tailfindr: Alignment-free poly(A) length measurement for Oxford Nanopore RNA and DNA sequencing

Adnan M. Niazi, Maximilian Krause, Kornel Labun, Yamila N. Torres Cleuren, Florian S. Müller, Eivind Valen

Valen Lab | Computational Biology Unit | University of Bergen | Norway

Read the full article RNA

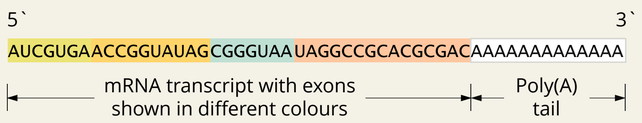
Download this poster

Highlights

- tailfindr is an R package
- It enables transcript isoform-specific poly(A)-tail length estimation
- It works on Oxford Nanopore RNA/DNA sequencing data

Background

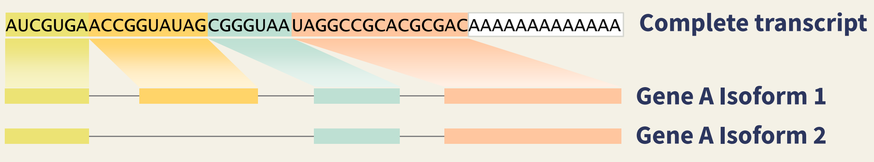
A poly(A) tail is a stretch of adenosines at 3`-end of mRNA transcripts; its length affects nuclear export, stability and translation of mRNA.



Illumina-based poly(A) tail profiling strategies do not allow for transcript isoform-specific poly(A) tail length assignment

Illumina sequencing captures only partial transcript information proximal to the poly(A) tail. A partial transcript may map to multiple isoforms, thereby, making isoform-specific poly(A)-tail length assignment difficult.

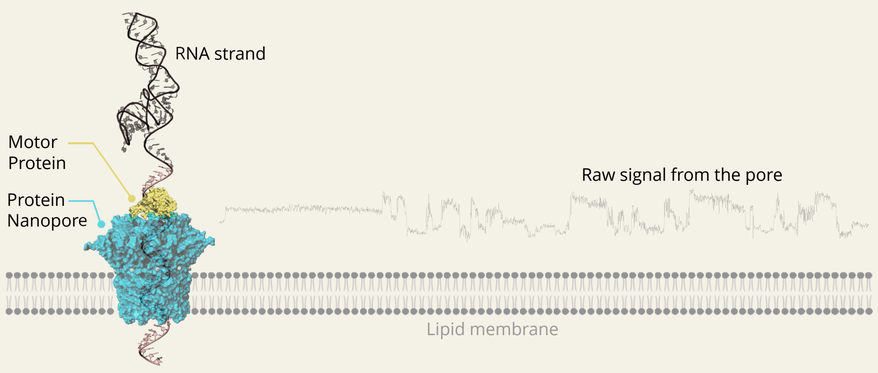
Transcript isoform-specific poly(A) tail length assignment is possible with long read sequencing



A single long read can capture the poly(A) tail and its associated full length transcript, which can then be mapped unambiguously to a single gene isoform. This makes transcript isoform-specific tail-length assignment possible.

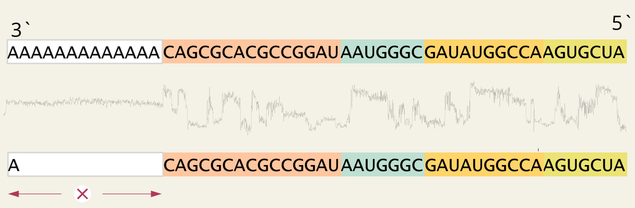
Nanopore sequencing can capture the raw signal for both the poly(A) tail and the associated full-length transcript

Nanopore sequencing detects characteristic changes in the voltage applied across a pore suspended in a membrane as an RNA/DNA molecule translocates through it. Homopolymer regions in the read are sequenced the same way as non-homopolymers regions, and the signal for the entire read can be captured.



Current basecallers cannot accurately basecall homopolymer regions

As there is no transition in signal from one base to another within the poly(A) tail, the basecallers cannot accurately basecall the poly(A) region.



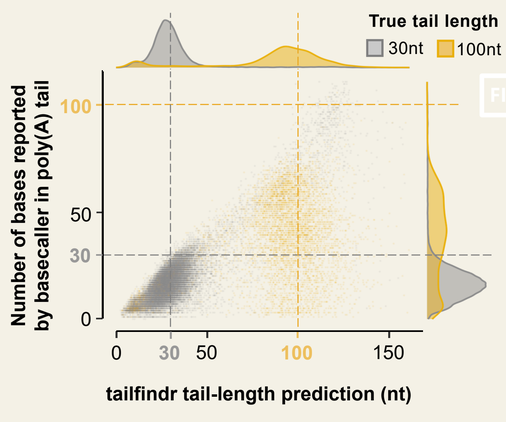


FIG Guppy's flipflop basecaller underestimates the poly(A) tail length especially on longer poly(A) tails, thereby, necessitating the use of tools such as tailfindr to correctly estimate poly(A) tail length.

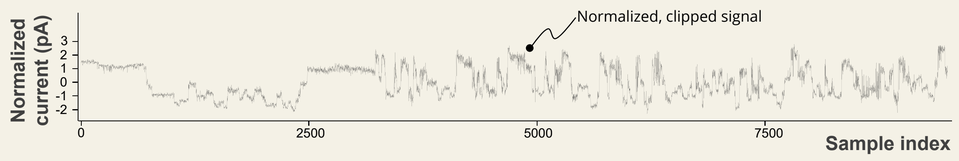
@adnaniazi2 github.com/adnaniazi

Method

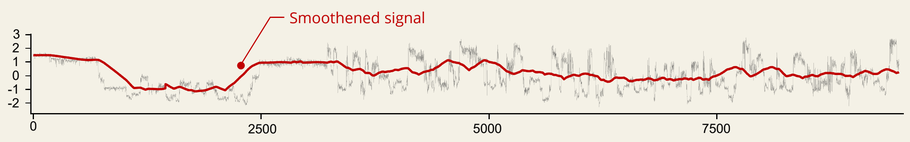
Oxford Nanopore Direct RNA Sequencing intrinsically contains the full poly(A) tail region of the sequenced transcripts. The adapter containing the motor protein (red) is ligated to the poly(A) tail (green) by splint oligo ligation. The poly(A) tail precedes the transcript (yellow) in sequencing direction (3' to 5').

The Algorithm

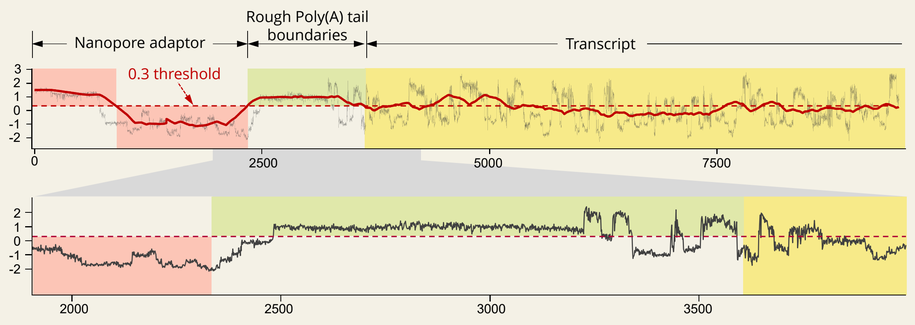
1 - Normalize raw signal
- Clip signal to ± 3 sd.



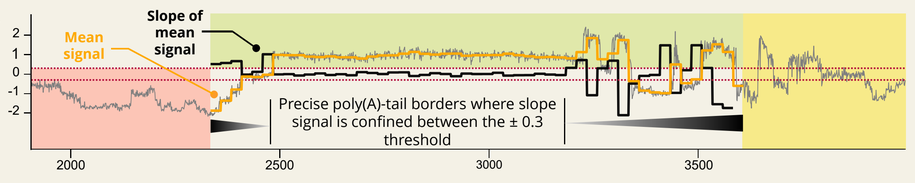
2 - Create a smoothened signal; window size 400; stride 1



3 - Segment smoothened signal using a threshold of 0.3
- Poly(A) tail containing segment follows the Nanopore adaptor segments. This segment defines the rough poly(A)-tail borders that need some refining.



4 - Within rough poly(A)-tail borders, compute:
1. mean signal (——) by averaging every 25 samples
2. slope (——) of the mean signal
- Precise poly(A) tail borders are where the slope signal is confined within the bounds of ± 0.3
- Normalize the tail length by a read-specific normalizer



Results

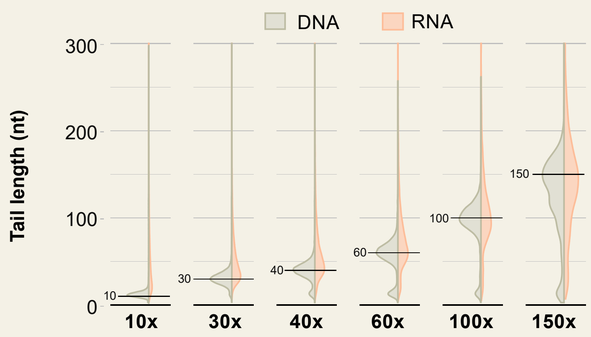


FIG DNA Sequencing approach result in less variable results than Direct RNA sequencing, due to a more robust nucleotide translocation rate

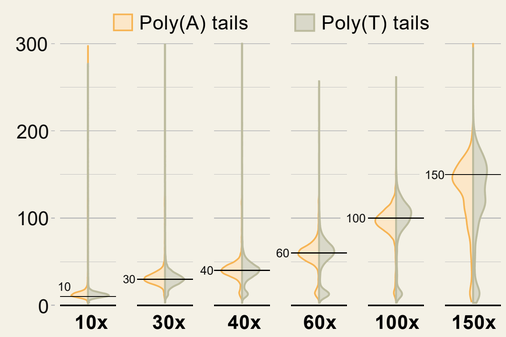
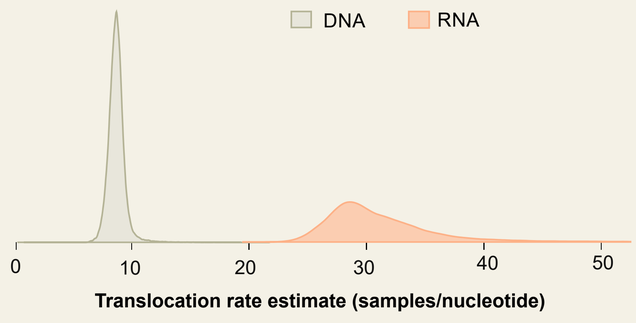


FIG tailfindr operates on ONT DNA sequencing approaches, and estimates both poly(A) and poly(T) tails of GFP PCR amplicons correctly.

Challenges ahead

The spread in the poly(A)-tail measurements is very large due to the spread in the translocation rate, which is stochastic in nature. This means that currently a large number of measurements (around 200-500) per transcript are required to get a robust estimate of that transcripts' poly(A) tail length profile.



Acknowledgment

Adnan Niazi has been supported by Oxford Nanopore Technologies with a travel grant.

adnan.niazi@uib.no www.valenlab.com