DEEP LEARNING AND DIRECT SEQUENCING OF LABELED RNA CAPTURES TRANSCRIPTOME DYNAMICS

Anonymous authors
Paper under double-blind review

ABSTRACT

In eukaryotes, genes produce a variety of distinct RNA isoforms, each with unique regulatory signals and resulting protein products. Assessing the metabolism of RNA isoforms is essential for unraveling gene regulatory mechanisms. However, this is impeded by current methods reliant on short-read sequencing, which are inadequate for differentiating between individual isoforms. Additionally, these methods cannot concurrently analyze RNA isoform metabolism and key regulatory elements of RNA stability, such as poly(A) tail and nucleotide modifications. Here, we metabolically label nascent RNA with 5-ethynyl uridine modification and employ direct RNA nanopore sequencing. We introduce RNAkinet, a deep convolutional and recurrent neural network, tailored to directly process electrical signals from nanopore sequencing for the detection of modified RNA molecules. RNAkinet effectively distinguishes between nascent and pre-existing RNA molecules and is generalizable to various cell types and organisms. By modeling RNA decay rates, RNAkinet allows reproducible identification of the kinetic parameters of individual RNA isoforms and facilitates efficient, integrated studies of RNA isoform metabolism and the regulatory elements that influence it.

1 MAIN

Unraveling the kinetics of RNA metabolism is critical for understanding gene regulation and cellular response to environmental cues. Methods for measuring RNA dynamics in mammalian cells are based on short-read RNA sequencing (RNA-Seq) following exposure of cells to nucleoside analogs that get incorporated into newly synthesized RNAs that are subsequently physically or bioinformatically separated (Baptista & Dölken, 2018; Boileau et al., 2021). However, established methods cannot confidently assign alternatively spliced isoforms, or collect single-molecule level information on transcription start site, poly(A) site usage and length, and post-transcriptional modifications, all of which affect RNA stability (Schoenberg & Maquat, 2012). Nanopore sequencing can measure these features through direct detection of RNA molecules (Garalde et al., 2018; Ibrahim et al., 2021; Hendra et al., 2022; Thi et al., 2022; Mulroney et al., 2023). A recent work (nano-ID) has shown the feasibility of combining nanopore sequencing and machine learning to identify metabolically labeled RNAs in a single nanopore direct RNA-Seq (dRNA-Seq) experiment following metabolic labeling (Maier et al., 2020). However, our results show that nano-ID does not generalize to other experiments to be widely applicable.

We tested nano-ID by culturing HeLa cells with 5-Ethynyl-uridine (5EU) for 24 hours (h) to maximize 5EU-labeled RNA and sequenced directly on a MinION device as previously described by Maier et al. (2020) (Table 1). Nano-ID has 2,590,493 parameters trained on fewer than 700,000 examples (Table 2, bottom). We found that Nano-ID did not generalize (Fig. 2A,B, Fig. 5A) and additionally, had low efficiency requiring 7 days of runtime and 800 Gb of memory per sample.

To address this gap, we aimed to develop a prediction tool, RNAkinet, that would a) accurately distinguish 5EU-labeled RNA, b) generalize to other datasets, c) minimize time/space requirements, d) depend solely on the sequencer raw electrical signal avoiding dependencies on basecalling or alignment, and e) quantify RNA kinetics. We used convolutional and recurrent layers to address the shortcomings of fully connected NNs, to reduce the parameter space and integrate long- and short-range interactions between electrical signals (Fig. 1A).

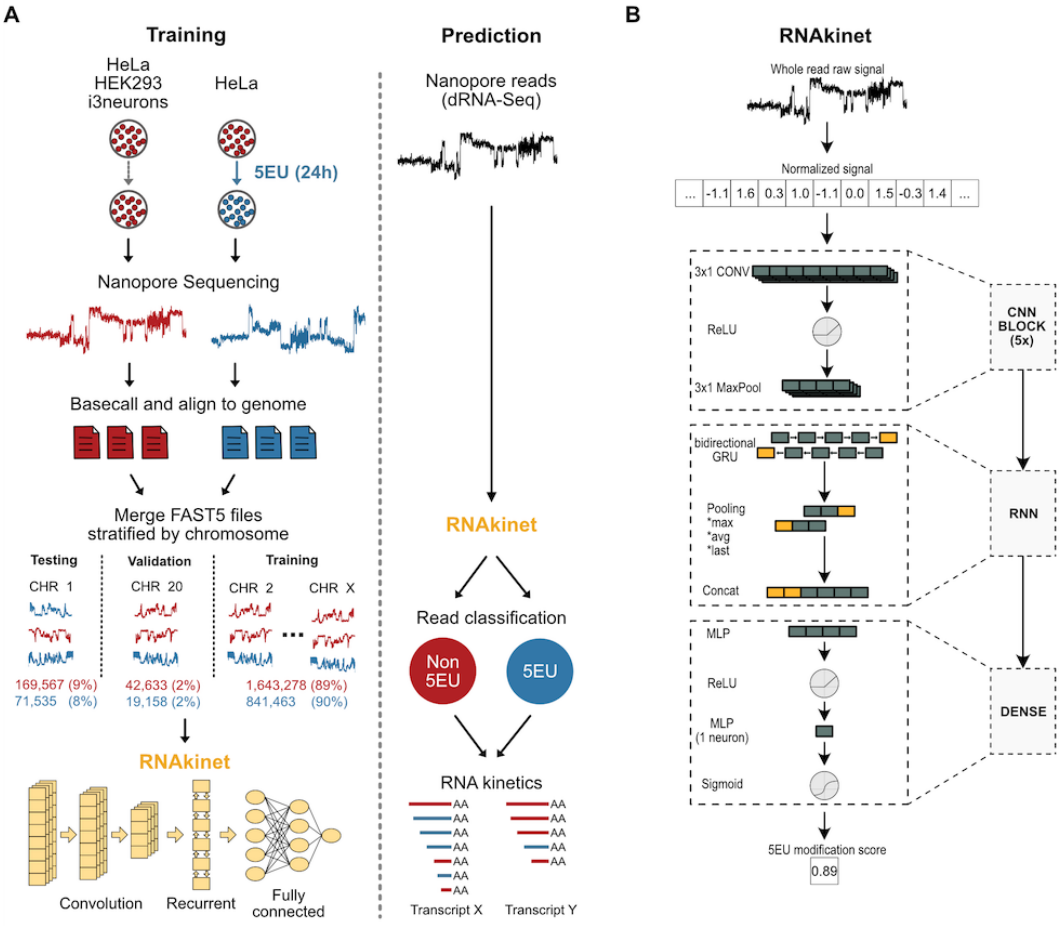

Figure 1: Data preparation and neural network design. A) Schematic of training and prediction workflow. Reads from chromosomes 1 and 20 were retained for testing and validation only. During prediction RNAkinet uses only the raw signal. B) Schematic representation of the architecture of RNAkinet.

We maintained dataset independence by allocating reads from chromosomes 1 and 20 exclusively for testing and validation, thus eliminating confounding sequence biases during performance evaluation (Fig. 1A). To enhance the domain representation, we also included negative samples from diverse dRNA-Seq experiments conducted by different experimenters on distinct cell lines (HeLa, HEK293T, iPSC-Neurons) (Table 1, 2).

NNs typically use padding to handle varying input lengths like electrical signals. However, this is infeasible in dRNA-Seq due to extremes in sequence size and possible alterations of the signal's true continuity (Fig. 5B). Segmentation in fixed length windows is also unsuitable because 5EU incorporation efficiency is low (2-3%) (Maier et al., 2020) and windows without labeled nucleotides would be common. To address this, we designed the NN to accept unpadded sequences of any length and shares learned features across entire sequences. Finally, we limited the input to only raw nanopore signal, thus maintaining independence of basecaller or alignment.

Analysis showed that RNAkinet robustly distinguished labeled RNAs (Fig. 6A). When testing on non-trained reads in chromosome 1 the model achieved a high area under curve (AUC) (Fig. 2A-C, Fig. 6B) indicating that a network involving convolution and recurrent characteristics with only 66,000 parameters can accurately capture 5EU signals in nanopore reads.

RNAkinet performed similarly across reads from all chromosomes indicating that it did not overfit (Fig. 2D, Fig. 6C-E). Furthermore, the content of uridines (where 5EU would be incorporated) in a

read had a negligible impact on performance (Fig. 2E). Shorter reads (0-1 kb) reached lower AUC than longer ones (Fig. 2F), indicating that RNAkinet's performance improves with longer reads.

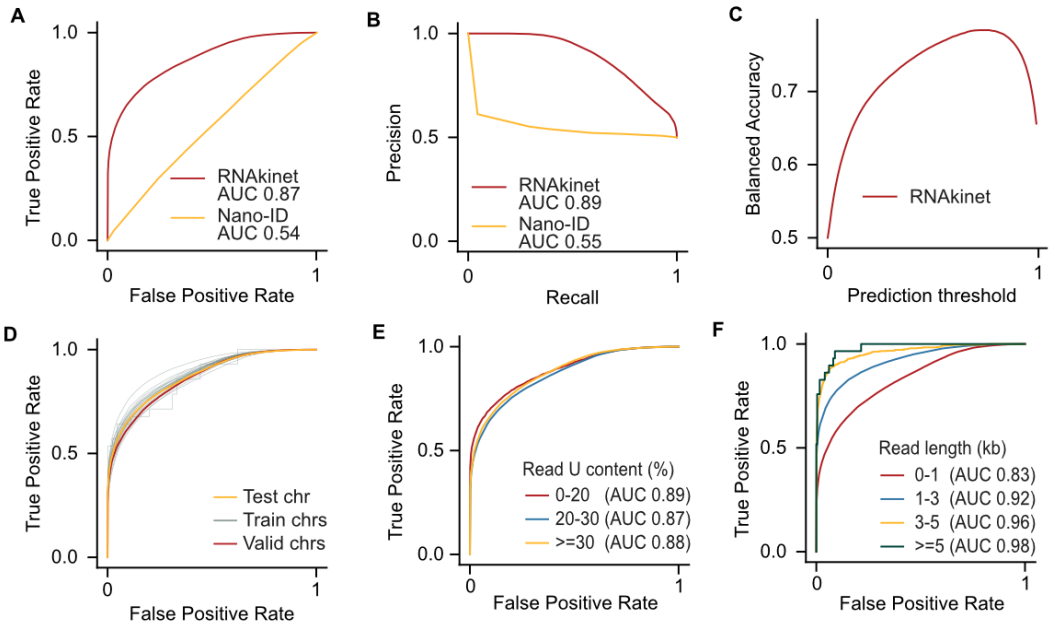

Figure 2: RNAkinet accurately classifies RNA molecules labeled with 5EU. A-B) ROC (A) and PR (B) on test data for RNAkinet and Nano-ID C) Balanced accuracy on test data for RNAkinet. D-F) ROC plot stratified by train, validation, and test chromosomes (D), percent of uridines in transcript (E) and read length (F).

Next, we tested if the model can generalize to independent datasets using data generated in K562 cells (Maier et al., 2020), another human cell line. We again isolated reads that aligned to chromosome 1 to avoid transcripts present in the training data. RNAkinet successfully distinguished labeled RNA molecules along all replicates of K562 cells (Fig. 7A) and had comparable performance to HeLa cells (Fig. 3A, B, Fig. 7B). A minor drop in performance was expected given the use of an earlier iteration of the sequencing kit (SQK-RNA001) for K562, not modeled during training (Maier et al., 2020). Again, stratification of reads by chromosome, U content or length resulted in minimal performance differences (Fig. 3C-E).

As labeling for 24 h is rarely used in canonical experimental settings, we explored the performance of RNAkinet on shorter labeling periods, where cells were subjected to heat shock for 60 min in the presence of 5EU (Maier et al., 2020). We reasoned that stress response genes, upregulated upon heat shock (Fig. 7C), should incorporate more 5EU. Indeed, the relative modification increase predicted by RNAkinet significantly correlated with expression change (Fig. 3F), especially for transcripts with high read support. Collectively, our results show that RNAkinet generalizes, identifies 5EU-labeled RNA molecules from short labeling periods and captures the dynamics of RNA metabolism across conditions.

We next interrogated the performance of RNAkinet in a different species. We cultured NIH/3T3 cells, a mouse fibroblast cell line, in the presence of 5EU for 2 h. Transcript half-lives were quantified and normalized based on the ratio of modified reads (Fig. 4A) (Russo et al., 2017). Here we assume that all reads at time 0 are non-modified. Reads that are predicted modified after two hours are considered a replacement for reads that decayed during this time. RNAkinet-quantified half-lives were reproducible (Fig. 4B) and significantly correlated with measured half-lives obtained from azide-bearing biotin tag isolation (Eisen et al., 2020) (Fig. 4C, D). Higher read support resulted in higher correlation indicating that deeper sequencing could further improve performance (Fig. 4E). We obtained similar results, arguably with lower correlation, for HeLa cells compared with measured half-lives from Tani et al. (2012) (Fig. 6D). Our results show that RNAkinet generalizes and can quantify RNA kinetics in diverse experimental settings.

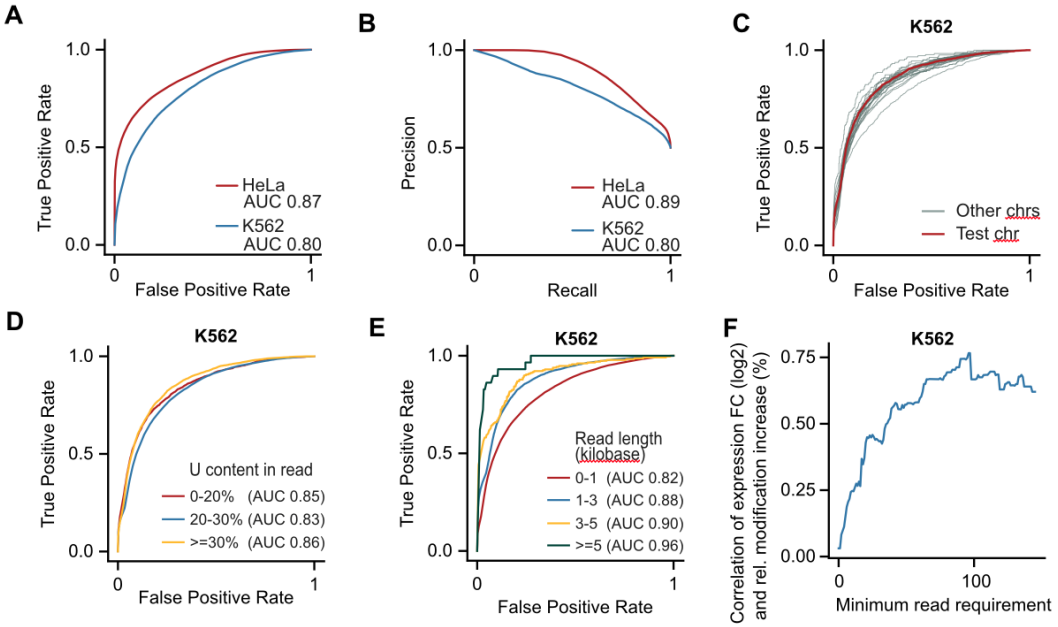

Figure 3: RNAkinet generalizes across cell lines and distinguishes nascent RNA molecules. A-B) ROC (A) and PR (B) plot for reads in chromosome 1 for HeLa and K562 cells. Data for HeLa are as in Fig. 2 and only shown here for comparison. C-E) ROC plot on K562 cells data stratified by chromosome (C), percent of uridines in transcript (D) and read length (E). F) Pearson's correlation coefficients of gene expression change and predicted 5EU modification rate for K562 cells subjected to heat shock for increasing levels of required read coverage per transcript.

Quantification of RNA dynamics is instrumental to understand cellular regulation. In this study, we developed RNAkinet, a tool to detect 5EU-labeled RNA molecules using only raw nanopore signals without dependencies on other software. Incorporation of convolutional layers dramatically reduced the parameter space and achieved high efficiency (Fig. 8A). Notably, RNAkinet generalizes to distinct experimental settings and quantifies isoform kinetics in alignment with established methods that lack this capability. This enables the association of RNA kinetics with post-transcriptional regulatory cues at single molecule resolution.

2 Online Methods

2.1 Cell culture and metabolic labeling

NIH3T3 cells (ATCC CRL-1658) were cultured at 37°C, 5% CO2, 90% humidity in Dulbecco's Modified Eagle Medium (Thermo Fisher Scientific) 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 Medium (Thermo Fisher Scientific) 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 by Brown et al. (2022). For metabolic labeling, cells were cultured in media containing 400 or 500 μM 5-Ethynyl-uridine (ThermoFisher) for 2 to 24 hours 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 (Ibrahim et al., 2021) 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

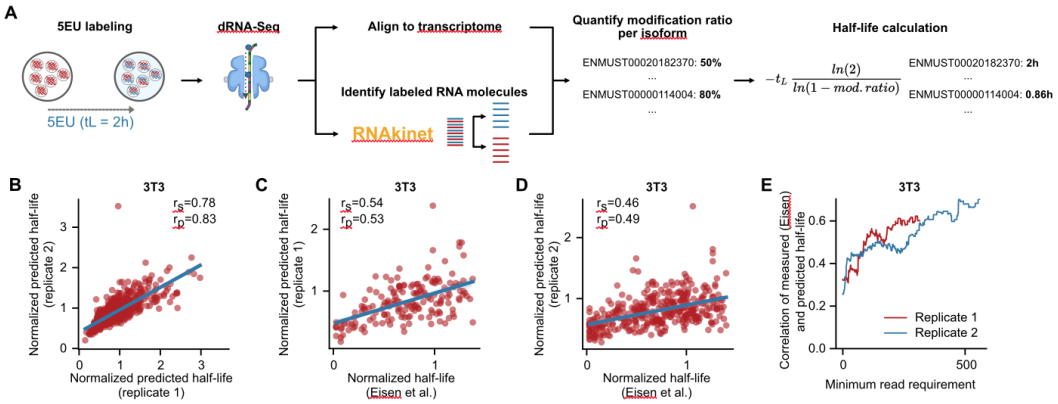

Figure 4: RNAkinet predicts RNA isoform kinetics across species. A) Schematic of RNAkinet pipeline to quantify transcript half-lives. B) Scatter plot and correlation coefficients of predicted half-lives for 2 independent biological replicates of 3T3 cells. C-D) Scatter plot and correlation coefficients of predicted of half-lives quantified in mouse 3T3 cells by Eisen et al. (2020) and RNAkinet for two biological replicates. E) Pearson's correlation coefficient between Eisen et al. (2020) and RNAkinet quantified half-lives for increasing levels of required read coverage per transcript.

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).

2.2 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 a chromosome they originated from. For evaluating precision, recall, and F1 metrics, which are sensitive to imbalanced data, the splits were up-sampled to have a balanced ratio of positives and negatives. The reads were basecalled with Guppy 6.4.8 , aligned to the Ensembl human genome (GRCh38) and transcriptome with Minimap2 (Li, 2018), 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 400000 raw signal values were discarded. The filtered reads were then normalized by median absolute deviation before being passed through the neural network.

2.3 Neural network design and training

To create a detection tool for 5EU modification, a convolutional neural network classifier was designed. It was trained with data from the training split described above and was optimized to perform a binary classification of raw nanopore signals into either modified or unmodified categories. The network was made up of multiple convolutional blocks extracting local patterns from the input signal, followed by a bidirectional recurrent layer with GRU units that aggregated the extracted information across the whole length of the sequence in both directions. These features were then pooled with max pooling, average pooling, and concatenated with hidden states from the last hidden states of the recurrent layers. To allow the network to accept signals of varying lengths without padding as an input, this pooling layer at the end of the network was utilized to aggregate information across the length dimension into a fixed-size vector. This vector was then fed into a small dense feed-forward network to predict the final 5EU modification score.

For training the network on sequences of variable length, a batch size of size 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 AUROC metric with threshold 0 and patience of 50 evaluation steps. The first 1000 learning steps were used as warm up steps where the learning rate was scheduled to linearly increase from 0 to the final learning rate of 0.001. To accelerate the network training, a single A100 Nvidia GPU was utilized. The implementation and training of the network were done using PyTorch (Paszke et al., 2019) and PyTorch Lightning frameworks. Snakemake (Mölder et al., 2021) 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.

2.4 Calculation of relative modification change

To accurately capture the change in gene expression between control and condition experiments using 5EU detection, a metric called relative modification increase was introduced. This metric used the percentage of modified reads of a given transcript from both the condition and control experiments and calculated a relative increase in the condition experiment compared to the control. A positive relative modification increase indicated a higher prevalence of modified reads under the experimental condition, while a negative value indicated a decrease.

$$Rel.mod.increase = \left(\left(\frac{mod\_perc_{\text{condition}}}{mod\_perc_{\text{control}}} - 1 \right) \times 100 \right)$$

where $mod\_perc_{\text{condition}}$ and $mod\_perc_{\text{control}}$ denote the percentage of reads of a given transcript predicted to be modified in the condition and control experiments. These are then divided and 1 is subtracted to represent increase, and result is multiplied by 100 to normalize it to a percentage.

3 Data availability

Sequencing data have been deposited in the Sequence Read Archive (SRA) under Bioproject accession: PRJNA1030003. The code for the analysis included in the manuscript has been deposited on Zenodo DOI: 10.5281/zenodo.10070389. The code for RNAkinet has been deposited on GitHub and will be available after double-blind review.

References

Marisa A P Baptista and Lars Dölken. Rna dynamics revealed by metabolic rna labeling and biochemical nucleoside conversions. *Nature Methods*, 15(3):171–172, Feb 2018. doi: https://doi.org/10.1038/nmeth.4608.

Etienne Boileau, Janine Altmüller, Isabel S Naarmann-de, and Christoph Dieterich. A comparison of metabolic labeling and statistical methods to infer genome-wide dynamics of rna turnover. *Briefings in Bioinformatics*, 22(6), Jul 2021. doi: https://doi.org/10.1093/bib/bbab219.

Anna-Leigh Brown, Oscar G. Wilkins, Matthew J. Keuss, Sarah E. Hill, Matteo Zanovello, Weaverly Colleen Lee, Alexander Bampton, Flora C. Y. Lee, Laura Masino, Yue A. Qi, Sam Bryce-Smith, Ariana Gatt, Martina Hallegger, Delphine Fagegaltier, Hemali Phatnani, Jia Newcombe, Emil K. Gustavsson, Sahba Seddighi, Joel F. Reyes, and Steven L. Coon. Tdp-43 loss and als-risk snps drive mis-splicing and depletion of unc13a. *Nature*, 603(7899):131–137, Mar 2022. doi: https://doi.org/10.1038/s41586-022-04436-3. URL `https://www.nature.com/articles/s41586-022-04436-3`.

Timothy Eisen, Stephen J Eichhorn, Alexander O Subtelny, Kathy J Lin, Sean E McGeary, Sumeet Gupta, and David P Bartel. The dynamics of cytoplasmic mrna metabolism. *Molecular Cell*, 77 (4):786–799.e10, Feb 2020. doi: https://doi.org/10.1016/j.molcel.2019.12.005.

Daniel R Garalde, Elizabeth A Snell, Daniel Jachimowicz, Botond Sipos, Joseph H Lloyd, Mark Bruce, Nadia Pantic, Tigist Admassu, Phillip James, Anthony Warland, Michael Jordan, Jonah Ciccone, Sabrina Serra, Jemma Keenan, Samuel Martin, Luke McNeill, E Jayne Wallace, Lakmal Jayasinghe, Chris Wright, and Javier Blasco. Highly parallel direct rna sequencing on an array of nanopores. *Nature Methods*, 15(3):201–206, Jan 2018. doi: https://doi.org/10.1038/nmeth.4577.

Christopher Hendra, Ploy N. Pratanwanich, Yuk Kei Wan, W. S. Sho Goh, Alexandre Thiery, and Jonathan Göke. Detection of m6a from direct rna sequencing using a multiple instance learning framework. *Nature Methods*, 19(12):1590–1598, Nov 2022. doi: https://doi.org/10.1038/s41592-022-01666-1.

Fadia Ibrahim, Jan Oppelt, Manolis Maragkakis, and Zissimos Mourelatos. Tera-seq: true end-to-end sequencing of native rna molecules for transcriptome characterization. *Nucleic Acids Research*, 49(20):e115–e115, Aug 2021. doi: https://doi.org/10.1093/nar/gkab713.

Heng Li. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics*, 34(18): 3094–3100, May 2018. doi: https://doi.org/10.1093/bioinformatics/bty191.

Kerstin C Maier, Saskia Gressel, Patrick Cramer, and Björn Schwalb. Native molecule sequencing by nano-id reveals synthesis and stability of rna isoforms. *Genome Res.*, 30(9):1332–1344, Sep 2020. doi: https://doi.org/10.1101/gr.257857.119.

Logan Mulroney, Ewan Birney, Tommaso Leonardi, and Francesco Nicassio. Using nanocompore to identify rna modifications from direct rna nanopore sequencing data. *Current Protocols*, 3(2), Feb 2023. doi: https://doi.org/10.1002/cpz1.683.

Felix Mölder, Kim Philipp Jablonski, Brice Letcher, Michael B. Hall, Christopher H. Tomkins-Tinch, Vanessa Sochat, Jan Forster, Soohyun Lee, Sven O. Twardziok, Alexander Kanitz, Andreas Wilm, Manuel Holtgrewe, Sven Rahmann, Sven Nahnsen, and Johannes Köster. Sustainable data analysis with snakemake. *F1000Research*, 10:33, Apr 2021. doi: https://doi.org/10.12688/f1000research.29032.2. URL `https://www.ncbi.nlm.nih.gov/pmc/articles/pmc8114187/`.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Emily Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lü Fang, and Junjie Bai. Pytorch: An imperative style, high-performance deep learning library. *arXiv (Cornell University)*, 32:8026–8037, Jan 2019.

Joseph Russo, Adam M Heck, Jeffrey Wilusz, and Carol J Wilusz. Metabolic labeling and recovery of nascent rna to accurately quantify mrna stability. *Methods*, 120:39–48, May 2017. doi: https://doi.org/10.1016/j.ymeth.2017.02.003.

Daniel R. Schoenberg and Lynne E. Maquat. Regulation of cytoplasmic mrna decay. *Nature Reviews Genetics*, 13(4):246–259, Mar 2012. doi: https://doi.org/10.1038/nrg3160.

Hidenori Tani, Rena Mizutani, Kazi Abdus Salam, Keiko Tano, Kenichi Ijiri, Ai Wakamatsu, Takao Isogai, Yutaka Suzuki, and Nobuyoshi Akimitsu. Genome-wide determination of rna stability reveals hundreds of short-lived noncoding transcripts in mammals. *Genome Research*, 22(5): 947–956, May 2012. doi: https://doi.org/10.1101/gr.130559.111. URL `https://pubmed.ncbi.nlm.nih.gov/22369889/`.

Anh Thi, Wei Joel, Pornchai Kaewsapsak, Louis Kok, Dominik Stanojevic, Hao Liu, Angelysia Cardilla, Albert Praditya, Zirong Yi, Mingwan Lin, Jong Ghut, Yin Ying Ho, Kai Lay, Yuanming Wang, Qixing Zhong, Jacki E Heraud-Farlow, Shuang Xue, Bruno Reversade, Carl R Walkley, and Ying Swan Ho. Direct identification of a-to-i editing sites with nanopore native rna sequencing. *Nucleic Acids Res.*, 19(7):833–844, Jun 2022. doi: https://doi.org/10.1038/s41592-022-01513-3.

A APPENDIX

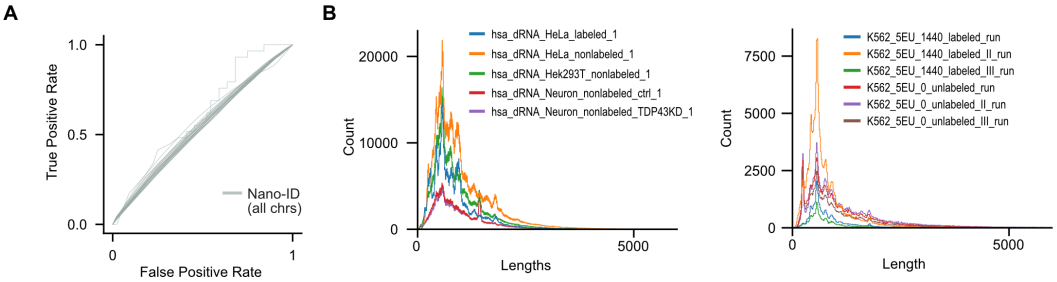

Figure 5: A) ROC plot of Nano-ID (Maier et al., 2020) on HeLa cells labeled with 5EU for 24 h. Reads are stratified by chromosome. B) Read length distribution for libraries used in training and testing.

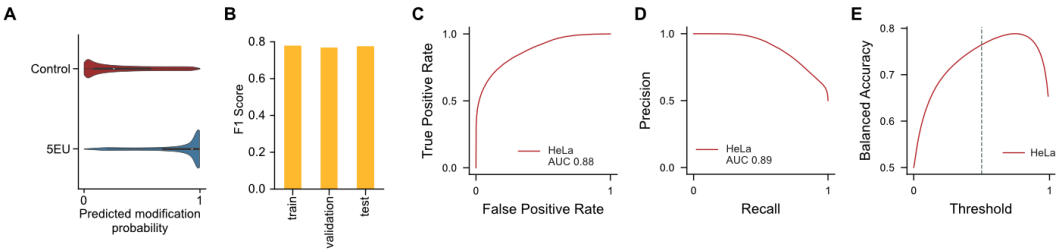

Figure 6: A) Predicted probability distribution for positive and negative samples used in training for RNAkinet B) Bar plot of F1 score on train, test, and validation data for RNAkinet. C-E) ROC (C), PR (D) and BA (E) plot for training data for RNAkinet.

Table 1: Description of dRNA-Seq datasets used in this work.

Dataset	**Use case in this work**	**Organism**	**Generated in**	**Cell line**
In-house HeLa	Positive and negative reads for training, validation, and testing	Homo Sapiens	This work	HeLa
In-house Neurons	Negative reads for training	Homo Sapiens	This work	iPSC-derived neurons
In-house HEK293T	Negative reads for training	Homo Sapiens	This work	HEK293T
In-house 3T3 5EU	Testing if model captures RNA decay rates in 3T3 cells	Mus Musculus	This work	3T3
In-house HeLa 2h 5EU	Testing if model captures mRNA decay rates in HeLa cells	Homo Sapiens	This work	HeLa
Maier K562 classification	Positive (24 hr labeled) reads for testing classification	Homo Sapiens	Maier et al. (2020)	K562
Maier K562 classification	Negative (non labeled) reads for testing classification	Homo Sapiens	Maier et al. (2020)	K562
Maier K562 heat shock	Test if RNAkinet captures differential expression during stress	Homo Sapiens	Maier et al. (2020)	K562

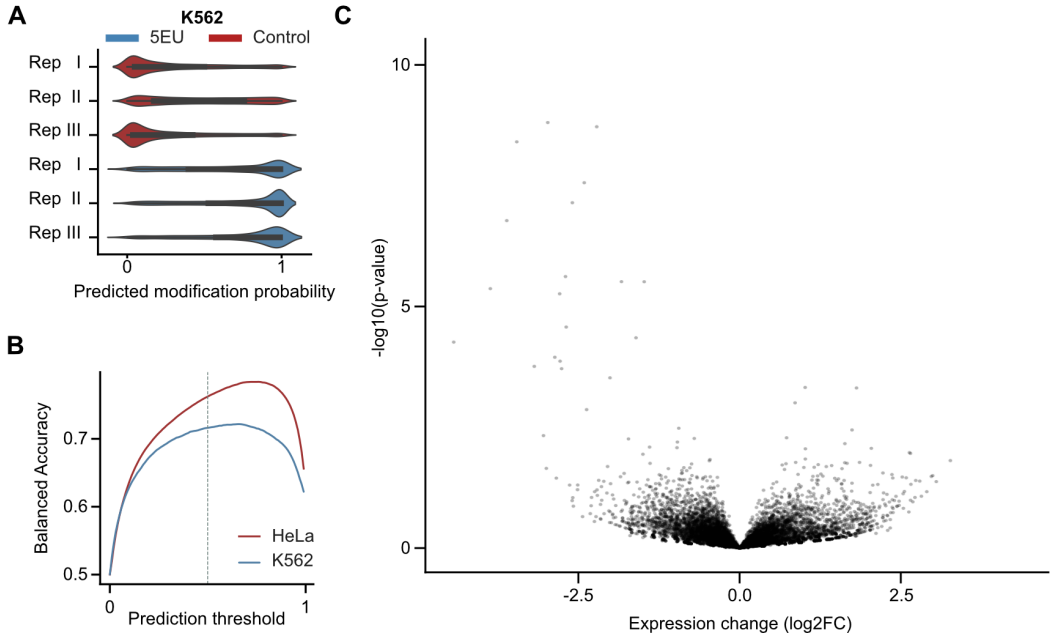

Figure 7: A) Predicted probability distribution on reads from chromosome 1 for K562 cells B) BA on reads from chromosome 1 of HeLa and K562 cells. Data for HeLa cells are the same as Fig. 2 and are only included here for comparison. The threshold used for inference is marked with a gray line. C) Volcano plot of isoform differential expression for heat shock against control cells.

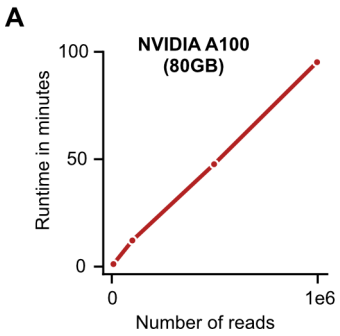

Figure 8: A) Scatter plot of RNAkinet runtime for inference and number of reads processed.

Table 2: Raw read counts per dataset

Experiment name	**Number of reads**
hsa_dRNA_HeLa_labeled_1	982840
hsa_dRNA_HeLa_nonlabeled_1	1878222
mmu_dRNA_3T3_labeled_1	620796
mmu_dRNA_3T3_labeled_2	1306329
hsa_dRNA_Hek293T_nonlabeled_1	1257185
hsa_dRNA_Neuron_nonlabeled_ctr1_1	517551
hsa_dRNA_Neuron_nonlabeled_TDP43KD_1	487902
hsa_dRNA_HeLa_5EU_2hr_1	1191933
hsa_dRNA_HeLa_5EU_2hr_2	1174797
hsa_dRNA_HeLa_5EU_2hr_3	1364193
20180514_1054_K562_5EU_1440_labeled_run	45461
20180514_1541_K562_5EU_1440_labeled_II_run	198289
20180516_1108_K562_5EU_1440_labeled_III_run	24821
20180327_1102_K562_5EU_0_unlabeled_run	144305
20180403_1102_K562_5EU_0_unlabeled_II_run	161941
20180403_1208_K562_5EU_0_unlabeled_III_run	109778