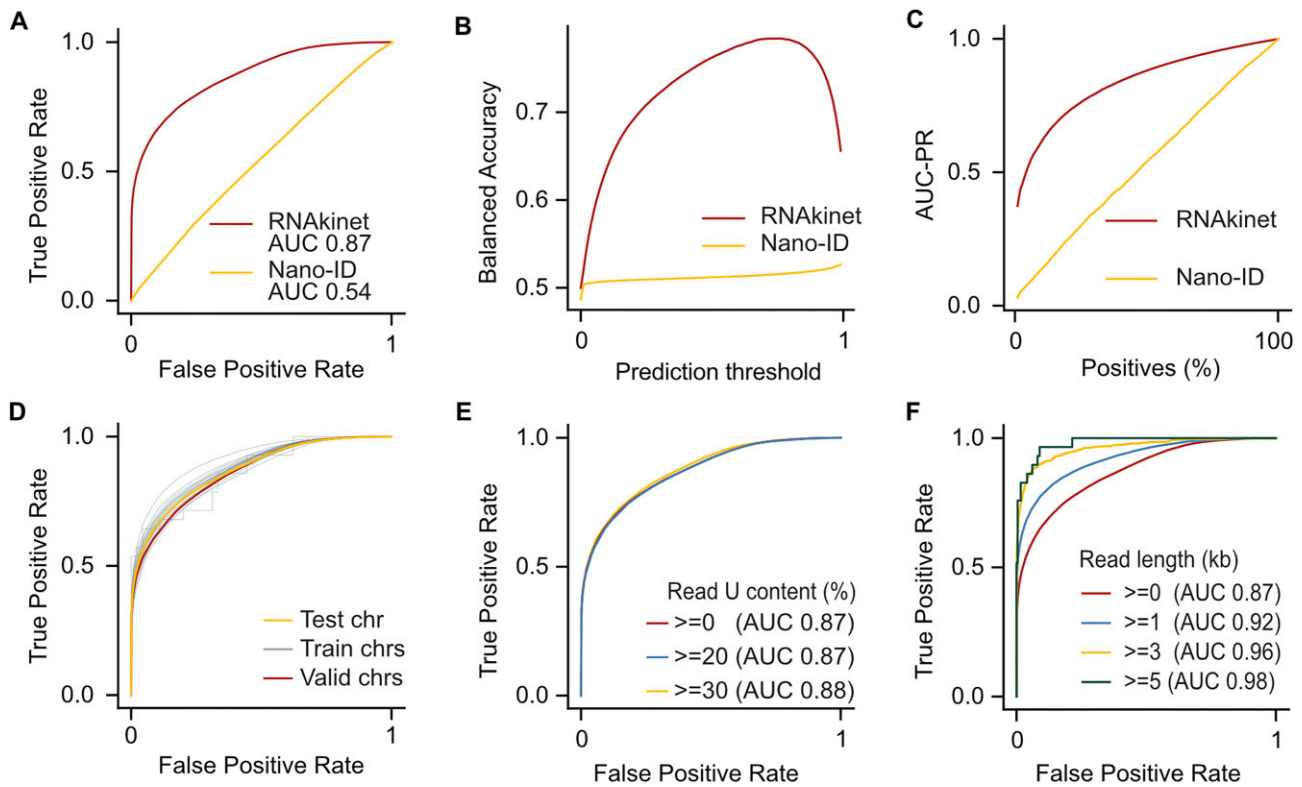


Figure 2. RNAkinet accurately classifies RNA molecules labeled with 5EU. (A–C) ROC (A), balanced accuracy (B) and area under the precision-recall curve for datasets of varying percent of positives (C) plots on test data for RNAkinet, and nano-ID (D–F) ROC plot stratified by train, validation and test chromosomes (D), percent of uridines in read (E) and read length (F).

know *a priori*. We therefore also tested the prediction performance across varying mixtures of positive and negative samples by down sampling the testing dataset to simulate various levels of class imbalance. Using these datasets, we calculated the area under the precision-recall curve (AUC-PR) (Figure 2C) and accuracy (Supplementary Figure S2A) for both RNAkinet and nano-ID. These results again showed that RNAkinet outperformed nano-ID across various positive/negative mixtures and could accurately capture 5EU signals in nanopore sequencing reads.

To further explore the training process and validate that RNAkinet does not overfit to reads in the training chromosomes, we calculated the AUC-ROC for individual chromosomes (Figure 2D). We found that RNAkinet performs similarly across reads from all chromosomes, independently of whether the reads come from training, testing, or validation chromosomes. This indicates that the training process is well controlled, preventing overfitting and thus increasing the generalization capacity of the trained model. To gain further insight into the model performance on varying read sequences, we stratified our testing reads based on U content and read length. Our data showed that the content of uridines had a negligible impact on performance with the AUC-ROC ranging from 0.89 to 0.87 (Figure 2E). In contrast, the read length appeared to have a larger effect on performance with inclusion of short reads (<1 kb) dropping performance to 0.87 AUC-ROC compared to higher than 0.92 AUC-ROC for longer reads (Figure 2F). Overall, these results indicate that RNAkinet accurately classifies 5EU-labeled molecules while its prediction performance improves with longer reads and is independent of the overall U content in the molecule.

RNAkinet generalizes across cell lines and distinguishes nascent RNA molecules

Our data had indicated that RNAkinet successfully avoided overfitting. We therefore wished to also evaluate whether it could generalize to completely independent datasets. To explore this, we analyzed previously generated nanopore data following labeling of K562 cells with 5EU for 24 h (positive) or control (negative) (24). Since this is also a human cell line, we again tested on reads that aligned to chromosome 1 to ensure that RNAkinet had not previously seen the corresponding sequence space and thus performance would not be confounded by read sequence. Our results showed that RNAkinet could successfully distinguish labeled RNA molecules along all replicates of K562 cells (Supplementary Figure S3A). Quantification of the AUC-ROC of the prediction showed that RNAkinet had comparable performance on K562 cells as in HeLa with a drop in performance from 0.87 AUC-ROC to 0.80 (Figure 3A). A drop in performance was expected given that the data in (24) were prepared with an earlier iteration of the ONT RNA sequencing kit (SQK-RNA001) the intricacies of which have not been modeled during training. Similarly, we found that RNAkinet had comparable balanced accuracy and AUC-PR, tested on varying positive/negative mixtures, on both datasets (Figure 3B, Supplementary Figure S3B). Again, we found that stratification of reads by chromosome, U content or read length resulted in minimal performance differences except for short RNAs (<1 kb) that again showed similarly reduced performance as in the original HeLa dataset (Figure 3C–E).

Our results have shown that RNAkinet could successfully identify 5EU-labeled RNA molecules from different cell types

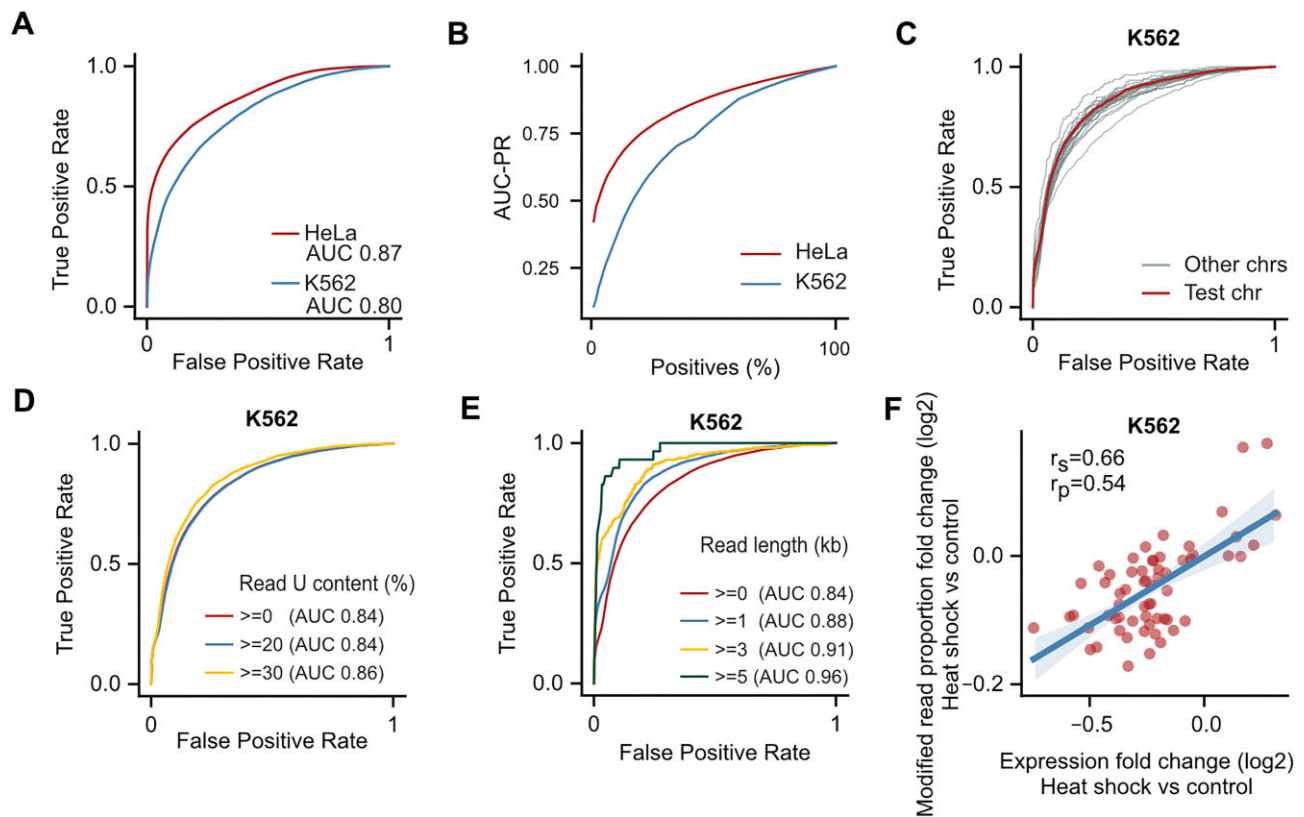


Figure 3. RNAkinet generalizes across cell lines and distinguishes nascent RNA molecules. (A–B) ROC (A) and area under the precision-recall curve for datasets of varying percent of positives (B) plots for reads in chromosome 1 for HeLa and K562 cells. Data for HeLa are as in Figure 2 and only shown here for reference. (C–E) ROC plot on K562 cells data stratified by chromosome (C), percent of uridines in read (D) and read length (E). (F) Scatter plot and correlation coefficients of gene expression and predicted mean-scaled 5EU modification proportion change for K562 cells subjected to heat shock.

in experiments that involve labeling for 24 h. However, such long labeling periods are rarely used under canonical experimental settings. We therefore explored if RNAkinet can be used to predict 5EU-labeled RNAs from shorter labeling periods. We reanalyzed data from (24) in which cells were subjected to heat shock at 42°C for 60 min at the presence of 5EU and performed differential expression analysis using DESeq2 (37) (Supplementary Figure S3C). We reasoned that stress response genes, upregulated upon heat shock, should incorporate more 5EU and thus be recognized as labeled in higher proportions than downregulated genes. As hypothesized, we found that the proportion of modified reads during heatshock was higher for upregulated genes, indicating that newly synthesized mRNAs of stress response genes do indeed incorporate 5EU at a higher rate (Figure 3F). We also found that RNA isoforms with higher read support had higher concordance with differential expression, indicating that higher sequencing depth improves modification prediction (Supplementary Figure S3D). Collectively, our results show that RNAkinet can generalize and can accurately identify 5EU-labeled RNA molecules even from short labeling periods and can be used to capture the dynamics of RNA isoform metabolism across conditions.

RNAkinet predicts RNA isoform kinetics across species

Next, we wished to interrogate the performance of RNAkinet in a different species and to explore whether it can be used to

quantify RNA kinetic parameters such as RNA half-life. To test this, we cultured NIH/3T3 cells, a mouse fibroblast cell line in the presence of 5EU for 2 h in biological duplicates. Following labeling, we harvested total RNA and performed dRNA-Seq. We then used RNAkinet to predict the 5EU labeling state of all individual reads and aggregated this information to quantify the total labeled and unlabeled read counts for each RNA isoform. Subsequently, we quantified the RNA isoform half-lives based on the ratio of modified over unmodified counts (Figure 4A) as previously shown (38) and described in the methods. Comparison of the two replicates showed high reproducibility with correlation reaching 0.78 for Pearson's and 0.83 for Spearman's coefficients (Figure 4B). To test these measurements against experimentally quantified half-lives, we re-analyzed data produced in (32) for the same cell line using a combination of metabolic labeling with 5EU and isolation azide-bearing biotin tags. Our results showed that RNA isoform half-lives quantified by RNAkinet were significantly correlated with measured half-lives for both replicates (Figure 4C, D) indicating high concordance between the methods. To explore whether read support for each RNA isoform affected half-life predictions, we quantified the concordance of the two methods along varying read count thresholds. Indeed, we found that higher read support resulted in higher correlation with measured half-lives indicating that higher sequencing depth further improves performance (Figure 4E). Collectively, our results show that RNAkinet generalizes to RNAs from other species and can quantify RNA half-lives in diverse experimental settings.

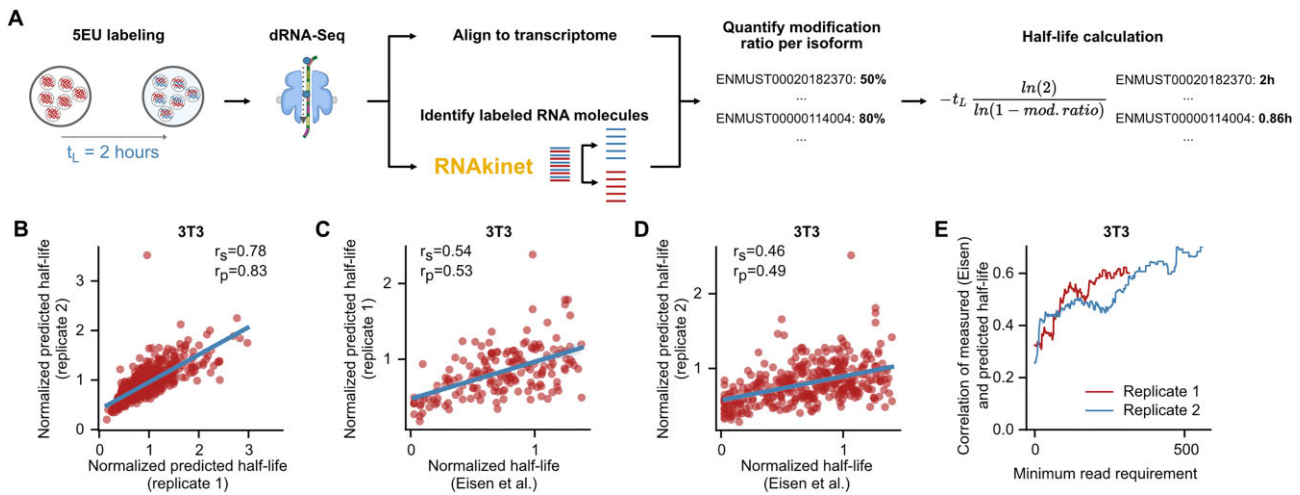


Figure 4. RNAkinet predicts RNA isoform kinetics across species. **(A)** Schematic of RNAkinet pipeline to predict and quantify RNA isoform half-lives. Cells are metabolically labeled with 5EU for 2 h followed by dRNA-Seq, nascent RNA identification with RNAkinet and RNA isoform half-life calculations (see Materials and methods). **(B)** Scatter plot and correlation coefficients of predicted half-lives by RNAkinet for two independent biological replicates of 3T3 cells. **(C, D)** Scatter plot and correlation coefficients of RNA isoform half-lives quantified in mouse 3T3 cells by (32) and those predicted by RNAkinet for two biological replicates. **(E)** Pearson's correlation coefficient between RNA isoform half-lives quantified in (32) and those predicted by RNAkinet for increasing levels of required read coverage per RNA isoform.

Discussion

The abundance of RNA within cells is determined by RNA kinetics, a finely controlled balance of synthesis, processing, and degradation rates, which ensures homeostatic maintenance and responsiveness to environmental signals. Therefore, measuring of RNA kinetics is crucial for deciphering cellular regulatory mechanisms in both normal and disease states. In this study, we metabolically labeled newly synthesized RNA with the nucleoside analog 5EU and developed RNAkinet, a machine learning tool to detect nascent, labeled, RNA molecules directly from raw nanopore signals. To our knowledge, RNAkinet is the only tool that identifies modified nucleotides directly from the raw signal showing the feasibility of this approach for modification detection. By being able to call the labeling status of individual RNA molecules, RNAkinet is expected to offer the unique ability to interrogate RNA kinetics concurrently with a multidimensional ensemble of RNA regulatory elements comprising of isoform usage, poly(A) tail length, alternative poly(A) site selection, and nucleotide modifications (12).

A prevalent caveat of machine learning is the potential limitation to generalize prediction performance to different systems. To address this, we adopted a thorough data preparation protocol aimed at achieving broad domain representation. Furthermore, we subjected RNAkinet to rigorous validation across a variety of distinct and completely independent experimental conditions, confirming its capability to generalize across various RNA sequencing kits, cell types and organisms. This also highlights the robustness of the training process which coupled with lack of external dependencies introduces a path to future developments and updates as nanopore sequencing technology advances. We also placed particular emphasis on model architecture to avoid overparameterization. By incorporating convolution and recurrent layers, RNAkinet captures long-range interactions inherent in the electrical signal while it avoids overfitting to training data and achieves high computational efficiency being able to process ~600 000 reads/h (Supplementary Figure S4A).

Finally, we show that RNAkinet can quantify isoform-level RNA degradation rates that significantly correlate with measurements from methods that lack isoform-level resolution. This feature is expected to enable the study of RNA kinetics at unprecedented resolution and to allow future associations with post-transcriptional regulatory cues at single molecule resolution and a novel set of study designs that will involve a holistic interrogation of RNA and gene regulation in cells.

Data availability

Sequencing data have been deposited in the Sequence Read Archive (SRA) under Bioproject accession: PRJNA1030003. Raw FAST5 files will be provided by the authors upon request. The code for the analysis and a complete snakemake pipeline for training the model has been deposited on Zenodo DOI: 10.5281/zenodo.10070389. The executable code and final model for RNAkinet has been deposited on GitHub (<https://github.com/maragkakislabs/rnakinet>).

Supplementary data

Supplementary Data are available at NARGAB Online.

Acknowledgements

We would like to thank Dr. Myong-Hee Sung for her gift of the 3T3 cells and Dr. Michael E. Ward for providing iPSC-derived i3Neurons. This work utilized the computational resources of the NIH HPC Biowulf cluster (<http://hpc.nih.gov>). Computational resources were also provided by the e-INFRA CZ project (ID:90254), supported by the Ministry of Education, Youth and Sports of the Czech Republic.

Author contributions: V.M. and J.M. developed RNAkinet and performed computational analysis. C.B. performed wet-lab experiments with assistance from M.J.P. and S.M. V.M. and M.M. interpreted the data assisted by J.M. and P.A.

V.M. and M.M. wrote the manuscript with feedback from all authors.

Funding

HORIZON-WIDERA-2022 [BioGeMT (ID: 101086768) to P.A.]; Intramural Research Program of the National Institute on Aging, National Institutes of Health [ZIA AG000696 and ZIA AG000446 to M.M.].

Conflict of interest statement

None declared.

References

- Tani,H., Mizutani,R., Salam,K.A., Tano,K., Ijiri,K., Wakamatsu,A., Isogai,T., Suzuki,Y. and Akimitsu,N. (2012) Genome-wide determination of RNA stability reveals hundreds of short-lived noncoding transcripts in mammals. *Genome Res.*, **22**, 947–956.
- Churchman,L.S. and Weissman,J.S. (2011) Nascent transcript sequencing visualizes transcription at nucleotide resolution. *Nature*, **469**, 368–373.
- Rabani,M., Levin,J.Z., Fan,L., Adiconis,X., Raychowdhury,R., Garber,M., Gnirke,A., Nusbaum,C., Hacohen,N., Friedman,N., et al. (2011) Metabolic labeling of RNA uncovers principles of RNA production and degradation dynamics in mammalian cells. *Nat. Biotechnol.*, **29**, 436–442.
- Kwak,H., Fuda,N.J., Core,L.J. and Lis,J.T. (2013) Precise maps of RNA polymerase reveal how promoters direct initiation and pausing. *Science*, **339**, 950–953.
- Duffy,E.E., Rutenberg-Schoenberg,M., Stark,C.D., Kitchen,R.R., Gerstein,M.B. and Simon,M.D. (2015) Tracking distinct RNA populations using efficient and reversible covalent chemistry. *Mol. Cell*, **59**, 858–866.
- Schwalb,B., Michel,M., Zacher,B., Frühauf,K., Demel,C., Tresch,A., Gagneur,J. and Cramer,P. (2016) TT-seq maps the human transient transcriptome. *Science*, **352**, 1225–1228.
- Herzog,V.A., Reichholf,B., Neumann,T., Rescheneder,P., Bhat,P., Burkard,T.R., Wlotzka,W., von Haeseler,A., Zuber,J. and Ameres,S.L. (2017) Thiol-linked alkylation of RNA to assess expression dynamics. *Nat. Methods*, **14**, 1198–1204.
- Schofield,J.A., Duffy,E.E., Kiefer,L., Sullivan,M.C. and Simon,M.D. (2018) TimeLapse-seq: adding a temporal dimension to RNA sequencing through nucleoside recoding. *Nat. Methods*, **15**, 221–225.
- Baptista,M.A.P. and Dölken,L. (2018) RNA dynamics revealed by metabolic RNA labeling and biochemical nucleoside conversions. *Nat. Methods*, **15**, 171–172.
- Boileau,E., Altmüller,J., Naarmann-de Vries,I.S. and Dieterich,C. (2021) A comparison of metabolic labeling and statistical methods to infer genome-wide dynamics of RNA turnover. *Brief. Bioinform.*, **22**, bbab219.
- Marasco,L.E. and Kornblihtt,A.R. (2023) The physiology of alternative splicing. *Nat. Rev. Mol. Cell Biol.*, **24**, 242–254.
- Schoenberg,D.R. and Maquat,L.E. (2012) Regulation of cytoplasmic mRNA decay. *Nat. Rev. Genet.*, **13**, 246–259.
- Garalde,D.R., Snell,E.A., Jachimowicz,D., Sipos,B., Lloyd,J.H., Bruce,M., Pantic,N., Admassu,T., James,P., Warland,A., et al. (2018) Highly parallel direct RNA sequencing on an array of nanopores. *Nat. Methods*, **15**, 201–206.
- Workman,R.E., Tang,A.D., Tang,P.S., Jain,M., Tyson,J.R., Razaghi,R., Zuzarte,P.C., Gilpatrick,T., Payne,A., Quick,J., et al. (2019) Nanopore native RNA sequencing of a human poly(A) transcriptome. *Nat. Methods*, **16**, 1297–1305.
- Price,A.M., Hayer,K.E., McIntyre,A.B.R., Gokhale,N.S., Abebe,J.S., Della Fera,A.N., Mason,C.E., Horner,S.M., Wilson,A.C., Depledge,D.P., et al. (2020) Direct RNA sequencing reveals m6A modifications on adenovirus RNA are necessary for efficient splicing. *Nat. Commun.*, **11**, 6016.
- Leger,A., Amaral,P.P., Pandolfini,L., Capitanich,C., Capraro,F., Miano,V., Migliori,V., Toolan-Kerr,P., Sideri,T., Enright,A.J., et al. (2021) RNA modifications detection by comparative Nanopore direct RNA sequencing. *Nat. Commun.*, **12**, 7198.
- Pratanwanich,P.N., Yao,F., Chen,Y., Koh,C.W.Q., Wan,Y.K., Hendra,C., Poon,P., Goh,Y.T., Yap,P.M.L., Chooi,J.Y., et al. (2021) Identification of differential RNA modifications from nanopore direct RNA sequencing with xPore. *Nat. Biotechnol.*, **39**, 1394–1402.
- Mulroney,L., Birney,E., Leonardi,T. and Nicassio,F. (2023) Using nanopore to identify RNA modifications from direct RNA nanopore sequencing data. *Curr Protoc*, **3**, e683.
- Liu,H., Begik,O., Lucas,M.C., Ramirez,J.M., Mason,C.E., Wiener,D., Schwartz,S., Mattick,J.S., Smith,M.A. and Novoa,E.M. (2019) Accurate detection of m6A RNA modifications in native RNA sequences. *Nat. Commun.*, **10**, 4079.
- Lorenz,D.A., Sathe,S., Einstein,J.M. and Yeo,G.W. (2020) Direct RNA sequencing enables m6A detection in endogenous transcript isoforms at base-specific resolution. *RNA*, **26**, 19–28.
- Begik,O., Lucas,M.C., Prysycz,L.P., Ramirez,J.M., Medina,R., Milenkovic,I., Cruciani,S., Liu,H., Vieira,H.G.S., Sas-Chen,A., et al. (2021) Quantitative profiling of pseudouridylation dynamics in native RNAs with nanopore sequencing. *Nat. Biotechnol.*, **39**, 1278–1291.
- Hendra,C., Pratanwanich,P.N., Wan,Y.K., Goh,W.S.S., Thiery,A. and Göke,J. (2022) Detection of m6A from direct RNA sequencing using a multiple instance learning framework. *Nat. Methods*, **19**, 1590–1598.
- Nguyen,T.A., Heng,J.W.J., Kaewsapsak,P., Kok,E.P.L., Stanojević,D., Liu,H., Cardilla,A., Praditya,A., Yi,Z., Lin,M., et al. (2022) Direct identification of A-to-I editing sites with nanopore native RNA sequencing. *Nat. Methods*, **19**, 833–844.
- Maier,K.C., Gressel,S., Cramer,P. and Schwalb,B. (2020) Native molecule sequencing by nano-ID reveals synthesis and stability of RNA isoforms. *Genome Res.*, **30**, 1332–1344.
- Brown,A.-L., Wilkins,O.G., Keuss,M.J., Hill,S.E., Zanovello,M., Lee,W.C., Bampton,A., Lee,F.C.Y., Masino,L., Qi,Y.A., et al. (2022) TDP-43 loss and ALS-risk SNPs drive mis-splicing and depletion of UNC13A. *Nature*, **603**, 131–137.
- Ibrahim,F., Oppelt,J., Maragkakis,M. and Mourelatos,Z. (2021) TERA-Seq: true end-to-end sequencing of native RNA molecules for transcriptome characterization. *Nucleic Acids Res.*, **49**, e115.
- Li,H. (2018) Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics*, **34**, 3094–3100.
- Howard,J. and Ruder,S. (2018) Universal language model fine-tuning for text classification. In: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Vol. 1, pp. 328–339.
- Paszke,A., Gross,S., Massa,F. and Lerer,A. (2019) Pytorch: an imperative style, high-performance deep learning library. *Adv. Neural Inf. Process. Syst.*, **32**, 8024–8035.
- Mölder,F., Jablonski,K.P., Letcher,B., Hall,M.B., Tomkins-Tinch,C.H., Sochat,V., Forster,J., Lee,S., Twardziok,S.O., Kanitz,A., et al. (2021) Sustainable data analysis with Snakemake. *F1000Res*, **10**, 33.
- Dölken,L., Ruzsics,Z., Rädle,B., Friedel,C.C., Zimmer,R., Mages,J., Hoffmann,R., Dickinson,P., Forster,T., Ghazal,P., et al. (2008) High-resolution gene expression profiling for simultaneous kinetic parameter analysis of RNA synthesis and decay. *RNA*, **14**, 1959–1972.
- Eisen,T.J., Eichhorn,S.W., Subtelny,A.O., Lin,K.S., McGeary,S.E., Gupta,S. and Bartel,D.P. (2020) The dynamics of cytoplasmic mRNA metabolism. *Mol. Cell*, **77**, 786–799.
- Vaswani,A., Shazeer,N., Parmar,N., Uszkoreit,J., Jones,L., Gomez,A.N., Kaiser,Ł. and Polosukhin,I. (2017) Attention is all you need. *Adv. Neural Inf. Process. Syst.*, **30**, 5998–6008.

34. Neumann,D., Reddy,A.S.N. and Ben-Hur,A. (2022) RODAN: A fully convolutional architecture for basecalling nanopore RNA sequencing data. *BMC Bioinf.*, **23**, 142.
35. Pagès-Gallego,M. and de Ridder,J. (2023) Comprehensive benchmark and architectural analysis of deep learning models for nanopore sequencing basecalling. *Genome Biol.*, **24**, 71.
36. LeCun,Y., Bengio,Y. and Hinton,G. (2015) Deep learning. *Nature*, **521**, 436–444.
37. Love,M.I., Huber,W. and Anders,S. (2014) Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.*, **15**, 550.
38. Russo,J., Heck,A.M., Wilusz,J. and Wilusz,C.J. (2017) Metabolic labeling and recovery of nascent RNA to accurately quantify mRNA stability. *Methods*, **120**, 39–48.