UNIVERSITY OF CALIFORNIA
SANTA CRUZ

IDENTIFICATION OF FULL-LENGTH TRANSCRIPT ISOFORMS USING NANOPORE SEQUENCING OF INDIVIDUAL RNA STRANDS

A dissertation submitted in partial satisfaction of the
requirements for the degree of

DOCTOR OF PHILOSOPHY

in

BIOMOLECULAR ENGINEERING AND BIOINFORMATICS

by

Logan Mulroney

December 2020

The Dissertation of Logan Mulroney
is approved:

Professor Mark Akeson, Chair

Professor Manuel Ares

Professor Rebecca DuBois

Professor Seth Rubin

Quentin Williams
Acting Vice Provost and Dean of Graduate Studies

Table of Contents

List of Figures

List of Tables

Abstract

Identification of full-length transcript isoforms using nanopore sequencing of individual RNA strands

by

Logan Mulroney

Before RNA can be sequenced using next generation sequencing (NGS) technologies, it is first converted into cDNA (RNA-Seq). In 2016 Oxford Nanopore Technologies released their direct RNA nanopore sequencing technology, circumventing the requirement for cDNA. The native RNA is sequenced continuously from the 3′ end through to the 5′ end. Two limitations of this approach are: ambiguity in discriminating between full-length and truncated reads; and the requirement for a known invariable 3′ end, such as the poly(A) tail.

In collaboration with New England Biolabs, we developed a technique to identify full-length native RNA nanopore reads by specifically labeling capped RNA 5′ ends with a nanopore detectable sequence. Using this strategy, we aimed to identify individual high-confidence full-length human mRNA isoform scaffolds among ∼4 million nanopore poly(A)-selected RNA reads. First, we exchanged the biological 5′ m^7G cap for a modified cap bearing a 45-nucleotide oligomer. This oligomer improved 5′ end sequencing and ensured identification of capped strands. Second, among these capped reads, we screened for 3′ ends consistent with documented polyadenylation sites. This gave 185,434 high-confidence mRNA scaffolds, including 4,262 that represented isoforms

absent from GENCODE. Most of these had transcription start sites internal to longer, previously identified mRNA isoforms. Combined with orthogonal data, these mRNA scaffolds provide decisive evidence for full-length mRNA isoforms.

In collaboration with the Ares lab, we developed a technique to label native RNA 3′ ends with polyinosine. This step permits sequencing adapters to be ligated to both poly(A) and non-poly(A) RNA in a single sequencing experiment. Polyinosine tails are not known to naturally occur and produce a recognizable signal in nanopore ionic current data. These two features make it ideal for adapting a variety of RNA types while preserving native RNA 3′ end sequence information. We implemented a Hidden Markov Model that identifies the polyinosine tail signal on the RNA 3′ ends with 98.46 % accuracy. This classifier can be used to filter the reads for a particular RNA 3′ end type (e.g. separate nascent RNA from mature mRNA).

I would not have been able to achieve my dream of becoming a scientist without the support of my parents, Kristin, Conor, and Tisha or my brother, Garrett. They all housed me when I needed a home, fed me when I needed a meal, celebrated the highs, and helped me through the lows.

This is as much their achievement as it is mine.

I love you all.

Acknowledgments

I would like to thank my advisor, Professor Mark Akeson, for the continuous support of my Ph.D research. Mark has been my strongest supporter and has pushed me to think critically about my own work. Professor Akeson has exemplified time and time again what a careful and thoughtful scientists looks like. His guidance has shaped my research and this thesis into something to be proud of, and his mentorship will influence my future scientific endeavors.

I would like to thank the rest of my thesis committee, Professor Manuel Ares, Professor Rebecca DuBois, and Professor Seth Rubin for their insights and encouragement during my graduate training.

My sincere thanks also goes to my many collaborators, Dr. Madalee Wulf, Dr. Laurence Ettwiller, Dr. Ivan R. Corrêa Jr, Dr. Ira Schildkraut, Dr. George Tzertzinis, Dr. John Buswell, Professor Manuel Ares, and Jenny Vo. Who gave advice and contributed to this research in enumerable ways. This body of research would not be possible without their contributions.

I thank Dr. David Bernick for the many years talking about research, working together on iGEM, and his friendship during my graduate career.

I especially thank: Jeff Nivala for mentoring me in protein expression and general molecular biology techniques; Andrew Smith for sharing his wisdom on RNA biology and the many discussions about our research; Miten Jain for the countless hours helping me with data analysis and his thoughtful comments; Hugh Olsen and

Robin AbuShumays for their constant encouragement and support during my graduate career. I also thank the many past and present lab mates in The Nanopore group and the Ares lab for being amazing people to work with. I consider all of you to be lifelong friends and colleagues. I could not have asked for a better group of people to work with.

I would also like to thank everyone in the BME and MCD departments. Being surrounded by people where we have casual conversations about data or papers over lunch or bread and tea has been a profound learning experience for me. The broad range of research topics that I've been able to see and talk about has broadened my own thinking.

I am most grateful to my family for their unconditional support throughout my life and graduate career. I love you all.

Chapter 1

Introduction

1.1 cDNA and RNA sequencing technologies

Nucleic acids have been a focus of biological research since DNA was discovered as the genetic material [1]. Fred Sanger made the first major advancement in sequencing technology in 1977 with chain terminating sequencing [2]. In the past five decades there have been many DNA sequencing technologies, many of which are no longer used [3]. The remaining next generation sequencing (NGS) technologies all read short DNA fragments that cannot be used to sequence RNA directly. Historically, the major classes of RNA (tRNA [4, 5], ribosomal RNA (rRNA) [6], mRNA [7, 8], and other non-coding RNA (ncRNA) [9]) have been probed using microarrays [10, 11].

Recently, use of reverse transcription [12, 13] to convert the RNA into cDNA has permitted NGS RNA sequencing. This process is called RNA-seq [14]. RNA-seq can generate data for low abundance transcripts, which has uncovered subtle variants

within all major classes of RNA. RNA-seq has been used to define gene boundaries [14], splice sites [14], and to understand gene expression and regulation [15].

The average human mRNA is 3392 nucleotides (nt) long [16]. Due to RNA-seq short read lengths (25-300 bp) [3], identifying both the transcription start site (TSS) and the transcription termination site (TTS) in the same read is difficult using RNA-seq [15]. Alternative TSS and TTS usage, and alternative splicing, increase the number of functional transcripts (isoforms) per gene [17]. Defining the TSS and TTS when there is more than one isoform per gene is further complicated by cDNA synthesis defects, such as incomplete cDNA synthesis [18] and internal priming [19].

There are techniques, such as CAGE [20], Oligo-capping [21], and $5'$ RACE [22], that identify TSS in NGS data. These techniques detect the TSS by adapting the $5'$ m^7G RNA cap [23], ligating adapters to deprotected $5'$ ends [24], or extending the RNA $5'$ end with a template switching oligonucleotide [22, 25]. The sensitivities and error profiles for detecting TSS varies among these techniques [26]. Low abundance or internal TSS are difficult to detect because of cDNA artifacts [18, 19, 27]. A long-read sequencing technology that captured both the TTS and TSS in a single read would resolve these uncertainties.

1.1.1 Long read sequencing technologies

Presently, there are two long read sequencing platforms: Single Molecule Real-Time sequencing (SMRT-seq) and Nanopore sequencing. Single Molecule Real-Time sequencing is a sequencing by synthesis platform that achieves long reads by measuring

fluorescent leaving groups during polymerization by individual polymerases anchored to a solid platform [28]. Single Molecule Real-Time sequencing typically acquires cDNA reads that range from 200 bp to 10 kb long at ∼1 % error rate [29]. Sequencing with the Single Molecule Real-Time sequencing platform has revealed new RNA TSS, TTS, and isoforms [30].

Oxford Nanopore Technologies (ONT) strand sequencing identifies nucleic acids directly [31] as they translocate through an angstrom scale pore and modulate the ionic current [32,33]. An enzyme translocase is coupled to the nanopore to regulate nucleotide translocation in discrete 2-20 ms steps [34–36]. Strands of nucleotide polymers can be sequenced as they existed in the cell without intermediate synthesis steps. This results in rapid sample to sequencer times and near real-time sequence acquisition. Additionally, because the sequencing device is a small USB powered device, samples can be sequenced on site, such as in West Africa [37] or on the International Space Station [38, 39].

Sequencing native RNA directly with nanopore technology circumvents cDNA intermediates, avoiding bias due to reverse transcription [18,19,27]. The ONT nanopore native RNA sequencing platform works by adapting the RNA 3′ end and sequencing continuously in the 3′ to 5′ direction along the RNA strand [40].

The standard ONT native RNA sequencing protocol targets poly(A) RNA [40]. This is accomplished using a splint adapter that selectively ligates to the RNA 3′ poly(A) tail. Poly(A) RNA is the most heterogeneous class of RNA, but only accounts for ∼5 % of the total RNA in a cell by mass [41]. There are many other types of RNA, such as rRNAs and tRNAs, which do not normally have poly(A) tails. These other RNA types

can be adapted by *in vitro* poly(A) tailing [42, 43] or using specific adapters for target 3′ ends [44].

In principle, identifying TTS in nanopore reads should be straightforward because a sequencing adapter is ligated to the mature mRNA 3′ end. However, some polyadenylated strands are degradation products and do not contain the original TTS [45], which can confound transcript interpretation.

TSS are more challenging to detect in nanopore RNA reads than are TTS. Close inspection of standard RNA nanopore sequences revealed that ∼10-15 nucleotides were missing from the 5′ ends [46]. The most likely explanation for this is that the enzyme-motor releases the RNA strand ∼12 nucleotides from the limiting aperture (Figure 1.1). This causes the terminal ∼12 nucleotides to translocate faster than the limit of detection. Furthermore, *in vivo*, *in vitro*, or *in silico* strand breaks will also prevent nanopore reads from reaching the TSS [46].

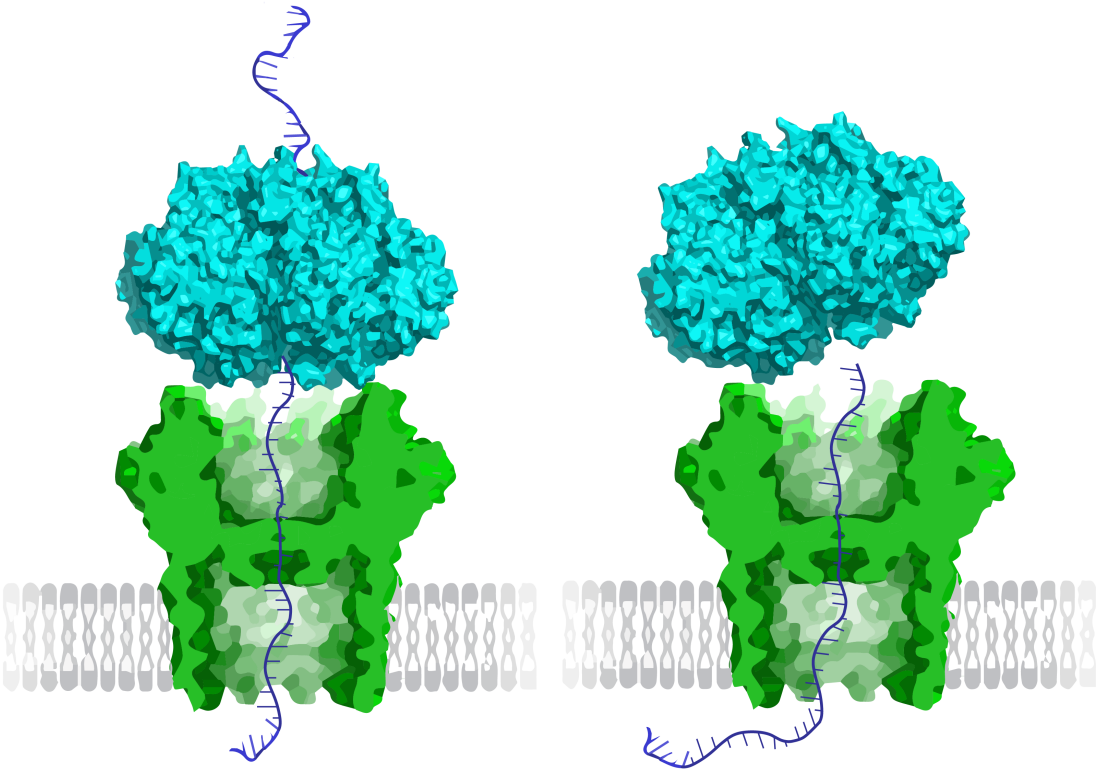

Figure 1.1 Diagram of nanopore/enzyme-motor/RNA complex. The first complex shows an enzyme regulated RNA strand translocating through the nanopore. The second complex shows the moment the enzyme releases the RNA strand and it transitions to voltage regulated translocation. Enzyme regulated RNA translocates ∼1 nt per 10 000 µs. Voltage regulated RNA translocates ∼1 nt per 1-10 µs [47]. RNA data are acquired at 3012 Hz, making the limit of detection ∼1 nt per 332 µs. When an RNA strand is released by the motor-enzyme, the remaining untranslocated terminal ∼11 nucleotides are driven through the nanopore by the applied voltage, which is one to two orders of magnitude faster than the limit of detection. (CsgG nanopore PDB ID: 4q79 [48]. SV40 Large T helicase PDB ID: 1SVM [49], modeled with DNA strand for illustration)

Identifying the transcription start and termination sites on individual native RNA strands would give a direct read out of RNA isoforms without relying on cDNA synthesis. Technology advancements that overcome these native nanopore RNA limitations could lead to a better understanding of isoform structure and function.

1.2 Research outline

My dissertation focuses on two research projects that are described in Chapters 2, 3, and 4. They are summarized below.

Chapter 2, *Identification of human poly(A) RNA isoform scaffolds using nanopore sequencing.* We aimed to identify individual high-confidence full-length human mRNA isoform scaffolds among ∼4 million nanopore poly(A)-selected RNA reads. First, we exchanged the biological 5′ m7G cap for a modified cap bearing a 45-nucleotide oligomer. This oligomer improved 5′ end sequencing and ensured identification of capped strands. Second, among these capped reads, we screened for 3′ ends consistent with documented polyadenylation sites. This gave 185,434 high-confidence mRNA scaffolds, including 4,262 that represented isoforms absent from GENCODE.

Chapter 3, *Optimizing 5′ cap-adaptation using* Saccharomyces cerevisiae *poly(A) RNA.* We optimized the 5′ cap-adaptation strategy using *S. cerevisiae* poly(A) RNA. The yeast transcriptome is suited for this because the m^7G cap is identical to the human m^7G cap, and because most yeast genes encode only one RNA isoform [50]. Using a Copper-free click reaction eliminated RNA degradation during the click step. Changing from Copper-catalyzed to Copper-free chemistry also improved the percentage of *S. cerevisiae* poly(A) RNA reads that were cap-adapted (13.4 % to 38.4 % respectively), and the read N50 length (692 nt to 744 nt respectively).

Chapter 4, *Detecting both poly(A) and non-poly(A) RNA with a generalized RNA nanopore sequencing strategy.* To preserve and capture native 3′ end nucleotide

sequence information, we developed a polyinosine tailing method that preserves the natural 3′ end structure while allowing the broader population of RNAs to be adapted for sequencing. We show that the inosine homopolymer produces a distinctive ionic current signature that allows it to be distinguished from a natural poly(A) tail. This signal was used to develop a classifier that identifies the presence of a poly(I) tail and estimate the tail length.

1.3 Individual contributions

The projects discussed in Chapters 2 and 3 were conceived and designed by myself, Mark Akeson, Madalee Wulf, Ira Schildkraut, George Tzertzinis, John Buswell, Miten Jain, Hugh Olsen, Ivan R. Corrêa Jr. and Laurence Ettwiller.

I performed all of the nanopore experiments, adapter detection optimization, and analyses that validated high-confidence full-length mRNA scaffolds using documented transcription start sites and polyadenylation sites. I am largely responsible for the final draft of a manuscript that will be submitted to Nature Biotechnology.

All RNA cap-adaptation was performed at New England Biolabs (NEB). The pipeline filtering for unannotated TSS was created by Laurence Ettwiller. Short read 5′ RACE was done by Ira Schildkraut and George Tzertzinis. The cap-adaptation enzymatic steps were performed by Ira Schildkraut and George Tzertzinis. The cap-adapter oligonucleotide was synthesized by John Buswell, and the "click" reaction was optimized and performed by Madelee Wulf and Ivan R. Corrêa Jr. Hugh Olsen and

Miten Jain helped with bioinformatic tool development. Mark Akeson oversaw and advised me on all aspects of this project.

The project in Chapter 4 was conceived by Manuel Ares. I performed most of the nanopore experiments, created the training data set of known inosine tail lengths, and developed a Hidden Markov Model to detect and classify the 3′ tail. I discovered that polyU polymerase could use ITP as a substrate.

Jenny Vo optimized the poly(I) tailing conditions and performed some of the nanopore experiments. Miten Jain assisted with the Hidden Markov Model used to classify the inosine-dependent nanopore ionic current.

Chapter 2

Identification of human poly(A) RNA isoform scaffolds using nanopore sequencing

2.1 Abstract

Nanopore sequencing devices read individual RNA strands directly. This facilitates identification of exon linkages and nucleotide modifications, however using conventional methods the $5'$ and $3'$ ends of mature mRNA cannot be identified unambiguously. This is due in part to the architecture of the nanopore/enzyme-motor complex, and in part to RNA degradation in vivo and in vitro. We identified individual full-length human mRNA isoform scaffolds among ∼4 million nanopore poly(A)-selected RNA reads. We

exchanged the biological 5′ m^7G cap for a modified cap bearing a 45-nucleotide oligomer. This oligomer improved 5′ end sequencing and ensured identification of capped strands that were then screened for 3′ ends consistent with a polyadenylation site. This gave 185,434 high confidence mRNA scaffolds, including 4,262 that represented isoforms absent from GENCODE. Most of these had transcription start sites internal to longer, previously identified mRNA isoforms. Combined with orthogonal data, these mRNA scaffolds provide decisive evidence for full-length mRNA isoforms.

2.2 Introduction

Most human genes encode multiple transcript isoforms. These isoforms are derived from alternative splicing, alternative transcription start sites (TSS), or alternative transcription termination sites (TTS). Together, alternative TSS and TTS account for most tissue dependent exon usage [17]. Identification of an RNA isoforms is difficult when either the TSS or TTS is unknown or positioned within the genomic region of a larger isoform [15], and internal isoforms are often omitted from transcriptome annotations [51]. Direct sequencing of nucleotides between the 5′ cap and 3′ poly(A) tail on individual RNA reads could reveal the isoform structure and associated modifications without error-prone computational tools [52–57].

Nanopore RNA sequencing is a single molecule technique that reads RNA directly rather than cDNA copies [40,46,58]. This avoids cDNA artifacts [27] and permits detection of RNA modifications, thus far including m^6A [40,46,58], pseudouridine [44],

inosine [46], and m^7G [44, 46]. Approximate Poly(A) tail lengths can also be discerned for those reads [46].

However, using standard protocols, nanopore direct RNA reads terminate before reaching the $5'$ end of captured molecules. This is because the enzyme that regulates translocation releases captured strands approximately 12 nucleotides from the pore limiting aperture [46]. Accurate identification of full-length isoforms is further complicated by RNA strand degradation that occurs in the cell, during sample preparation, or *in silico* [46]. A possible marker for full-length reads would be the $5'$ m^7G cap that is found in most eukaryotic mRNA [23].

Two groups [58, 59] independently employed an enzymatic decapping and ligation strategy [21] to document the $5'$ ends of nanopore RNA reads that are capped by m^7G. Here we introduce an alternative chemo-enzymatic method wherein an RNA oligonucleotide adapter is chemically attached to a cap analogue replacing the native m^7G cap. To demonstrate the utility of this strategy, we acquired 574,091 cap-adapted reads from human GM12878 poly(A) RNA. Among these, 185,434 high confidence scaffolds aligned to protein coding genes and had $3'$ ends consistent with polyadenylation sites. We systemically validated these scaffolds using orthogonal data for transcription initiation.

2.3 Results

2.3.1 Nanopore cap-adaptation strategy

The chemo-enzymatic cap-adaptation strategy is diagrammed in Figure 2.1a. First, we used yeast scavenger decapping enzyme (yDcpS) [60] to remove the m^7G cap from poly(A)-enriched RNA, leaving 5′-diphosphate ends [61]. Second, the 5′-diphosphate RNA strands were recapped with 3′-azido-ddGTP using Vaccinia capping enzyme (VCE) [62] (Supplemental Figure 2.7). Third, the 3′-azido recapped RNA strands were covalently attached to a dibenzocyclooctyne-amine (DBCO) reactive group on the 3′ end of a 45 nt long RNA oligonucleotide adapter using specific Copper-free 'click' chemistry [63, 64] (Supplemental Figure 2.8).

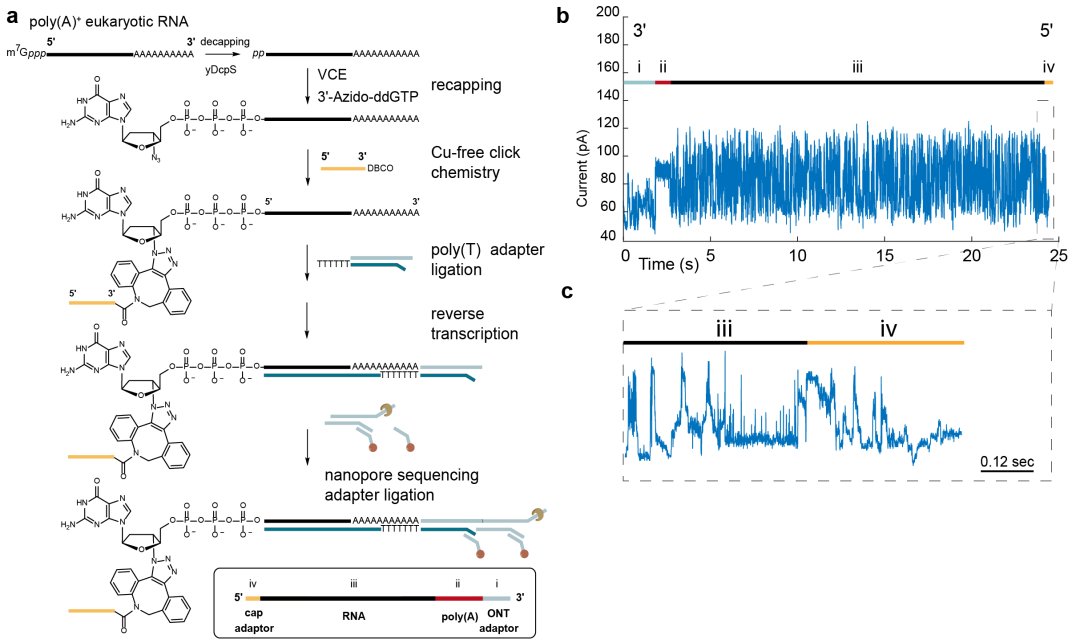

Figure 2.1 5′ cap-adaptation strategy. (**a**) Adaptation and library preparation workflow. (**b**) Representative ionic current trace for a cap-adapted full-length RNA read is shown for the thymidine phosphorylase gene (TYMP). The trace begins with ionic current associated with the ONT adapter (i). This is followed by a monotonic ionic current associated with the 3′ poly(A) tail (ii) and then a variable ionic current associated with the RNA transcript nucleotides (iii). The final segment is an ionic current signature characteristic of the 45 nt RNA cap-adapter (iv). (**c**) An approximately one second window centered on the boundary between the ionic current associated with the 5′ end of the transcript (iii) and a characteristic adapter ionic current trace (iv).

A typical cap-adapted ionic current trace for human Thymidine Phosphorylase (TYMP) is shown in Figure 2.1b. Following strand capture, a variable ionic current is caused by translocation of the ONT 3′ adapter (i). This is followed by a monotonic ionic current associated with the 3′ poly(A) tail (ii) and then a variable ionic current with a bottle brush appearance associated with a sequence of RNA nucleotides (iii). The trace terminated with an ionic current signature characteristic of the 45 nt RNA cap-adapter

(iv). This signature indicated that individual strands were read through the original 5′ end (Figure 2.1c). We used a sequence-based barcode identification software (Porechop `https://github.com/rrwick/Porechop`) to detect the adapter on individual nanopore reads (see Methods).

2.3.2 Optimizing 5′ cap-adaptation using *Saccharomyces cerevisiae* poly(A) RNA

We optimized the 5′ cap-adaptation strategy using *S. cerevisiae* poly(A) RNA. The yeast transcriptome is suited for this because the m^7G cap is identical to the human m^7G cap, and because most yeast genes encode only one RNA isoform [50].

Initially, we used a Copper-catalyzed click reaction for the 5′ adaptation step (Supplementary Methods), however RNA degradation was unacceptable as measured by RNA integrity number (RIN) [65] (Table 2.1). As an alternative, we implemented a Copper-free chemistry step based on a strain-promoted click reaction (Figure 2.1a, and Methods) [64, 66]. This eliminated RNA degradation during the click step (Table 2.1). Changing from Copper-catalyzed to Copper-free chemistry also improved the percentage of *S. cerevisiae* poly(A) RNA reads that were cap-adapted (13.4 % to 38.4 % respectively), and the read N50 length (692 nt to 744 nt respectively).

Table 2.1 Effect of Copper-catalyzed and Copper-free click reactions on RNA integrity and nanopore read quality. The RIN was measured from total RNA after enzyme treatment and purification for each step of the cap-adaptation process using an Agilent RNA 6000 Nano Kit (mean ± SD for n = 2 experiments). Percent cap-adapted is the percent of poly(A) RNA nanopore reads identified by Porechop as cap-adapted. The read N50 is where half of the total bases sequenced are in reads of that length or longer.

	No Treatment	yDcpS	VCE	Copper-catalyzed Click Adaptation	Copper-Free Click Adaptation
RIN	9.5 ± 0.1	9.4 ± 0.4	8.1 ± 0.2	6.7 ± 0.2	8.1 ± 0.7
Percent cap-adapted	-	-	-	13.4 %	38.4 %
N50	957 nt	-	-	692 nt	744 nt

2.3.3 Applying cap-adaptation to human poly(A) RNA transcripts

Having optimized the 5′ cap-adaption chemistry, we applied the Copper-free strategy to poly(A) RNA isolated from GM12878 cells, a model human B-lymphocyte cell line. We acquired 4 million reads that went through the cap-adaptation process (we refer to this population as 'treated reads' in the text that follows). We identified 574,091 of treated reads as cap-adapted (14.3 %) (see Methods, Supplementary Figure 2.6, and Chapter 3 for a detailed description of the optimization).

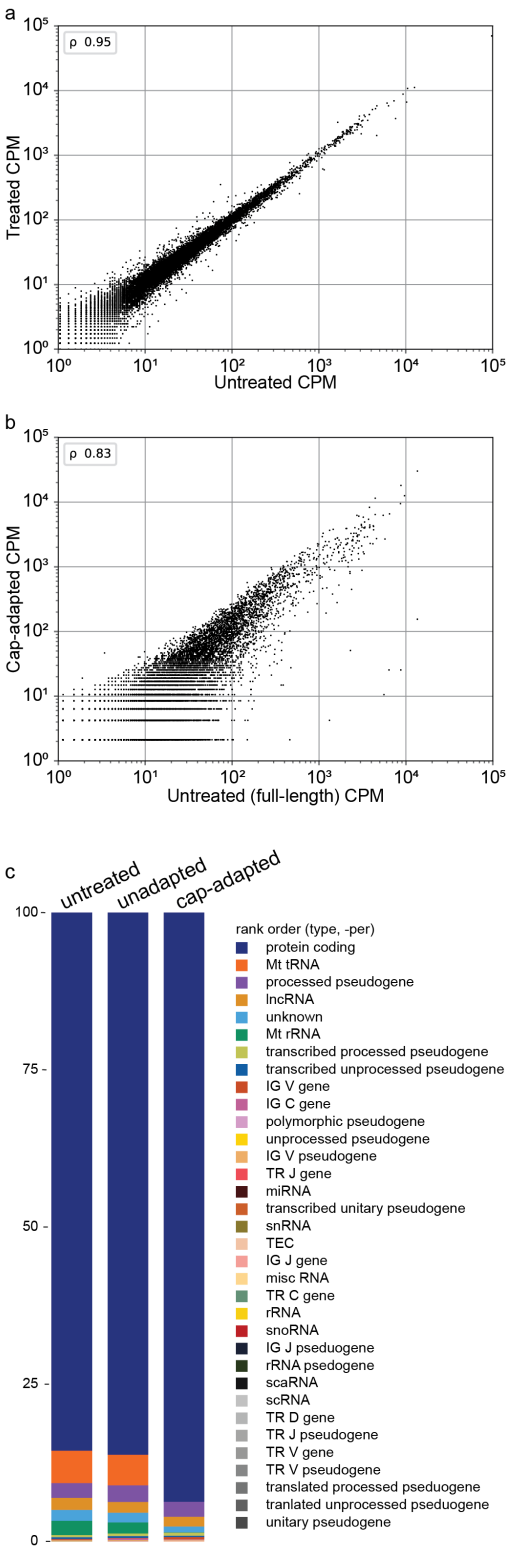

Figure 2.2 Comparison of RNA read counts between treated and untreated samples. Number of transcripts per gene counts per million (CPM) correlation plots (**a**) between untreated and treated samples and (**b**) between full-length untreated and cap-adapted samples. The Spearman's r (ρ) was calculated for each. (**c**) Percent of RNA by class for untreated, unadapted, and cap-adapted reads. All class percentages are in Supplementary Table 2.3.

The N50 value for the cap-adapted reads was 1301 nt, which was shorter than the N50 value for an untreated control (1614 nt) (Supplementary table 2.2). Given this difference, we were concerned that cap-adaptation adversely affected human RNA transcript recovery. To test this, we compared the number of transcript copies per gene for the untreated vs the treated samples. The Spearman rank correlation score was very strong (0.95), indicating that the cap-adaptation process did not substantially affect RNA strand recovery (Figure 2.2a). We then compared the number of transcript copies per gene for the cap-adapted vs full-length untreated samples, using a previously described definition for full-length [44, 46] (see Methods). The Spearman rank correlation score was also very strong, but lower (0.83) (Figure 2.2b).

Among untreated, treated, and cap-adapted reads, at least 85 % of aligned nanopore reads corresponded to protein coding genes (Figure 2.2c). Mitochondrial RNA reads accounted for ∼5 % of the treated and untreated reads. By comparison, mitochondrial RNA reads were underrepresented among the cap-adapted reads. This made sense because mitochondrial mRNA 5′ ends usually bear a 5′ monophosphate, or they are capped by NAD^+ and NADH [67], which the cap-adaptation strategy does not recognize [61].

2.3.4 5′ cap-adaptation improves identification of human poly(A) RNA transcription start sites

The cap-adaptation strategy was designed to identify m^7G capped RNA 5′ ends and improve base calling near those ends. If successful, we predicted that cap-

adapted nanopore reads would be enriched for 5′ ends proximal to TSS annotated by GENCODE [51]. This prediction was substantiated (Figure 2.3a). For example, we found that 99 % of cap-adapted reads 5′ ends were within 300 nt of an annotated TSS compared to 77 % for untreated reads (Figure 2.3a).

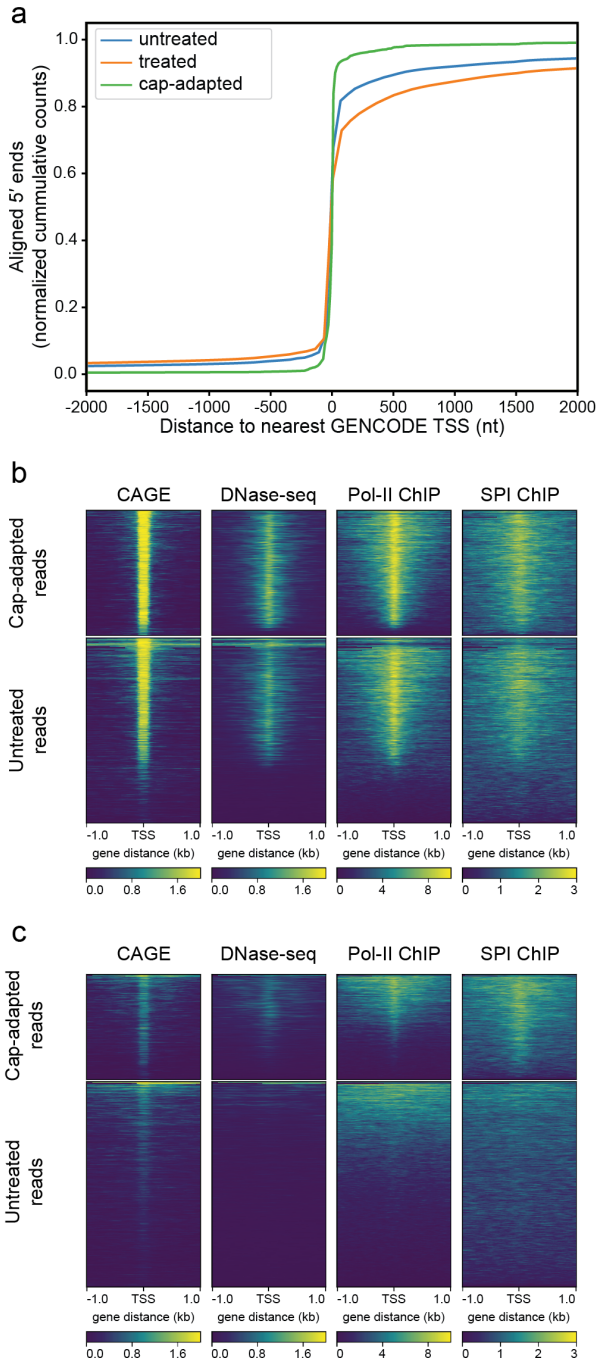

Figure 2.3 Correspondence between RNA nanopore read 5′ ends and orthogonal TSS evidence. (**a**) Nucleotide distance of RNA 5′ ends from annotated GENCODE TSS. The x-axis is the number of nucleotides between a nanopore read 5′ end and the closest TSS annotated in GENCODE v.32. Negative numbers are upstream (5′) from the TSS; positive numbers are downstream (3′) from the TSS. The y-axis is the cumulative number of RNA nanopore reads at a given distance of their 5′ end from an annotated GENCODE TSS. Values are normalized as a fraction of total counts for a given treatment. (**b**) Comparison of all poly(A) RNA nanopore read 5′ ends to orthogonal TSS markers. Each plot is a heatmap where the x-axis is a ±1 kb window centered on the 5′ end of each RNA nanopore read. Each row in the y-axis is an individual read. The color intensity is the read depth normalized signal (CAGE, DNase-seq) or fold change over control for each position (POLR2 and SPI1). The top plots are cap-adapted reads, the bottom plots are untreated reads. (**c**) Comparison of unannotated RNA nanopore read 5′ ends to orthogonal TSS markers. Unannotated 5′ ends are defined as reads where the 5′ end is aligned more than 300 nucleotides from an annotated TSS. The number of reads in each plot were down sampled to 9,116 reads (The number of unannotated cap-adapted reads).

We compared the nanopore read 5′ ends to other markers of transcription ini-

tiation including: DNase-seq (a measure of open chromatin); PolII Chip-seq (a measure of where polII binds to genomic DNA); SPiI ChiP-seq (a measure for an immune-cell-specific transcription factor binding to genomic DNA); and CAGE (Cap Analysis of Gene Expression) (Figure 2.3b). A majority of cap-adapted read 5′ ends overlapped with these other markers of transcription initiation.

It is noteworthy that a few of the cap-adapted reads had 5′ ends that mapped to regions that do not correspond to annotated TSS, suggesting potential alternative TSS. To test this postulate, we selected cap-adapted reads where the 5′ end mapped farther than 300 nt away from any GENCODE TSS [51] and were defined as having a high confidence 5′ end (see Methods). We found 9,116 (1.6 %) such reads corresponding to 1,915 genes. The majority of these newly identified 5′ ends overlapped with other the same genomic markers of TSS (Figure 2.3c). There is overlap between the unannotated cap-adapted read 5′ ends, albeit less than all of the cap-adapted reads. In comparison, there was no distinguishable overlap between the orthogonal markers of TSS with untreated read 5′ ends that aligned further than 300 nt from any annotated TSS. Although neighboring orthogonal markers of TSS are not definitive proof that these unannotated TSS were functional, they do increase confidence that they were bonafide TSS [26].

2.3.5 Documentation of full-length mRNA scaffolds

A core aim of this study was to facilitate mature mRNA isoform identification using individual full-length transcripts as scaffolds. This required identification of cap-adapted nanopore reads that aligned to protein coding genes, and that correctly

identified both the 5′ and 3′ ends of the mRNA.

Among the 294,107 high confidence cap-adapted nanopore 5′ capped RNA reads, 257,721 aligned to protein coding genes documented by GENCODE [51]. These were screened for the presence of poly(A) tails using nanopolish-polya [46], which resulted in 195,222 reads. To filter for mRNA, we documented reads that had 3′ ends that aligned within -60 to +10 nt of annotated polyadenylation sites [68]. This resulted in 185,434 individual full-length mRNA scaffolds (Figure 2.4). These scaffolds corresponded to 7,794 protein coding genes. Per gene transcript coverage ranged from 1-to-3,176. Among the 9,116 reads with unannotated TSS, 4,262 were associated with full-length mRNA scaffolds.

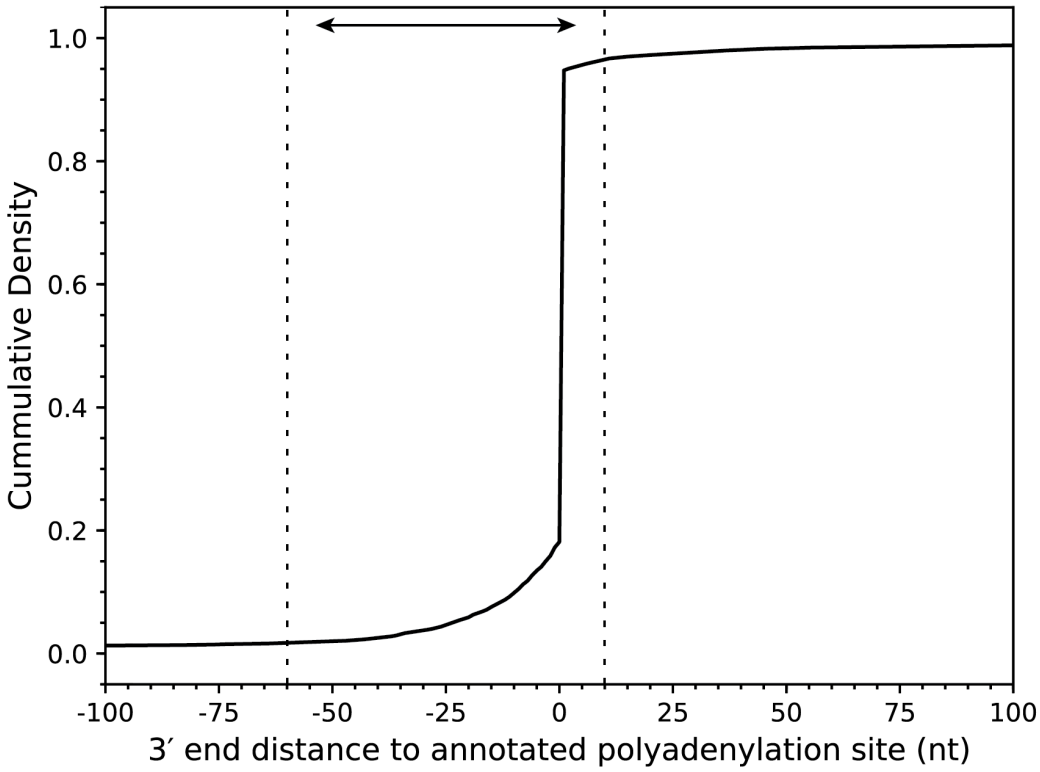

Figure 2.4 Distance of mRNA scaffold 3′ ends from known polyadenylation sites. The x-axis is the number of nucleotides between the aligned 3′ end of each mRNA scaffold and the closest annotated polyadenylation site [68]. Negative numbers are upstream (5′) from the TSS; positive numbers are downstream (3′) from the TSS. The y-axis is the cumulative number of RNA nanopore reads at a given distance of their 3′ end from an annotated polyadenylation site. Values are normalized as a fraction of total counts for the mRNA scaffolds. The dashed lines are at -60 nt and +10 nt from an annotated polyadenylation site [68].

We then performed a statistical measure of confidence for each mRNA scaffold using mapping quality scores [69]. These mapping quality scores for minimap2 range from zero (equal probability that the scaffold aligned to more than one position in the reference genome) to 60 ($\sim$1x10^{-6} probability that the alignment was in the wrong position). Among the 185,434 mRNA scaffolds, 145,661 (78.6 %) had mapping quality

scores of 60. There were 3,487 (1.9 %) mRNA scaffolds with mapping quality scores of zero. Among the 4,262 mRNA scaffolds for unannotated isoforms, 4,048 (95.0 %) had mapping quality scores of 60. There were 27 (0.6 %) mRNA scaffolds with mapping quality scores of zero. By comparison, the untreated reads had 71.2 % of the reads with a mapping quality score of 60.

2.3.6 Use of high confidence mRNA scaffolds to define novel human mRNA isoforms

We proposed that high confidence RNA scaffolds could help define RNA isoforms at sufficient precision to warrant further detailed biological experimentation. The following example illustrates a pipeline we used to characterize an unannotated mRNA isoform.

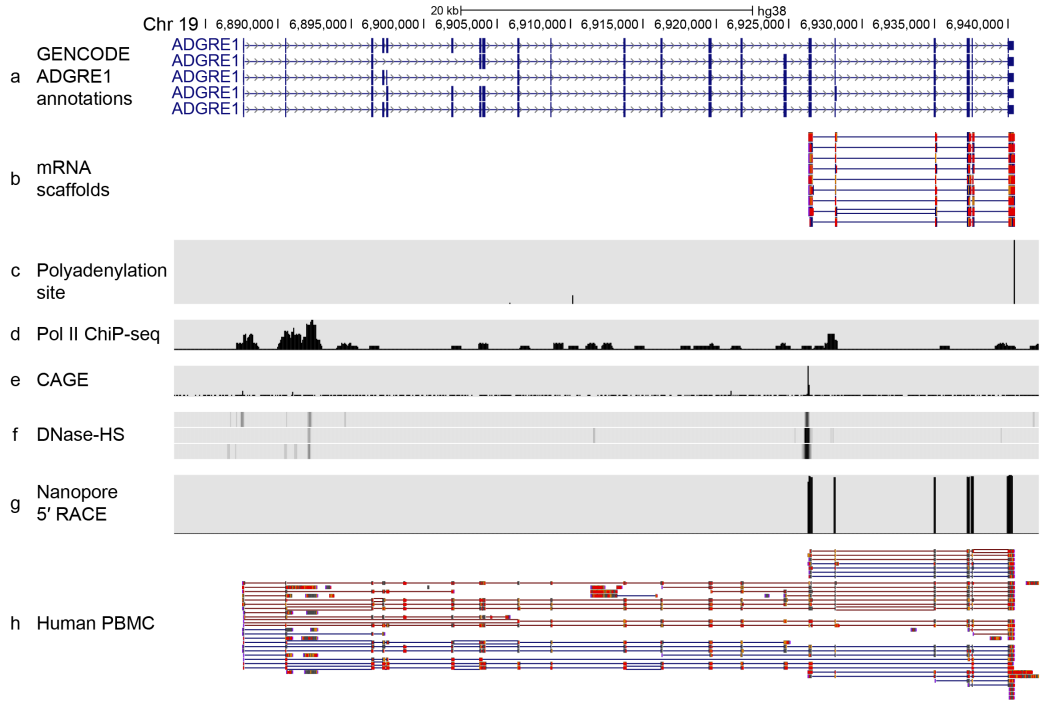

Figure 2.5 mRNA scaffolds predict an unannotated ADGRE1 isoform. (**a**) GENCODE v32 annotations for ADGRE1 mRNA isoforms. (**b**) mRNA scaffolds for an unannotated ADGRE1 isoform. (**c**) Polyadenylation sites annotated by the Poly(A)Site 2.0 atlas [68]. GM12878 specific data from: (**d**) Pol II ChiP-seq sites [70]; (**e**) CAGE sites for the positive strand [71]; (**f**) Three replicate DNase-HS tracks [72]. (**g**) Read coverage from full-length nanopore 5′ RACE sequencing. (**h**) Human peripheral blood mononuclear cell (PBMC) nanopore cDNA reads [73]. Red and blue lines indicated forward and reverse alignments respectively.

Adhesion G protein-coupled receptor E1 (ADGRE1) is a class II adhesion GPCR that is expressed in differentiated cells in the human myeloid lineage [74]. ADGRE1 is often used as a biomarker for macrophages, however, its function is unknown [74]. The five annotated human isoforms (Figure 2.5a) encode proteins with extracellular EGF-like binding domains and 7-transmembrane domains [75].

In our high confidence data set, each of nine individual mRNA scaffolds aligned

to a proposed ∼1,100 nt long unannotated isoform of ADGRE1 (Figure 2.5b). This proposed isoform had a TSS that was internal to the annotated ADGRE1 isoforms. The scaffolds included six previously documented exons, that together encoded an in-frame ORF consistent with a protein composed of transmembrane domains 3-to-7 of the annotated ADGRE1 receptors. The extracellular amino terminus and transmembrane domains 1 and 2 were absent in the isoform predicted by the nine scaffolds.

The expected median identity is 87 % for nanopore RNA sequencing reads [46]. Consequently, additional information would be needed to establish a high confidence isoform based on a single nanopore mRNA scaffold. The steps we used to substantiate the candidate ADGRE1 isoform were:

i) Confirmation that each of the nine scaffolds had a high mapping quality score (60), and that there was a poly(A) site proximal to the 3′ ends (Figure 2.5c) [68];

ii) use of orthogonal GM12878 data [72] to support or refute the proposed isoform. We found that PolII ChiP-seq, CAGE, and DNase-seq data all supported the proposed unannotated ADGRE1 isoform (Figure 2.5d-f respectively);

iii) 5′ RACE for the full-length proposed unannotated ADGRE1 isoform revealed amplicons with identical length and exon composition as the RNA nanopore scaffolds (Figure 2.5g);

iv) confirmation that the proposed ADGRE1 isoform was expressed in human primary tissue and was not an artefact specific to the immortalized GM12878 cell line. To this end, we examined long-read cDNA sequencing data from primary human peripheral blood mononuclear cells (PBMC) [73]. Seven of ∼fifty reads that aligned

to ADGRE1 were identical to the isoform identified by the nanopore mRNA scaffolds (Figure 2.5h).

2.3.7 Discussion

In this study, we describe a strategy that uses individual nanopore reads to define high confidence human mRNA scaffolds. These scaffolds include the 5′ m^7G cap, the 3′ polyadenylation site, and the series of covalently linked nucleotides in between at 87 % identity. A majority of these scaffolds had a mapping quality score of Q60. Most of these (95 %) confirmed isoforms previously annotated in GENCODE v32. There were also 4,262 mRNA scaffolds that were not annotated in GENCODE v32. Most of these scaffolds had undocumented TSS that were internal to known mRNA isoforms.

This strategy includes a new chemo-enzymatic method to specifically adapt 5′ capped RNA strands. The RNA oligonucleotide component of the cap-adapter permitted both identification of the biological 5′ end, and sequencing approximately six nucleotides that were systematically missed using the conventional ONT RNA sequencing protocol. Due to the nature of the cap-adapter linkage, approximately five nucleotides are still missed at the 5′ end of each strand.

The 5′ capping procedure described in this study can attach the oligonucleotide adapter to triphosphate or diphosphate 5′ termini as well as m^7G 5′ termini [76]. Therefore it was conceivable that the poly(A) RNA dataset contained transcripts produced by RNA polymerases other than polII. As a test, we screened for pre-processed ribosomal RNA which bears triphosphate 5′ ends. We found that only 11 transcripts out of

574,091 total cap-adapted reads aligned to ribosomal RNA genes.

In this study, we used GENCODE v32 [51] to identify annotated and unannotated mRNA isoforms in our high confidence mRNA scaffold data. This gene model set is the method of choice for high-throughput RNA data set analysis. RefSeq is an alternative gene model set that is often used for human genetics [77]. The unannotated ADGRE1 exemplar was absent in both gene models. However, in other cases we found annotations in RefSeq that matched isoforms from our nanopore data that were absent in GENCODE. Two examples are Profilin 1 (PFN1) and Voltage Dependent Anion Channel 1 (VDAC1). We recommend comparing proposed unannotated isoforms against both of these gene model sets.

We substantiated a proposed in-frame ADGRE1 isoform using four criteria that could be broadly applied. Two of these criteria were evaluated in a few hours using standard alignment visualization software (e.g. the UCSC Genome Browser) and orthogonal data for the GM12878 transcriptome. A third criterion required an additional experiment using full-length $5'$ RACE amplicons that were sequenced using nanopores. This was completed in approximately three days. The fourth criterion (expression in human primary tissue) was achieved in a few days through consultation with a colleague with expertise in immune cell transcriptomics. Unambiguous proof that this and other proposed mRNA isoforms are translated by the ribosome will require protein evidence.

These high confidence mRNA scaffolds provide unambiguous information that is absent from conventional native RNA nanopore reads. An example is illustrated for Diacylglycerol O-Acyltransferase 1 (DGAT1) (Supplementary Figure 2.9). DGAT1

encodes a multi-pass transmembrane protein that catalyzes the conversion of diacylglycerol and fatty acyl CoA to triacylglycerol. There are six annotated isoforms in GENCODE and two annotated isoforms in RefSeq. Among thirty aligned untreated nanopore reads, two reads had 5′ exons that are not documented by GENCODE nor RefSeq. In neither case was it possible to determine if the 5′ ends represented a mature mRNA transcript or a truncation product. Importantly, one of these presumptive isoforms was also observed among the high confidence mRNA scaffolds. This confirmed connectivity between an m^7G cap, the unannotated first exon, seventeen exons present in known isoforms, and a confirmed poly(A) tail.

We anticipate a number of improvements in the technology going forward, some of which depend on platform improvements by ONT. For example, the percent identity of ONT direct RNA base calls has remained at ∼87 % since the technology was introduced in 2016. By comparison, percent identity for DNA base calls has increased from ∼66 % in 2014 [31] to ∼95 % in 2020 (unpublished UCSC data). Also, ONT native RNA nanopore sequencing throughput could be increased. Currently RNA throughput is typically 1-to-2 million reads per MinION flow cell compared to 5-to-10 million reads per MinION flow cell for cDNA.

Other improvements could be implemented by the research community:

i) The current 5′ end adapter includes a PEG spacer that causes the ONT motor enzyme to slip and thus miss ∼five nucleotides at the 5′ end of each strand. An adapter structure that allowed for sequencing of those five nucleotides would be useful, especially because the N1 and N2 positions are often modified [23];

ii) splice-aware long-read mapping errors can cause short exons to be improperly mapped into introns or sequence clipping issues. Future improvements to splice-aware long-read mapping algorithms will help resolve these;

iii) a fraction (14 %) of the total treated human poly(A) RNA nanopore reads were cap-adapted. Incomplete adaptation of m^7G caps is an unlikely cause for the low yield [61] (Supplementary Figures 2.7 and 2.8). A likely reason for this is RNA truncations caused by RNA strand breaks (either biological or *in vitro*) [46]. A second possible reason for RNA read truncations is nanopore software errors triggered in part by electrical noise [46]. This is suboptimal because the nanopore platform is capable of sequencing very long polynucleotides (over 2 million nt for DNA [78] and up to 13,753 nt for an mRNA scaffold in this study (a golgin B1 transcript). *In vitro* RNA strand breaks could be addressed by reducing the number of RNA processing steps, and software errors could be addressed by eliminating electronic noise that prematurely truncate RNA reads [46]. This is important because a comprehensive picture of human transcriptomes will require at least an order of magnitude more total poly(A) reads than achieved in this study, and a similar increase in high confidence mRNA scaffolds;

iv) nanopore direct RNA sequencing has been used to report nucleotide modifications [40, 44, 46, 58]. For example, base miscalls relative to canonical training data in our ADGRE1 aligned reads (Supplementary Figure 2.10) strongly suggest an unannotated pseudouridine in the stop codon, which is known to cause translation read through [79]. Ionic current signal-based RNA modification detection would increase the utility of high confidence scaffolds.

2.4 Methods

The following methods are part of the same project as described in Chapter 3. Duplicated methods will be referenced to the corresponding section in Chapter 3 where appropriate.

2.4.1 GM12878 cell tissue culture and RNA isolation

GM12878 cells were cultured the same as previously described [46]. Briefly, GM12878 cells (passage 6) were cultured in RPMI medium (Invitrogen #21870076) supplemented with 15 % non heat-inactivated FBS (Lifetech #12483020) and 2 mM L-Glutamax (Lifetech #35050061). Cells were expanded to 9 × T75 flasks (45 mL of medium in each) and centrifuged for 10 min at 100 × g (4 °C), washed in 1/10th volume of PBS (pH 7.4), and combined for homogeneity. The cells were then evenly split between 8 × 15 mL tubes and pelleted at 100 g for 10 min at 4 °C. The cell pellets were then snap frozen in liquid nitrogen and immediately stored at −80 °C.

2.4.2 Isolation of total GM12878 RNA

GM12878 RNA was isolated the same as previously described [46]. Briefly, 4 mL of TRI-Reagent (Invitrogen AM9738) was added to a frozen pellet of 5×10^7 GM12878 cells and vortexed immediately. This sample was incubated at room temperature for 5 min. $CHCl_3$ (chloroform, 200 µl) was added per ml of sample, vortexed, incubated at room temperature for 5 min, vortexed again, and centrifuged for 10 min at 12,000 g (4 °C). The aqueous phase was pooled in a LoBind Eppendorf tube and com-

bined with an equal volume of isopropanol. The tube was mixed, incubated at room temperature for 15 min, and centrifuged for 15 min at 12,000 g (4 °C). The supernatant was removed, the RNA pellet was washed with 750 µl 80 % ethanol and then centrifuged for 5 min at 12,000 g (4 °C). The supernatant was removed. The pellet was air-dried for 10 min, resuspended in nuclease-free water (100 µl final volume), quantified, and either stored at −80 °C or processed further by poly(A) purification.

2.4.3 GM12878 poly(A) RNA purification

Poly(A) RNA was purified from GM12878 total RNA using NEXTflex poly(A) beads (Bioo Scientific, NOVA-512980) following the manufacturer's instructions. We used 50 µl of beads per 100 µg of total RNA. GM12878 poly(A) RNA was aliquoted and stored at −80 °C.

2.4.4 Isolation of total *S. cerevisiae* S288C RNA

See Chapter 3 Method 3.4.3 for details.

2.4.5 *S. cerevisiae* S288C Yeast Poly(A) Isolation

See Chapter 3 Method 3.4.4 for details.

2.4.6 Decapping and recapping of RNA samples

RNA was decapped and recapped with 3′-azido ddGTP (Cu-free click) according to methods previously described [61]. Briefly, decapping of 1.5-6 µg poly(A) RNA is

performed with 300 units yDcps (NEB) in 1X yDcpS reaction buffer (10 mM Bis-Tris-HCl pH 6.5, 1 mM EDTA) in 50 µl total volume for 1 h at 37 °C. The de-capped RNA is purified with the Zymo RNA Clean and Concentrate Kit (Zymo Research #R1013) with the standard protocol (recovers RNA > 17 nt) eluted in 30 µl.

Recapping the 5′ end of 1.5-6 µg of the decapped poly(A) RNA with 3′ azido-ddGTP is performed in 60 µl total volume with 6 µl Vaccinia capping enzyme (NEB M2080S), 6 µl *E. coli* Inorganic Pyrophosphatase (NEB M0361S) and final concentration of 0.5 mM 3′ azido-ddGTP (Trilink #N-4008), and 0.2 mM SAM for 30 min at 37 °C. The RNA is then purified with the Zymo RNA Clean and Concentrate Kit as above.

2.4.7 Synthesis of 3′-DBCO-45mer RNA

The 45-nucleotide 3′-DBCO RNA oligomer (CUCUUCCGAUCUACACUCU UUCCCUACACGACGCUCUUCCGAUCU) was synthesized by coupling the 3′-NH2 RNA oligomer with a DBCO-sulfo-NHS ester (Glen Research, #50-1941). The 3′-NH2 RNA synthesis was performed on an ABI 394 DNA synthesizer (Applied Biosystems) starting with 3′-PT-amino-modifier C3 CPG (Glen Research, #20-2954) and UltraFast RNA TBDMS RNA amidites (Glen Research: Bz-A-CE #10-3003, Ac-C #10-3015, Ac-G-CE #10-3025, and U-CE #10-3030). The oligonucleotide was deprotected according to the manufacturer's protocol using ammonium hydroxide/methylamine and purified using a Glen-Pak RNA purification cartridge (Glen Research, #60-6100) followed by PAGE purification. The purified 3′-NH2 RNA was dissolved in 5 mL of 0.1 M sodium borate (pH 8.3). Then 2.5 mL of a 20 mM solution of DBCO-sulfo-NHS ester in DMSO

was added and stirred for 1.5 h at room temperature. The reaction was then dissolved in 0.1 M TEAB (up to 35 mL) and purified by C8 RP-HPLC (Higgins Analytical) using 0.1 M TEAB and acetonitrile as the mobile phase. The 3′-DBCO RNA oligonucleotide was concentrated and re-purified by PAGE and desalted using a Clarity-RP desalting cartridge (Phenomenex, #8B-S041-HBJ).

2.4.8 Adaptation of recapped Poly(A) RNA

Azido-ddGTP recapped RNA (1-2 µg) was concentrated briefly on a SpeedVac vacuum concentrator (Savant) to reduce the volume to approximately 5-10 µl. Copper-free Click Chemistry reactions were performed in a total volume of 50 µl, containing 25 % v/v PEG 8000 (NEB, #B1004) and 20 % v/v acetonitrile (Sigma-Aldrich, #271004) in 0.1 M sodium acetate buffer, pH 4 (10X, Alfa Aesar, #J60104) and 10 mM EDTA (50x, Invitrogen, #15575-038). Azido-ddGTP recapped RNA and the 3′-DBCO RNA adapter (200 nmol, final concentration of 4 µM) were added and shaken for 2 h at room temperature. Then, acetonitrile was removed by brief concentration on a SpeedVac, and the adapted RNA recovered using RNA Clean & Concentrator (Zymo Research, #R1013) following the protocol to separate large RNA (desired) from small RNA (excess adapter).

2.4.9 Copper-catalyzed click chemistry of RNA adapter

See Chapter 3 Method 3.4.6 for details.

2.4.10 Full-length ADGRE1 cDNA synthesis and sequencing

Full-length cDNA for 5′ RACE sequencing was made using the 5′ RACE Protocol using the Template Switching RT Enzyme Mix (NEB, #M0466) following the manufacturer's instructions. ADGRE1 cDNA was reverse transcribed from total GM12878 RNA using a template switching oligo (TSO) (GCTAATCATTGCAAGCAGTGGTAT CAACGCAGAGTACATrGrGrG) a poly(dT) reverse transcription primer. ADGRE1 cDNA was PCR amplified using a forward primer (CATTGCAAGCAGTGGTATCA AC) and a gene-specific reverse primer (AACAAGGGCAGGAGAAAACAAAATGG TAG) with Q5 Hot Start High-Fidelity 2X Master Mix (NEB, #M0494S). cDNA was prepared for sequencing using the barcoded NBD 104 expansion of the SQK-LSK109 protocol following the manufacturer's recommendations and sequenced using a Flongle flow cell. Ionic current traces were basecalled with MinKnow real-time base calling using the high-accuracy model.

2.4.11 MinION RNA sequencing

Poly(A) RNA (500-775 ng) were prepared for nanopore direct RNA sequencing generally following the ONT SQK-RNA002 kit protocol, including the optional reverse transcription step recommended by ONT. Instead of using Superscript III, as in the ONT protocol, Superscript IV (Thermo Fisher, #18091050) was used for reverse transcription. RNA sequencing on the MinION was performed using ONT R9.4 flow cells and the standard MinKNOW protocol (48 h sequencing script) as recommended by ONT, with

one exception. We collected bulk phase continuous data files for 2 h of sequencing and then restarted the sequencing runs after the two hours of initial sequencing.

2.4.12 Base calling, filtering and alignments

ONT Guppy workflow (version 3.0.3+7e7b7d0 configuration file rna_r9.4.1_70bps_hac.cfg) was used for base calling direct RNA. NanoFilt (version 2.5.0) [80] was used to classify reads as pass if the pre-read average Phred-score threshold was greater than or equal to 7 and fail if less than 7. Porechop (version 0.2.4) was used to identify the 5′ adapter sequence (`https://github.com/rrwick/Porechop`). We used barcode_diff 1 and barcode_threshold 70 or 74 for S288C or GM12878 reads, respectively. Porechop was run twice, once where the adapters were trimmed and once without adapter trimming for use with the TSS filtering pipeline. The barcode search sequence, TCCCTACACGACGCTCTTCCGA, was added to the end of the adapter list in the adapters.py. Reads were then aligned to the appropriate reference, sacCer3 or GRCh38, using minimap2 [81] (version 2.16-r922). Human minimap2 parameters: –secondary=no -ax splice -k14 -uf. Yeast alignment parameters: -ax splice -k10 -G2000 -uf.

2.4.13 Porechop optimization analysis

See Chapter 3 Method 3.4.10 for overall details. The same approach was used in this chapter to optimize the Porechop parameters for GM12878 data (Supplementary Figure 2.6). We found a barcode threshold of 74 to be optimal for GM12878 poly(A) RNA.

2.4.14 TSS filtering pipeline

The TSS filtering pipeline is available on github (`https://github.com/mitenjain/dRNA_capping_analysis`).

Cap-adapted reads identified by Porechop were mapped to the human genome (GRCh38.p3.genome.fa) using Minimap2 (version 2-2.9) with the following parameters: –secondary=no -ax splice -k14 -uf. Remaining secondary alignments were removed (using samflag -F 2048). The number of soft and hard clipped bases were used to filter false positive cap-adapted reads, and miss-mapped reads. Reads where the adapter sequence remained that contained 23 or fewer soft or hard clipped bases were removed (Porechop false positives). Reads where the adapter sequence was trimmed and contained 15 or more soft or hard clipped bases at the $5'$ end of the read were removed (miss-mapping). Together, these two filtering steps removed approximately half of the cap-adapted reads identified by Porechop. In order to identify only unannotated alternative TSS, reads that had $5'$ ends aligned within 300 bases from both upstream and downstream GENCODE v32 annotated TSSs were removed. The final filtering step was to remove reads that did not align to a GENCODE annotation.

2.4.15 General Data manipulation

General sequencing data manipulations were done using bedtools [82] and samtools [83].

2.5 Supplementary Information

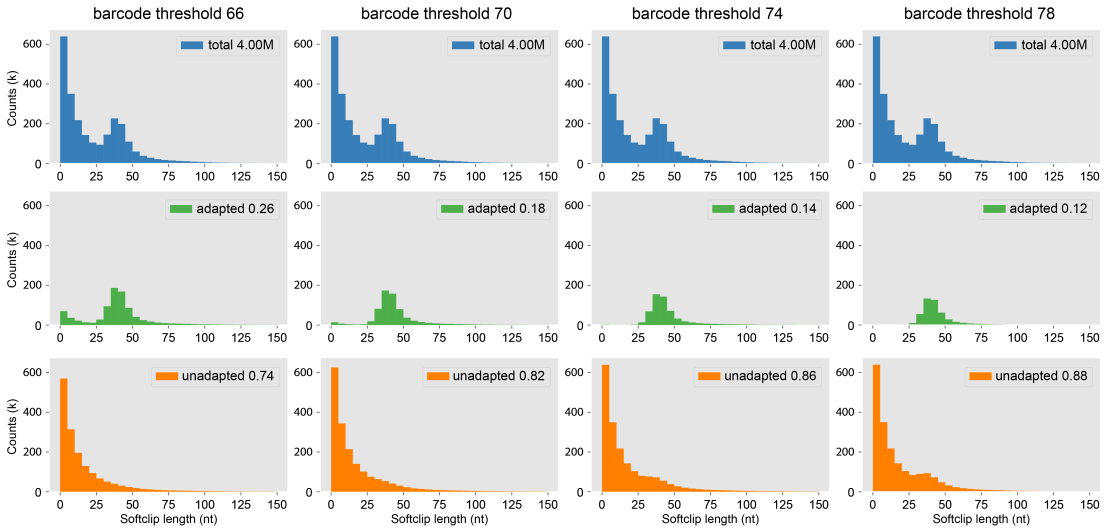

Figure 2.6 Optimization of Porechop parameters using GM12878 data. The barcode threshold is the minimum specificity required for determining if the 5′ end of the nanopore read matched the adapter sequence. The adapter sequence does not exist in the genome, and thus will be soft clipped from the 5′ end of the alignment. We selected a barcode threshold which first, minimized the cap-adapted reads with zero soft clipped bases and second minimized the unadapted reads with ∼40 nt of 5′ end soft clipped bases. The x-axis of each plot represents the number of soft clipped bases from the 5′ end of each nanopore read. The y-axis of each plot represents the number of reads (in thousands) for a given soft clip length. The plots in each column represent data analyzed for a given barcode threshold value. The top row (blue) are the soft and hard clip lengths for 4 million reads. The middle row (green) shows the 5′ end soft clip lengths for cap-adapted reads, as identified by Porechop for a given barcode threshold. The bottom row (orange) are the reads where Porechop did not identify the adapter sequence. The proportion of reads represented in each plot is denoted in the upper right hand corner. We found 74 was the optimal barcode threshold.

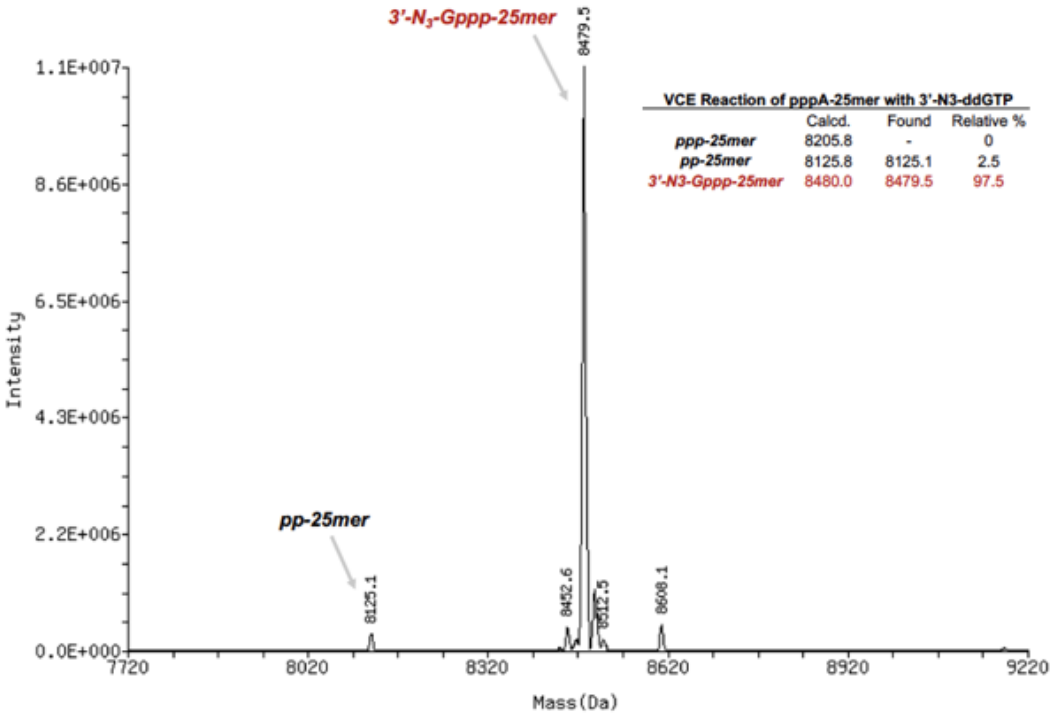

Figure 2.7 Vaccinia Capping Enzyme (VCE) caps RNA with 3′-azido-ddGTP. Deconvoluted ESI-MS spectra of a synthetic 25-nucleotide 5′-triphosphate RNA oligomer (ppp-25mer) capped with 3′-azido-ddGTP. Liquid Chromatography tandem Mass Spectrometry (LC-MS/MS) was performed on a Vanquish Horizon UHPLC System coupled with a Thermo Q-Exactive Plus mass spectrometer operating under negative electrospray ionization mode (–ESI). MS data acquisition was performed in the scan mode. ESI-MS raw data were deconvoluted using Promass HR (Novatia). The composition of each peak was determined by comparison with calculated average atomic mass. The results show nearly complete oligomer capping after 60 min incubation with VCE (see Methods for capping conditions).

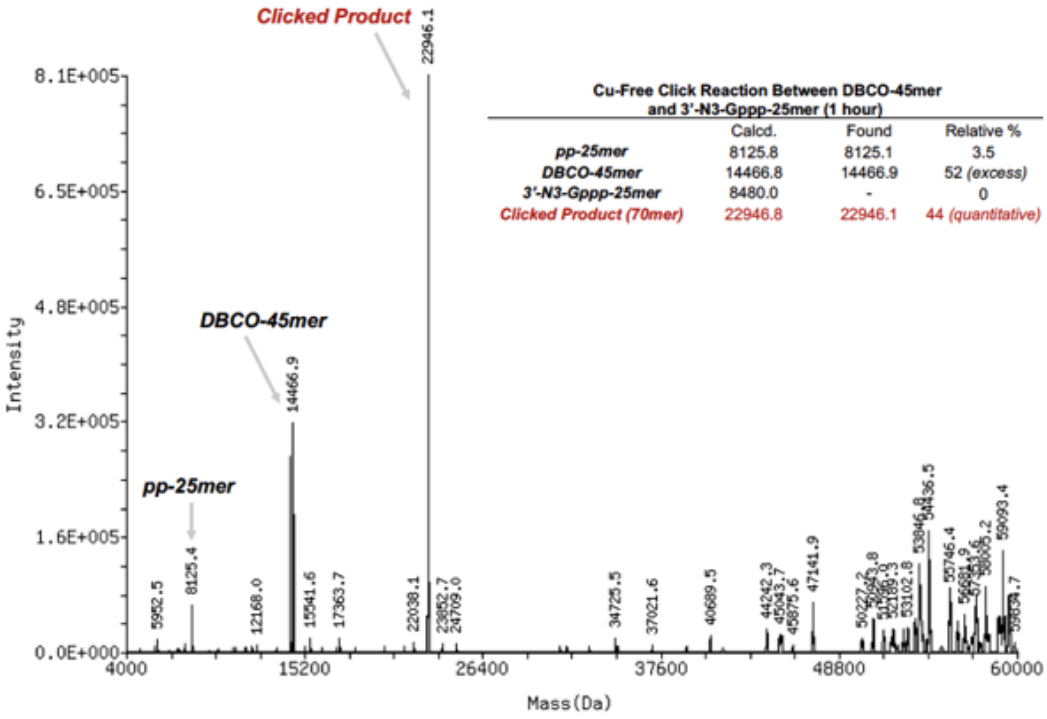

Figure 2.8 Deconvoluted ESI-MS spectra of the synthetic 25-nucleotide azido-ddGTP capped RNA oligomer from Supplementary Figure 2.7 coupled with the 3′-DBCO RNA adapter (DBCO-45mer). LC-MS/MS and spectral deconvolution were performed as described in Supplementary Figure 2.7. The composition of each peak was determined by comparison with calculated average atomic mass. The results show that after 60 min the azido-ddGTP capped RNA is entirely consumed forming the desired adapted RNA ("clicked" product). Excess of unreacted 3′-DBCO adapter and some remaining 5′-diphosphate RNA (pp-25mer) were also detected.

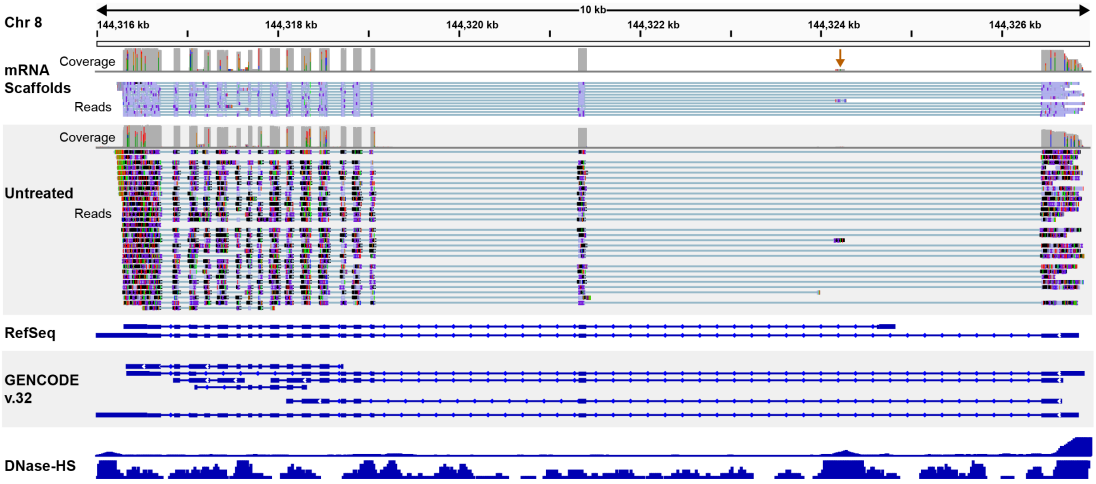

Figure 2.9 Evidence for an unannotated Diacylglycerol O-Acyltransferase 1 (DGAT1) isoform is supported by a single high-confidence mRNA scaffold. The row entitled mRNA scaffolds includes 12 aligned reads in the 3′-to-5′ orientation. Most of these aligned to a GENCODE v.32 annotated isoform. One of these mRNA scaffolds (orange arrow) corresponds to an unannotated first exon of a proposed unannotated DGAT1 isoform. This unannotated isoform is also observed among the untreated reads. However, untreated reads lack strong evidence of a mature mRNA 5′ end because they are not cap-adapted. The first exon of the proposed unannotated isoform is consistent with open chromatin revealed by the DNase-HS data.

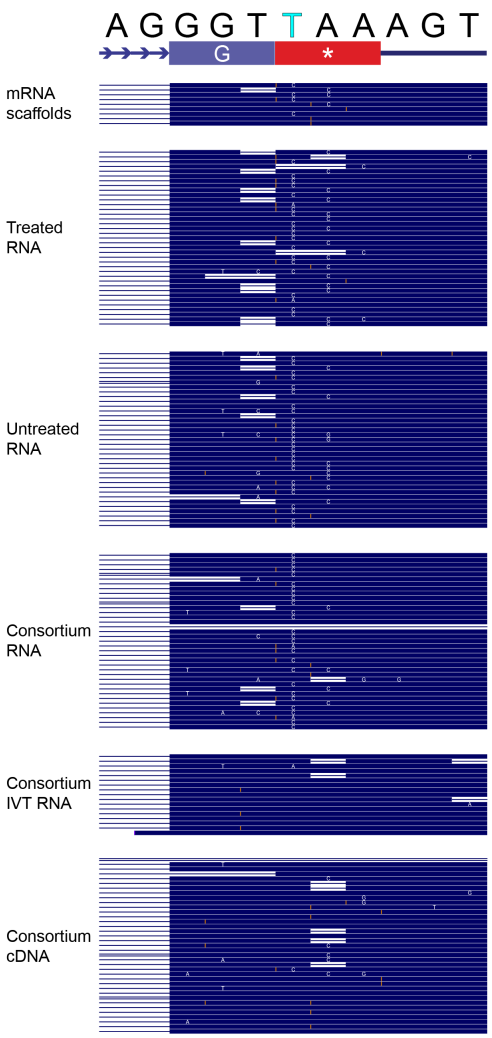

Figure 2.10 Nanopore evidence for a pseudouridine in the stop codon of ADGRE1. The top row is HG38 chr19:6,940,022-6,940,032 which corresponds to eleven nucleotides in the last exon of ADGRE1. The G (blue background) in the second row is a glycine of the ADGRE1 gene product. The * (red background) is the canonical stop codon for ADGRE1. A pseudouridine at the first nucleotide of that stop codon can promote ribosome read through in other genes [79]. Rows entitled mRNA scaffolds, treated RNA, and untreated RNA represent nanopore reads from this study. Consortium RNA represents nanopore reads of biological RNA from a prior study [46]. Consortium IVT RNA represents nanopore reads for *in vitro* transcripts that are composed exclusively of canonical nucleotides [46]. Consortium cDNA represents nanopore reads of amplicons derived from the consortium GM12878 poly(A) RNA [46]. The dark blue pattern is where nanopore base calls match the reference sequence. White horizontal lines are nucleotide deletions in the nanopore reads. Orange vertical lines represent nucleotide insertions in the nanopore reads. White letters represent base calls that disagree with the reference sequence. The vertical line of C miscalls at the second T of the reference sequence (aqua) is characteristic of a pseudouridine [44]. The absence of C miscalls for the Consortium IVT RNA data confirm that the C miscalls are not due to sequence context errors. The Consortium cDNA alignments confirm that the C miscalls are not due to base substitutions.

Table 2.2 Native RNA nanopore sequencing statistics for GM12878 poly(A) RNA.

Sample	Pass reads	N50	Fraction adapted
Untreated 1	1,332,182	1,572	0.0000
Untreated 2	2,497,808	1,691	0.0000
Untreated Pooled	**3,829,990**	**1,615**	**0.0000**
Treated 1	640,221	1,334	0.1106
Treated 2	919,701	1,131	0.1021
Treated 3	797,451	1,212	0.1596
Treated 4	1,706,866	1,189	0.1653
Treated Pooled	**4,064,239**	**1,207**	**0.1413**

Table 2.3 RNA types identified from untreated, unadapted, and cap-adapted reads

RNA type	untreated	unadapted	cap-adapted	RNA type	untreated	unadapted	cap-adapted
protein_coding	988063	1954644	278718	rRNA	157	91	2
Mt_tRNA	59686	110205	6	snoRNA	143	104	8
processed_pseudogene	26917	59970	7014	TEC	128	249	35
Mt_rRNA	26325	39885	154	misc_RNA	91	69	16
lncRNA	22023	38564	4642	TR_C_gene	64	252	2
unknown	20129	35120	2823	IG_J_gene	52	433	12
transcribed_processed_pseudogene	4011	9612	1447	unitary_pseudogene	14	25	1
transcribed_unprocessed_pseudogene	3035	6087	631	rRNA_pseudogene	5	3	0
IG_V_gene	1540	1818	1087	translated_processed_pseudogene	5	7	1
unprocessed_pseudogene	670	1463	189	TR_V_gene	4	6	1
IG_C_gene	554	6050	26	scRNA	3	2	0
polymorphic_pseudogene	317	905	426	TR_V_pseudogene	3	2	1
snRNA	271	93	16	translated_unprocessed_pseudogene	3	8	6
miRNA	252	549	37	IG_J_pseudogene	2	15	1
IG_V_pseudogene	227	1027	112	TR_D_gene	2	0	0
TR_J_gene	175	117	81	TR_J_pseudogene	2	0	0
transcribed_unitary_pseudogene	172	395	31	scaRNA	1	2	0

2.6 Chapter 2 Acknowledgments

Mark Diekhans processed and helped analyze ENCODE CAGE data, and coordinated discussions with GENCODE staff. Guillermo Chacaltana generated the nanopore full-length ADGRE1 cDNA data. Chris Vollmers provided the long-read PBMC cDNA data. Jonathan Mudge and Irwin Jungreis gave advice on GENCODE annotations. Kristof Tigyi cultured the GM12878 cells. Niki Thomas and Robin Abu-Shumays edited drafts of the manuscript. We acknowledge the ENCODE Consortium and the following encode data producers for their assistance: J. Michael Cherry, Stan-

ford; Gregory Crawford, Duke; and John Stamatoyannopoulos, University of Washington. We downloaded call sets from the ENCODE portal [72] (`https://www.encodeproject.org/`) with the following identifiers: ENCFF743ULW, ENCFF093VXI, ENCFF066VBS, and ENCFF969DFL, ENCSR000EMT, ENCSR000EJD, ENCSR000EMT, ENCSR000CKA, ENCSR000CJZ, ENCSR000CJY.

Chapter 3

Optimizing 5′ cap-adaptation using *Saccharomyces cerevisiae* poly(A) RNA

3.1 Introduction

In Chapter 2, we presented a strategy for documenting individual high confidence human mRNA scaffolds based on nanopore sequencing. This strategy was based in part on replacing the biological RNA m^7G cap with a synthetic cap. In this chapter, we describe in detail how we optimized cap-adaptation to improve throughput of full-length, intact, biological RNA nanopore reads. As stated previously, we used *S. cerevisiae* poly(A) RNA in these experiments because the m^7G cap is identical to the

human m^7G cap, and because most yeast genes encode only one RNA isoform [50].

3.2 Results

3.2.1 5′ cap-adaption using Copper-catalyzed click chemistry

Our initial strategy for adapting the 5′ end of mRNA strands is outlined in Figure 3.1. In some respects it is identical to Figure 2.1. Importantly, however, the cap analog is a 3′ propargyl GTP (Figure 3.1) vs a 3′-azido-ddGTP (Figure 2.1). Additionally, the cap-adapter bears an azide group on the 3′ end of the RNA oligonucleotide (Figure 3.1) in place of a dibenzocyclooctyne-amine (DBCO) (Figure 2.1). These two components are required for a Copper-catalyzed click reaction.

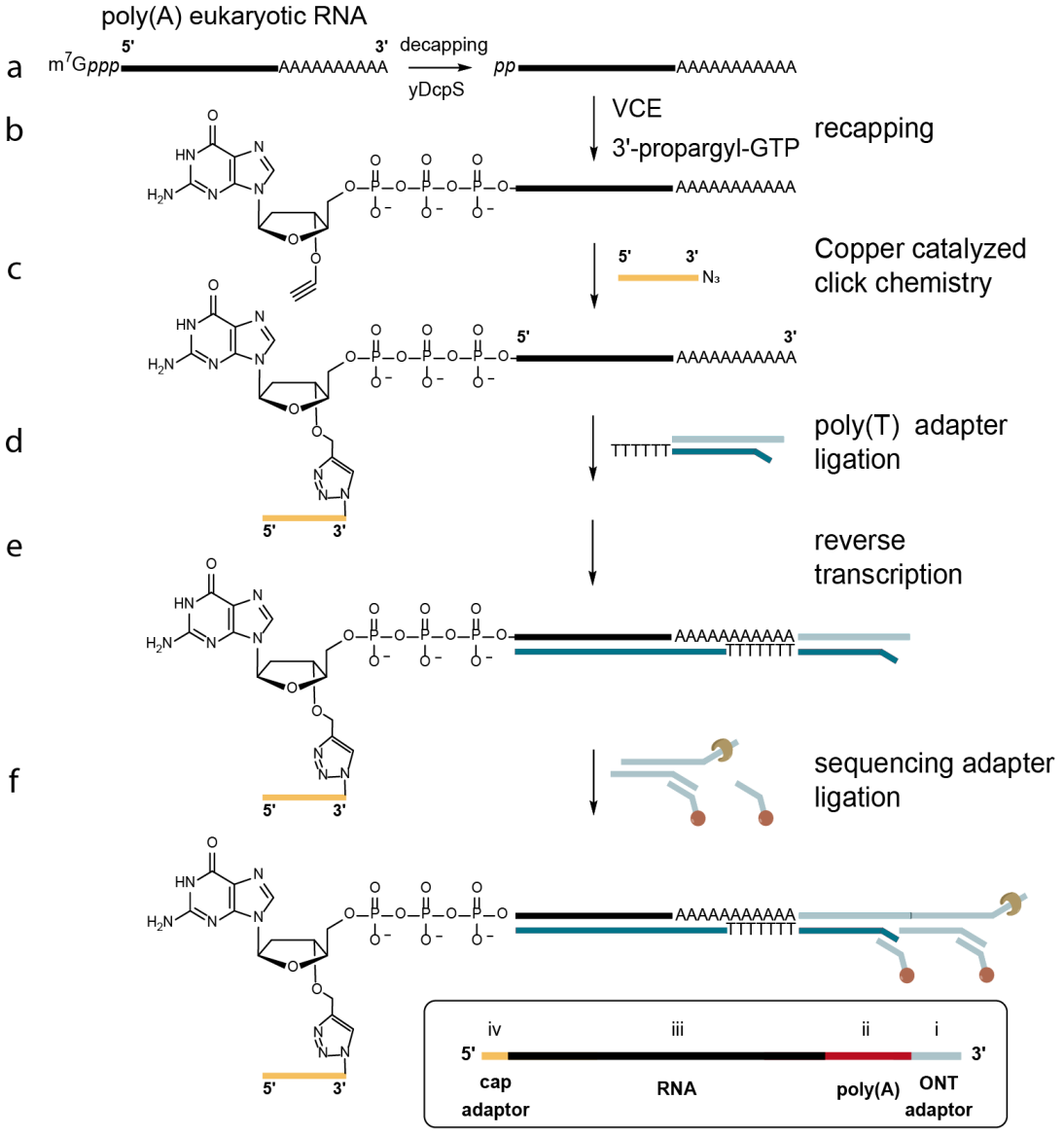

Figure 3.1 Schematic of the 5′ cap-adaption workflow using Copper-catalyzed click chemistry. Steps a, d, e, and f are identical to steps described in Figure 2.1. Step b differs from the analogous step in Figure 2.1 because VCE recapping uses a 3′ -propargyl-GTP. Step c also differs from Figure 2.1 because the 3′ azide covalent attachment is Copper-catalyzed.

We tested this Copper-catalyzed strategy by adapting the triphosphorylated

5′ end of a Gaussia princeps Luciferase (GLuc) IVT transcript. GLuc is a 801 nt RNA that approximates an average eukaryotic mRNA transcript. When we sequenced this preparation on a MinION flow cell, we acquired 558,133 reads with Phred-scale quality value threshold of 7 or greater (see Methods). Among these reads, we observed a fraction with an ionic current signature on the 5' end that was absent from the control (Figure 3.2). To determine if that signature corresponded to covalently attached adapters, we used Porechop to detect the adapter sequence. Porechop (`https://github.com/rrwick/Porechop`) is an open source software package designed to read nanopore barcodes (see Methods). We found that 41,581 of 558,133 reads (7.45 %) had the 5′ cap-adapter 'barcode' sequence. By comparison, zero out of 810,225 reads were identified with the cap-adapter sequence in the control sample.

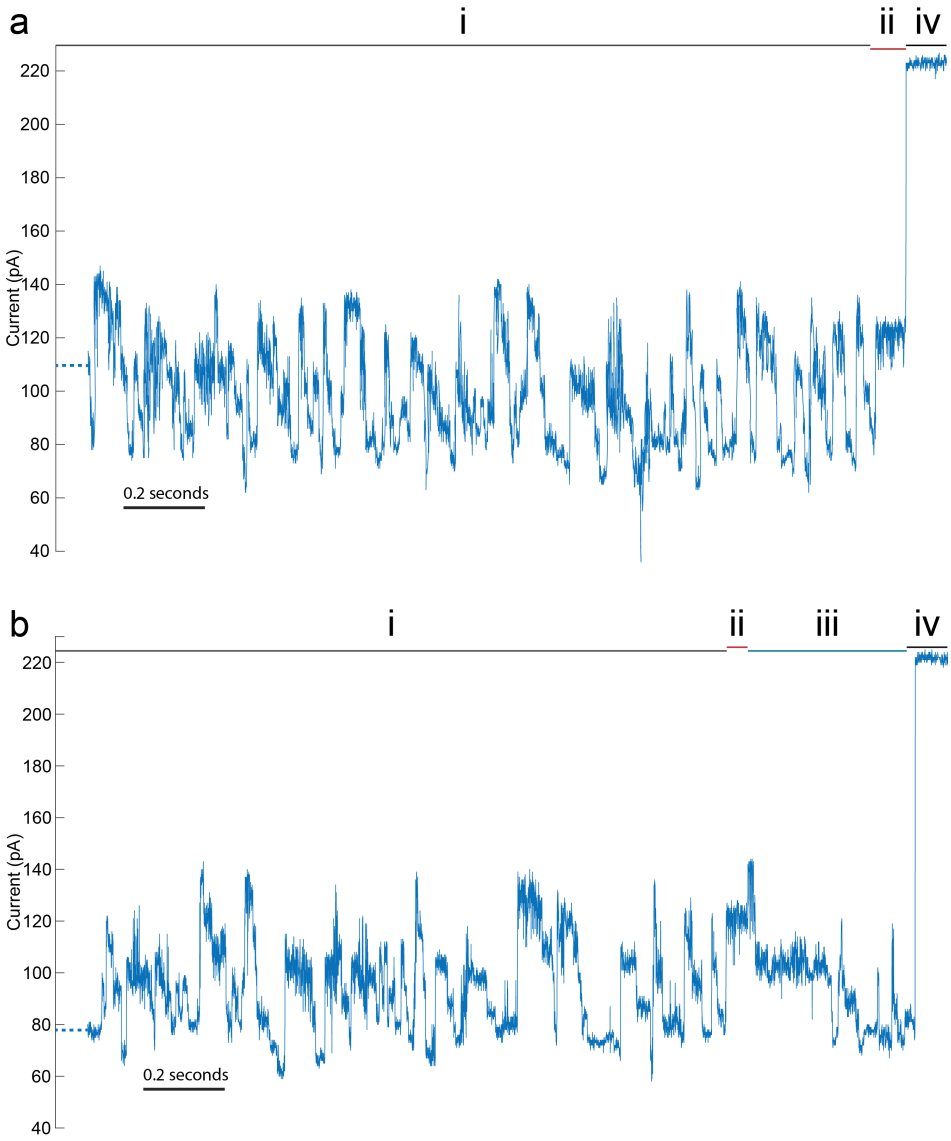

Figure 3.2 Representative GLuc ionic current traces with or without Copper-catalyzed click cap-adaptation. (**a**) A two second window of an ionic current trace associated with the 5′ end of GLuc without a cap-adapter. This trace is a segment of the full-length strand translocation event. (i) represents the nucleotides translocating until the terminal state. (ii) Is the last state identified in the trace before the molecule exits the nanopore. (iv) is the open channel current. (**b**) A two second window of an ionic current trace associated with the 5′ end of GLuc with a cap-adapter. (i) and (ii) are the same as in panel a. (iii) is the ionic current associated with the cap-adapter. (iv) is the same as in panel a. The 5′ end of both traces is on the right.

We aligned the cap-adapted reads to a GLuc reference (Figure 3.3). Predictably, the cap-adapted reads had more sequence coverage proximal to the 5′ end of GLuc (Figure 3.3b) when compared to control (Figure 3.3a). There were numerous skipped basecalls in the alignments at the 5' end. One possible explanation was that the adapter sequence was too short. As a test, we increased the RNA oligonucleotide adapter length from 33 nt to 45 nt. This resulted in an increase from 7.45 % to 16.65 % of cap-adapted reads, however the frequency of inserts was unchanged.

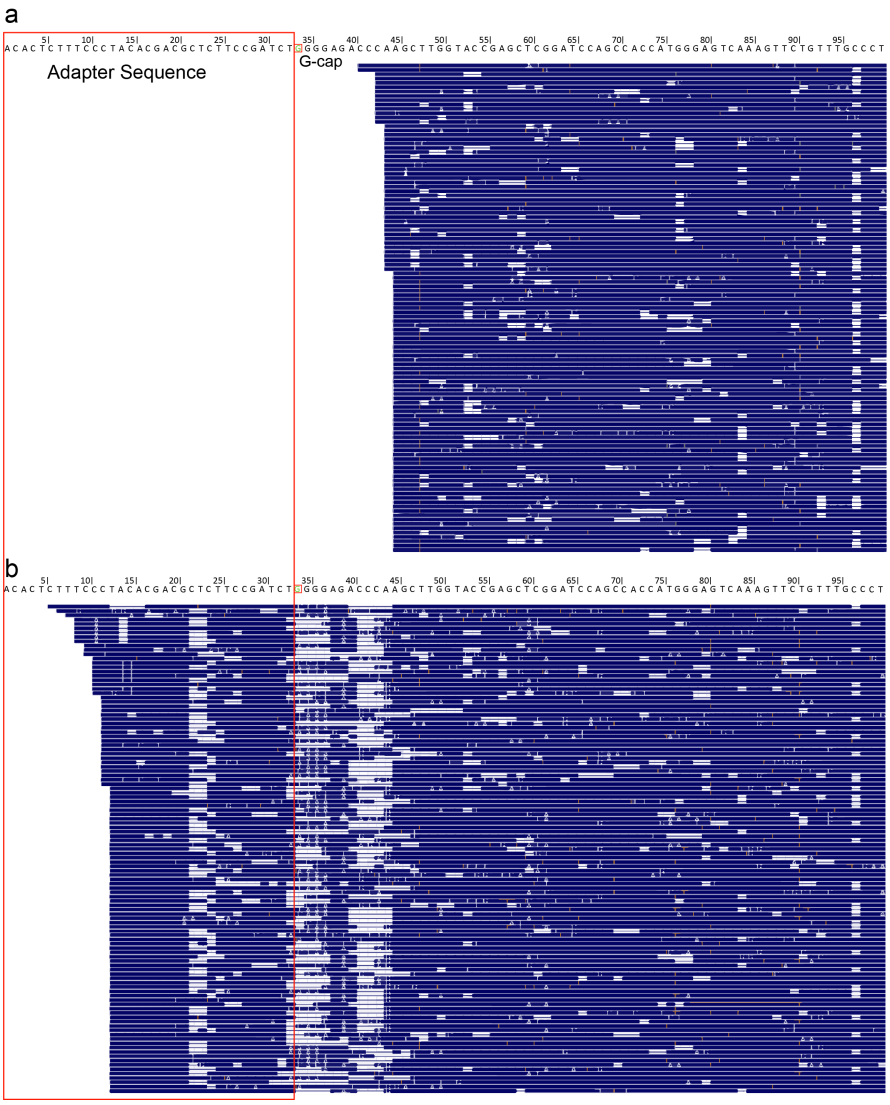

Figure 3.3 Sequencing of the GLuc RNA 5' end is improved by cap-adaptation. (**a**) Nanopore GLuc RNA reads that were not cap-adapted. (**b**) Nanopore GLuc RNA reads that were cap-adapted. Horizontal stripes are individual nanopore strand sequences. Blue represents a match to the reference sequence (top). White represents gaps in the alignment for each nanopore read. Orange vertical bars represent insertions in the alignments. Letters represent basecalls different from the reference. The red box denotes the 33 nt cap-adapter sequence. The green G in the red box is the 3' G of the cap-adapter that replaced the terminal m^7G of the GLuc RNA substrate.

3.2.2 Testing the Copper-catalyzed click cap-adaptation strategy using *S. cerevisiae* S288C poly(A) RNA

Having tested and improved cap-adaption, we next tested this strategy using biological poly(A) RNA from *S. cerevisiae*. As stated previously, *S. cerevisiae* m^7G cap is identical to the human m^7G cap, and most *S. cerevisiae* genes encode only one isoform [50]. We isolated poly(A) RNA from *S. cerevisiae* S288C total RNA and performed the cap-adaption treatment process with five replicates, called 'treated' in the text that follows (see Methods). Each treated sample was nanopore sequenced in addition to four control samples that did not undergo the cap-adaption process (called 'untreated' in the text that follows). We acquired 10.3 million treated reads (Table 3.1). The per sample throughput for the treated samples was comparable to the untreated samples.

Table 3.1 *S. cerevisiae* sequencing statistics.

Sample	**Click reaction**	**pass reads**	**read N50**	**Fraction adapted**
Untreated 1	None	1,391,596	531	0.0000
Untreated 2	None	1,755,395	1,092	0.0000
Untreated 3	None	3,22,192	1,034	0.0000
Untreated 4	None	861,210	947	0.0000
Untreated Pooled	**None**	**7,230,375**	**957**	**0.0000**
Treated 1	Copper-catalyzed	2,576,122	798	0.2382
Treated 2	Copper-catalyzed	3,933,777	676	0.0889
Treated 3	Copper-catalyzed	1,848,281	722	0.1451
Treated 4	Copper-catalyzed	1,382,164	575	0.0696
Treated 5	Copper-catalyzed	600,225	435	0.0618
Treated Pooled	**Copper-catalyzed**	**10,340,569**	**692**	**0.1340**
Treated 1	Copper-free	1,128,595	737	0.3347
Teared 2	Copper-free	3,799,042	755	0.4138
Treated 3	Copper-free	1,275,112	715	0.3407
Treated Pooled	**Copper-free**	**6,202,749**	**745**	**0.3844**

The default Porechop parameters are not optimal for a complex biological poly(A) RNA sample, because the sequence adjacent to the cap-adapter is diverse. To account for this, we optimized the Porechop parameters to reduce false positives (see

Methods). We found that a barcode threshold of 70 was optimal for specifically detecting the cap-adapter on the 5′ ends of *S. cerevisiae* RNA (Figure 3.4).

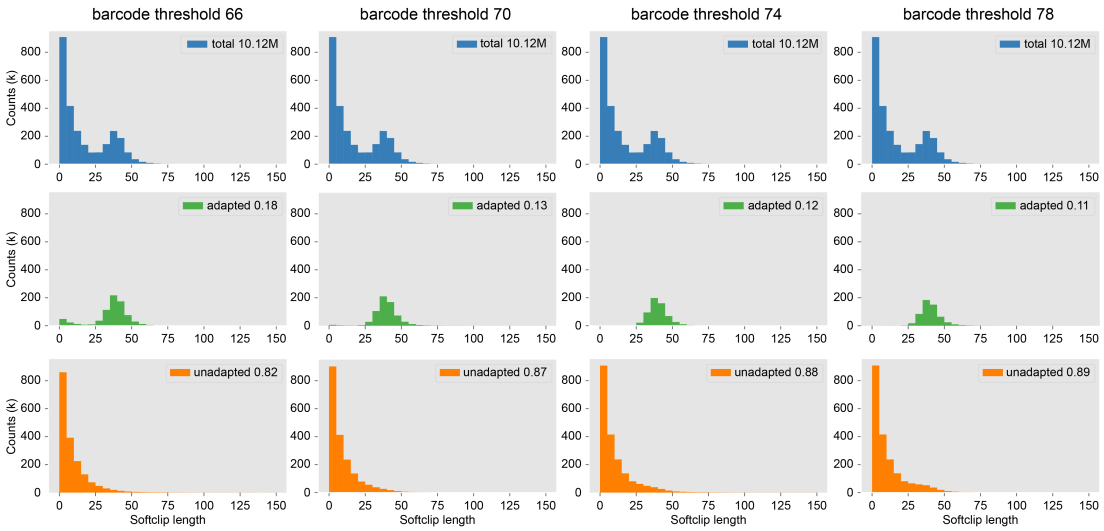

Figure 3.4 Optimization of Porechop parameters using Copper-catalyzed click using *S. cerevisiae data*. The barcode threshold is the minimum specificity required for determining if the 5′ end of the nanopore read matched the adapter sequence. The adapter sequence does not exist in the genome, and thus will be soft clipped from the 5′ end of the alignment. The x-axis of each plot represents the number of soft clipped bases from the 5′ end of each nanopore read. The y-axis of each plot represents the number of reads (in thousands) for a given soft clip length. The plots in each column represent data analyzed for a given barcode threshold value. The plots in the top row (blue) are the soft clipped lengths for all 10 million reads. The plots in middle row (green) are the 5′ end soft clipped lengths for cap-adapted reads, as identified by Porechop for a given barcode threshold. The plots in the bottom row (orange) are the 5′ end soft clipped lengths for reads that Porechop did not identify the adapter sequence for a given barcode threshold. Each plot denotes the proportion of the total reads that were analyzed in the plot. We selected a barcode threshold which first minimized the cap-adapted reads with zero soft clipped bases and second minimized the unadapted reads with 40 nt of 5′ end soft clipped bases. We found 70 was the optimal barcode threshold.

Having optimized the Porechop conditions, we found that 13.4 % of the treated reads were cap-adapted (Table 3.1). This was comparable to the proportion of cap-adapted reads we found for GLuc (16.65 %). However, the treated reads N50 [84] value

(692 nt) was lower than the untreated reads N50 (957 nt). This can be seen by the high density of shorter cap-adapted reads compared to the untreated reads (Figure 3.5a).

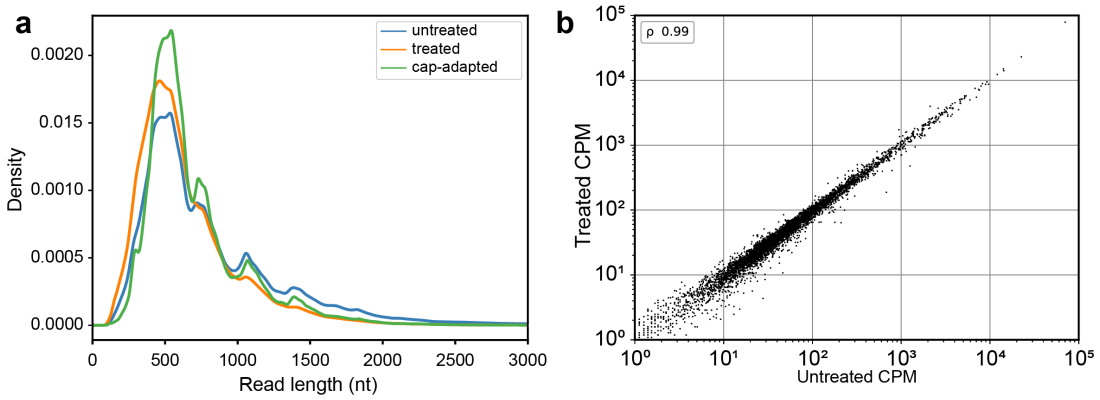

Figure 3.5 Comparison among *S. cerevisiae* S288C untreated, Copper-catalyzed click treated, and Copper-catalyzed click cap-adapted nanopore reads. (**a**) Read length distribution for pooled untreated (blue), treated (orange), and cap-adapted (green) reads. The x-axis is the read length in nucleotides. The y-axis is the density of the Gaussian distributions of read lengths. (**b**) Number of transcripts per gene for untreated vs treated samples. Axes are counts per million (CPM) plotted on a $\log_{10}$ scale. The Spearman's r (ρ) was 0.99.

Based on these N50 values, we were concerned that the treatment process adversely affected RNA transcript recovery. To test this, we compared the number of transcript copies per gene for treated and untreated reads. The Spearman correlation (ρ) was 0.99, indicating a very strong correlation between the treated and untreated samples (Figure 3.5b).

Although the transcript gene capture is largely unaffected, there was still a read length bias. The treated and cap-adapted reads were enriched for shorter reads when compared to the untreated reads (Figure 3.5a). A possible explanation for the lower N50 value is that the adaptation process caused RNA strand breaks. Even if the

RNA cap was successfully adapted, an RNA strand break will separate the cap-adapted end from the ONT adapted 3′ end.

To test if there was RNA degradation, we performed an RNA integrity number (RIN) analysis [65] on *S. cerevisiae* total RNA after each step of the cap-adaptation process (see Methods). The Copper-catalyzed click reaction had the largest impact with a RIN score of 6.7 (see also Table 2.1), which was unacceptably low.

The most obvious reagent in the Copper-catalyzed click reaction that could cause RNA degradation was the Copper (II) [85]. Divalent metals are known to cause RNA degradation by hydrolyzing the phosphodiester backbone [86]. Efforts have been made to keep the RNA strands intact with Copper-catalyzed click reactions, however RNA degradation has not been eliminated [87].

As an alternative we switched to a Copper-free click reaction [88, 89]. This involved changing the cap analogue to a 3′ azido ddGTP and adding a 3′ Dibenzocyclooctyne-amine (DBCO) to the 45 nt RNA oligonucleotide (as described in Figure 3.1). We found that this chemistry eliminated the RNA degradation observed using the Copper-catalyzed reaction (see Table 2.1).

3.2.3 Testing the Copper-free click reaction for poly(A) RNA nanopore sequencing

We were concerned that the bulkier DBCO group might affect RNA translocation through nanopores. To test this, we cap-adapted the GLuc RNA control using the Copper-free chemistry. We acquired 312,530 reads from one treated sample, which

was normal for IVT RNA samples. Therefore DBCO appeared to have no detrimental effect. In fact, there was a 2-fold improvement in yield of cap-adapted reads (31.43 %) relative to Copper-catalyzed method (16.65 %).

We wanted to ensure that the Copper-free chemistry was suitable for eukaryotic poly(A) RNA analysis using nanopores. Therefore, we acquired 6.2 million *S. cerevisiae* poly(A) RNA reads from three technical replicates and performed the following:

i) We identified 2,384,308 (38.44 %) cap-adapted reads using Porechop parameters optimized for *S. cerevisiae* (Figure 3.6 and Table 3.1). This was a 2.87 fold increase in cap-adapted reads compared to the Copper-catalyzed alternative (13.4 %);

ii) We determined that the N50 value for *S. cerevisiae* cap-adapted reads was higher for Copper-free chemistry (744 nt) than for Copper-catalyzed chemistry (692 nt) (Figure 3.7a);

iii) We determined that the transcript per gene counts for treated samples and untreated samples strongly correlated (ρ 0.90) (Figure 3.7b).

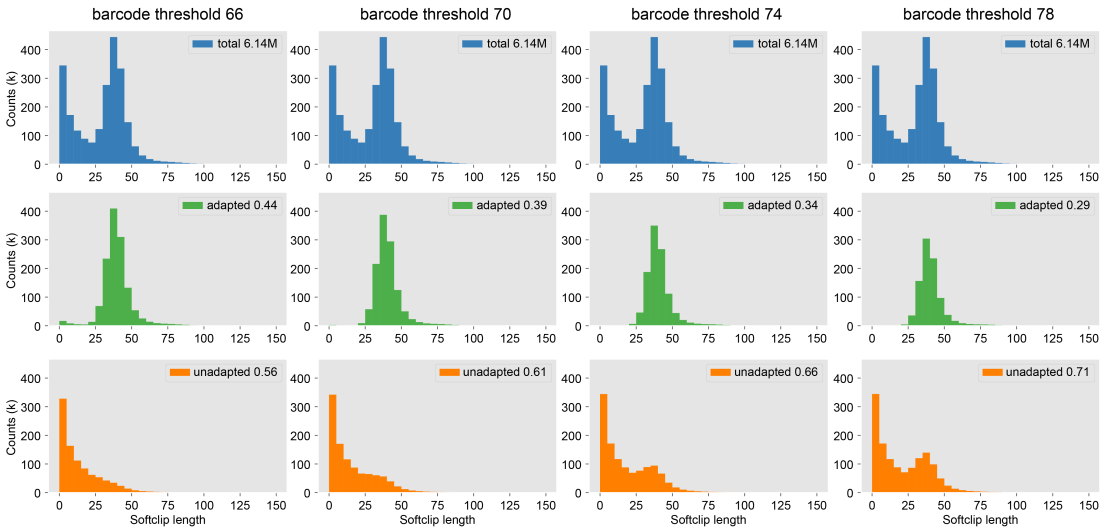

Figure 3.6 Optimization of Porechop parameters using Copper-free click using *S. cerevisiae data*. The barcode threshold is the minimum specificity required for determining if the 5′ end of the nanopore read matched the adapter sequence. The adapter sequence does not exist in the genome, and thus will be soft clipped from the 5′ end of the alignment. The x-axis of each plot represents the number of soft clipped bases from the 5′ end of each nanopore read. The y-axis of each plot represents the number of reads (in thousands) for a given soft clip length. The plots in each column represent data analyzed for a given barcode threshold value. The plots in the top row (blue) are the soft clipped lengths for all 6 million reads. The plots in middle row (green) are the 5′ end soft clipped lengths for cap-adapted reads, as identified by Porechop for a given barcode threshold. The plots in the bottom row (orange) are the 5′ end soft clipped lengths for reads that Porechop did not identify the adapter sequence for a given barcode threshold. Each plot denotes the proportion of the total reads that were analyzed in the plot. We selected a barcode threshold which first minimized the cap-adapted reads with zero soft clipped bases and second minimized the unadapted reads with 40 nt of 5′ end soft clipped bases. We found 70 was the optimal barcode threshold.

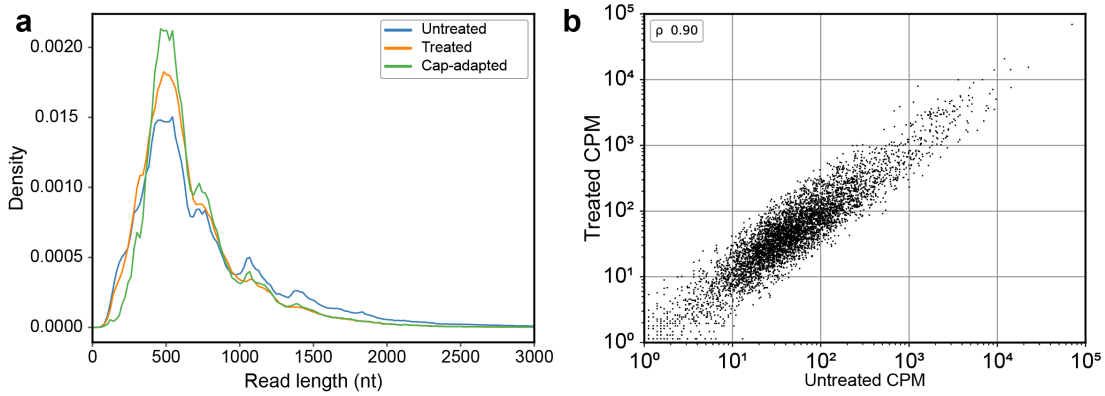

Figure 3.7 Comparison among *S. cerevisiae* S288C untreated, Copper-free treated, and Copper-free cap-adapted reads. (**a**) Read length distribution for pooled untreated (blue), treated (orange), and cap-adapted (green) reads. The x-axis is the read length in nucleotides. The y-axis is the density of the Gaussian distributions of read lengths. (**b**) Number of transcripts per gene for untreated vs copper-free-free treated samples. Axes are counts per million (CPM) plotted on a $\log_{10}$ scale. The Spearman's r (ρ) was 0.90.

3.3 Conclusions

Together, these results demonstrate that Copper-free click chemistry adaptation of poly(A) RNA is substantially better than the Copper-catalyzed alternative as measured by throughput, RNA integrity, and 5' cap-adaptation efficiency. This was true for model synthetic RNA, and for a mixed population of eukaryotic poly(A) RNA.

3.4 Methods

The following methods are part of the same project as described in Chapter 2. Duplicated methods will be referenced to the corresponding section in Chapter 2 where appropriate.

3.4.1 Synthesis of 3′ DBCO-45mer RNA.

See Chapter 2 Method 2.4.7 for details.

3.4.2 Synthesis of 3′-Azido RNA adapter

The 45-nucleotide 3′-azido RNA oligomer (CUCUUCCGAUCUACACUCUU UCCCUACACGACGCUCUUCCGAUCU) was synthesized on an ABI 394 DNA synthesizer (Applied Biosystems) starting with 3′-alkyne modifier Serinol CPG (BaseClick, #BCA-02) and UltraFast RNA TBDMS RNA amidites (Glen Research: Bz-A-CE #10-3003, Ac-C #10-3015, Ac-G-CE #10-3025, and U-CE #10-3030). The oligonucleotide was deprotected according to the manufacturer's protocol using ammonium hydroxide/methylamine and purified using a Glen-Pak RNA purification cartridge (Glen Research, #60-6100) followed by PAGE purification. The oligonucleotide was further purified by PAGE followed by a desalting step on RP-HPLC (C-8 Higgins Analytical) using 0.1 M TEAB and acetonitrile as the mobile phase. The purified oligonucleotide was converted to 3′-azido in a total volume of 889.2 µl, containing 25 % v/v DMSO in 0.2 M triethylammonium acetate buffer, pH 7 as follows (unless other specified, final concentrations are given): 100 µM oligomer, 20 mM N3-PEG1-N3 (BroadPharm, #BP-20908) and 500 µM ascorbic acid were combined and the solution briefly degassed with nitrogen. 44.4 µl of a 10 mM Copper(II)-TBTA complex in 55 % aq. DMSO (500 µM final concentration) (Lumiprobes, #21050) was added and the solution briefly degassed with nitrogen. The reaction stirred for 3 h at room temperature in absence of light. The

reaction was then dissolved in 0.1 M TEAB (up to 35 mL) and purified by C8 HPLC (Higgins Analytical) using 0.1 M TEAB and acetonitrile as the mobile phase to yield the 3′-azido RNA adapter.

3.4.3 Isolation of Total *S. cerevisiae* S288C RNA.

Total RNA was purified from *Saccharomyces cerevisiae* S288C. The *S. cerevisiae* were grown in 1 L YPD media (1 % yeast extract, 2 % peptone, 2 % dextrose) at 30 °C. The cells were pelleted and resuspended in cold 10 mM EDTA. The cells were again pelleted and resuspended in 5 ml of 50 mM sodium acetate (pH 5.5), 10 mM EDTA, 1 % SDS. 5 ml of acid-phenol:chloroform (5 ml, ThermoFisher #(AM9720) was added, and the mixture was vortexed. The mixture was incubated in a 65 ℃ water bath with brief vortexing every 5 min for a total incubation time of 30 min. The mixture was placed on ice for 10 min, and the phases separated by centrifugation. The upper phase was collected, and an equal volume of chloroform was added. The mixture was vortexed again, and the phases separated by centrifugation. The upper phase was collected and 0.1 volume of 3 M sodium acetate pH 5.3 was added. An equal volumes of isopropanol was mixed into the solution and RNA was precipitated at −20 °C. The resulting RNA precipitate was dissolved in 5 ml of TE. The RNA was reprecipitated by adding 0.25 volumes of 1 M sodium acetate pH 5.5 and 2.5 volumes of ethanol and incubated for 60 min at −20 °C. The total RNA was pelleted and redissolved in TE.

3.4.4 *S. cerevisiae* S288C Poly(A) Isolation.

Poly(A) RNA was isolated from 2 mg of total *S. cerevisiae* RNA using the poly(A) Spin mRNA Isolation Kit (NEB S1560). After a single round of isolation the RNA was precipitated by adding glycogen and 2.5 volumes of ethanol. The polyA RNA pellet was dried and resuspended in 1 mM Tris-HCl pH 7.5, 0.1 mM EDTA.

3.4.5 Decapping and recapping of RNA samples.

See Chapter 2 Method 2.4.6 for details.

3.4.6 Copper-catalyzed click chemistry of RNA adapter

Copper-catalyzed click chemistry reactions were performed in a total volume of 10 µl, containing 25 % v/v DMSO in 0.2 M triethylammonium acetate buffer, pH 7 as follows (unless other specified, final concentrations are given): 0.5 µM propargyl capped RNA, 4 µM 3′-azido RNA adapter and 500 µM ascorbic acid were combined and the solution briefly degassed with nitrogen. 0.5 µl of a 10 µM Copper(II)-TBTA complex in 55 % aq. DMSO (500 µM final concentration) (Lumiprobes, #21050) was added and the solution briefly degassed with nitrogen. The reaction shaken overnight at room temperature in absence of light. The adapted RNA was recovered using RNA Clean & Concentrator (Zymo Research, #R1013).

3.4.7 Copper-free click chemistry of RNA and adaptor

See Chapter 2 Method 2.4.8 for details.

3.4.8 MinION RNA sequencing

See Chapter 2 Method 2.4.11 for overall details. The optional reverse transcription step in the ONT library preparation was skipped for GLuc RNA.

3.4.9 Basecalling, filtering and alignments

See Chapter 2 Method 2.4.12 for overall details. Methods specific to this chapter are as follows. GLuc RNA reads were aligned to the GLuc reference sequence (3.4.12) using minimap2 (version 2.16-r922) -ax map-ont parameters.

3.4.10 Porechop optimization

The Porechop barcode_threshold parameter was evaluated using the number of soft or hard clipped bases on the $5'$ end of each alignment [83, 90]. A soft or hard clipped end is a portion of the alignment masked to maximize the alignment score. This feature of minimap2 allows for reads to be aligned to a reference even when a portion of the read's ends (which tend to be more error prone) can't be aligned to the reference. The Porechop parameter, barcode_threshold, is the minimum proportion of the adapter sequence length which must have a perfect match in the $5'$ most 150 bases of the read sequence, or the minimum proportion of the adapter sequence which the $5'$ most 150 bases of the read sequence must match. The barcode_threshold was tested from 66 to 80 and the total reads, cap-adapted, and unadapted reads were aligned to the reference genome using minimap2. Porechop parameters were chosen that first minimized the number of cap-adapted reads with zero soft and hard clipped bases on the $5'$ end, and

second, that minimized the number of unadapted reads with soft and hard clipped bases of ∼40. We found that a barcode threshold value of 70 was optimal for *S. cerevisiae* S288C Poly(A) for both click reactions.

3.4.11 *In vitro* transcription of synthetic poly(A) GLuc RNA.

A 809 nucleotide transcript of *Gaussia* luciferase was generated by *in vitro* transcription using HiScribe™ T7 Quick High Yield RNA Synthesis Kit (NEB E2050S) following the manufacture's directions. The DNA template was generated by PCR from the plasmid pCMV-GLuc-2 (NEB N8081S) with the LongAmp Taq 2X Master mix (NEB, M0287S) and the following PCR primers: The forward primer incorporated the T7 promoter: 5′ - TCGAAATTAATACGACTCACTATAGGGAGACCCAA - 3′ and the reverse primer was used to add a 3′ terminal tail of 125 A residues: 5′ - (T_{125})ACAGTAAGAATTATTTCTAGACACAC - 3′.

3.4.12 GLuc *in vitro* transcript sequence

GGGAGACCCAAGCTTGGTACCGAGCTCGGATCCAGCCACCATGGGAGTCA
AAGTTCTGTTTGCCCTGATCTGCATCGCTGTGGCCGAGGCCAAGCCCACCG
AGAACAACGAAGACTTCAACATCGTGGCCGTGGCCAGCAACTTCGCGACC
ACGGATCTCGATGCTGACCGCGGGAAGTTGCCCGGCAAGAAGCTGCCGCT
GGAGGTGCTCAAAGAGATGGAAGCCAATGCCCGGAAAGCTGGCTGCACCA
GGGGCTGTCTGATCTGCCTGTCCCACATCAAGTGCACGCCCAAGATGAAG
AAGTTCATCCCAGGACGCTGCCACACCTACGAAGGCGACAAAGAGTCCGC

ACAGGGCGGCATAGGCGAGGCGATCGTCGACATTCCTGAGATTCCTGGGT

TCAAGGACTTGGAGCCCATGGAGCAGTTCATCGCACAGGTCGATCTGTGT

GTGGACTGCACAACTGGCTGCCTCAAAGGGCTTGCCAACGTGCAGTGTTC

TGACCTGCTCAAGAAGTGGCTGCCGCAACGCTGTGCGACCTTTGCCAGCA

AGATCCAGGGCCAGGTGGACAAGATCAAGGGGGCCGGTGGTGACTAAGCG

GCCGCAATAAAATATCTTTATTTTCATTACATCTGTGTGTTGGTTTTTTGT

GTGTCTAGAAATAATTCTTACTGTAAAAAAAAAAAAAAAAAAAAAAAAAA

AAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAA

AAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAA

3.4.13 General Data manipulation

General sequencing data manipulations were done using bedtools [82] and samtools [83].

3.5 Chapter 3 Acknowledgments

Hugh Olsen, Miten Jain, Robin AbuShumays, and Mark Akeson read and provided feedback on versions of this chapter. Cap-adaptation work was done at New England Biolabs.

Chapter 4

Detecting both poly(A) and non-poly(A) RNA with a generalized RNA nanopore sequencing strategy

4.1 Abstract

The standard ONT protocol for direct RNA sequencing is designed to capture RNA strands ending in a poly(A) tail. Sequencing non-poly(A) RNA is possible using custom adapters or by adding a 3′ end poly(A) to RNA 3′ ends before preparing the sample for sequencing. There are several natural post-transcriptional processes that

add adenosine nucleotides to RNA 3′ ends, and thus *in vitro* tailing with adenosines obscures important biological information in the sample. To preserve and capture native 3′ end nucleotide sequence information, we developed a polyinosine tailing method that preserves the natural 3′ end sequence while allowing the broader population of RNA strands to be adapted for sequencing. We show that the inosine homopolymer produces a distinctive ionic current signature that allows it to be distinguished from a native poly(A) tail. This signal was used to develop a classifier that identifies the presence of a poly(I) tail and estimate it's length in nucleotides.

4.2 Introduction

A cell's transcriptome contains information about gene structure, function and regulation [91]. The transcriptome has been interrogated by various methods, such as microarrays [11] and next generation sequencing techniques (NGS) [92–94]. For these techniques, the RNA is typically reverse transcribed into cDNA before analysis. The conversion of RNA into cDNA has biases and removes information, such as modified bases [18, 19]. Long read sequencing of cDNA allows for a direct measure of isoform structure [95–98] which is difficult to achieve using short read sequencing [15]. Oxford Nanopore Technologies has recently adapted the MinION to sequence RNA directly [40, 44, 46, 58].

Characteristics of individual RNA strands can be documented using nanopores because each strand is sequenced continuously in the 3′-to-5′ orientation. However,

using standard protocols, only poly(A) RNA can be adapted for sequencing. Due to this limitation, other methods have been developed to sequence non-poly(A) RNA using the nanopore platform. One method uses a custom adapter that is specific to the 3′ end of target RNA [44, 99]. This works well if the 3′ end is known and invariable, as is the case for rRNA and tRNA. Adapter ligation efficiency and 3′ end variability will limit RNA adaptation [100, 101].

To overcome these limitations, a poly(A) tail can be enzymatically added to RNA 3′ ends in a sample [43,102]. This is advantageous because standard ONT adapters can be used for sequencing. However, this obscures the presence and length of biological 3′ poly(A) tails in the sample.

We developed another approach that retains native 3′ end end bases while simultaneously preparing the strand for sequencing adapter ligation and sequencing. We used the *Schizosaccharomyces pombe* polyU polymerase, Cid-1 (PUP) [103], to extend RNA 3′ ends with an inosine homopolymer tail. We found that this poly(I) tail had a recognizable nanopore ionic current signal, which was distinct from poly(A) and encoded RNA signals. We used this ionic current difference to develop a Hidden Markov model (MarginAi) that detected inosine tails at 98.46 % accuracy. MarginAi estimates of poly(I) and poly(A) tail lengths were comparable to estimates using nanopolish-polya [46].

4.3 Results

4.3.1 Characterizing polyinosine ionic current signals on RNA 3′ ends

It was uncertain if polyinosine (poly(I)) tails would be distinguishable from the encoded RNA sequence or poly(A) tail in nanopore ionic current data. To test this, we constructed RNA substrates with known poly(I) tails. To ensure that the poly(I) tails were entirely composed of inosine nucleotides, we ordered a synthetic 5′ phosphorylated 15 nt poly(I) oligomer that was confirmed using mass spectrometry (Figure 4.1a and Methods). The dominant peak of the mass spectrum (5265.99 g/mol) matched closely with the predicted mass (5,268.89 g/mol).

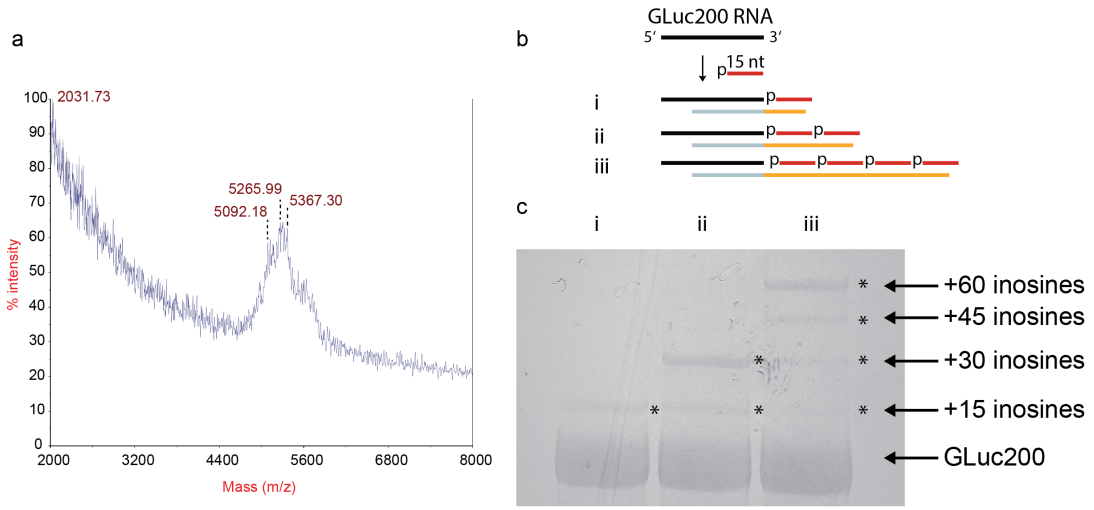

Figure 4.1 Preparation of poly(I) tailed RNA standards. (**a**) Mass spectrometry analysis of a synthetic inosine 15mer. The highest peak has the expected mass for the inosine 15mer oligomer. (**b**) Preparation of inosine-tailed RNA standards. Red bars represent 5′ phosphorylated inosine 15mer. Blue bars represent 10 complementary nucleotides to the GLuc200 3′ end. Orange bars represent poly(dC) oligomers, which are 10 nt (i), 25 nt (ii), and 55 nt (iii). (**c**) Denaturing PAGE of GLuc200 RNA ligated to one-to-four copies of the inosine 15mer. The lane labels correspond to the labels in panel **b**. The longest band from each lane was excised from the gel for nanopore sequencing.

This 15mer poly(I) oligomer was ligated to the 3′ end of a synthetic 200 nt RNA transcript (GLuc200). To generate poly(I) tails of different lengths, we ligated one-to-four of these 15mer poly(I) oligomers to the 3′ end of GLuc200 using a DNA splint adapter with three different poly(dC) lengths to guide and enhance assembly (Figure 4.1b and Methods). The resulting ligation products were size selected (+15 inosines, +30 inosines, and +60 inosines) by gel purification (Figure 4.1c and Methods). These ligation reactions were repeated for the same synthetic RNA bearing a 44 nt 3′ poly(A) tail (GLuc200A44).

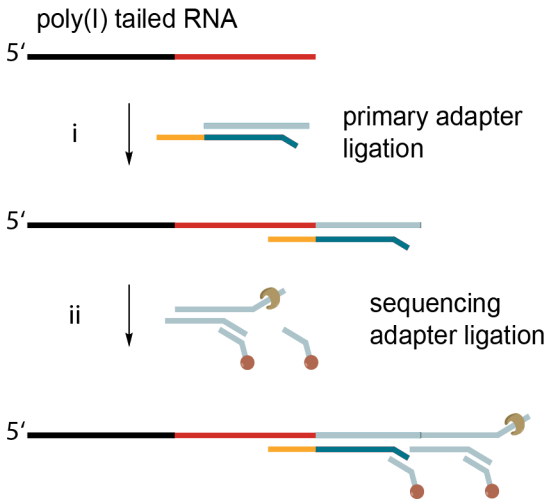

Figure 4.2 Preparation of poly(I) tailed RNA for nanopore sequencing. The black bar represents the GLuc200 sequence. The red bar represents the poly(I) tail. (i) Ligation of the poly(I) tailed-GLuc200 RNA is facilitated by hybridization of the poly(I) tail with a DNA adapter bearing a poly(dC) 10mer overhang (orange). (ii) The ONT sequencing adapter, bearing the pre-bound motor enzyme, is ligated to the primary adapter. The fully adapted RNA is then loaded onto a MinION flow cell.

Each sample was adapted for nanopore sequencing using a modified splint bearing 10 cytosines in place of 10 thymines on the standard ONT adapter (Figure 4.2 and Methods). The poly(I)-tailed constructs, and control samples without poly(I), were each sequenced on a dedicated nanopore flow cell.

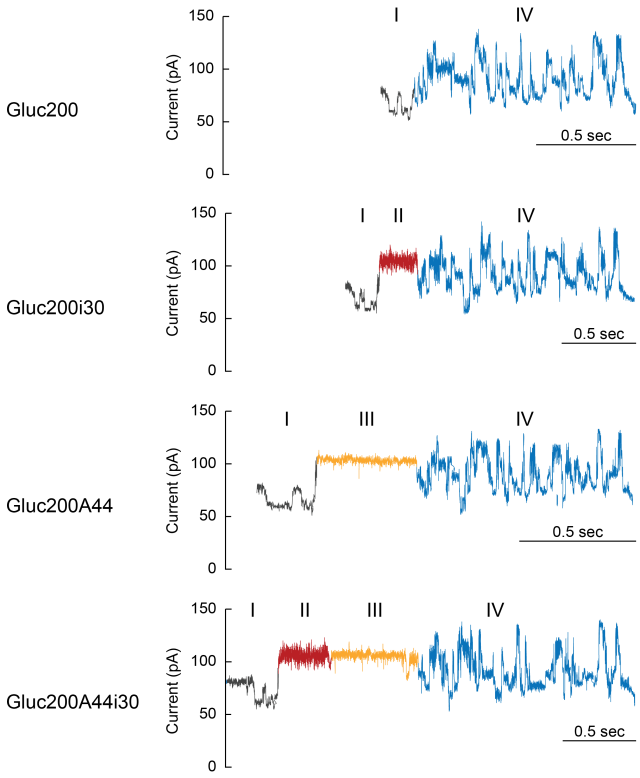

Figure 4.3 Representative ionic current traces for GLuc200 RNA bearing four different 3′ tails. The y-axis is ionic current in pA. Along the x-axis, the RNA is translocated and read 3′ to 5′ from left to right. The four ionic current traces are segments of full-length RNA reads. Top to bottom the traces are GLuc200 bearing: no tail; a 30 nt poly(I) tail; a 44 nt poly(A) tail; and both a 30 nt poly(I) tail and a 44 nt poly(A) tail. The four major regions in each ionic current trace are the ONT adapter (grey, I), the poly(I) tail (red, II), the poly(A) tail (gold, III), and GLuc200 nucleotides (blue, IV).

There was a visually distinguishable ionic current signature present on the 3′ ends of poly(I)-tailed RNA strand reads that was absent from the control RNA strand reads (Figure 4.3). This was due to a much higher mean current variance for the poly(I) segment relative to the poly(A) segment (σ^2 =15.1 ± 6.55 pA and 4.30 ± 3.49 pA, respectively). Surprisingly, the poly(I) mean current (μ = 111.98 ± 3.62 pA) was nearly identical to the poly(A) mean current (μ = 111.85 ± 2.06 pA). Because of

this, the ONT basecaller treated the poly(I) segment as if it were part of the poly(A) segment.

4.3.2 Modeling polyinosine tails with MarginAi

Hidden Markov Models (HMMs) [104, 105] are used to model sequential data that have an observed state and a hidden state. They are often used to model nanopore data that are inherently sequential and have observable ionic currents that correspond to an unknown (hidden) nucleotide in the translocating strand. HMMs applied to mean ionic currents have been used for early versions of nanopore basecallers (`https://github.com/nanoporetech/scrappie`). HMMs have also been used to identify modified DNA nucleotides [106], and to estimate RNA poly(A) tail lengths [46] using nanopore data. However, as stated above, HMMs based on mean current cannot distinguish between poly(I) and poly(A).

As an alternative, we used both μ and σ^2 to segment the RNA ionic current. To this end, we built a two stage HMM (MarginAi) based on YAHMM and PyPore frameworks [107]. The first stage of MarginAi used μ to assign ionic current segments to states (Figure 4.4a). These states represent physical regions of the RNA strand including: the ONT adapter; the poly(A) and poly(I) tails; and the GLuc200 RNA. The hidden state emission values for this first stage were based on a widely used program nanopolish-polya [46] with two modifications (see Methods). The second stage of MarginAi used only σ^2 of the homopolymer segments to classify the RNA tail as poly(A) (Figure 4.4b i), poly(I) (Figure 4.4b ii), or both (Figure 4.4b iii).

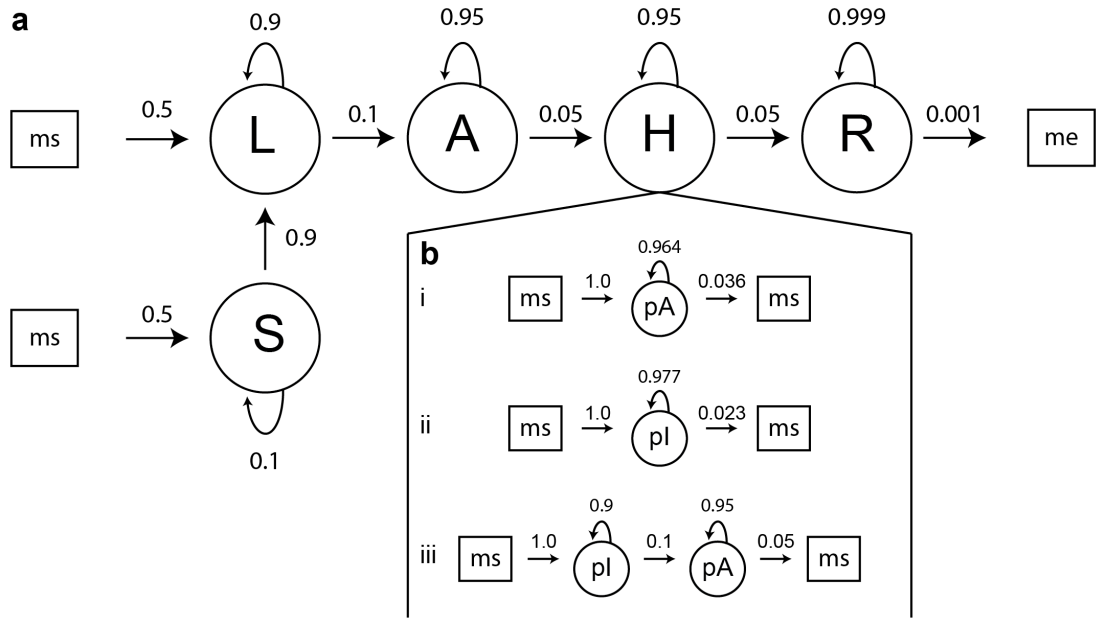

Figure 4.4 Two stage Hidden Markov Model schematic for classifying 3′ tail types. (**a**) Stage one model schematic. Squares represent the model start (ms) and model end (me) states. Circles represent match states and the arrows represent transitions from one state to another state with the probability of that transition. The stage one model uses the ionic current mean to assign ionic current segments to states representing physical features of the RNA molecule (adapter, homopolymer tail, and read). (**b**) Stage two model schematic. The squares, circles, and lines represent the same types of features as in panel a. The stage two model uses the mean current variance of segments assigned to the homopolymer state by the stage one model. The emission values for each match state are described below

(a) Stage 1: segment means

(S) START: $\mathcal{N}(\mu = 70.2737, \sigma^2 = 3.7743)$

(L) LEADER: $\mathcal{N}(\mu = 110.973, \sigma^2 = 5.237)$

(A) ADAPTER: $0.874 \times \mathcal{N}(\mu = 79.347, \sigma^2 = 8.3702)+$

$0.126 \times \mathcal{N}(\mu = 63.3126, \sigma^2 = 2.7464)$

(H) HOMOPOLYMER: $\mathcal{N}(\mu = 108.883, \sigma^2 = 3.257)$

(R) READ: $0.346 \times \mathcal{N}(\mu = 79.679, \sigma^2 = 6.966)+$

$0.654 \times \mathcal{N}(\mu = 105.784, \sigma^2 = 16.022)$

(b) Stage 2: segment variance

i. Poly(A) only

pA: $\mathcal{N}(\mu = 4.3025, \sigma^2 = 3.4873)$

ii. Poly(I) only

pI: $\mathcal{N}(\mu = 15.1075, \sigma^2 = 6.5471)$

iii. Poly(I) + Poly(A)

pI: $\mathcal{N}(\mu = 14.1075, \sigma^2 = 6.5471)$

pA: $\mathcal{N}(\mu = 4.3025, \sigma^2 = 3.4873)$

MarginAi was tested using 100,000 randomly selected nanopore reads from each GLuc200 sample (Table 4.1). Among these samples, on average 27,278 reads passed segmentation and model alignment. When poly(I) tails were present, they were detected on average 98.46 % of the time. However, MarginAi miss-classified 20.03 % of poly(A) tails as containing a poly(I) segment. This is likely due to a high variance current at the transition from the ONT adapter to poly(A) (see Figure 4.3 GLuc200A44 between regions I and III). MarginAi also miss-classified poly(A) + poly(I) tails as only poly(I)-tailed 12.86 % of the time. This miss-classification occurred more for the 15 nt poly(I) tail than for the 30 nt and 60 nt poly(I) tails. MarginAi correctly classified the 30 nt poly(I) tail more accurately than all other samples.

Table 4.1 RNA training data tail type classification by MarginAi and nanopolish-polyi. The alternating grey and white columns are the classification results from RNA strands bearing a given inosine tail length. These are no inosine tail, 15 nt inosine tail, 30 nt inosine tail or 60 nt inosine tail. These are labeled, i0, i15, i30, and i60 respectively. There are two columns within the grey or white columns, denoting which program was used for the classification. The different poly(I) tails were added to two different RNA transcripts, GLuc200 and GLuc200A44. There are three possible tail types, poly(I) only (pI), poly(A) only (pA), and both tails (pApI). The rows entitled Tail identification proportion are the proportion of the total reads that were classified as a given tail type. The rows entitled Tail identification counts are the number of reads classified as each class.

Proportion of tail types identified

		i0		i15		i30		i60	
RNA type	tail call	MarginAi	Nanopolish	MarginAi	Nanopolish	MarginAi	Nanopolish	MarginAi	Nanopolish
GLuc200	pi	N/a	N/a	0.9502	0.9954	0.9744	0.9965	0.9825	0.9978
	pA	N/a	N/a	0.0102	0.0026	0.0066	0.0014	0.0088	0.0009
	pApi	N/a	N/a	0.0396	0.002	0.0189	0.002	0.0086	0.0014
GLuc200A44	pi	0.0515	0.1631	0.1829	0.3134	0.0965	0.227	0.1064	0.702
	pA	0.7997	0.7614	0.0054	0.0012	0.0035	0.0015	0.0016	0.0007
	pApi	0.1488	0.0755	0.8117	0.6854	0.9	0.7715	0.892	0.2972

Counts of tail types identified

		i0		i15		i30		i60	
RNA type	tail call	MarginAi	Nanopolish	MarginAi	Nanopolish	MarginAi	Nanopolish	MarginAi	Nanopolish
GLuc200	pi	N/a	N/a	31429	18710	10131	17286	24542	18700
	pA	N/a	N/a	337	48	69	25	221	16
	pApi	N/a	N/a	1311	38	197	35	215	26
GLuc200A44	pi	1689	3509	5151	5773	2642	4320	3308	15148
	pA	26207	16383	152	22	95	28	49	16
	pApi	4877	1624	22854	12625	24645	14679	27721	6414

The length of the poly(A) tail can influence trafficking and stability of RNA in the cell [108, 109]. Therefore, we implemented a rudimentary poly(A) tail length

estimator as part of MarginAi (see Methods). This is a similar approach as used by nanopolish-polya [46]. Unfortunately, we found that the poly(I) and poly(A) nucleotide lengths were combined by stage one of MarginAi and by nanopolish-polya.

4.3.3 Poly(I) tailing RNA 3′ ends with S. pombe Cid-1 polyU polymerase

We tested enzymatic polyinosine tailing by S. pombe Cid-1 polyU polymerase (PUP), using a 510 nt synthetic RNA substrate. At 37 °C and 0.5 mM ITP, we observed nucleotide additions within one minute (Figure 4.5). By 40 minutes, the RNA substrate was quantitatively extended by 50 nucleotides.

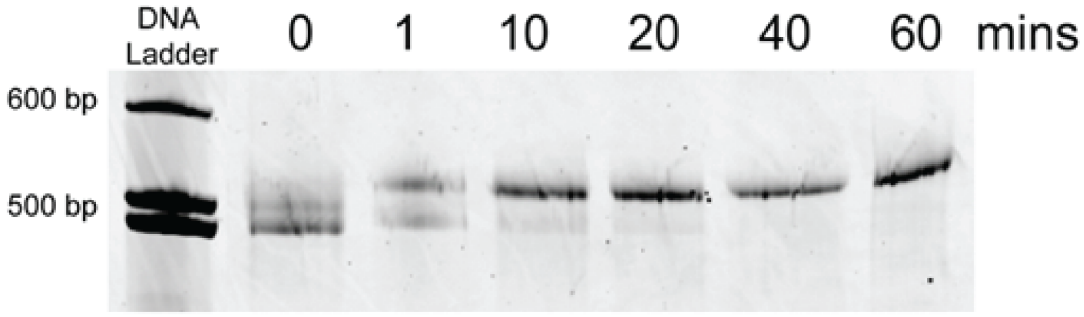

Figure 4.5 Time course of inosine extensions by *S. pombe* Cid-1 polyU polymerase. Each lane is an RNA sample taken at 0-to-60 minutes. The lower band at time 0 minutes is the untailed 510 nucleotide RNA substrate. The discrete upper band at times 1-to-60 minutes is the substrate extended by ∼50 nucleotides. The image is a denatured PAGE gel stained with sybr gold. The marker is DNA.

When used the same enzymatic strategy to poly(I)-tail GLuc200A44, we found that the associated nanopore ionic current trace contained the same high current vari-

ance segment as observed for the GLuc200A44 synthetic poly(I) tail (Figure 4.6a and b respectively). The read coverage for the poly(I)-tailed GLuc200A44 was comparable to the untailed control (Figure 4.6c and d respectively). This indicated that polymerase extension did not adversely affect RNA read throughput.

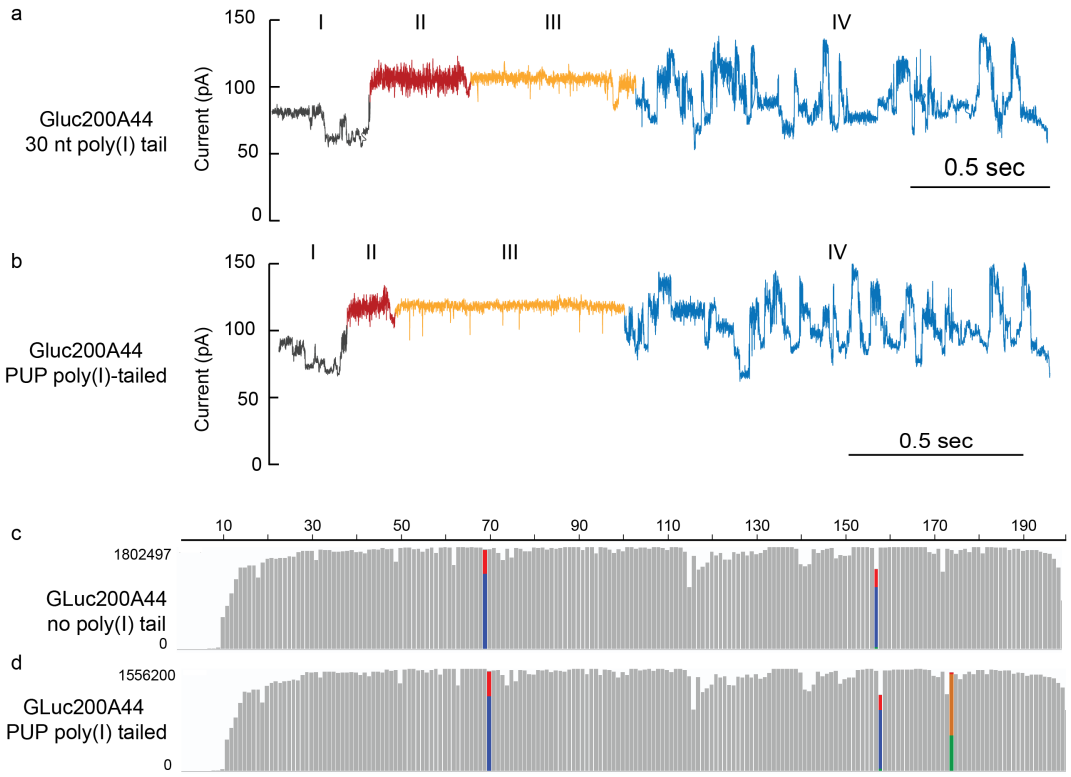

Figure 4.6 GLuc200A44 poly(I)-tailed by polyU polymerase nanopore sequencing. (**a**) Untailed GLuc200A44 representative ionic current trace segment. The y-axis is ionic current in pA. Along the x-axis, the RNA is translocated and read 3′ to 5′ from left to right. This ionic current trace of GLuc200A44 bearing a 30 nt poly(I) tail is a segment of the full-length RNA read. The four major regions in each ionic current trace are the ONT adapter (grey, I), the poly(I) tail (red, II), the poly(A) tail (gold, III), and GLuc200 nucleotides (blue, IV). (**b**) Nanopore ionic current trace of GLuc200A44 poly(I)-tailed by PUP. The axes are the same as in a. (**c**) GLuc200 sequencing coverage. Coverage is determined by the number of reads that have an aligned base at that position. Lower coverage at a particular position either means fewer reads aligned to that position, or that position was skipped in reads. If the proportion of reads aligned at a position differs from the reference by more than 20 %, the coverage bar is colored according to the proportion of bases aligned to that position. Otherwise the coverage bar is grey. (**d**) PUP poly(I)-tailed GLuc200A44 sequencing coverage. The coverage in d is determined as in c.

4.4 Conclusion

We demonstrated that 3′ inosine tails on RNA have a distinguishable nanopore ionic current signature that can be accurately identified using a new HMM termed MarginAi. We have further shown that polyU polymerase can append inosine nucleotides on the 3′ end of model mRNA strands which can be detected by nanopore sequencing. The major difficulty in this approach is that the mean ionic current for poly(A) and poly(I) are nearly identical. We anticipate that alternative non-canonical RNA nucleotides might yield measurably different mean ionic currents vs poly(A), and thus resolve this technical problem.

4.5 Methods

4.5.1 Poly(I)-tailing

Inosine homopolymers were added to the 3′ end of RNA molecules by first resuspending the RNA in 0.1 mM EDTA to a final volume of 2.95 µl. The RNA is denatured at 95 °C for 2 min then placed on ice for 2 min. The RNA is added to a reaction containing 4 mM ITP, 50 mM NaCl, 13.5 mM MgCl2, 1 mM DTT, BSA 500 µg ml^{-1}, pH 7.9, and 1 µl of polyU polymerase (PUP) (NEB #M0337S) in a final volume of 7.5 µl and incubated at 37 °C for 1 h. The RNA was purified using SPRIselect Reagent (Beckman Coulter #B23318). The reaction was resuspended with 1.8x volume of SPRIselect Reagent and incubated at room temp for 10 min. The beads were pelleted on a magnet and the supernatant was decanted. The beads were washed with 70 % ethanol three

times, then air dried until visibly matte. The beads were resuspended in 11 µl of water and incubated in 10 min at room temperature, pelleted, then eluate was transferred to a new tube.

4.5.2 Poly(I) 15mer ligations

To prepare GLuc200A44i15, GLuc200A44i30, GLuc200A44i60, GLuc200i15, GLuc200i30 and GLuc200i60 samples, 15 pmol of GLuc200 RNA with 30 pmol of the appropriate bottom splint adapter ($C_{10}T_{10}$, $C_{25}T_{10}$, or $C_{55}T_{10}$) (IDT) for GLuc200A44 or (C_{10}CCTAAGAGCAAGAAGAAG, C_{25}CCTAAGAGCAAGAAGAAG, or C_{55}CC TAAGAGCAAGAAGAAG) (IDT) for GLuc200), 1.4 nmol of a synthetic 5′ p-15mer inosine homopolymer (Stanford PAN facility) in 10 mM tris pH 8.0, 1 mM EDTA, and 50 mM NaCl in 6 µl reaction volume was heated to 55 °C and slow cooled to 16 °C in 25 min. One µl 10X T4 ligation reaction buffer (NEB B0202S) and 2,000 units T4 DNA ligase (NEB M0202T) were added to each reaction and brought to 10 µl volume with water then incubated at 16 °C overnight. 2X RNA loading dye (NEB N0362) was added to each sample and denatured at 95 °C for 5 min before loading into a 10 % acrylamide gel and ran for 3.5 h at 28 W. The gel excision was performed by post-staining with 1X SYBR gold in TBE and visualized on a UV transilluminator while cutting with a razorblade. The samples were eluted from the gel slice using a D-tubeTM Dialyzer Midi MWCO 3.5 kDa (Millipore Sigma 71507) in 850 µl of 1x TAE buffer for at least 90 min at 130 V. The electro-eluted samples were precipitated with 85 µl 0.3 M NaOAc (pH 5.2) and 850 µl isopropanol at −20 °C overnight. The samples were centrifuged at 4,000

g for 30 min, supernatant decanted and the pellets were washed with 70 % ethanol twice with subsequent centrifugations at 16,000 g for 15 min. The pellets were air dried for 15 min and resuspended in 10 µl nuclease free water with yields between 25-100 ng per sample. The libraries for each ligation product were prepared following ONT's direct RNA nanopore sequencing library preparation with up to 50 ng of RNA without the optional reverse transcription step.

4.5.3 MinION Library Preparation

RNA (500-775 ng) were prepared for nanopore direct RNA sequencing generally following the ONT SQK-RNA001/SQK-RNA002 kit protocol. Poly(A) RNA were adapted with ONT's RTA adapter. Poly(I) RNA were adapted with a custom adapter duplex in place of the RTA adapter. The custom adapter duplex was made by mixing 30 pmol of the top 5′-pGGCTTCTTCTTGCTCTTAGGTAGTAGGTTC-3′ (IDT) and bottom 5′-CCTAAGAGCAAGAAGAAGCCCCCCCCCCCC-3′ oligonucleotides (IDT) in (50 mM Tris pH 8, 50 mM NaCl, 1 mM EDTA) and heated to 55 °C and slow cooled to 23 °C over 25 min. The optional reverse transcription reaction was used for all biological samples, but Superscript IV (Thermo Fisher) was used for reverse transcription instead of Superscript III, as in the ONT protocol. RNA sequencing on the MinION was performed using ONT R9.4 flow cells and the standard MinKNOW protocol script RNA002 recommended by ONT, with one exception. We collected bulk phase raw files for 2 h of sequencing. After 2 h the runs were restarted normally.

4.5.4 Basecalling

We used the ONT Guppy flipflop workflow (version 3.0.3+7e7b7d0 using configuration file "rna_r9.4.1_70bps_hac.cfg") for basecalling direct RNA. NanoFilt (version 2.5.0) [80] was used to classify reads as pass if the pre-read average Phred-score threshold was greater than or equal to 7 and fail if less than 7. A custom python2.7 script was used to convert the U's to T's in the fastq files.

4.5.5 Alignments

We used minimap2 [81] recommended parameters to map the RNA pass reads to the GLuc reference, pCMV-GLuc 2 Control Plasmid neb # N8081.

4.5.6 Classification

MarginAi can be found here (`https://github.com/mitenjain/marginAi`). Briefly, the continuous ionic current traces were segmented using an ONT publicly available basecaller, scrappie. The Nanopolish module, dump-initial alignment [110], is used as a python accessible wrapper for the scrappie segmentation function. Both stages of the HMM were built from tRNApore (`https://github.com/mitenjain/tRNApore`) using the PyPore (`https://github.com/jmschrei/PyPore`) and YAHMM (`https://github.com/jmschrei/yahmm`) frameworks [107].

The stage one emission values and transition probabilities for each state were based on the HMM implemented by nanopolish-polya [46]. The 'cliff' state was removed and the homopolymer state Gaussian mean was adjusted for values identified in

GLuc200A44 data collected by this study. The ionic current segments are aligned to the model, and assigned to states using the Viterbi path [111].

The stage two HMM emission values and transition probabilities were hand curated using 10 reads of each tail type (poly(A), pol(A), poly(I) + poly(A)). The values were hand adjusted and evaluated using 30 reads of each tail type. The segments corresponding to the tail region were aligned to each of the three different models. The model alignment with the highest score was chosen as the tail classification.

The homopolymer state length in nucleotides is estimated by first summing the duration of segments assigned to the Read state. Second, the guppy base called sequence length is divided by the duration of the Read segments to calculate an average nucleotides per second RNA strand translocation rate. Third, the sum of the Homopolymer segment duration is multiplied by the RNA strand translocation rate.

4.5.7 General Data manipulation

General sequencing data manipulations were done using bedtools [82] and samtools [83].

4.6 Chapter Acknowledgments

Jenny Vo optimized the PUP inosine tailing protocol and created the gel image in Figure 4.5. The inosine 15mer was synthesized and mass spectrum was done by the PAN oligo facility at Stanford. Miten Jain advised on all aspects of the RNA ionic

current signal processing. Robin AbuShumays and Hugh Olsen read drafts of this thesis chapter.

Bibliography

[1] Oswald T Avery, Colin M MacLeod, and Maclyn McCarty. Studies on the chemical nature of the substance inducing transformation of pneumococcal types: induction of transformation by a desoxyribonucleic acid fraction isolated from pneumococcus type iii. *The Journal of experimental medicine*, 79(2):137–158, 1944.

[2] Frederick Sanger, Steven Nicklen, and Alan R Coulson. DNA sequencing with chain-terminating inhibitors. *Proceedings of the national academy of sciences*, 74(12):5463–5467, 1977.

[3] James M Heather and Benjamin Chain. The sequence of sequencers: The history of sequencing DNA. *Genomics*, 107(1):1–8, 2016.

[4] Mahlon B Hoagland, Mary Louise Stephenson, Jesse F Scott, Liselotte I Hecht, and Paul C Zamecnik. A soluble ribonucleic acid intermediate in protein synthesis. *Journal of Biological Chemistry*, 231(1):241–257, 1958.

[5] Koji Tamura. Origins and early evolution of the tRNA molecule. *Life*, 5(4):1687–1699, 2015.

[6] Harry F Noller. On the origin of the ribosome: coevolution of subdomains of tRNA and rRNA. *COLD SPRING HARBOR MONOGRAPH SERIES*, 24:137–137, 1993.

[7] Sydney Brenner, François Jacob, and Matthew Meselson. An unstable intermediate carrying information from genes to ribosomes for protein synthesis. *Nature*, 190(4776):576–581, 1961.

[8] François Gros, Howard Hiatt, Walter Gilbert, Chuck G Kurland, RW Risebrough, and James D Watson. Unstable ribonucleic acid revealed by pulse labelling of *Escherichia coli*. *Nature*, 190(4776):581–585, 1961.

[9] John S Mattick and Igor V Makunin. Non-coding RNA. *Human molecular genetics*, 15(suppl_1):R17–R29, 2006.

[10] Stephen P Fodor, J Leighton Read, Michael C Pirrung, Lubert Stryer, A Tsai Lu, and Dennis Solas. Light-directed, spatially addressable parallel chemical synthesis. *science*, 251(4995):767–773, 1991.

[11] M Schena, D Shalon, R W Davis, and P O Brown. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science (New York, N.Y.)*, 270(5235):467–70, oct 1995.

[12] Satoshi Mizutani, David Boettiger, and Howard M Temin. A DNA-dependent DNA polymerase and a DNA endonuclease in virions of Rous sarcoma virus. *Nature*, 228(5270):424–427, 1970.

[13] David Baltimore. Viral RNA-dependent DNA polymerase: RNA-dependent DNA polymerase in virions of RNA tumour viruses. *Nature*, 226(5252):1209–1211, 1970.

[14] Zhong Wang, Mark Gerstein, and Michael Snyder. RNA-Seq: a revolutionary tool for transcriptomics. *Nature reviews genetics*, 10(1):57–63, 2009.

[15] Ana Conesa, Pedro Madrigal, Sonia Tarazona, David Gomez-Cabrero, Alejandra Cervera, Andrew McPherson, Michal Wojciech Szcześniak, Daniel J Gaffney, Laura L Elo, Xuegong Zhang, and Ali Mortazavi. Erratum to: A survey of best practices for RNA-seq data analysis. *Genome biology*, 17(1):181, aug 2016.

[16] Allison Piovesan, Maria Caracausi, Francesca Antonaros, Maria Chiara Pelleri, and Lorenza Vitale. GeneBase 1.1: a tool to summarize data from NCBI gene datasets and its application to an update of human gene statistics. *Database*, 2016, 2016.

[17] Alejandro Reyes and Wolfgang Huber. Alternative start and termination sites of transcription drive most transcript isoform differences across human tissues. *Nucleic acids research*, 46(2):582–592, 2018.

[18] Saurabh Agarwal, Todd S Macfarlan, Maureen A Sartor, and Shigeki Iwase. Sequencing of first-strand cDNA library reveals full-length transcriptomes. *Nature communications*, 6(1):1–12, 2015.

[19] Haridha Shivram and Vishwanath R Iyer. Identification and removal of sequencing artifacts produced by mispriming during reverse transcription in multiple RNA-seq technologies. *RNA*, 24(9):1266–1274, 2018.

[20] Toshiyuki Shiraki, Shinji Kondo, Shintaro Katayama, Kazunori Waki, Takeya Kasukawa, Hideya Kawaji, Rimantas Kodzius, Akira Watahiki, Mari Nakamura, Takahiro Arakawa, et al. Cap analysis gene expression for high-throughput analysis of transcriptional starting point and identification of promoter usage. *Proceedings of the National Academy of Sciences*, 100(26):15776–15781, 2003.

[21] Maruyama Kazuo and Sugano Sumio. Oligo-capping: a simple method to replace the cap structure of eukaryotic mrnas with oligoribonucleotides. *Gene*, 138(1-2):171–174, 1994.

[22] Michael A Frohman, Michael K Dush, and Gail R Martin. Rapid production of full-length cDNAs from rare transcripts: amplification using a single gene-specific oligonucleotide primer. *Proceedings of the National Academy of Sciences*, 85(23):8998–9002, 1988.

[23] Yasuhiro Furuichi. Discovery of m^7g-cap in eukaryotic mRNAs. *Proceedings of the Japan Academy, Series B*, 91(8):394–409, 2015.

[24] Maruyama Kazuo and Sugano Sumio. Oligo-capping: a simple method to replace the cap structure of eukaryotic mRNAs with oligoribonucleotides. *Gene*, 138(1-2):171–174, 1994.

[25] YY Zhu, EM Machleder, A Chenchik, R Li, and PD Siebert. Reverse transcriptase template switching: A smart™ approach for full-length cDNA library construction. *Biotechniques*, 30(4):892–897, 2001.

[26] Xian Adiconis, Adam L Haber, Sean K Simmons, Ami Levy Moonshine, Zhe Ji, Michele A Busby, Xi Shi, Justin Jacques, Madeline A Lancaster, Jen Q Pan, et al. Comprehensive comparative analysis of 5′-end RNA-sequencing methods. *Nature methods*, 15:505–511, 2018.

[27] Nicola Minshall and Anna Git. Enzyme-and gene-specific biases in reverse transcription of RNA raise concerns for evaluating gene expression. *Scientific Reports*, 10(1):1–7, 2020.

[28] John Eid, Adrian Fehr, Jeremy Gray, Khai Luong, John Lyle, Geoff Otto, Paul Peluso, David Rank, Primo Baybayan, Brad Bettman, et al. Real-time DNA sequencing from single polymerase molecules. *Science*, 323(5910):133–138, 2009.

[29] Shanika L Amarasinghe, Shian Su, Xueyi Dong, Luke Zappia, Matthew E Ritchie, and Quentin Gouil. Opportunities and challenges in long-read sequencing data analysis. *Genome biology*, 21(1):1–16, 2020.

[30] Hagen Tilgner, Fereshteh Jahanbani, Tim Blauwkamp, Ali Moshrefi, Erich Jaeger, Feng Chen, Itamar Harel, Carlos D Bustamante, Morten Rasmussen, and Michael P Snyder. Comprehensive transcriptome analysis using synthetic long-read sequencing reveals molecular co-association of distant splicing events. *Nature biotechnology*, 33(7):736, 2015.

[31] Miten Jain, Ian T Fiddes, Karen H Miga, Hugh E Olsen, Benedict Paten, and Mark Akeson. Improved data analysis for the MinION nanopore sequencer. *Nature methods*, 12(4):351–356, 2015.

[32] John J Kasianowicz, Eric Brandin, Daniel Branton, and David W Deamer. Characterization of individual polynucleotide molecules using a membrane channel. *Proceedings of the National Academy of Sciences*, 93(24):13770–13773, 1996.

[33] Mark Akeson, Daniel Branton, John J Kasianowicz, Eric Brandin, and David W Deamer. Microsecond time-scale discrimination among polycytidylic acid, polyadenylic acid, and polyuridylic acid as homopolymers or as segments within single RNA molecules. *Biophysical journal*, 77(6):3227–3233, 1999.

[34] Kate R Lieberman, Gerald M Cherf, Michael J Doody, Felix Olasagasti, Yvette Kolodji, and Mark Akeson. Processive replication of single DNA molecules in a nanopore catalyzed by phi29 DNA polymerase. *Journal of the American Chemical Society*, 132(50):17961–17972, 2010.

[35] Gerald M Cherf, Kate R Lieberman, Hytham Rashid, Christopher E Lam, Kevin Karplus, and Mark Akeson. Automated forward and reverse ratcheting of DNA in a nanopore at 5-å precision. *Nature biotechnology*, 30(4):344–348, 2012.

[36] Elizabeth A Manrao, Ian M Derrington, Andrew H Laszlo, Kyle W Langford, Matthew K Hopper, Nathaniel Gillgren, Mikhail Pavlenok, Michael Niederweis, and Jens H Gundlach. Reading DNA at single-nucleotide resolution with a mutant MspA nanopore and phi29 DNA polymerase. *Nature biotechnology*, 30(4):349–353, 2012.

[37] Joshua Quick, Nicholas J Loman, Sophie Duraffour, Jared T Simpson, Ettore Severi, Lauren Cowley, Joseph Akoi Bore, Raymond Koundouno, Gytis Dudas, Amy Mikhail, et al. Real-time, portable genome sequencing for Ebola surveillance. *Nature*, 530(7589):228–232, 2016.

[38] Sarah L Castro-Wallace, Charles Y Chiu, Kristen K John, Sarah E Stahl, Kathleen H Rubins, Alexa BR McIntyre, Jason P Dworkin, Mark L Lupisella, David J Smith, Douglas J Botkin, et al. Nanopore DNA sequencing and genome assembly on the International Space Station. *Scientific reports*, 7(1):1–12, 2017.

[39] Aaron S Burton, Sarah E Stahl, Kristen K John, Miten Jain, Sissel Juul, Daniel J Turner, Eoghan D Harrington, David Stoddart, Benedict Paten, Mark Akeson, et al. Off earth identification of bacterial populations using 16S rDNA nanopore sequencing. *Genes*, 11(1):76, 2020.

[40] 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, Javier Blasco, Stephen Young, Denise Brocklebank, Sissel Juul, James Clarke, Andrew J Heron, and Daniel J Turner. Highly parallel direct RNA sequencing on an array of nanopores. *Nature Methods*, 15(3):201–206, mar 2018.

[41] Harvey Lodish, Arnold Berk, S Lawrence Zipursky, Paul Matsudaira, David Baltimore, and James Darnell. Processing of rRNA and tRNA. *Molecular Cell Biology*, page 1184, 2000.

[42] Felix Grünberger, Robert Knüppel, Michael Jüttner, Martin Fenk, Andreas Borst, Robert Reichelt, Jörg Soppa, Sebastien Ferreira-Cerca, and Dina Grohmann. Nanopore-based native RNA sequencing provides insights into prokaryotic transcription, operon structures, rRNA maturation and modifications. *bioRxiv*, 2019.

[43] Heather L Drexler, Karine Choquet, and L Stirling Churchman. Splicing kinetics and coordination revealed by direct nascent RNA sequencing through nanopores. *Molecular cell*, 77(5):985–998.e8, mar 2020.

[44] Andrew M. Smith, Miten Jain, Logan Mulroney, Daniel R. Garalde, and Mark Akeson. Reading canonical and modified nucleobases in 16S ribosomal RNA using nanopore native RNA sequencing. *PLOS ONE*, 14(5):e0216709, may 2019.

[45] Shimyn Slomovic, Ella Fremder, Raymond HG Staals, Ger JM Pruijn, and Gadi Schuster. Addition of poly (A) and poly (A)-rich tails during RNA degradation in the cytoplasm of human cells. *Proceedings of the National Academy of Sciences*, 107(16):7407–7412, 2010.

[46] Rachael E. Workman, Alison D. Tang, Paul S. Tang, Miten Jain, John R. Tyson, Roham Razaghi, Philip C. Zuzarte, Timothy Gilpatrick, Alexander Payne, Joshua Quick, Norah Sadowski, Nadine Holmes, Jaqueline Goes de Jesus, Karen L. Jones, Cameron M. Soulette, Terrance P. Snutch, Nicholas Loman, Benedict Paten, Matthew Loose, Jared T. Simpson, Hugh E. Olsen, Angela N. Brooks, Mark Akeson, and Winston Timp. Nanopore native RNA sequencing of a human poly(A) transcriptome. *Nature Methods*, 16(12):1297–1305, dec 2019.

[47] David Deamer, Mark Akeson, and Daniel Branton. Three decades of nanopore sequencing. *Nature biotechnology*, 34(5):518–524, 2016.

[48] Baohua Cao, Yan Zhao, Yongjun Kou, Dongchun Ni, Xuejun Cai Zhang, and Yihua Huang. Structure of the nonameric bacterial amyloid secretion channel. *Proceedings of the National Academy of Sciences*, 111(50):E5439–E5444, 2014.

[49] Dahai Gai, Rui Zhao, Dawei Li, Carla V Finkielstein, and Xiaojiang S Chen. Mechanisms of conformational change for a replicative hexameric helicase of SV40 large tumor antigen. *Cell*, 119(1):47–60, 2004.

[50] David Botstein and Gerald R Fink. Yeast: an experimental organism for 21st century biology. *Genetics*, 189(3):695–704, 2011.

[51] Adam Frankish, Mark Diekhans, Anne-Maud Ferreira, Rory Johnson, Irwin Jungreis, Jane Loveland, Jonathan M Mudge, Cristina Sisu, James Wright, Joel Armstrong, et al. GENCODE reference annotation for the human and mouse genomes. *Nucleic acids research*, 47(D1):D766–D773, 2019.

[52] Adam Roberts, Harold Pimentel, Cole Trapnell, and Lior Pachter. Identification of novel transcripts in annotated genomes using RNA-Seq. *Bioinformatics*, 27(17):2325–2329, 2011.

[53] Aziz M Mezlini, Eric JM Smith, Marc Fiume, Orion Buske, Gleb L Savich, Sohrab Shah, Sam Aparicio, Derek Y Chiang, Anna Goldenberg, and Michael Brudno. iReckon: simultaneous isoform discovery and abundance estimation from RNA-seq data. *Genome research*, 23(3):519–529, 2013.

[54] Jingyi Jessica Li, Ci-Ren Jiang, James B Brown, Haiyan Huang, and Peter J Bickel. Sparse linear modeling of next-generation mRNA sequencing (RNA-Seq) data for isoform discovery and abundance estimation. *Proceedings of the National Academy of Sciences*, 108(50):19867–19872, 2011.

[55] Mihaela Pertea, Geo M Pertea, Corina M Antonescu, Tsung-Cheng Chang, Joshua T Mendell, and Steven L Salzberg. StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nature biotechnology*, 33(3):290, 2015.

[56] Sam Kovaka, Aleksey V Zimin, Geo M Pertea, Roham Razaghi, Steven L Salzberg, and Mihaela Pertea. Transcriptome assembly from long-read RNA-seq alignments with StringTie2. *Genome Biology*, 20(1):1–13, 2019.

[57] Alison D Tang, Cameron M Soulette, Marijke J van Baren, Kevyn Hart, Eva Hrabeta-Robinson, Catherine J Wu, and Angela N Brooks. Full-length transcript characterization of SF3B1 mutation in chronic lymphocytic leukemia reveals downregulation of retained introns. *Nature communications*, 11(1):1–12, 2020.

[58] Matthew T Parker, Katarzyna Knop, Anna V Sherwood, Nicholas J Schurch, Katarzyna Mackinnon, Peter D Gould, Anthony JW Hall, Geoffrey J Barton, and Gordon G Simpson. Nanopore direct RNA sequencing maps the complexity of arabidopsis mRNA processing and m^6a modification. *eLife*, 9, 2020.

[59] Feng Jiang, Jie Zhang, Qing Liu, Xiang Liu, Huimin Wang, Jing He, and Le Kang. Long-read direct RNA sequencing by 5′-cap capturing reveals the impact of Piwi on the widespread exonization of transposable elements in locusts. *RNA biology*, 16(7):950–959, 2019.

[60] Hudan Liu, Nancy D Rodgers, Xinfu Jiao, and Megerditch Kiledjian. The scavenger mRNA decapping enzyme DcpS is a member of the HIT family of pyrophosphatases. *The EMBO journal*, 21(17):4699–4708, 2002.

[61] Madalee G Wulf, John Buswell, Siu-Hong Chan, Nan Dai, Katherine Marks, Evan R Martin, George Tzertzinis, Joseph M Whipple, Ivan R Corrêa, and Ira Schildkraut. The yeast scavenger decapping enzyme DcpS and its application for *in vitro* RNA recapping. *Scientific reports*, 9(1):1–9, 2019.

[62] SA Martin, E Paoletti, and B Moss. Purification of mRNA guanylyltransferase and mRNA (guanine-7-) methyltransferase from vaccinia virions. *Journal of Biological Chemistry*, 250(24):9322–9329, 1975.

[63] Fabian Muttach and Andrea Rentmeister. One-pot modification of 5′-capped RNA based on methionine analogs. *Methods*, 107:3–9, 2016.

[64] Pengjuan Zhang, Haohao Fu, Shuangli Du, Fengchao Wang, Jie Yang, Wensheng Cai, and Dingbin Liu. Click RNA for rapid capture and identification of intracellular microRNA targets. *Analytical chemistry*, 91(24):15740–15747, 2019.

[65] Andreas Schroeder, Odilo Mueller, Susanne Stocker, Ruediger Salowsky, Michael Leiber, Marcus Gassmann, Samar Lightfoot, Wolfram Menzel, Martin Granzow, and Thomas Ragg. The RIN: an RNA integrity number for assigning integrity values to RNA measurements. *BMC molecular biology*, 7(1):3, 2006.

[66] Montserrat Shelbourne, Xiong Chen, Tom Brown, and Afaf H El-Sagheer. Fast copper-free click DNA ligation by the ring-strain promoted alkyne-azide cycloaddition reaction. *Chemical Communications*, 47(22):6257–6259, 2011.

[67] Jeremy G Bird, Urmimala Basu, David Kuster, Aparna Ramachandran, Ewa Grudzien-Nogalska, Atif Towheed, Douglas C Wallace, Megerditch Kiledjian, Dmitry Temiakov, Smita S Patel, et al. Highly efficient 5′ capping of mitochondrial RNA with NAD^+ and NADH by yeast and human mitochondrial RNA polymerase. *Elife*, 7:e42179, 2018.

[68] Christina J Herrmann, Ralf Schmidt, Alexander Kanitz, Panu Artimo, Andreas J Gruber, and Mihaela Zavolan. Polyasite 2.0: a consolidated atlas of polyadenylation sites from 3′ end sequencing. *Nucleic acids research*, 48(D1):D174–D179, 2020.

[69] Heng Li, Jue Ruan, and Richard Durbin. Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome research*, 18(11):1851–1858, 2008.

[70] Artem Barski, Suresh Cuddapah, Kairong Cui, Tae-Young Roh, Dustin E Schones, Zhibin Wang, Gang Wei, Iouri Chepelev, and Keji Zhao. High-resolution profiling of histone methylations in the human genome. *Cell*, 129(4):823–837, 2007.

[71] Sarah Djebali, Carrie A Davis, Angelika Merkel, Alex Dobin, Timo Lassmann, Ali Mortazavi, Andrea Tanzer, Julien Lagarde, Wei Lin, Felix Schlesinger, et al. Landscape of transcription in human cells. *Nature*, 489(7414):101–108, 2012.

[72] Carrie A Davis, Benjamin C Hitz, Cricket A Sloan, Esther T Chan, Jean M Davidson, Idan Gabdank, Jason A Hilton, Kriti Jain, Ulugbek K Baymuradov, Aditi K Narayanan, et al. The encyclopedia of dna elements (ENCODE): data portal update. *Nucleic acids research*, 46(D1):D794–D801, 2018.

[73] Roger Volden, Theron Palmer, Ashley Byrne, Charles Cole, Robert J Schmitz, Richard E Green, and Christopher Vollmers. Improving nanopore read accuracy with the R2C2 method enables the sequencing of highly multiplexed full-length single-cell cDNA. *Proceedings of the National Academy of Sciences*, 115(39):9726–9731, 2018.

[74] Lindsey A Waddell, Lucas Lefevre, Stephen J Bush, Anna Raper, Rachel Young, Zofia M Lisowski, Mary EB McCulloch, Charity Muriuki, Kristin A Sauter, Emily L Clark, et al. ADGRE1 (EMR1, F4/80) is a rapidly-evolving gene expressed in mammalian monocyte-macrophages. *Frontiers in immunology*, 9:2246, 2018.

[75] Andrew J McKnight and Siamon Gordon. The EGF-TM7 family: unusual structures at the leukocyte surface. *Journal of leukocyte biology*, 63(3):271–280, 1998.

[76] Bo Yan, George Tzertzinis, Ira Schildkraut, and Laurence M Ettwiller. ReCappable Seq: Comprehensive determination of transcription start sites derived from all RNA polymerases. *BioRxiv*, page 696559, 2019.

[77] Nuala A O'Leary, Mathew W Wright, J Rodney Brister, Stacy Ciufo, Diana Haddad, Rich McVeigh, Bhanu Rajput, Barbara Robbertse, Brian Smith-White, Danso Ako-Adjei, et al. Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic acids research*, 44(D1):D733–D745, 2016.

[78] Alexander Payne, Nadine Holmes, Vardhman Rakyan, and Matthew Loose. BulkVis: a graphical viewer for Oxford nanopore bulk FAST5 files. *Bioinformatics*, 35(13):2193–2198, 2019.

[79] Hironori Adachi and Yi-Tao Yu. Pseudouridine-mediated stop codon readthrough in *S. cerevisiae* is sequence context–independent. *RNA*, 26(9):1247–1256, 2020.

[80] Wouter De Coster, Svenn D'Hert, Darrin T Schultz, Marc Cruts, and Christine Van Broeckhoven. NanoPack: visualizing and processing long-read sequencing data. *Bioinformatics*, 34(15):2666–2669, 2018.

[81] Heng Li. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics*, 34(18):3094–3100, 2018.

[82] Aaron R Quinlan and Ira M Hall. Bedtools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, 26(6):841–842, 2010.

[83] Heng Li, Bob Handsaker, Alec Wysoker, Tim Fennell, Jue Ruan, Nils Homer, Gabor Marth, Goncalo Abecasis, and Richard Durbin. The sequence alignment/map format and samtools. *Bioinformatics*, 25(16):2078–2079, 2009.

[84] Jason R Miller, Sergey Koren, and Granger Sutton. Assembly algorithms for next-generation sequencing data. *Genomics*, 95(6):315–327, 2010.

[85] James J Butzow and Gunther L Eichhorn. Interactions of metal ions with polynucleotides and related compounds. IV. Degradation of polyribonucleotides by zinc and other divalent metal ions. *Biopolymers: Original Research on Biomolecules*, 3(1):95–107, 1965.

[86] Qiaoyu Hu, Vindi M Jayasinghe-Arachchige, Joshua Zuchniarz, and Rajeev Prabhakar. Effects of the metal ion on the mechanism of phosphodiester hydrolysis catalyzed by metal-cyclen complexes. *Frontiers in chemistry*, 7:195, 2019.

[87] Eduardo Paredes and Subha R Das. Click chemistry for rapid labeling and ligation of RNA. *ChemBioChem*, 12(1):125–131, 2011.

[88] Ellen M Sletten and Carolyn R Bertozzi. From mechanism to mouse: a tale of two bioorthogonal reactions. *Accounts of chemical research*, 44(9):666–676, 2011.

[89] Xinghai Ning, Jun Guo, Margreet A Wolfert, and Geert-Jan Boons. Visualizing metabolically labeled glycoconjugates of living cells by copper-free and fast huisgen cycloadditions. *Angewandte Chemie International Edition*, 47(12):2253–2255, 2008.

[90] Temple F Smith, Michael S Waterman, et al. Identification of common molecular subsequences. *Journal of molecular biology*, 147(1):195–197, 1981.

[91] Rohan Lowe, Neil Shirley, Mark Bleackley, Stephen Dolan, and Thomas Shafee. Transcriptomics technologies. *PLOS Computational Biology*, 13(5):e1005457, may 2017.

[92] Ugrappa Nagalakshmi, Zhong Wang, Karl Waern, Chong Shou, Debasish Raha, Mark Gerstein, and Michael Snyder. The transcriptional landscape of the yeast genome defined by RNA sequencing. *Science (New York, N.Y.)*, 320(5881):1344–9, jun 2008.

[93] Ali Mortazavi, Brian A Williams, Kenneth McCue, Lorian Schaeffer, and Barbara Wold. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nature Methods*, 5(7):621–628, jul 2008.

[94] Ryan Lister, Ronan C. O'Malley, Julian Tonti-Filippini, Brian D. Gregory, Charles C. Berry, A. Harvey Millar, and Joseph R. Ecker. Highly integrated single-base resolution maps of the epigenome in Arabidopsis. *Cell*, 133(3):523–536, may 2008.

[95] Kin Fai Au, Vittorio Sebastiano, Pegah Tootoonchi Afshar, Jens Durruthy Durruthy, Lawrence Lee, Brian A Williams, Harm van Bakel, Eric E Schadt, Renee A

Reijo-Pera, Jason G Underwood, et al. Characterization of the human ESC transcriptome by hybrid sequencing. *Proceedings of the National Academy of Sciences*, 110(50):E4821–E4830, 2013.

[96] Donald Sharon, Hagen Tilgner, Fabian Grubert, and Michael Snyder. A single-molecule long-read survey of the human transcriptome. *Nature biotechnology*, 31(11):1009, 2013.

[97] Hagen Tilgner, Fereshteh Jahanbani, Tim Blauwkamp, Ali Moshrefi, Erich Jaeger, Feng Chen, Itamar Harel, Carlos D Bustamante, Morten Rasmussen, and Michael P Snyder. Comprehensive transcriptome analysis using synthetic long-read sequencing reveals molecular co-association of distant splicing events. *Nature biotechnology*, 33(7):736, 2015.

[98] Michael B Clark, Tomasz Wrzesinski, Aintzane B Garcia, Nicola AL Hall, Joel E Kleinman, Thomas Hyde, Daniel R Weinberger, Paul J Harrison, Wilfried Haerty, and Elizabeth M Tunbridge. Long-read sequencing reveals the complex splicing profile of the psychiatric risk gene CACNA1C in human brain. *Molecular psychiatry*, 25(1):37–47, 2020.

[99] Andrew M Smith, Robin Abu-Shumays, Mark Akeson, and David L Bernick. Capture, Unfolding, and Detection of Individual tRNA Molecules Using a Nanopore Device. *Frontiers in bioengineering and biotechnology*, 3:91, 2015.

[100] Sam EV Linsen, Elzo de Wit, Georges Janssens, Sheila Heater, Laura Chapman, Rachael K Parkin, Brian Fritz, Stacia K Wyman, Ewart de Bruijn, Emile E Voest, et al. Limitations and possibilities of small RNA digital gene expression profiling. *Nature methods*, 6(7):474–476, 2009.

[101] Fanglei Zhuang, Ryan T Fuchs, Zhiyi Sun, Yu Zheng, and G Brett Robb. Structural bias in T4 RNA ligase-mediated 3′-adapter ligation. *Nucleic acids research*, 40(7):e54–e54, 2012.

[102] Felix Grünberger, Robert Knüppel, Michael Jüttner, Martin Fenk, Andreas Borst, Robert Reichelt, Winfried Hausner, Jörg Soppa, Sébastien Ferreira-Cerca, and Dina Grohmann. Exploring prokaryotic transcription, operon structures, rRNA maturation and modifications using Nanopore-based native RNA sequencing. *bioRxiv*, page 2019.12.18.880849, may 2020.

[103] Olivia S Rissland, Andrea Mikulasova, and Chris J Norbury. Efficient RNA polyuridylation by noncanonical poly(A) polymerases. *Molecular and cellular biology*, 27(10):3612–24, may 2007.

[104] R. L. Stratonovich. Conditional Markov Processes. *Theory of Probability & Its Applications*, 5(2):156–178, jan 1960.

[105] Leonard E. Baum and J. A. Eagon. An inequality with applications to statistical estimation for probabilistic functions of Markov processes and to a model for ecology. *Bulletin of the American Mathematical Society*, 73(5):360–363, sep 1967.

[106] Jared T Simpson, Rachael E Workman, P C Zuzarte, Matei David, L J Dursi, and Winston Timp. Detecting DNA cytosine methylation using nanopore sequencing. *Nature Methods*, 14(4):407–410, apr 2017.

[107] Jacob Schreiber and Kevin Karplus. Analysis of nanopore data using hidden Markov models. *Bioinformatics*, 31(12):1897–1903, 2015.

[108] Diana F Colgan and James L Manley. Mechanism and regulation of mRNA polyadenylation. *Genes & development*, 11(21):2755–2766, 1997.

[109] Aimee L Jalkanen, Stephen J Coleman, and Jeffrey Wilusz. Determinants and implications of mRNA poly (A) tail size–does this protein make my tail look big? In *Seminars in cell & developmental biology*, volume 34, pages 24–32. Elsevier, 2014.

[110] Nicholas J Loman, Joshua Quick, and Jared T Simpson. A complete bacterial genome assembled *de novo* using only nanopore sequencing data. *Nature Methods*, 12(8):733–735, aug 2015.

[111] A. Viterbi. Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. *IEEE Transactions on Information Theory*, 13(2):260–269, apr 1967.