*Applied Comparative Genomics*, 2019

|  |
| --- |
| Poly(A)-Tail Length Estimation from long ONT sequencing: Motivation and Tool Comparison  Remy Schwab1,2 and Luchao Qi1,2  1Department of Biomedical Engineering, Johns Hopkins University Whiting School of Engineering  2Team RemChao, Johns Hopkins University  Abstract  **Motivation:** Polyadenylation is a co-transcriptional process in which an adenosine monophosphate molecule is added to the 3’ end of mRNA. These molecules, which consist solely of ade-nines, are known as Poly(A) tails. On mRNAs, the poly(A) tail protects the mRNA molecule from enzymatic degradation in the cytoplasm and aids in transcription termination, export of the mRNA from the nucleus, and translation. Also, poly(A) lengths (the number of bases that make up the molecule) are believed to have some important biological implications. Along with the advent of native RNA sequencing, two tools, nanopolish and tailfindr, have proposed methods for estimating the lengths of Poly(A) tails using the Oxford Nanopore platform. Our project aims to compare these tools.  **Results:** The tool tailfindr can work on single or multi-fast5 file reads and support data that has been basecalled with Albacore or Guppy. It also supports data that has been basecalled using the newer ‘flipflop’ model. However, it requires some metadata table and it leads to incompatibility with tailfindr.  **Contact:** [rschwab6@jhmi.edu](mailto:rschwab6@jhmi.edu), [lqi9@jhu.edu](mailto:lqi9@jhu.edu) |

# Introduction

Polyadenylation is the addition of a poly(A) tail to a [messenger RNA](https://en.wikipedia.org/wiki/Messenger_RNA). The poly(A) tail consists of multiple [adenosine monophosphates](https://en.wikipedia.org/wiki/Adenosine_monophosphate); in other words, it is a stretch of RNA that has only [adenine](https://en.wikipedia.org/wiki/Adenine) bases. In [eukaryotes](https://en.wikipedia.org/wiki/Eukaryote), polyadenylation is part of the process that produces mature [messenger RNA](https://en.wikipedia.org/wiki/Messenger_RNA) (mRNA) for [translation](https://en.wikipedia.org/wiki/Translation_(biology)). It, therefore, plays a critical role in RNA metabolism including stability, enhanced translation, nuclear export and miRNA mediated gene regulation. The process of polyadenylation begins as the [transcription](https://en.wikipedia.org/wiki/Transcription_(genetics)) of a [gene](https://en.wikipedia.org/wiki/Gene) [terminates](https://en.wikipedia.org/wiki/Transcription_(genetics)#Termination). The [3'-most](https://en.wikipedia.org/wiki/Directionality_(molecular_biology)) segment of the newly made pre-mRNA is first cleaved off by a [set of proteins](https://en.wikipedia.org/wiki/Protein_complex); these proteins then synthesize the poly(A) tail at the RNA's 3' end.

Poly(A) tails are necessary for translocation of mRNA to the cytoplasm and directly influence a transcript’s stability and translation. While the importance of poly(A) tails is somewhat understood, what effect the length of the poly(A) tail has is not. The initial poly(A) tail length after polyadenylation is reported to be about 250nt however this becomes dynamic after it has left the nucleus and the data so far is difficult to unify. For example, even though only a short poly(A) tail is needed to prevent degradation, hyper-adenylated transcripts degrade quickly in the nuclease. Additionally, there has been evidence showing that long tails during embryonic development are positively correlated with translational efficiency but in *C. elegans* shorter tails are shown to be more actively translated.

Early attempts to measure poly(A) tail length were low-throughput and were transcript-specific. High throughput methods enabled transcriptome wide resolution but were severely limited by short read lengths and technical artifacts related to PCR. It is becoming increasingly clear that while the technologies of today may be capable of providing population-level sequencing to both researchers and clinicians, key limitations remain. From a technological perspective, accuracy and coverage across the genome are still problematic, particularly for GC-rich regions and long homopolymer stretches. In addition, the short reads lengths produced by most current platforms severely limit our ability to accurately characterize large repeat regions, many indels and structural variation, leaving significant portions of the genome opaque or inaccurate. Given the limitations and biases of different platforms, it is also likely that accurate genome sequencing will use a combination of technologies.

The Oxford Nanopore platform is able to sequence mRNA molecules. A protein nanopore is set in an electrically resistant polymer membrane. An ionic current is passed through the nanopore by setting a voltage across this membrane. If an analyte passes through the pore or near its aperture, this event creates a characteristic disruption in current. Measurement of that current makes it possible to identify the molecule in question. The long reads lengths and library preparation protocol allow for sequencing of the full transcript. However, it is very difficult to sequence repetitive regions with Oxford Nanopore and so two tools (nanopolish and tailfindr) have attempted to reconcile this.

# Methods

Nanopolish is a suite of algorithms for analyzing Oxford Nanopore Sequencing data.  It can calculate an improved consensus sequence for a draft genome assembly, detect base modifications (m6a, m5c), call SNPs and indels with respect to a reference genome and more. Also, it has a component, nanopolish polya, to estimate poly(A) tail lengths.

This method combines a hidden markov model (HMM) with an estimator of the translocation rate of the read through the pore. The foundation of the algorithm involves separating the raw “squiggle” into states which correspond to the states that make up the HMM. These states are denoted as follows:

* Start (S): an optional state appearing before the “leader”
* Leader (L): the sequencing adaptor, attached to, and sequenced prior to the RT splint adaptor
* Adaptor (A): RT splint adaptor which is directly attached to the poly(A) tail
* PolyA (P): the polyadenylated region of the read
* Cliff (C): a state to model brief sequencing artifacts within the poly(A) region of the read
* Transcript (T): The coding sequence

Below is a breakdown into these states from an example squiggle:

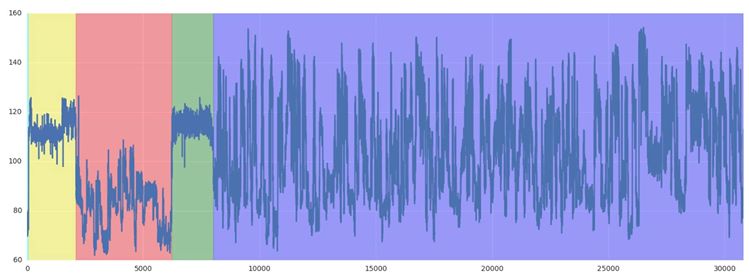


Figure x: Example Segmented Squiggle. Note the read is fed through the pore in the 3’-5’ direction so the last (purple) segment is the transcript. The poly(A) tail segment is shown in green.

And below is a figure showing an example of the HMM:

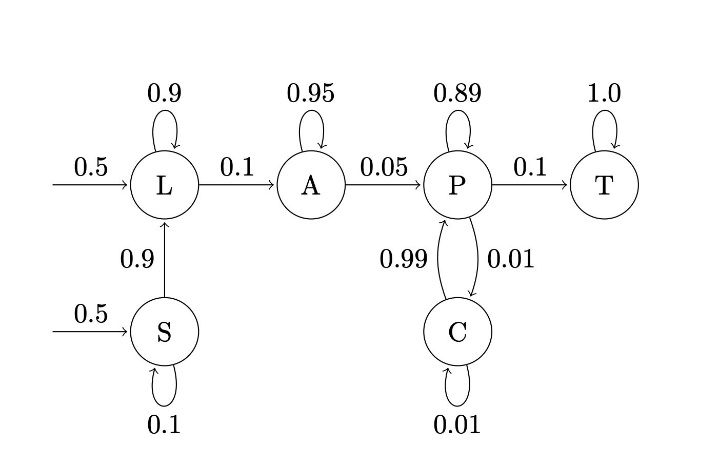


Figure x: Example of hidden markov model with states: start (S), leader (L), adaptor (A) and polyA (P). As shown in fig.2, the transition probabilities control the way the hidden state at time t is chosen given the hidden state at time.

The tailfindr algorithm works as follows:

* Read fast5 file
* Z-normalize raw data
* Clip signal values exceeding 3 sd
* Smoothen by sliding window
* Define poly(A)-containing segment by thresholding smoothened signal by 0.3 (this defines the rough poly(A) boundaries
* Compute mean of every 25 samples of clipped signal within rough poly(A) boundaries
* Compute slope between every two consecutive points of mean signal
* Get precise poly(A) boundaries by shrinking rough poly(A) boundaries until slope signal is confined in +/- 0.3 range
* Normalize tail length by read specific translocation rate

This is shown in the below figure:

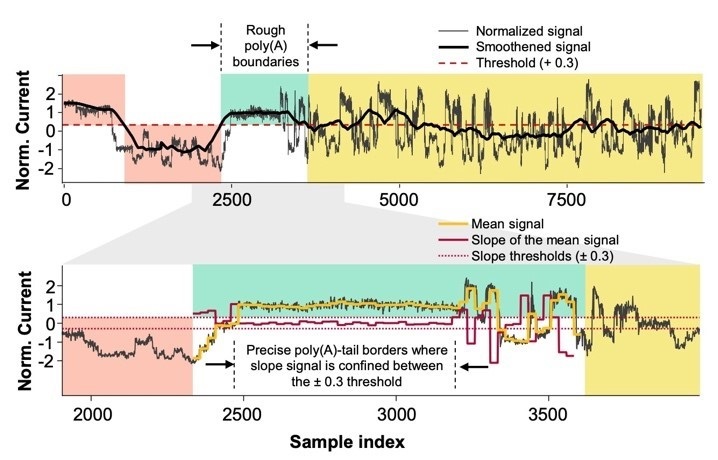


Figure x:

# Results

The dataset we are trying to apply these tools to is the Oxford Nanopore poly(A) standards (linked [here](https://www.ebi.ac.uk/ena/data/view/PRJEB28423)). However, despite the fact that these files are already available with basecalls, they are not compatible with tailfindr because they require some metadata table. Additionally, basecalling is a very computationally heavy task and requires the proprietary software “albacore” or “guppy”. I am currently looking into a potential software “hack” around this.

# Discussion

Blah blah….

Acknowledgements

Thank you to Roham Razaghi from the Timp lab!

References

1. Jalkanen, A. L., S. J. Coleman and J. Wilusz (2014). "Determinants and implications of mRNA poly(A) tail size - Does this protein make my tail look big?" Seminars in Cell & Developmental Biology **34**: 24-32.
2. Tailfindr
3. Nanopolish
4. ONT standards dataset
5. Direct RNA Seq