Estimating polyadenylated tail lengths from direct RNA sequencing signals

Nanopore RNA Consortium

1 Model Description

To estimate the number of bases in the polyadenylated tails of mRNA reads, we developed a predictive model that combines a hidden markov model with an estimator of the translocation rate through the pore. The hidden markov model uses linear-chain state transitions to perform a segmentation of the raw sequencing signal of an mRNA read; the estimated translocation rate is used in conjunction with the segmentation to estimate the tail length, which we elaborate upon below.

In the rest of this supplementary note, we follow Oxford Nanopore Technologies' nomeclature in referring to the sequential raw current measurement values corresponding to a read sequenced via the direct RNA protocol [1] as its *squiggle*, and individual values of a squiggle as *samples*; squiggles are oriented in the direction of time, i.e. in the 3'-to-5' orientation with respect to the strand. We simultaneously refer to a given sequenced mRNA molecule and its sequence of nucleotides as a *read*. Atypically, every mRNA read that we consider below is assumed to be oriented in the 3'-to-5' direction; this is to match the orientation of the direct RNA protocol, which sequences reads in the 3'-to-5' direction.

1.1 Signal Segmentation via Hidden Markov Model

A hidden markov model, which we call the *Segmentation HMM*, is used to segment the squiggle of a read into distinct *regions* appearing sequentially. Biologically, each sequenced read consists of a sequencing adapter (which we call the *leader* region), the RT splint adapter (which we call the *adapter* region), the *polyadenylated tail*, and the coding *transcript*, respectively, from 3' to 5' [1]. The segmentation HMM contains one state for each of these regions connected sequentially via linear chain state transitions. We additionally include two states to handle "jumps" in the squiggle that are due to idiosyncrasies specific to nanopore sequencing, which we explain below.

We assume each state has an associated emission distribution and treat the raw samples of a squiggle as realizations from one of these distributions, dependent on a latent state. For a squiggle $\vec{s} = (s_1, \ldots, s_n)$ with associated latent states $\vec{h} = (h_1, \ldots, h_n)$ — where each h_i is a label representing a region of the read — we have that

$$\forall i: \; s_i \sim p(s|h_i) = \epsilon_i(s),$$

where $\epsilon_i(\cdot)$ is the emission distribution for state h_i. In our HMM, we use Gaussian, Gaussian mixture, and uniform distributions to model emissions. We use the Viterbi algorithm to infer $\vec{h}$ from any given $\vec{s}$.

Prior to running the Viterbi algorithm, we apply a global linear rescaling on all samples of the squiggle to remove per-read variations from the base model. The coefficients of the linear transformation[1] are estimated individually for each read using the same procedure as in [2]. Following [2], we refer to a segmentation of a squiggle $\vec{s}$ into a sequence

$$\vec{e} = (\langle \mu_1, \sigma_1, \delta_1 \rangle, \ldots, \langle \mu_K, \sigma_K, \delta_K \rangle)$$

of contiguous samples (called *events*) as the *event sequence* associated to the squiggle. Samples associated to a single event approximately correspond to a 5-mer residing in the pore at the time of sampling. The event sequence associated to a squiggle is determined by a segmentation algorithm[2] provided by Oxford Nanopore.

[1] The linear rescaling is implemented as a part of the *SquiggleRead* class in nanopolish: https://github.com/jts/nanopolish.

[2] https://github.com/jts/nanopolish/blob/master/src/thirdparty/scrappie/event_detection.c#L268

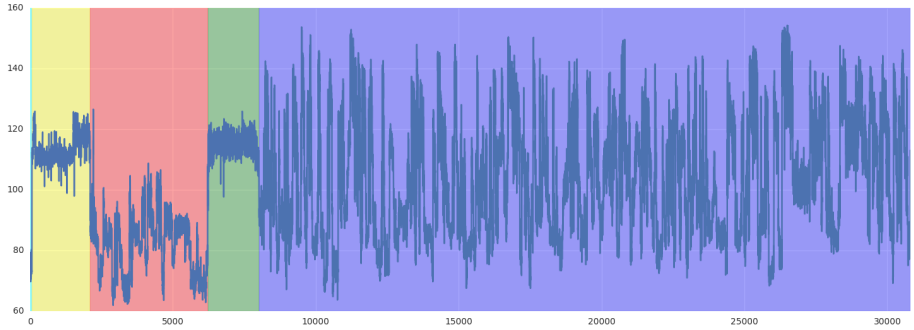

Figure 1: An example of a squiggle segmentation generated by the hidden markov model. Distinct regions, from left to right: *start* (cyan), *leader* (yellow), *adapter* (red), *poly(A) tail* (green), and *transcript* (purple). Two samples flagged as "cliffs" can be observed in the poly(A) tail.

To fit the emission distributions, we use a two-stage bootstrapped procedure where manually-tuned emissions were used in an initial HMM before fitting emissions on the samples of the passing segmentations via maximum likelihood; this is elaborated below in the subsection on the emission distributions. We devote the rest of this section to explaining the state transitions and the emission distributions of the segmentation HMM in further detail.

1.1.1 State Transitions

The hidden states of the Segmentation HMM have the following names (single-letter label in parentheses) and interpretations:

- *START (S)*: an optional state appearing before the *LEADER* segment.
- *LEADER (L)*: the sequencing adapter attached to, and sequenced prior to, the RT splint adapter.
- *ADAPTER (A)*: the RT splint adapter sequence attached to the polyadenylated region as a part of the direct RNA sequencing protocol.
- *POLYA (P)*: the polyadenylated region of a read.
- *CLIFF (C)*: a state that models brief sequencing artifacts within the polyadenylated region.
- *TRANSCRIPT (T)*: the coding sequence of a read.

The states L, A, P, and T are connected via one-way transitions in a linear chain, representing their biologically-expected order of appearance in an mRNA squiggle. *START* is an optional state to account for a short open-pore signal that appears in some reads before the *LEADER* segment. *CLIFF* is a state that models sequencing errors that appear in the *POLYA* region; these are short, sparse regions within the *POLYA* region, occurring for < 10 samples at a time and typically representing $< 1\%$ of the length of the *POLYA* region, that would otherwise cause a mis-segmentation if not modelled. We observed that erroneous 1-sample artifacts of atypically high or low current level caused the segmentation HMM to fail unless we added a *CLIFF* state to model them. As the number of samples in each of the four regions represented by states L, A, P, T is typically fairly large — on the order of thousands of raw samples per region — the weight on the self-loop of each state is much higher than that of a transition to the next state. We set the probability of a self-loop for L to 0.9, for A to 0.95, for P to 0.89, and for T to 1.0, since the latter represents the final region of a read in the 3'-to-5' direction. A full diagram of the state transitions is provided in Figure 2.

1.1.2 Emission Distributions

Emissions are modelled with Gaussian, uniform, and Gaussian mixture distributions. The following emission distributions are used:

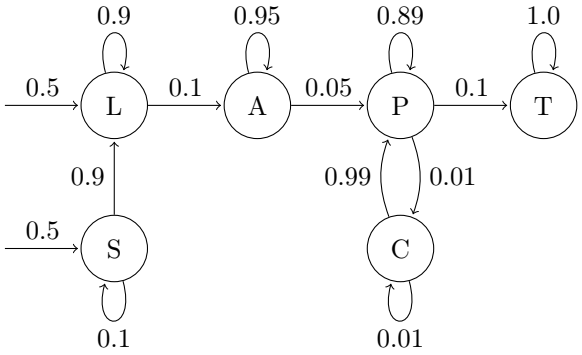

Figure 2: The state transitions of the segmentation HMM. Edges without an origin node on the left indicate the initial state probabilities.

- *START*: $\mathcal{N}(\mu = 70.2737, \sigma^2 = 3.7743)$
- *LEADER*: $\mathcal{N}(\mu = 110.973, \sigma^2 = 5.237)$
- *ADAPTER*: $0.874 \times \mathcal{N}(\mu = 79.347, \sigma^2 = 8.3702) + 0.126 \times \mathcal{N}(\mu = 63.3126, \sigma^2 = 2.7464)$
- *POLYA*: $\mathcal{N}(\mu = 108.883, \sigma^2 = 3.257)$
- *CLIFF*: $\mathcal{U}([70.0, 140.0])$
- *TRANSCRIPT*: $0.346 \times \mathcal{N}(\mu = 79.679, \sigma^2 = 6.966) + 0.654 \times \mathcal{N}(\mu = 105.784, \sigma^2 = 16.022)$

Emission distributions were fitted with a two-stage bootstrapped approach. For each region of the squiggle corresponding to a state, we made an initial estimate of the mean current level and variance of the current levels, and ran the segmentation HMM on each read using these as the parameters of initial emission distributions, before manually filtering the resulting segmentations based on quality. Sample values from each of the S, L, A, P, T regions were aggregated from each of the filtered segmentations, and Gaussians were fitted via maximum likelihood estimation to each squiggle region to obtain the above emission distributions, while each Gaussian mixture was fitted via 100 iterations of expectation-maximization. The number of Gaussian components in each mixture distribution was chosen to be equal to the number of observed peaks in the kernel density estimate of the sample data for each region. The uniform emission distribution for the *CLIFF* state was not fitted with this approach; the upper and lower limits for the uniform distribution were chosen based on manually-tuned observed upper and lower bounds for all samples across all datasets.

1.2 Estimation of the Polyadenylated Tail Length

Fix a read R. Given a segmentation

$$\langle L_0, A_0, P_0, T_0 \rangle$$

of a squiggle

$$\vec{s} = (s_1, \ldots, s_n)$$

with associated events

$$\vec{e} = (\langle \mu_1, \sigma_1, \delta_1 \rangle, \ldots, \langle \mu_K, \sigma_K, \delta_K \rangle),$$

where each component of the segmentation represents the starting index of its respective region — e.g. s_{P_0} is the first sample in the poly(A) tail — we compute an estimate of the number of nucleotides in the poly(A) region by multiplying the duration of time spent in the poly(A) region by the *read rate*, the rate at which the nucleotides of a read translocate through the pore during sequencing. The translocation rate of a read varies as it is being sequenced; hence we instead use the reciprocal of the median event duration as a proxy for a uniform sequence read rate. We found that using the median event duration gave poly(A) tail length estimates that were more robust to read rate differences across different reads than other read-level summary statistics such as the mean event duration.

Our estimator of the polyadenylated tail length is given by

$$\hat{n}_{p(A)} := \frac{|T_0 - P_0|}{\rho \cdot \text{med}(\vec{\delta})} - 5,$$

where:

- $\hat{n}_{p(A)}$ is the estimated number of nucleotides in the polyadenylated region of the read;
- $\text{med}(\vec{\delta}) = \text{med}(\{\delta_i\}_{i=1}^{K})$ is the median event duration from events in the read, in seconds;
- $|T_0 - P_0|$ is the number of samples in the polyadenylated region, as indicated by the segmentation;
- ρ is the sample rate (in $\frac{\text{samples}}{\text{sec}}$) of the nanopore sequencer, i.e. the number of current level samples observed per second; and
- a constant term is subtracted from the quotient term to adjust for the k-mer size associated to the event sequence (in our case, 5).

The sample rate ρ is a fixed constant set by the nanopore sequencer hardware whereas the median event duration differs for each read.

1.3 Reproducibility

The polyadenylated tail length estimator is implemented in the `polya` subprogram of nanopolish:

https://github.com/jts/nanopolish

The analyses performed on the datasets in the accompanying paper may be reproduced by running the associated pipeline, implemented as a Makefile:

https://github.com/paultsw/polya_analysis

1.4 Software

The pipeline referred to in the previous subsection makes use of `albacore` version 2.3.3, `samtools` version 1.9, `minimap2` version 2.12, and `nanopolish` version 10.2. Plotting scripts in the pipeline were developed in `python` version 3.4.6 and `R` version 3.4.4, with `ggplot2` version 2.2.1.

2 References

1. Garalde, D., et al. Highly parallel direct RNA sequencing on an array of nanopores. *Nature Methods* **15**, pages 201-206 (2018).
2. Simpson, J.T., Workman, R.E., Zuzarte, P.C., David, M., Dursi, L.J., Timp, W. Detecting DNA cytosine methylation using nanopore sequencing. *Nature Methods* **14**, pages 407-410 (2017).

SUPPLEMENTARY TABLES

Supplementary Table 1 Yield and read alignment statistics for native RNA and 1D cDNA. General read length statistics (first 5 rows) were calculated using NanoStat package [1]. Aligned reads refers to reads aligned against GENCODE v27 using minimap2. Mean Aligned % identity was calculated using scripts described in Quick *et al.* [2]

	Native RNA Pass	1D cDNA Pass
Reads	10,302,647	15,152,101
Bases (Gb)	10.61	14.13
Mean Read Length	1,030	933
Median Read Length	771	780
Read Length N50	1,334	1,072
Mean Aligned % Identity	86.1	85.0
Median Aligned % Identity	86.6	85.5
Mean Aligned Read Length	987	791
Median Aligned Read Length	726	643
Longest Aligned Read Length	21,608	9,969
Flowcells Used	30	12

1. De Coster, W., D'Hert, S., Schultz, D. T., Cruts, M. & Van Broeckhoven, C. NanoPack: visualizing and processing long-read sequencing data. *Bioinformatics* **34**, 2666–2669 (2018).
2. Quick, J., Quinlan, A. R. & Loman, N. J. A reference bacterial genome dataset generated on the MinION\texttrademark portable single-molecule nanopore sequencer. *Gigascience* **3**, 1–6 (2014).

Supplementary Table 2 Native RNA reads by gene. 9.7 million individual pass native RNA reads were aligned to genes in GENCODE v27 using minimap2 (splice aware setting). 20,289 separate genes were identified in these alignments. (Attached)

Supplementary Table 3 Native RNA reads by isoform assignment. 9.7 million individual pass native RNA reads were aligned to isoforms in GENCODE v27 using minimap2 (splice aware setting). 64,241 separate isoforms were identified in these alignments. (Attached)

Supplementary Table 4 Mapping statistics for RNA and cDNA data aligned to both GRCh38 and GENCODE v27 using minimap2 version 2.1.

	RNA	cDNA
GRCh38	19405001 + 0 in total (QC-passed reads + QC-failed reads) 9057321 + 0 secondary 45033 + 0 supplementary 0 + 0 duplicates 19044282 + 0 mapped (98.14% : N/A) 0 + 0 paired in sequencing 0 + 0 read1 0 + 0 read2 0 + 0 properly paired (N/A : N/A) 0 + 0 with itself and mate mapped 0 + 0 singletons (N/A : N/A) 0 + 0 with mate mapped to a different chr 0 + 0 with mate mapped to a different chr (mapQ>=5)	28313010 + 0 in total (QC-passed reads + QC-failed reads) 12617114 + 0 secondary 543795 + 0 supplementary 0 + 0 duplicates 28028806 + 0 mapped (99.00% : N/A) 0 + 0 paired in sequencing 0 + 0 read1 0 + 0 read2 0 + 0 properly paired (N/A : N/A) 0 + 0 with itself and mate mapped 0 + 0 singletons (N/A : N/A) 0 + 0 with mate mapped to a different chr 0 + 0 with mate mapped to a different chr (mapQ>=5)
GENCODE v27	33875291 + 0 in total (QC-passed reads + QC-failed reads) 23266864 + 0 secondary 305780 + 0 supplementary 0 + 0 duplicates 33312595 + 0 mapped (98.34% : N/A) 0 + 0 paired in sequencing 0 + 0 read1 0 + 0 read2 0 + 0 properly paired (N/A : N/A) 0 + 0 with itself and mate mapped 0 + 0 singletons (N/A : N/A) 0 + 0 with mate mapped to a different chr 0 + 0 with mate mapped to a different chr (mapQ>=5)	50261192 + 0 in total (QC-passed reads + QC-failed reads) 34327618 + 0 secondary 781473 + 0 supplementary 0 + 0 duplicates 49240961 + 0 mapped (97.97% : N/A) 0 + 0 paired in sequencing 0 + 0 read1 0 + 0 read2 0 + 0 properly paired (N/A : N/A) 0 + 0 with itself and mate mapped 0 + 0 singletons (N/A : N/A) 0 + 0 with mate mapped to a different chr 0 + 0 with mate mapped to a different chr (mapQ>=5)

Supplementary Table 5 Kmer coverage for nanopore native RNA reads aligned to GENCODE isoforms. The read sequences were filtered by length and only reads that covered 90% or more of the respective reference sequence were chosen. Expected kmer counts were calculated from the set of reference sequences, and observed kmer counts were calculated from the set of read sequences. (Attached)

Supplementary Table 6 Kmer coverage for nanopore cDNA reads aligned to GENCODE isoforms. The read sequences were filtered by length and only reads that covered 90% or more of the respective reference sequence were chosen. Expected kmer counts were calculated from the set of reference sequences and observed kmer counts were calculated from the set of read sequences. (Attached)

Supplementary Table 7 *MT-CO1* reads for which signal was recovered from either the start or end of the original read file. Reads were mapped using minimap2 (standard parameters). Only the subset of reads for which read mappings were improved are shown. (Attached)

Supplementary Table 8 FLAIR and GENCODE isoform and gene statistics using sensitive and stringent read assignment criteria.

	Total Isoforms	Total Genes	Number of overlapping genes with GENCODE-stringent RNA
GENCODE-sensitive cDNA	79760	24681	12761
GENCODE-stringent cDNA	28408	12659	10151
GENCODE-sensitive RNA	62284	20621	13169
GENCODE-stringent RNA	28302	13169	13169
FLAIR-sensitive RNA	53067	12298	10748
FLAIR-stringent RNA	33984	10793	9816

Supplementary Table 9 Native RNA isoform numbers for the FLAIR-sensitive and FLAIR-stringent sets. The table includes the total number of isoforms and the number of unannotated isoforms by category.

Category	Sensitive	Stringent
Total isoforms	53,067	33,984
Unannotated	31,990	17,116
Unannotated-novel combination of annotated junctions	15,832	7.961
Unannotated-contains intron retention	7,025	2,281
Unannotated-contains novel exon	2,504	1,180

Supplementary Table 10 Number of isoforms detected as a function of sampling depth. Native RNA reads from the total population (8.16 M) were subsampled in 10% increments. The number of isoforms detected per subsample are tabulated for each isoform set. These data are plotted in **Supplementary Figure 10**.

Sampling depth		Number of isoforms detected at a given sampling depth			
Percent	Number of reads	FLAIR-sensitive	FLAIR-stringent	GENCODE-sensitive	GENCODE-stringent
10	815616	14086	9192	28046	11870
20	1631232	20802	13505	36909	15857
30	2446848	26189	17026	47275	18508
40	3262464	30951	20283	46980	20546
50	4078080	35347	22719	50609	22384
60	4893695	39705	25462	53470	23822
70	5709311	43186	27773	56075	25129
80	6524927	46880	30105	58294	26206
90	7340543	50558	32080	60450	27339
100	8156159	53067	33984	62284	28302

Supplementary Table 11 Summary of allele-specificity data for reads containing at least 2 haplotype-informative variants.
The columns are organized as follows:

1. Ensembl gene ID
2. Total reads for that gene (read > gene assignment was done using the output of FLAIR)
3. Portion of reads originating from maternal allele
4. Portion of reads originating from paternal allele
   [note : maternal and paternal may not add to 1, as some reads were not assigned]
5. Chromosome
6. Gene Assignment by Binomial Test, (p=0.01) (Maternal / Paternal / Unassigned)
   [note : inclusive of all isoforms of the gene]
7. Gene Symbol
8. Annotation for whether gene is on autosome or allosome

(Attached)

Supplementary Table 12 Unique isoforms expressed from each of the parental alleles. (Attached)

Supplementary Table 13 Estimate of poly(A) lengths for a synthetic enolase control transcript bearing different known poly(A) tail lengths.

Statistic	10x	15x	30x	60x	60xN	80x	100x
Read count	27477	23000	18680	29823	91930	175162	59207
Median absolute deviation (mad)	5	5.2	6.25	9.28	12.52	12.01	22.87
Mean	14.57	19.93	37.48	72.82	64.13	102.74	173.21
Median	11.43	17	32.89	62.63	56.26	82.13	108.68
Mode	5.96	14	31	58	59	77	93
Percent within 2 mad of expected	79.97	80.51	78.1	76.52	72.06	74.78	66.84
Percent within 2 stdv of expected	97.53	97.63	97.24	96.76	97.44	95.97	92.54
Standard deviation (stdv)	15.3	17.07	25.35	48.61	54.52	80.06	173.38

Supplementary Table 14 Statistics for poly(A) tail length of GENCODE-sensitive genes with greater than 500 reads. (Attached)