Ninetails: a comprehensive tool for investigation of non-adenosine residues within poly(A) tails from direct RNA sequencing data

N. Gumińska[1], K. Matylla-Kulińska[2], W. Orzeł[1], P. Krawczyk[1], S. Mroczek[1,2], A. Dziembowski[1]

1. International Institute of Molecular and Cell Biology in Warsaw, Poland
2. Institute of Genetics and Biotechnology, Faculty of Biology, University of Warsaw, Poland

Highlights

- Oxford Nanopore direct RNA sequencing (DRS) facilitated by the **Ninetails** package is the only method for the analysis of the composition of poly(A) tails, unaffected by reverse transcription and/or PCR amplification
- **Ninetails** leverages a neural network trained for the detection of non-A residues within the poly(A) tails in DRS data
- **Ninetails** can be used for visual inspection of DRS read signals

Nanopore DRS as a method of investigating nucleotide composition of poly(A) tails

Nowadays it became clear that mRNA poly(A) tail composition is more diverse than previously anticipated. A variety of enzymes (TENTs) can post-transcriptionally decorate poly(A) tails with non-A nucleotides. Several methods are available for detection and profiling of non-A residues in poly(A) tails (**Tab. 1**).

Tab. 1. Comparison of available non-A residues assessment methods.

method	Illumina (TAIL-seq)	PacBio (FLAM-seq, PAIso-seq)	Nanopore (Ninetails)
assessed molecule	cDNA	cDNA	RNA
entire tail analysis	NO (terminal 10-30nt)	YES	YES
tail length limitation	<250 nt	NO	NO
immune to amplification bias	NO	NO	YES

However, the results obtained via the sequencing of a synthetic copy rather than the molecule of interest (PacBio, Illumina) are inconsistent in terms of the prevalence of C or G, nor are they immune to the amplification bias. Whereas, the ground truth of the nucleotide content of nascent poly(A) tails can be asserted by Oxford Nanopore direct RNA sequencing.

Nanopore sequencing records the signal for the full-length transcript

The MinION device detects changes in the ionic current as the RNA molecule threads through the pore in 3' to 5' direction. Thus, an entire RNA molecule can be captured, including the complete poly(A) tail, regardless of its length (**Fig. 1**).

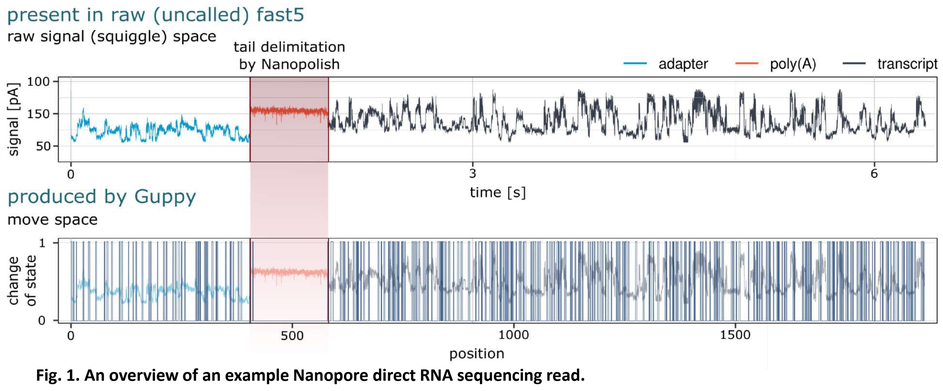

Fig. 1. An overview of an example Nanopore direct RNA sequencing read.

Non-A residues can be recognized within the poly(A) tail

The homopolymer stretches cannot be accurately basecalled due to the uniformity of raw signal between adjacent nucleotides. Nevertheless, the presence of non-A residues in the poly(A) region changes the ionic current significantly enough to be spotted even by naked eye (**Fig. 2**).

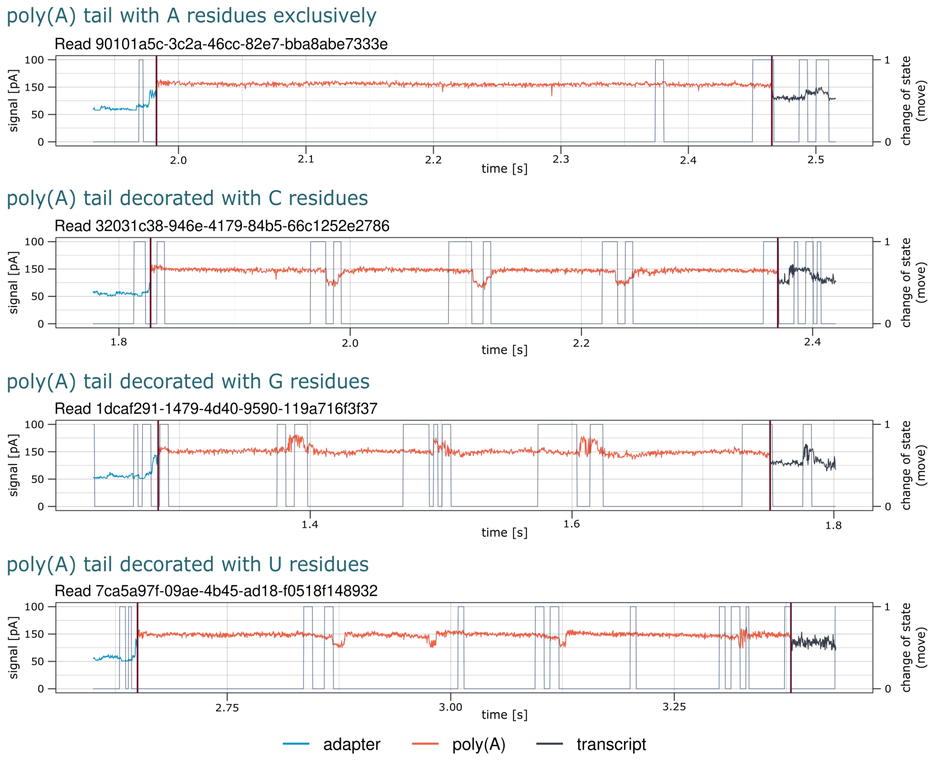

Fig. 2. Examples of Nanopore signals corresponding to various nucleotide compositions of poly(A) tails.

Computer vision can be leveraged to recognize patterns in Nanopore data

Gramian angular field (GAF) transformation is one of the most popular time series imaging algorithms. It was proven to be effective in classification of various types of biological data (e.g. EMG and EEG signals). In GAF, the time series (raw signal) is represented in a polar coordinate system instead of the Cartesian coordinates (**Fig. 3**).

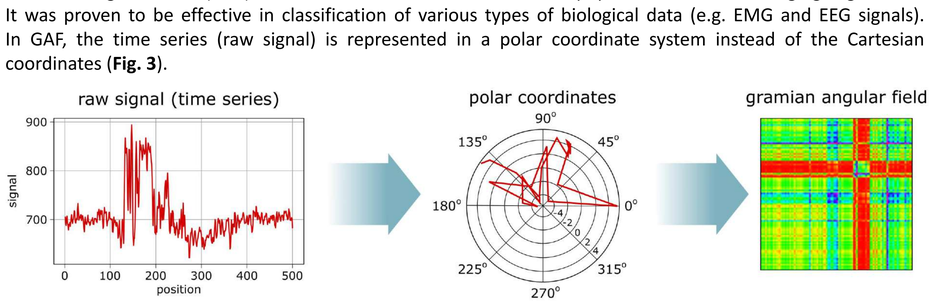

Fig. 3. An overview of the proposed encoding map of the gramian angular field.

Ninetails detects non-adenosine residues in poly(A) tails with high precision & accuracy

First, the raw signal and move table corresponding to the tail region delimited by Nanopolish are extracted. Then the signal is screened with the z-score algorithm, which allows to select fragments displaying two crucial features: (I) significant signal deviation and (II) state transition between the k-mers (move) recorded by Guppy basecaller. Preselected signal chunks are then transformed into 2-channel gramian angular fields (GASF + GADF) and classified by the VGG-based 2D convolutional neural network (**Fig. 4**).

To train and test our model we constructed a set of *in vitro* transcribed sequins based on Renilla luciferase. We designed poly(A) tails composed of 60 residues: either blank (exclusively A-containing), or decorated with a single C, G or U in the center of the tail. The final products were sequenced on MinION device using SQK-RNA002 chemistry.

Our model was trained on a dataset of 37 760 signals in total (9 440 signals per class). The data was split into the 80% training, 10% testing, 10% validation sets. Final architecture and hyperparameters were finetuned based on the model performance. This was further assessed with another independent batch of synthetic and biological data (2 948 114 signals from the *S. cerevisiae*). As a result, our model achieved following performance (**Tab. 2**).

Tab. 2. Model statistics by class.

parameter \ class	A	C	G	U
sensitivity	0.9199	0.9236	0.9434	0.8486
specificity	0.9798	0.9555	0.9653	0.9752
pos pred value	0.9173	0.8736	0.9114	0.9300
neg pred value	0.9804	0.9740	0.9783	0.9432
balanced accuracy	0.9498	0.9395	0.9543	0.9119

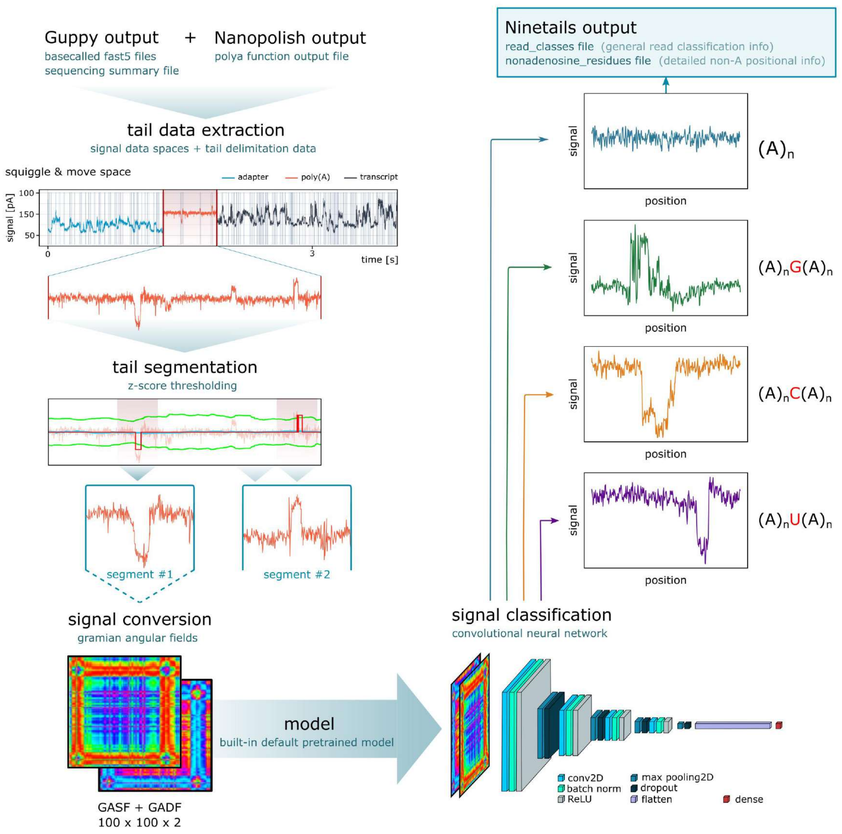

Fig. 4. An overview of the Ninetails pipeline.

Moderna mRNA-1273 vaccine has composite poly(A) tail

Our study revealed that mRNA-1273 has a poly(A) tract of ~100 As often ending with an unexpected perturbation between the poly(A) tail and the adaptor used for library preparation. Such a signal indicates the terminal non-adenosine residues which likely are the remnants of restriction cleavage of the DNA template (**Fig. 6**).

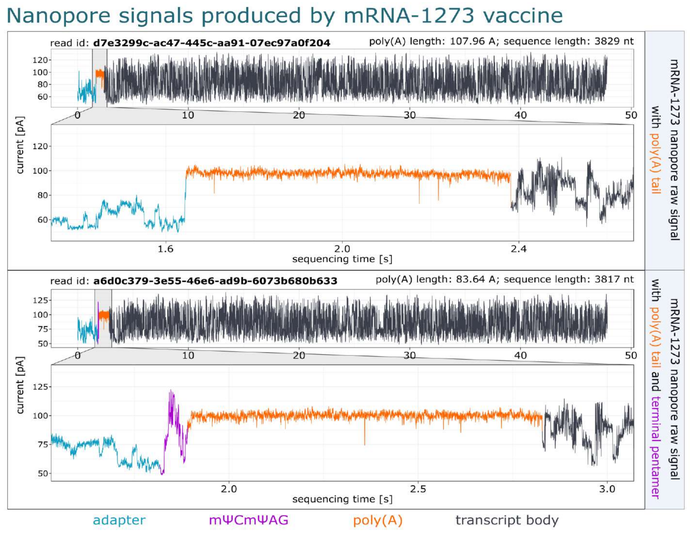

Fig. 6. Representative raw signals from DRS showing mRNA-1273 without and with 3' terminal mΨCmΨAG, top and bottom panels, respectively (Krawczyk et al., 2022).

In murine macrophages, vaccine mRNA can be stable and even undergo poly(A) tail elongation. We noticed that 24 h after transfection, the poly(A) tails of vaccine mRNAs were re-adenylated by ~20 As on average, reaching up to 200 As. Whereas the mean poly(A) length returned to the initial ~100 As 72 h after delivery. This dynamics was observed only in tails lacking the 3' terminal pentamer, (**Fig. 7.**).

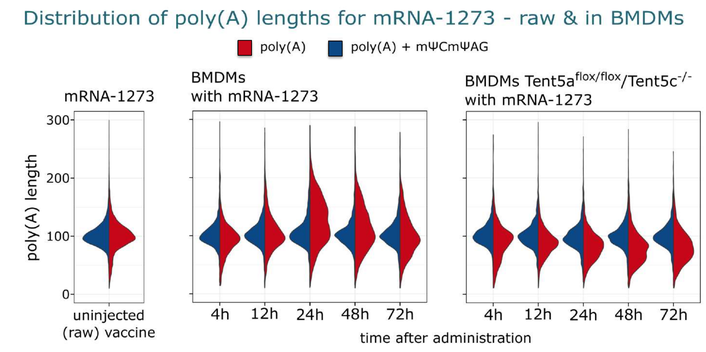

Fig. 7. Distribution of poly(A) lengths for raw mRNA-1273 and up to 72h after delivery to murine BMDMs (Krawczyk et al., 2022)

The investigation of mRNA-1273 with the Ninetails pipeline has revealed higher variety of nucleotide content than just the presence and/or absence of 3' terminal pentamer. Notably, in some reads, signature corresponding to the mΨCmΨAG instead of being adjacent directly to the adaptor, was separated from it by the stretch of As – a hallmark of re-adenylation. Furthermore, in many instances it was completely absent. Instead, the algorithm detected scattered non-adenosine residues (single or multiple per tail) (**Fig. 8.**).

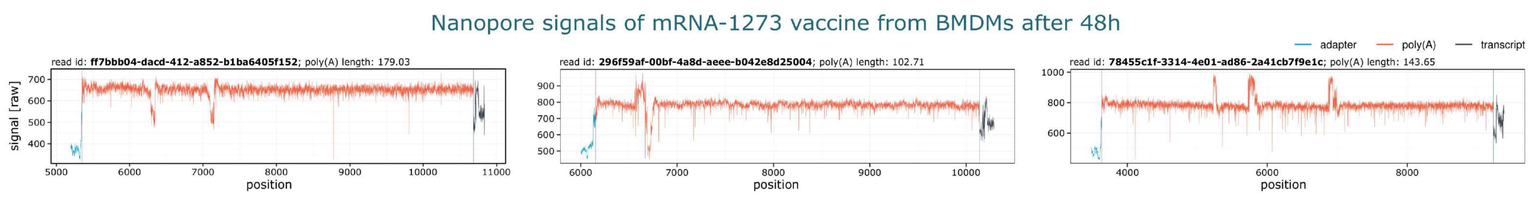

Fig. 8. Examples of raw signals of mRNA-1273 decorated with various non-As. None of them contains mΨCmΨAG directly at the poly(A) tail terminus.

In our previous study, TENT5A/TENT5C non-canonical poly(A) polymerases were indicated as factors responsible for cytoplasmic re-adenylation of mRNA-1273 tails *in vivo*. We observe that in wild-type macrophages (BMDMs), the elongation of mRNA-1273 tails is coupled with the significant enrichment of non-adenosine residues. No such effect is visible in cells devoid of TENT5A/TENT5C (**Fig. 9**). Interestingly, mRNA-1273 exhibits a higher percentage of reads with non-A nucleotides than endogenous mouse transcripts. We speculate that the presence of non-adenosine nucleotides may play a role in stabilizing mRNA-1273 transcripts (thus elongating their half-lives). If this is indeed the case, unlocking this regulatory potential of poly(A) tails could not only lead to better understanding of (synthetic) mRNA turnover, but also contribute to the development of better mRNA therapeutics.

Poly(A) composition of mRNA-1273 vaccine changes dynamically in murine macrophages after delivery

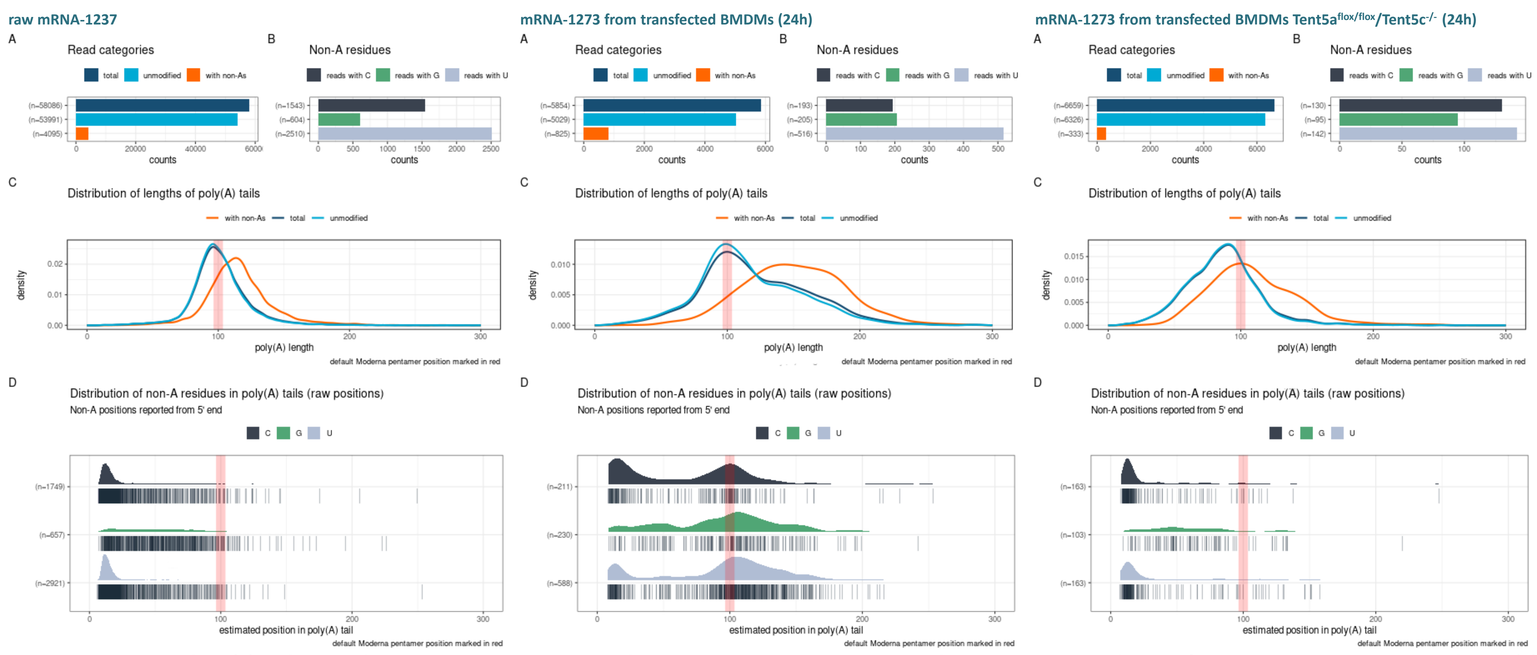

Fig. 9. Detailed survey of poly(A) tail features of raw mRNA-1273 and mRNA-1273 from murine macrophages 24h post transfection, wild & devoid of TENT5A/C, respectively. General read classification (A). Composition of reads with given residue (B). Distribution of poly(A) tail lengths (C). Distribution of non-adenosine residues within the poly(A) tails of mRNA-1273.

Visit Ninetails' repository:

Read our preprint on mRNA-1273:

MOSaIC project received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement no 810425