Title

Using polyinosine tailing to capture RNA 3' ends by Direct-RNA Nanopore sequencing

Author

Vo, Jenny

Publication Date

2020

UNIVERSITY OF CALIFORNIA
SANTA CRUZ

USING POLYINOSINE TAILING TO CAPTURE RNA 3' ENDS BY DIRECT-RNA NANOPORE SEQUENCING

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

MASTER OF SCIENCE

in

MOLECULAR, CELL AND DEVELOPMENTAL BIOLOGY

by

Jenny Mai Vo

June 2020

The Thesis of Jenny Mai Vo is
approved:

Professor Manuel Ares Jr., chair

Professor Joshua Arribere

Professor Jeremy Sanford

Quentin Williams
Acting Vice Provost and Dean of Graduate Studies

Table of Contents

List of Figures and Tables

Abstract

Jenny Mai Vo

Using polyinosine tailing to capture RNA 3’ ends by nanopore sequencing

Transcriptome is a word that describes the information contained in all the RNAs in a cell or cell population. Capturing the transcriptome requires RNA sequencing, giving a snapshot on the biological processes happening in the organism or cell. A popular sequencing platform for sequencing the transcriptome is Next Generation Sequencing (NGS) from Illumina, a short-read, RT-PCR based system that sequences RNA sample through amplified cDNA which can cause misrepresentation of the starting RNA population. An alternative to NGS is Direct-RNA Nanopore Sequencing from Oxford Nanopore Technologies (ONT), a long-read sequencing platform that directly sequences the entire length of the RNA transcript. The standard library preparation protocol for direct RNA sequencing is focused on capturing RNAs containing a poly(A) tail. Sequencing of non-poly(A) RNAs is possible through use of custom adaptors designed for a conserved 3' sequence, which limits the population of RNAs that can be sequenced in a sample or by adding poly(A) using poly(A) polymerase which confounds the native 3' end sequence of RNAs. Here, I describe a new technique for Direct-RNA Nanopore sequencing, which modifies the 3′ ends of RNAs with inosine-tails allowing for sequencing of diverse RNA samples. My work demonstrates the versatility and potential of applying the poly(I) sequencing method to diverse transcriptomes, providing a new sequencing method that may enable new RNA discoveries.

Dedication or Acknowledgements

I would like to thank Dr. Manny Ares, for the opportunity to do research, and the mentorship and guidance to help me learn and grow closer to becoming a real scientist. I would also like to thank Dr. Mark Akeson, for giving me the resources to fund the project and I would also like to thank my thesis committee, Dr. Josh Arribere and Dr. Jeremy Sanford for their input and feedback on my thesis. I would especially thank the people who have contributed to this project, Logan Mulroney and Dr. Jen Quick-Cleveland who were not just collaborators on this project but also great mentors with their patience and guidance. Logan especially, the biggest contributor to this project, which would not exist without his discovery, great ideas and his hard work has helped move the project forward. He has mentored me since the beginning and has always given the time to help me no matter what, for which I am truly thankful. I would also like to thank my lab mates who are not just my colleagues, but also my friends. They have been there for me intellectually and emotionally, and I have learned a great deal from each of them. Their support has helped with my achievements and my confidence, and my relationships with them has had a significant positive impact on me. I would lastly like to thank everyone who has supported me, whom all have let me know of my worth and capabilities. They have helped me in all aspects of my life, and I would not be where I am today without each of them.

Introduction

The transcriptome consists of all the RNA molecules within a cell or cell population, which has an abundance of information that reflects the state and regulation of processes within the cell. By elucidating the expression, regulation, and function of genes, we can better understand the biology of an organism and disease (1).

A common method of studying the transcriptome is through Next Generation Sequencing (NGS) by Illumina, which analyzes RNA through sequencing-bysynthesis (SBS). SBS synthesizes new strands of modified cDNA through multiple cycles of nucleotide addition on a flowcell using 3' blocked fluorescently labeled nucleotides, allowing for detection of fluorescence for the canonical nucleotides A, C, G, and T, each attached to a different dye. The color of the fluoresce indicates the base being added, which is recorded with each cycle to determine the sequence of a DNA molecule, creating a "read". The core steps of preparing RNA for sequencing using NGS includes fragmentation of long RNA transcripts, ligation of adapters containing unique barcodes for differentiating RNA samples on the 5' and 3' ends of the RNA for hybridization onto the flowcell, reverse transcription, PCR, optional unique molecular identifiers for managing PCR bias, and subsequent wash steps for excess adapter removal from the final library (2).

NGS has a low error rate in sequence determination and has a high throughput of reads. NGS has relatively low cost by allowing for multiple samples to be sequenced in the same run by the use of barcodes for each RNA sample (3).

Most RNA transcripts are at least 500nt or longer, but no more than ~300 bases can be read for NGS, requiring fragmentation in order to cover most of the transcript. Computational reconstruction of the transcript is done by aligning the fragments to a reference genome. This can introduce difficulties in interpretation, such as identifying alternative splicing events, ambiguities in isoform identification, quantification, and mapping to the genome. Artifacts like template switching, where a synthesized DNA strand re-associates onto a different template, can produce chimeric cDNAs that do not represent any RNA in the sample (4). During reverse transcription, the reverse transcriptase may jump to a different part of the RNA strand leading to false identification of gene fusion events and spliced isoforms. The reverse transcriptase is also unable to transfer most RNA modification information onto the cDNA strand, or is actually inhibited by RNA modifications, preventing study of many RNA modifications. PCR also introduces biases by preference in amplification of shorter products resulting in number of reads not correlating to actual abundance of transcripts (2-4).

Direct-RNA sequencing from Oxford Nanopore Technologies (ONT) allows for sequencing of the native RNA without cDNA or amplification, which eliminates many biases inherent in NGS technologies, and preserves information about each RNA transcript in the raw current measurements. Direct-RNA Nanopore sequencing also allows for sequencing of the full length of RNA transcripts. Sequencing the full length of an RNA transcript is advantageous for studying the transcriptome. For example, isoforms are transcribed from the same chromosomal locus as their canonical counterparts but have different transcription start sites by alternative RNA processing. ONT's full length sequencing allows for better interpretation of different transcription start sites compared to NGS, as fragmentation generates many read ends which can obscure these sites. Sequencing the full length of the RNA is important for detecting these isoforms and eases the proper alignments of these reads to the genome.

In Nanopore sequencing, negatively charged polynucleotides are driven towards the nanopores embedded onto a membrane through an electrical potential (5), and are translocated using a protein motor that controls the rate of movement using ATP hydrolysis. As the RNA moves through the pore, a string of 5-6 nucleotides (kmer) occupies the narrowest part of the pore at a given time (6). Each k-mer produces its own unique signal, and changes in the electrical current are produced by each kmer as each passes through the pore in sequence. Base-calling programs use these electric current measurements to interpret the sequence of the molecule based on each k-mer measurement. Since ONT's Direct-RNA sequencing allows for sequencing of the native RNA without cDNA or amplification, RNA modifications can be detected and the population of genes can be determined by the number of reads. Direct-RNA sequencing has been used by several groups to investigate properties of RNA, for example the detection of modifications (7-9), structure analysis (10, 11), and discovery of novel isoforms (12, 13), and estimation of poly(A) tail length (8).

One limitation of ONT’s standard direct-RNA sequencing library preparation method is that it requires a poly(A) enriched sample and a poly(T) adapter for RNA capture at the 3’ ends, so that only the polyadenylated or poly(A)+ fraction of the transcriptome can be observed (5). Some groups have used different techniques to bypass this limitation, such as polyadenylation of non-adenylated RNA (14) and creation of a custom adapter for the conserved 3' end sequence of RNA (7). Despite these adaptations, a more convenient protocol would have a generalized approach for capturing RNA species independent of their 3' ends.

I have developed a new technique that does not require prior knowledge of 3' end sequences by the usage of a universal adapter for all RNA samples and the method does not obscure the native 3' end sequence. My technique utilizes Cid-1, a poly(U) polymerase that can create poly-inosine tails at the 3’ ends of many RNAs that lack a poly(A) tail. Due to chemical differences between inosine and the four common nucleobases, I can distinguish the inosine homopolymer signal from that of native RNA sequences. So far, 3' tails of non-canonical nucleotides are not known to exist in nature, enabling the native 3' end of every sequenced RNA molecule to be determined. Inosine is a derivative of guanine lacking a 2-amino group, making it a good candidate for universal capture with a poly(C) adapter. I have optimized and characterized the I-tailing activity of Cid-1 on polyadenylated and nonpolyadenylated RNAs and found that the poly(I) method has equivalent ability to capture mRNA to ONT’s standard poly(A) sequencing, while also capturing other RNA populations in total RNA and ribosomal RNA depleted samples. I have successfully sequenced chromatin-associated RNA via RNA Polymerase II capture and have been able to detect nascent and processed RNA still bound to the polymerase. The poly(I) method has demonstrated versatility in sequencing a diverse population of RNA molecules and enhances the potential for expanding the scope of RNA discoveries.

Results:

Cid-1 poly(U) polymerase can efficiently add inosines to the 3' ends of different RNA molecules

To develop methods for adding modified nucleotides to the 3' ends of RNA I tested commercial preparations of two well-studied enzymes, the S. pombe poly(U) polymerase Cid-1 (15), and S. cerevisiae poly(A) polymerase (PAP) (16). I first determined that these commercial preparations had the expected nucleotide adding specificities by incubating them with various rNTPs and a 24-nucleotide oligomer of adenosine (A24) under the reported conditions (15, 16). This test (Fig 1A and 1B) showed that each commercial preparation behaved as originally described. Briefly, Cid-1 from New England Biolabs efficiently added U and A, but only poorly added C or G (Fig 1A), in agreement with previous work (15). PAP from ThermoFisher efficiently added A, and to a much lesser extent G, but only poorly added U or C, as observed previously (16). I also examined the ability of each enzyme to use inosine triphosphate (ITP, a purine similar to G) to make poly-inosine tails (Figure 1). I found that Cid-1 added long stretches of inosine to A24 (Fig 1A, lane 6) whereas PAP only added a few residues (Fig 1B, lane 6). I conclude that, the commercial preparations faithfully reproduce the published activity of each enzyme using the four standard rNTPs (15, 16). Furthermore Cid-1 efficiently added the modified base inosine to the A24 substrate, whereas PAP only did so inefficiently.

To test these enzymes with substrates more representative of those found in natural samples, I prepared model mRNAs based on MYL6 mRNA from human with

(MYL6(A+)) or without (MYL6(A-)) a pre-existing 40 nucleotide poly(A) tail, to represent polyadenylated mRNAs and non-polyadenylated RNAs. The respective 476nt and 407nt length of this RNA allowed for ease in observation of poly(I) tailing on polyacrylamide gels, and the 40nt poly(A) tail length is within range of 40-60nt median poly(A) tail length estimates for yeast (17). As expected, Cid-1 used U and A to make long tails on both model RNAs (Fig 1C). Unexpectedly, Cid-1 uniformly added about 50 inosine residues to a majority of MYL6(A+) molecules, but also produced some molecules with long (>200 nt) tails (Fig 1C, lane 6). In contrast, Cid-1 only inefficiently added long tails (i. e. it appears to act processively on just a few molecules of MYL6(A-)) (Fig 1C, lane 9).

In comparison, PAP efficiently added poly(A) to both MYL6(A+) and MYL6(A-) RNAs, but only poorly added U to either (Fig 1D, compare lanes 3 and 4 to lanes 6 and 7). The presence of a poly(A) tail on MYL6(A+) appeared to allow PAP to add short heterogenous stretches of I (Fig1D , lane 5) but without a poly(A) tail MYL6(A-) was much less efficiently used (Fig 1D, lane 8). When I tested Cid-1 and PAP using a shorter model mRNA with 44A residues (Gluc200(A+)) or without a poly(A) tail (GLuc200(A-)), I saw similar results with A and U, but Cid-1 did not add a +50 inosine tail to this RNA (Suppl Fig 1A, lane 5), similar to the result with A24 (Fig 1A, lane 6), suggesting that an RNA of a certain size may be necessary for producing the +50 product. I conclude that the presence of a pre-existing poly(A) tail significantly promotes the use of an RNA as a substrate for both enzymes, except that

RNAs below a certain size or with other as yet unidentified features may be poor substrates.

Given that Cid-1 added a +50 I-tail on a model mRNA but inefficiently added I-tails on complex RNAs (Fig 1C lane 5 and 8, Suppl Fig 1 lane 8), I tested the ability of Cid-1 to add modified U residues to a model mRNA (Fig 1E). I considered that its native uridylation activity might favor the incorporation of modified Us on a wider variety of RNAs. I set up reactions with modified-UTPs alone, and with a mixtures of modified-UTP and UTP. Cid-1 was unable to use 2'-O-methyl-UTP whether or not UTP was added in the reaction (Fig1E, lane 3 and 4), suggesting that it is a competitive inhibitor of UTP addition, likely occupying the binding site but blocking catalysis. Surprisingly, Cid-1 also did not use ΨTP, but worked well with a mixture (Fig1E, lanes 5 and 6). Similar results were found with NmeΨTP and NmeΨTP:UTP (Fig 1E, lanes 11 and 12), suggesting that these modified UTPs do not bind well to the UTP binding site and thus do not act as competitive inhibitors. Both 5meUTP and 4thioUTP can be used by Cid-1, where it produced long tails with or without the addition of UTP (Fig1E lanes 7, 8, 9, 10). It is surprising that ΨTP and NmeΨTP appeared to not to be used by Cid-1 given that it used 5meUTP. All three modified bases differ at the same position in the base (in pseudouridine the N1 position occupies the same location as the C5 of uridine). The enzyme therefore seemed sensitive to the flipped orientation of pseudouridine analogs. I conclude that the ability for Cid-1 to use certain modified UTPs to add polymers of uncommon nucleotides to the ends of natural RNAs illustrates additional potential for development of the methodology for nanopore RNA sequencing.

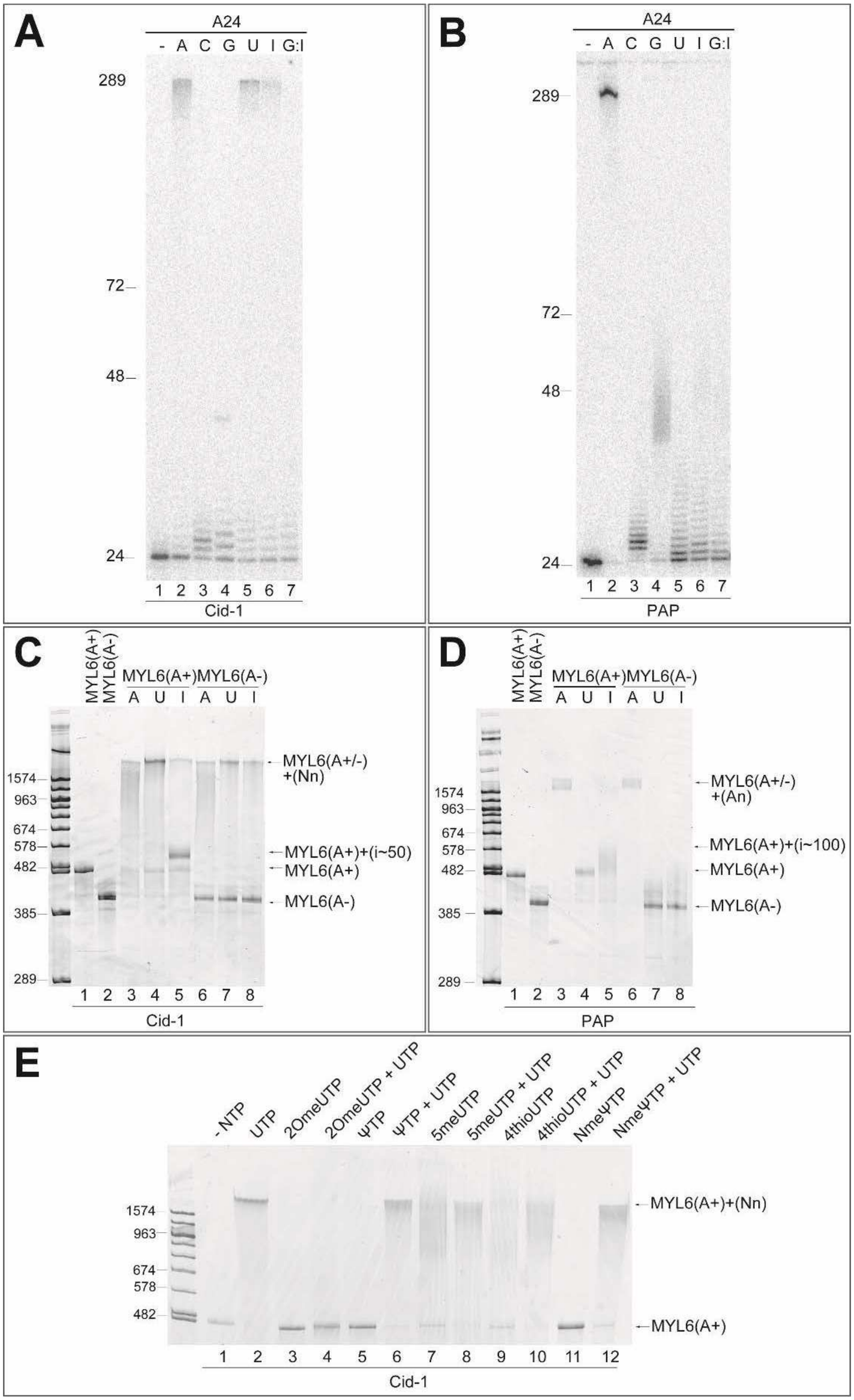

A
A24
- A C G U I G:I
289
72
48
24
1 2 3 4 5 6 7
Cid-1
B
A24
- A C G U I G:I
289
72
48
24
1 2 3 4 5 6 7
PAP
C
MYL6(A+)
MYL6(A-)
MYL6(A+) MYL6(A-)
A U I A U I
1574
963
674
578
482
385
289
MYL6(A+/-) +(Nn)
MYL6(A+)+(i~50)
MYL6(A+)
MYL6(A-)
1 2 3 4 5 6 7 8
Cid-1
D
MYL6(A+)
MYL6(A-)
MYL6(A+) MYL6(A-)
A U I A U I
1574
963
674
578
482
385
289
MYL6(A+/-) +(An)
MYL6(A+)+(i~100)
MYL6(A+)
MYL6(A-)
1 2 3 4 5 6 7 8
PAP
E
-NTP
UTP
2OmeUTP
2OmeUTP + UTP
ΨTP
ΨTP + UTP
5meUTP
5meUTP + UTP
4thioUTP
4thioUTP + UTP
NmeΨTP
NmeΨTP + UTP
1574
963
674
578
482
MYL6(A+)+(Nn)
MYL6(A+)
1 2 3 4 5 6 7 8 9 10 11 12
Cid-1

Figure 1: A test of commercial preparations of two enzymes for their ability to add nucleotide homopolymers to the 3’ end of different RNAs. Incorporation of rNMPs onto the 3' end of ^{32}P-5' end-labeled A24 RNA using (Panel **A**) Cid-1 Poly(U) polymerase from New England Biolabs and (Panel **B**) Poly(A) polymerase from ThermoFisher. Incorporation of AMP, UMP, and IMP onto the 3' ends of MYL6(A+) and MYL6(A–) using (Panel **C**) Cid-1 Poly(U) Polymerase from New England Biolabs and (Panel **D**) Poly(A) Polymerase from ThermoFisher. (Panel **E**) Using Cid1, incorporation of modified-UTPs and mixtures of modified-UTPs and UTP onto the 3' ends of MYL6(A+)

Optimization of the Cid-1 inosine tailing reaction for Nanopore library preparation

Based on our results to this point, I decided to focus our effort on the enzyme I thought would be most broadly useful in tailing a complex natural RNA sample with inosine for analysis in the Nanopore. Although PAP was able to add a short heterogenous inosine tail to poly(A)+ RNAs, the uniform +50 inosine tails added to most poly(A)+ RNA molecules by Cid-1 seemed to be an advantage in creating representative libraries for direct RNA sequencing in nanopore systems. Neither enzyme seemed particularly good at adding inosine distributivity to all poly(A)– RNAs in a sample. However, Cid-1 was able to add very long tails to a larger fraction of poly(A)– molecules, whereas PAP may be able to add a few inosines to a small fraction of molecules. Based on this I pursued optimization and validation of the inosine tailing reaction catalyzed by Cid-1.

To evaluate the capacity of the Cid-1 reaction to modify amounts of RNA necessary for Direct-RNA Nanopore library construction, I incubated MYL6(A+) with Cid-1 and ITP under different conditions, and measured production of the ~50 Itailed product (Fig 2A and B). I found that Cid-1 reacted immediately after being introduced to the reaction mixture ("0" time) and the reaction was complete by 40 minutes. Under optimal conditions, 1.2 pmol of RNA ends was reacted with 2 units of Cid-1 from NEB (Fig 2B) to convert nearly all input molecules to the ~+50 form (Fig 2B).

To confirm for future studies that this property is a feature of native Cid-1 and not generated by an unknown step in the commercial preparation of the enzyme, I cloned, expressed and FPLC purified a recombinant Cid-1 truncated 26 nucleotides from the C-terminus (vCID1) in E. coli and found that the addition of ~50 inosines to MYL6(A+) is a natural property of the enzyme (Figure 2C, Suppl Figure 2). I conclude that Cid-1 has the potential for use in Direct-RNA Nanopore sequencing libraries for adding modified homopolymers at the 3' ends.

To precisely measure the length of the ~+50 extension of RNA catalyzed by Cid-1 I labeled various inosine-tailed (and untailed control) molecules at their 3' ends with 32P-pCp using RNA ligase, and then digested them with RNAseA, which cuts only after pyrimidines, leaving homopurine polymers like the poly(A) tail, or a poly(I) tail intact (Figs 2D, E and F). For example, MYL6(A+) contains two Gs at its 3' end to which the 40 nucleotide poly(A) tail is added, creating an RNAseA resistant product that is exactly 42 nucleotides in length (Fig 2D). I measured the RNAseA digested I-tailed MYL6(A+) product and estimated that the unreacted MYL6(A+) tail is 43 nucleotides (including the ligated 32P-pCp -labeled residue) long, whereas the Itailed MYL6(A+) is calculated to be 93.4 nucleotides. Thus, the calculated difference of 50.4 nucleotides indicates that most mRNA molecules carry 50 inosine residues after the Cid-1 reaction.

The non-polyadenylated RNAs MYL6(A–) and yeast 5.8S rRNA did not acquire a uniform 50 nt poly(I) extension; rather, a heterogeneous long I-tail was added (Fig 2C and data not shown). Accordingly, RNaseA digestion of 32P-pCp labeled Cid-1 product generated a ladder of RNAse-resistant products that extended far up the gel. I conclude that poly(A)+ RNAs generally acquire a uniform 50 nt inosine extension whereas the non-polyadenylated RNAs in a sample acquire inosine tails of a wide variety of lengths from a few to >1000 nt (Figs 2D and F).

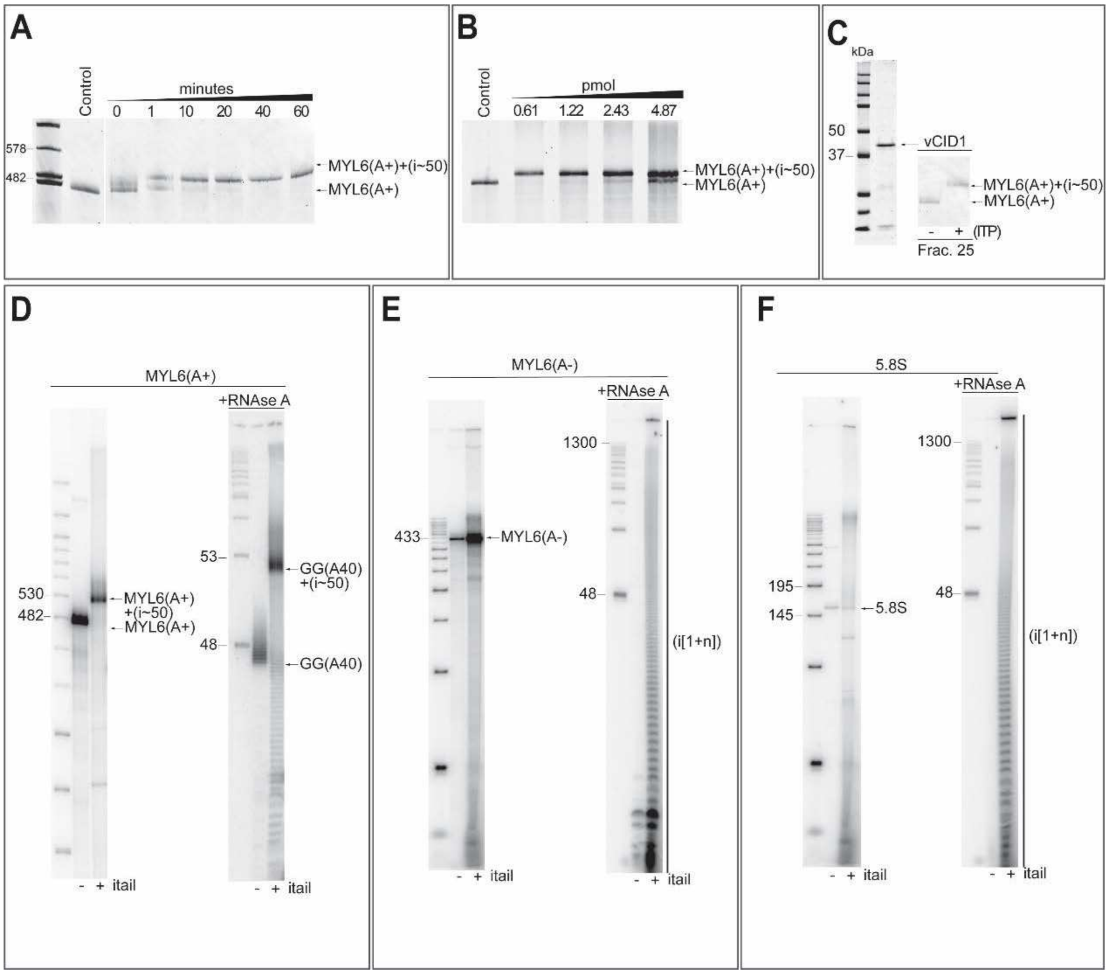

A
Control
minutes
0
1
10
20
40
60
578
482
MYL6(A+)+(i~50)
MYL6(A+)
B
Control
pmol
0.61
1.22
2.43
4.87
MYL6(A+)+(i~50)
MYL6(A+)
C
kDa
50
37
vCID1
MYL6(A+)+(i~50)
MYL6(A+)
- + (ITP)
Frac. 25
D
MYL6(A+)
+RNAse A
530
482
MYL6(A+)
+(i~50)
MYL6(A+)
53
48
GG(A40)
+(i~50)
GG(A40)
- + itail
- + itail
E
MYL6(A-)
+RNAse A
433
MYL6(A-)
1300
48
(i[1+n])
- + itail
- + itail
F
5.8S
+RNAse A
195
145
5.8S
1300
48
(i[1+n])
- + itail
- + itail

Figure 2. Cid-1 can uniformly tail microgram amounts of poly(A) RNA with ~50 inosine residues. Using MYL6(A+) (Panel **A**) Time-course of 0-60 minutes for the rate of the reaction (Panel **B**) and input titration. (Panel **C**) SDS-PAGE of purified vCID1 and I-tailing activity on MYL6(A+) (Panels **D-F**) RNase A digestion of inosine-tailed and untailed ^{32}P-pCp-labeled MYL6(A+), MYL6(A-) and 5.8S.

Inosine tails generate a distinct signal in the nanopore during sequencing

Adapting an RNA molecule for direct RNA sequencing by nanopore requires hybridization of the target RNA to a DNA splint, followed by ligation of the RNA 3' end to one strand of the DNA adaptor (8). To capture inosine-tailed RNA molecules for analysis in the nanopore, I used a custom adaptor with a poly(C) segment of 10 residues that would base pair with the inosine tail, to promote ligation of the RNA 3' end to the nanopore sequencing adaptor (Fig 3A). This allowed us to obtain signals from inosine-tailed molecules by direct nanopore sequencing (Fig 3B and C).

I hypothesized that tailing with non-standard (other than A, G, C, or U) RNA nucleotides might produce electronic signals in the pore that would differ from those produced by polymers of the standard nucleotides, in particular the poly(A) tails found on many natural RNAs. To determine whether this is the case, I modified ONT’s standard Direct-RNA Nanopore sequencing method using the custom cytosine splint adapter ligation (Fig 3A) and created sequencing libraries from synthetic and natural RNA samples. To identify distinct poly(I) signals appended to signals from a known RNA sequence, I prepared control samples of GLuc200(A+) and GLuc200(A) with or without a splint-ligated 30 nt poly(I) homopolymer (see Methods), created libraries using the appropriate adaptor oligonucleotide, then sequenced the libraries in the Nanopore using poly(C) adapters. The raw current trace of a single representative molecule from each library is shown in (Fig 3B). Direct RNA sequencing in the ONT nanopore format used here threads the 3' end of the RNA into the pore first, and the current (in picoamperes) across the pore is shown on the y-axis with time displayed on the x-axis, thus tracing the current as the molecule transits the pore from the 3' to the 5' end over time. For convenience the part of the trace corresponding to the adaptor is in gray (segment I), and the GLuc200 sequence is in light blue (segment IV), poly(A) is in purple (segment III) and poly(I) is in dark blue (segment II). The top panel shows GLuc200(A-) with no I-tail, showing that after the adaptor sequence, a trace consistent with an RNA of complex sequence follows immediately. GLuc200(A+) on the other hand shows a monotonic signal at slightly above 100 pA in this experiment before the appearance of the complex sequence trace. The inosinetailed molecule Gluc200(A-)i30 also shows a monotonic signal between the adaptor and the complex sequence trace at just above 100 pA, yet this appears distinct from that observed on Gluc200(A+). This distinction is even more clear in the Gluc200(A+)i30 molecule to which the inosine tail is added 3' to the poly(A) tail: following the adaptor the monotonic inosine signal is followed by the monotonic poly(A) signal, and then finally the complex sequence trace of Gluc200. Although the poly(I) and poly(A) signals have a similar mean current of just above 100 pA, the variation in mean current is substantially different, with poly(I) currents wavering across an approximately 20 pA window, and poly(A) providing more narrow variation. In addition, there is a peculiar drop (marked as an arrow) in the current amplitude during the transition from the poly(I) to the poly(A) homopolymeric segments.

To explore how the distinct poly(I) signal may be useful in characterizing biological samples, I tailed poly(A) RNA enriched samples of yeast BY4741 RNA and prepared libraries using the poly(C) adaptor. For comparison, I made standard libraries for poly(A) RNAs using the standard “RTA adapter” provided by ONT. Shown as an example is the TDH3 gene and 25S rRNA from Saccharomyces cerevisiae. When comparing the raw current traces of the poly(I) tailed and the poly(A) reads, a distinctive raw current segment is observed at the transition of the 3' end of the poly(A) homopolymer signal and the adapter that is not seen in any of the raw current traces of the standard poly(A) libraries. (Figure 3C). The placement of the new signal indicates that this is a modification of the RNA from the incorporation of the inosine homopolymer. The poly(I) signal is observed to have the same mean current amplitude but has more variation in a broader range of current amplitudes of the poly(A) signal, and there is a peculiar drop in the current amplitude in the transition of the poly(I) to the poly(A) signal found in all the reads of the poly(I) libraries. I conclude that an inosine homopolymer has a distinct trace on the raw current signal that can be distinguished from the poly(A) tail, therefore not obscuring the length of the pre-existing poly(A) tail and also differentiating poly(A) RNA from non-poly(A) RNA in sequencing data.

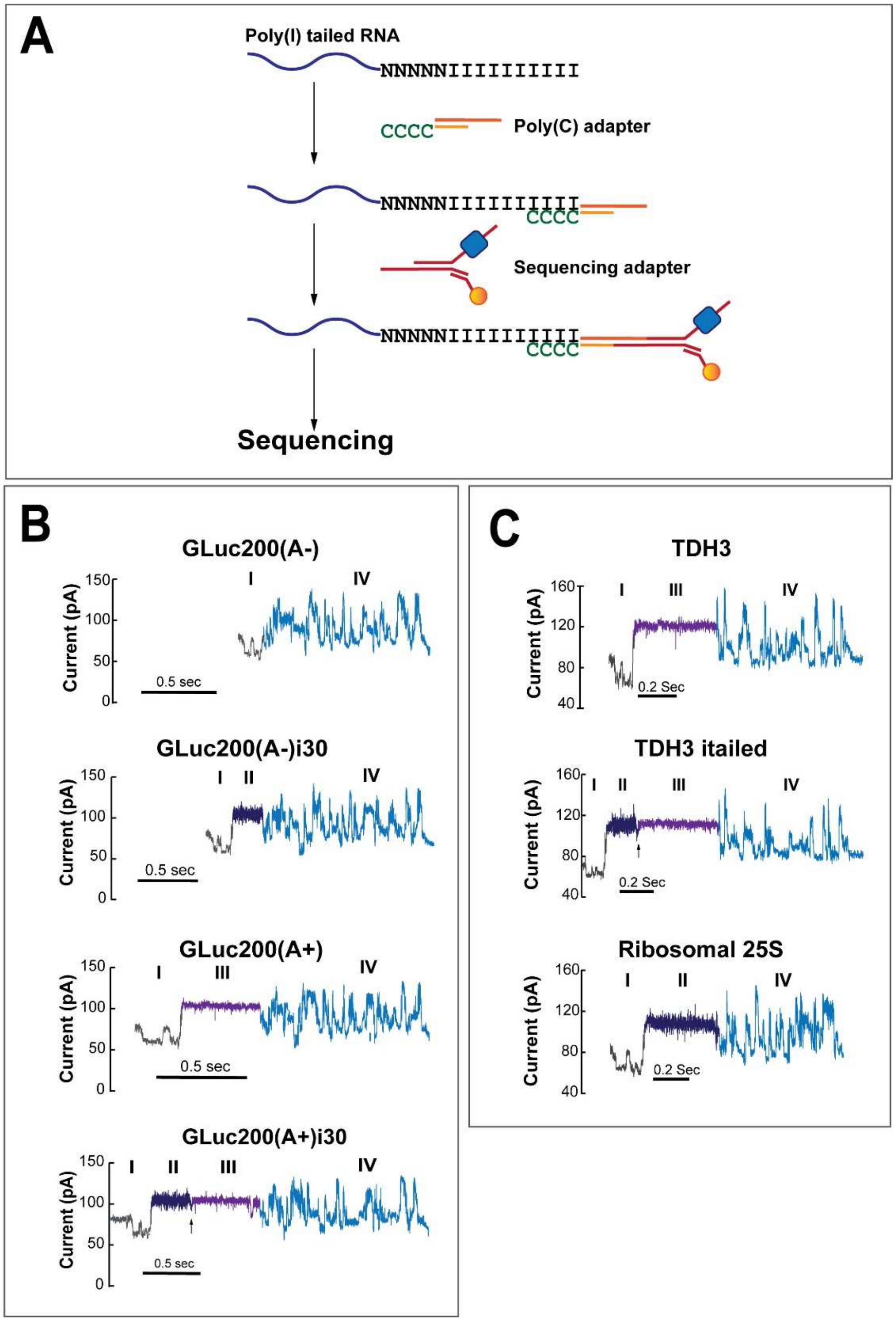

A
Poly(I) tailed RNA
NNNNNIIIIIIIIIII
CCCC
Poly(C) adapter
NNNNNIIIIIIIIIII
CCCC
Sequencing adapter
NNNNNIIIIIIIIIII
CCCC
Sequencing
B
GLuc200(A-)
Current (pA)
150
100
50
0
0.5 sec
I
IV
GLuc200(A-)i30
Current (pA)
150
100
50
0
0.5 sec
I
II
IV
GLuc200(A+)
Current (pA)
150
100
50
0
0.5 sec
I
III
IV
GLuc200(A+)i30
Current (pA)
150
100
50
0
0.5 sec
I
II
III
IV
C
TDH3
Current (pA)
160
120
80
40
0.2 Sec
I
III
IV
TDH3 itailed
Current (pA)
160
120
80
40
0.2 Sec
I
II
III
IV
Ribosomal 25S
Current (pA)
160
120
80
40
0.2 Sec
I
II
IV

Figure 3. Inosine tails generate a distinct signal in the nanopore during sequencing. (Panel **A**) Library preparation method for poly(I) sequencing. (Panels **B-C**): Example current traces with signals **I.** Sequencing Adapter **II.** Poly(I) **III.** Poly(A) **IV.** 3' native sequence of the RNA (Panel **B**) Raw current traces of control samples Gluc200 and GLuc200A44 containing unligated and ligated 30nt inosine homopolymer tail (Panel **C**) Example traces of mRNA TDH3 found in poly(A) and poly(I) sequencing, and non-adenylated 25S RNA in poly(I) sequencing.

Detection and abundance estimates for poly(A) mRNAs using I-tailing are equivalent to the standard library method

To evaluate that the poly(I) method has bias that is distinct from direct-RNA nanopore sequencing with respect to ONT's standard poly(A), I did a regression analysis comparing the gene coverage of poly(A) controls and poly(I)-tailed poly(A) (poly(A)poly(I)) libraries of yeast BY4741 RNA using their respective reads-permillion (RPM), and normalized reads-per-million without ribosomal RNA reads (RPM -rRNA). To first determine the reproducibility between poly(A) control samples (Control 1, 2, 3), regression analysis of the (RPM) and (RPMs -rRNA) of each pairing (Control 1 vs Control 2, Control 1 vs Control 3, Control 2 vs Control 3) were done for complete comparisons (Figure 4A). The plot indicates a R^2 of 0.957 for both (RPM) and (RPM -rRNA) of Control 1 vs Control 2, with a Spearman's Correlation of 0.95 (Figure 4A). When comparing Control 1 to Control 2, there is a R^2 of 0.961 for both (RPM) and (RPM -rRNA) with a Spearman's Correlation of 0.9527. For samples that contain the same poly(A) enrichment preparations (Control 2 and Control 3), there is a R^2 of 0.998 and a Spearman's Correlation of 0.9916 for both (RPM) and (RPM -rRNA). In conclusion, ONT's standard poly(A) library preparation is very reproducible when applied repeatedly to the same polyA-selected sample and can detect small differences between different poly(A) sample preparations. It appears unaffected by the presence of small amount of rRNA because removing rRNA reads from the analysis has little effect on the correlation coefficients.

To evaluate the reproducibility of the poly(I) method, two samples of poly(A)poly(I) with different poly(A) isolation preparations were compared (Fig 4B). The RPM of both samples were found to have a correlation at R^2 at 0.918 and 0.967 for (RPM) and (RPM -rRNA) respectively with a Spearman Correlation of 0.904 and 0.905. The values for the rRNAs are situated at the bottom of the diagonal, indicating that there are more rRNAs in the poly(A)poly(I) #4 than in the poly(A)poly(I) #6. The significant change of the R^2 when the rRNA reads are omitted indicate that more rRNAs are captured in the poly(I) method in comparison to ONT's standard poly(A), and the high value of 0.967 indicates that the runs are similar and comparable to the poly(A) controls in terms of reproducibility.

To examine the comparability of the poly(I) method to ONT's standard poly(A) method, reads pooled from poly(A) Control samples were compared with reads pooled from the poly(A)poly(I) samples (Fig 4C). An R^2 of 0.0618 and high Spearman Correlation of 0.9486 was found when comparing the RPM values of both pools. However, when comparing the (RPM -rRNA) of both sample types I see that the R^2 increases to 0.959 and the Spearman Correlation increases slightly to 0.955. The rRNA reads captured in the poly(A)poly(I) RNA has a dramatic effect on the overall correlation of the samples, but an only slightly changed Spearman's Correlation indicates that there is still a positive association among the non-rRNAs. This indicates that the poly(I) method allows for the capture of significantly more rRNA than ONT's poly(A) method for sequencing, as expected.

To compare poly(I) sequencing with to ONT's poly(A) sequencing method, libraries from the same samples were compared (Figure 4C). Similarly to Figure 4B, there is an extremely low R^2 of 0.0147 in the (RPM) when all the genes are accounted for but an R^2 of 0.92 was found with (RPM -rRNA) with little change in the Spearman's correlation at 0.9213 and 0.9275 respectively.

Since poly(I) sequencing decreases the throughput of the reads compared to ONT's poly(A) (Appendix Sequencing Table), I wanted to see if the number of reads obtained in poly(A)poly(I) sequencing runs had an effect on reproducibility. The poly(A)poly(I) runs that obtained the most reads were pooled together and analyzed with the pooled poly(A) controls (Figure 4D). Many rRNA reads were obtained the poly(A)poly(I) pool, as indicated by the R^2 value of 0.0688 of the (RPM) comparisons, and 0.962 for the (RPM -rRNA). In general, the libraries had a better association than the samples with the sample poly(A) preparation. It appears that samples with a lower number of reads in the individual runs produce less reproducible results.

Even though Poly(I) sequencing generates ~10-20% of the number of reads compared to poly(A) sequencing (Appendix: Sequencing Table), I see that even with the low number of reads the proportion of the gene populations in the sample is similar to what is seen in ONT poly(A) sequencing. The lower R^2 values of the snoRNA RPMs indicate that with fewer reads per genes, less reproducible results are obtained.

To determine how the analysis of the nanopore signals compares with our knowledge that Cid-1 is adding short (+50) I-tails onto mRNAs, my collaborator Logan Mulroney and I estimated the poly(I) extension and the natural poly(A) tail length of the mRNAs using Nanopolish (18), a commonly used software package that can estimate poly(A) tail lengths using ONT nanopore sequencing data. we have found that Nanopolish is unable to distinguish the poly(I) from the poly(A) signal in each read, and thus only provides an estimate of the length of the sum of both homopolymers. To determine how the software estimates the length of the poly(I) tails I compared the median tail lengths estimated by Nanopolish of the same mRNAs from the standard poly(A)-sequenced to the poly(I)-tailed poly(A) libraries that have at a minimum of 3 reads for each poly(A) or poly(A)poly(I) run. On a plot, I ordered the genes by the sum of the reads found for each gene, going from least number of reads to highest number reads on the x-axis. On the y-axis, I plotted the Nanopolish-called median tail length estimate for the genes of the poly(I) tailed reads (in blue) and the genes of the control poly(A) reads (in black) (Fig 4E). I determined that this arrangement is appropriate, as the regression analysis showed that the number of reads is about the same in both samples. The majority of the poly(A) control median homopolymer estimates are in the 40-60 nucleotide range, with an average of 42 nucleotides as found previously for yeast mRNAs (17). For the poly(A)poly(I) library reads, Nanopolish calls an average of ~27 nucleotide increase for all the reads, with the majority of the plots ranging from 50-75 nucleotides with an average of 70 nucleotides. For lower expressed genes, there is a higher variability in median added length, and there are a few genes with tails that extend up to ~100-125 nucleotides in total length. In more highly expressed genes, I see a shift to a tighter distribution where most of the reads are in the lower ranges stated above.

To validate this analysis of tail length distributions on the overall mRNA population and to observe the reproducibility of poly(I) tail lengths on genes, I looked at the average and standard deviation of median tail lengths on genes with the highest and lowest coverage found for both sample types. With poly(A)poly(I) tailed (in light grey) and the poly(A) (in dark grey) (Fig 4G) and the mean of Nanopolish median tail lengths are plotted on the y-axis. I found that across these top ten genes, Nanopolish calls these to have an averaged estimated poly(A)-tail of ~35 nucleotides, with a tight distribution where the standard deviation across all samples is between 1-3 nucleotides across the libraries, with an averaged standard error of 1.3. In the poly(A)poly(I) samples, Nanopolish calls the majority of the poly(A)poly(I) tails with an estimated length of ~65 nucleotides, and a much higher standard deviation between 17 to 26 nucleotides across each gene and an average standard error of 8.5. The averaged difference between the Nanopolish estimates after adding poly(I) is 30 nucleotides rather than our determination by biochemical methods of 50 nucleotides.

In the lower expressed genes, I see that there is more variability and longer poly(A) tail lengths from 35-48 nucleotides in length with an average of 46 nucleotides, with standard deviations ranging between 2-12 nucleotides, and a standard error between 1-6 nucleotides with an average of 3.7 nucleotides. There was a wide range of poly(I) tail lengths for these genes, with estimations between 60-122 nucleotides. This replicates the pattern of homopolymer tail lengths found in Figure 4F, where there was more variability but longer tail lengths in lower expressed genes compared to the shorter and more uniform tail length distributions in highly expressed genes.

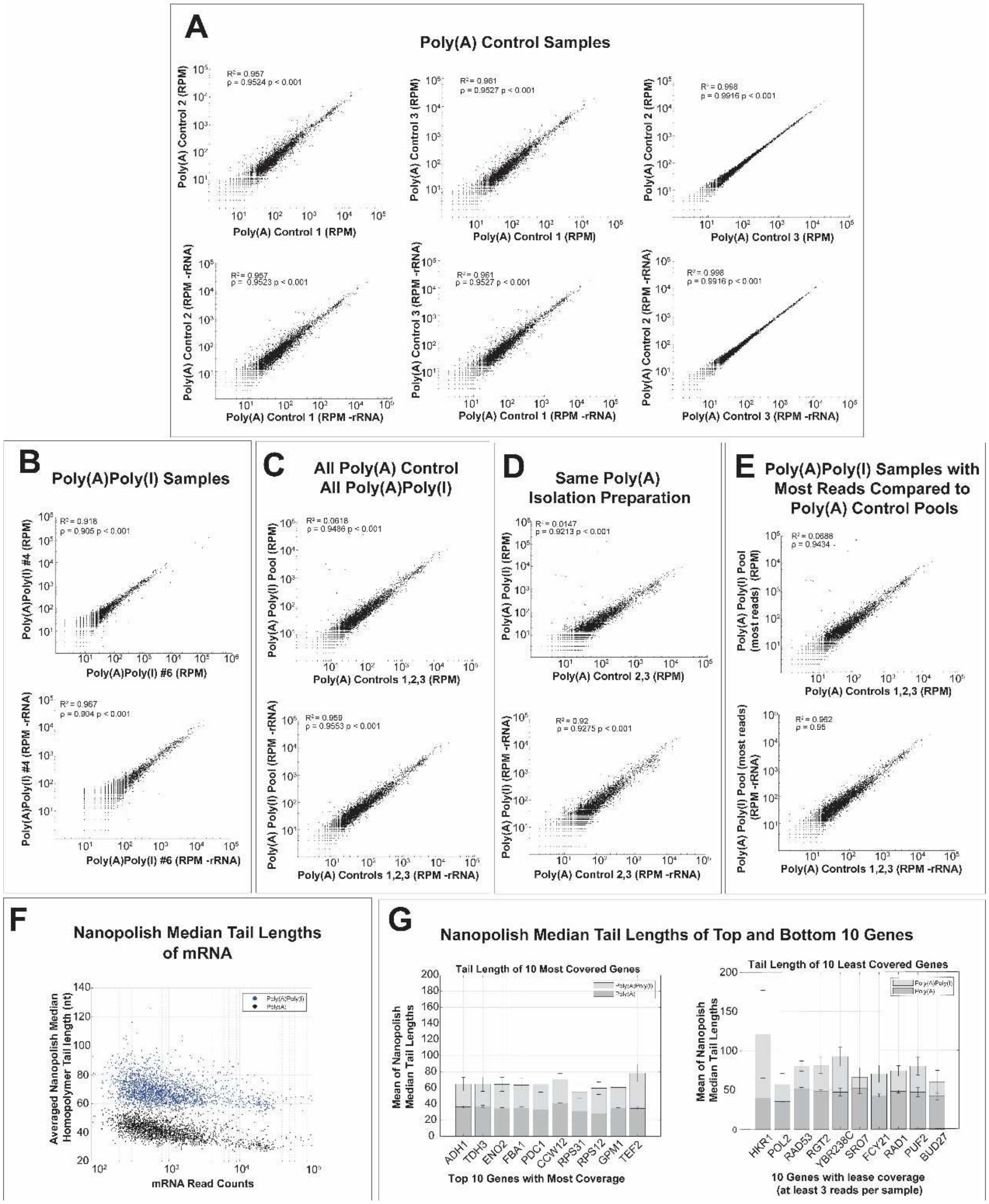

A
Poly(A) Control Samples
R² = 0.957
ρ = 0.9524 p < 0.001
Poly(A) Control 2 (RPM)
Poly(A) Control 1 (RPM)
R² = 0.961
ρ = 0.9527 p < 0.001
Poly(A) Control 3 (RPM)
Poly(A) Control 1 (RPM)
R² = 0.998
ρ = 0.9916 p < 0.001
Poly(A) Control 2 (RPM)
Poly(A) Control 3 (RPM)
R² = 0.957
ρ = 0.9523 p < 0.001
Poly(A) Control 2 (RPM -rRNA)
Poly(A) Control 1 (RPM -rRNA)
R² = 0.961
ρ = 0.9527 p < 0.001
Poly(A) Control 3 (RPM -rRNA)
Poly(A) Control 1 (RPM -rRNA)
R² = 0.998
ρ = 0.9916 p < 0.001
Poly(A) Control 2 (RPM -rRNA)
Poly(A) Control 3 (RPM -rRNA)
10¹ 10² 10³ 10⁴ 10⁵
B
Poly(A)Poly(I) Samples
R² = 0.918
ρ = 0.905 p < 0.001
Poly(A)Poly(I) #4 (RPM)
Poly(A)Poly(I) #6 (RPM)
R² = 0.967
ρ = 0.904 p < 0.001
Poly(A)Poly(I) #4 (RPM -rRNA)
Poly(A)Poly(I) #6 (RPM -rRNA)
C
All Poly(A) Control
All Poly(A)Poly(I)
R² = 0.0618
ρ = 0.9486 p < 0.001
Poly(A) Poly(I) Pool (RPM)
Poly(A) Controls 1,2,3 (RPM)
R² = 0.959
ρ = 0.9553 p < 0.001
Poly(A) Poly(I) Pool (RPM -rRNA)
Poly(A) Controls 1,2,3 (RPM -rRNA)
D
Same Poly(A)
Isolation Preparation
R² = 0.0147
ρ = 0.9213 p < 0.001
Poly(A) Poly(I) (RPM)
Poly(A) Control 2,3 (RPM)
R² = 0.92
ρ = 0.9275 p < 0.001
Poly(A) Poly(I) (RPM -rRNA)
Poly(A) Control 2,3 (RPM -rRNA)
E
Poly(A)Poly(I) Samples with
Most Reads Compared to
Poly(A) Control Pools
R² = 0.0688
ρ = 0.9434 -
Poly(A) Poly(I) Pool (most reads) (RPM)
Poly(A) Controls 1,2,3 (RPM)
R² = 0.962
ρ = 0.95
Poly(A) Poly(I) Pool (most reads) (RPM -rRNA)
Poly(A) Controls 1,2,3 (RPM -rRNA)
F
Nanopolish Median Tail Lengths
of mRNA
Averaged Nanopolish Median Homopolymer Tail length (nt)
20 40 60 80 100 120 140
Poly(A)Poly(I)
Poly(A)
mRNA Read Counts
10² 10³ 10⁴ 10⁵
G
Nanopolish Median Tail Lengths of Top and Bottom 10 Genes
Tail Length of 10 Most Covered Genes
Mean of Nanopolish Median Tail Lengths
0 20 40 60 80 100 120 140 160 180 200
Poly(A)Poly(I)
Poly(A)
ADH1 TDH3 ENO2 FBA1 PDC1 CCW12 RPS31 RPS12 GPM1 TEF2
Top 10 Genes with Most Coverage
Tail Length of 10 Least Covered Genes
Mean of Nanopolish Median Tail Lengths
0 50 100 150 200
Poly(A)Poly(I)
Poly(A)
HKR1 POL2 RAD53 RGT2 YBR238C SRO7 FCY21 RAD1 PUF2 BUD27
10 Genes with lease coverage
(at least 3 reads per sample)

Figure 4: Detection and abundance estimates for poly(A) mRNAs using I-tailing are equivalent to the standard library method. (Panel **A**) R^2 value plots of poly(A) controls (Panel **B**) R^2 value plots of poly(A) controls compared to poly(A)poly(I) libraries. (Panel **C**) R^2 value plots of poly(A) controls compared to poly(A)poly(I) that contain the same poly(A) sample preparation. (Panel **D**) R^2 value plots of the three poly(A)poly(I) samples with the highest coverage compared to the poly(A) controls (Panel **E)** Plot of median Nanopolish estimated tail length of mRNA encoding genes in poly(I) and poly(A) sequencing method (Panel **D**) Averaged Nanopolish median tail lengths found for top and bottom 10 covered genes found in poly(A) and poly(A)poly(I) samples.

Detection of non-polyadenylated RNAs is a product of representation in the sample and efficiency of tailing

As shown in Figure 1, the ability for Cid-1 to extend both polyadenylated and non-polyadenylated RNA molecules with inosine makes it promising for using Cid-1 poly(I)-tailing to sequence a sample with both poly(A) and non-poly(A) 3' ends. To first test the ability for Cid-1 to I-tail a sample containing a mixed RNA species, I Itailed yeast BY4741 Total RNA with a MYL6(A+) spike in control, and compared it to a un-I-tailed Total RNA sample on a denaturing polyacrylamide gel (Figure 5A). In the un-I-tailed Total RNA lane, I can distinctly see four bands of ribosomal RNAs. Seen on the gel is the 18S from the small 40S subunit (1800 nucleotides), the three ribosomal RNAs from the large 60S subunit: 25S (3396 nucleotides), 5.8S (158 nucleotides), and the 5S (121 nucleotides) (19), and tRNA (76-90 nucleotides) (20). In the I-tailed lanes, Ribosomal 25S and 18S were observed to have an upward shift in size after I-tailing, and 5.8S can no longer be seen. There is an increase of products seen between the large ribosomal RNAs and under the I-tailed 18S, and there were no distinguishable changes in 5S or the tRNAs. When observing the 3' end structure of the rRNAs (Fig B), the 5S has only one nucleotide (U) free from the 3' end, with a strong double helix interaction with the 5' end. While the 3' end of the 5.8S interacts with the 5' end of the 25S, it has three nucleotides free at its 3' end (UUU) with the 5' end of the 25S containing three free nucleotides (GUU). The 18S and 25S ha their 3'ends with seven (AUCAUUA) and two nucleotides (GU) respectively, with their 3' ends generally free. In conclusion, for ribosomal RNAs Cid-1 requires a 3'-end that is free with at least two nucleotides that is accessible for Cid-1 interaction. Cid-1 may require some single stranded regions in its RNA substrate that are not available in tRNA.

To test the poly(I) sequencing method on a mixed RNA sample, I sequenced yeast BY4741 Total RNA, and various ribosomal-depleted RNA in the hopes of obtaining a biological sample that enriches for non-rRNA reads. Since rRNA makes up the majority of total RNA, I was interested in seeing if I can sequence ribosomaldepleted RNA in order to better study the various non-rRNA species found in Total RNA. I tested three different rRNA depletion methods (Fig 4C and 4E). I tested the RiboMinus Transcriptome Isolation kit (Thermofisher), that uses a biotin-labeled locked nucleic acid (LNA) probe that is specific for large rRNA capture. I also tested pre-treatment with Terminator 5'-phosphate dependent exonuclease (Epicentre) for processive digestion of 5' monophosphate ends but not 5'-triphosphate, 5'cap, or 5'hydroxyl groups. Finally, I developed a new technique that involves oligo blocking of the 3' end of the rRNA to prevent poly(I) extension. To see the number of ribosomal reads in each sample and to qualitatively see the effect of each rRNA-depletion method on the rRNAs, I aligned the reads onto the genome browser to the sacCer3 genome (21) with the reads aligning from the 5' (left) to the 3' (right), with the number on the top left of each panel indicating the number of reads at that height. The yeast genome contains more than 150 copies of the repeats encoding for the rRNAs on the RDN1 locus. Shown in the figure is only one copy of the repeat unit. Since the

5S produces few reads, it is not shown. The height of each ribosomal-depleted sample is normalized to the Total RNA reads to provide an estimate of the efficiency of rRNA depletion obtained from each method (Fig 5B). Since the 3' end of the RNA enters the nanopore first, each read begins at the 3' end location where the adaptor has been added. However, in Direct-RNA Nanopore sequencing not all of the RNAs are read to their 5' ends, and some may have been broken, resulting many reads that do not span the entire gene body. This may be caused by current spikes during sequencing (8). This explains the characteristic upward slope in coverage going from 5' to 3' (Fig 5B).

In our Total RNA sample, I saw that rRNAs accounted for 73% of the reads (Appendix Sequencing Table). There are more reads for the 18S rRNA compared to the 25S, fewer reads for the 5.8S, and the coverage shows the characteristic upward slope found in Direct-RNA Nanopore sequencing. In the RiboMinus sample, a striking finding was the high percentage of 41% of 5.8S reads. However, RiboMinus was found to be efficient in depleting the large rRNAs 18S and 25S, with an only slight increase in 5S reads. Terminator treatment before library construction creates reads that show the 5' exonuclease activity of the enzyme, with significant 5' loss for most reads, but only partial digestion of the rRNAs near the 3' ends. In the center of the 25S gene body is a GGGG sequence that may somehow promote ligation to the polyC adaptor to the RNA at that location, resulting in reads that begin at that site (this is also visible in other libraries). Terminator overall is found to be efficient at depleting rRNAs, but less efficient than RiboMinus at depleting 18S or 25S. Another method, 3' oligo blocking, was found to have a slight enrichment of 18S reads, with

34% total reads in comparison to 29% seen in the Total RNA. Since the 3' blocking oligos block the mature 3’ end of the 18S, more reads were detected that started at the downstream site within the ITS, where the 3' ends of incompletely processed 18S rRNA lie. The effect of the GGGG sequence in the center of the 25S gene body that was seen in the terminator reads can also be seen here. Since there was no oligo designed for blocking that site, I see reads that start at that site and extends to the 5' end of the RNA. A summary of the rRNA read count findings in percentage found in each sample (Appendix Table 3), indicated I get close to the expected number of rRNA reads within a total RNA sample and can observe the effect of each ribosomal depletion method on Total RNA.

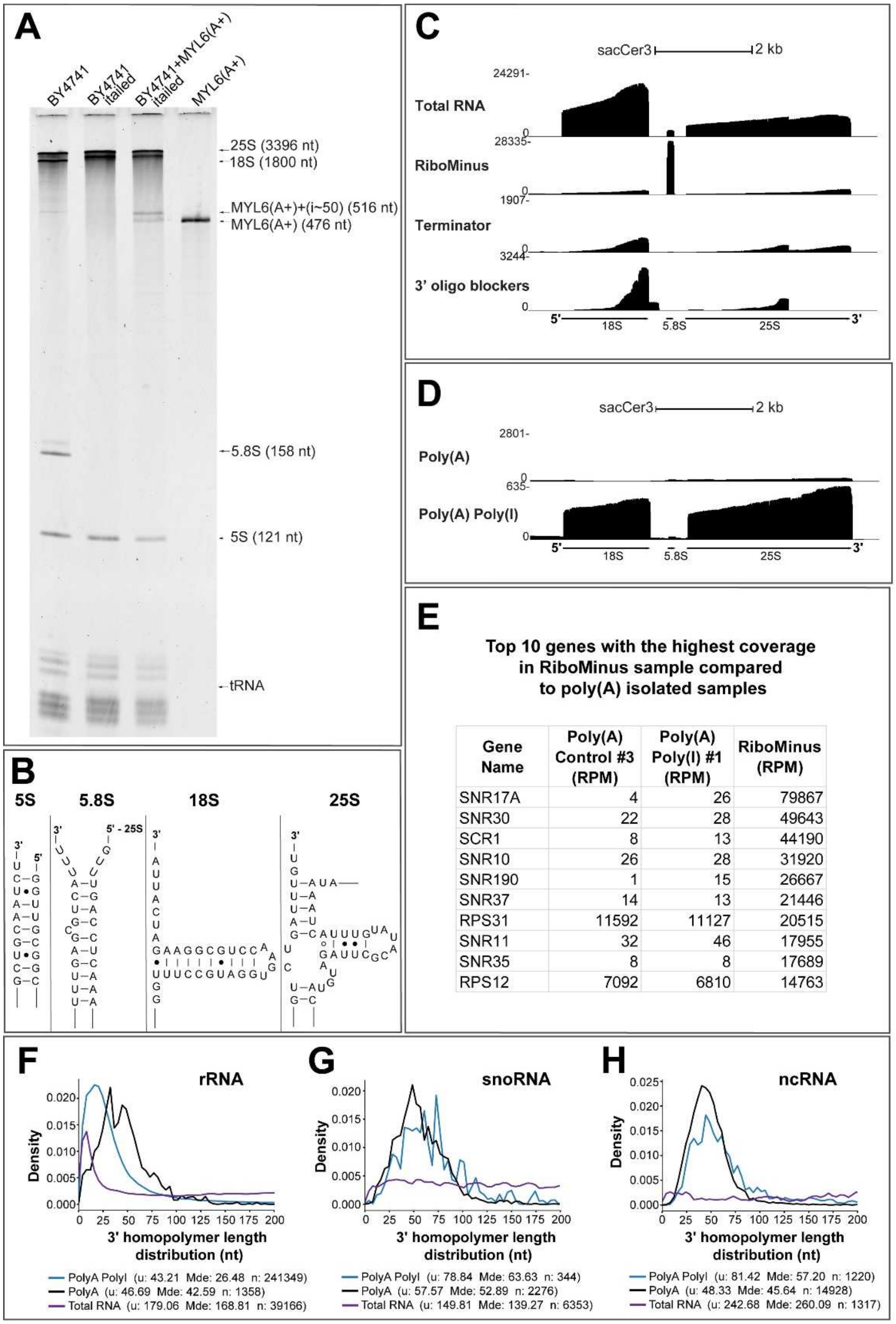
Since the adapter splint in Direct-RNA Nanopore sequencing only allows for sequencing of RNAs with 3' ends that can hybridize to it, I should not be able to see non-polyadenylated RNAs in a standard poly(A) sequencing run. When using the poly(I) method on poly(A)-selected RNA, I should be able to pick up residual nonpolyadenylated RNAs in the sample that are missed by the standard library construction method. To test this, I aligned the reads of a typical standard poly(A) sequencing library and a poly(I)-tailed poly(A) library to the genome browser to the same tandem repeat of Figure 5B on the RDN1 locus (Fig 5C). In the poly(A) libraries, I typically saw less < 0.1% of the reads contributing to aligned rRNA reads (Appendix Sequencing Table). However, when the poly(A) RNA was I-tailed, I generally found ~4-8% of the reads contributing to rRNA (Appendix Sequencing

Table), which is within the range for contaminating rRNA in a poly(A) enriched sample according to the manufacturer.

In conclusion the poly(I) sequencing method is able to capture nonpolyadenylated RNA within a sample, provided they can be I-tailed. Regarding the ribosomal depletion methods I tested, I find that every method has a different effect on the remaining rRNAs still left in the sample. RiboMinus is the best for 18S and 25S depletion but does not deplete 5.8S rRNA, whereas 3' oligo blockers are effective against 5.8S rRNA, and terminator is decent at depleting all rRNAs but leaves substantial amounts of 3' ends undigested leading to their presence in the libraries.

When comparing the read numbers of individual genes for each rRNA depleted sample, RiboMinus was found to have the highest coverage of non-mRNAs (data not shown). To determine the enrichment of non-mRNAs, the top ten genes with the highest coverage in the RiboMinus sample was compared to a typical poly(A) and poly(A)poly(I) sample (Fig 5E). Majority of the genes with the highest coverage in RiboMinus were snoRNAs, and there were some mRNAs and ncRNAs. When comparing these reads to the poly(A) isolated samples, most of the genes were enriched by a few thousand-fold, indicating that the poly(I) method can detect nonpoly(A) RNA.

In Figure 2, I noticed a pattern of Cid-1 I-tailing on non-polyadenylated RNAs where there is a heterogeneous distribution of inosine tail lengths. To ensure that Cid1 is tailing not just the polyadenylated forms of non-coding RNAs, I examined the homopolymer tail length distribution from 0-200nt on rRNA (Fig 5E), snoRNA (Fig 5F), and ncRNAs (Fig 5G) from the poly(A) (black) , poly(I)-tailed poly(A) (blue), and total RNA samples (purple), with the homopolymer tail lengths plotted against the x-axis, and the fraction of the reads with that containing that size on the y-axis as "density" (density of 1 = 100% of all reads from that class) (Figure 5E-F). When observing the rRNA, the majority of the tail lengths for poly(I)-tailed poly(A) samples were shorter than the poly(A) control samples. In the total RNA sample, I found that a large number of reads contain homopolymer tail lengths in the ~25nt range and an even distribution from ~25-200nt (Figure 5E). snoRNA and other ncRNA had similar distributions of homopolymer tail lengths (Fig 5F, 5G). The poly(I) tailed RNA was found to be overall ~25nt larger than the poly(A) tail, but in the total RNA sample there is an even distribution of tail lengths from ~10nt to 200nt at a similar density. In conclusion, polyadenylated and non-polyadenylated RNAs are also being I-tailed, and they create a long stretch of I-tails with a heterogenous distribution. This supports our finding in Figure 2, that Cid-1 creates long stretches of I-tails on non-poly(A) RNA and only extends a short stretch of poly(I) on poly(A) RNA.



Top 10 genes with the highest coverage in RiboMinus sample compared to poly(A) isolated samples

Gene Name	Poly(A) Control #3 (RPM)	Poly(A) Poly(I) #1 (RPM)	RiboMinus (RPM)
SNR17A	4	26	79867
SNR30	22	28	49643
SCR1	8	13	44190
SNR10	26	28	31920
SNR190	1	15	26667
SNR37	14	13	21446
RPS31	11592	11127	20515
SNR11	32	46	17955
SNR35	8	8	17689
RPS12	7092	6810	14763

Figure 5: Detection of non-polyadenylated RNAs is a product of representation in the sample and efficiency of tailing (Panel **A**) Cid-1 I-tailed Total BY4741 Yeast RNA with MYL6 control. rRNA read alignments in (Panel **B**) 3'-end structures of ribosomal RNAs 5S, 5.8S, 18S, and 25S. (Panel **C**) Ribosomal depletion methods and (Panel **D)** Comparison of RPM values of top ten most covered genes in RiboMinustreated Total RNA to poly(A) samples (Panel **E**) Graph of % of rRNA reads in ribosomal depletion methods (Panels **F-H**) Tail length distribution on (**F**) rRNA (**G**) snoRNA **(H)** ncRNA

Initial examination of nascent transcript structure using I-tailing of chromatin associated RNA

Knowing that the poly(I) method works well on non-poly(A) RNA, I was interested in sequencing a sample with mostly non-poly(A) and minimal poly(A). I was interested in sequencing RNA transcripts still bound to RNA Polymerase II (RNAPII). RNAPII transcribes pre-mRNA and several snRNAs (22) and RNAs bound to RNA Polymerase II are considered to be nascent, as they are just coming to existence. Most of these transcripts bound to RNAPII should be unprocessed, but it is known that co-transcriptional splicing does happen (23). Since RNAPII transcribes the RNA starting at the 5' end, the 3' end of the RNA indicates the location of the RNAPII on the DNA as it was interrupted during transcription. Nascent RNA should give information about RNAPII progression on the gene body during transcription and the rate of co-transcriptional splicing (23).

To obtain these nascent RNAs, I purified chromatin-associated RNAs by isolating chromatin from a yeast strain CKY2647 that produces a recombinant RNAPII that contains a fused small biotinylated affinity tag AviTag™ (24), and used streptavidin beads to capture the recombinant RNAPII and the chromatin associated to it. After purifying the RNA from this sample, I sequenced using the poly(I) method (Fig 6A). While isolating pure nascent RNA would be ideal, it is difficult to purify pure nascent RNA from lysed cells from the endogenous biotinylated RNA and other contaminating RNAs that may stick onto streptavidin beads. Considering these factors, I referenced our sample as “chromatin-associated” RNA. Unfortunately, the low throughput of the poly(I) method and the high contaminants of non-RNAPII transcripts makes examining the nascent RNA difficult. However, I found transcripts that were most likely halted during RNAPII transcription (Fig 4B and 4C). To better inspect the sample for nascent transcripts, the chromatin-associated RNA was compared to the RiboMinus-treated Total RNA sample shown in figure 5C. The RiboMinus-treated total RNA was found to have the highest number of non-rRNA transcripts of the rRNA-depleted samples. Shown is an example of a gene IMD4, with reads aligned to the genome browser (Fig 4B). IMD4 which contains an intronic snoRNA snR54 gene that demonstrates RNAPII tracking and has evidence of cotranscriptional splicing. In this chromatin-associated sample, I see that the 3' ends of reads span the gene body, with a few spliced reads, and a few transcripts of the snR5 gene. When comparing the chromatin-associated IMD4 to the reads found in RiboMinus-treated Total-RNA (Fig 6C), I found that the IMD4 in this sample contains a few reads that span the gene body and short reads that cluster at the 3' end. These short reads are most likely degradation products, and almost all of the few reads obtained in the sample have 3' ends at the poly(A) site. In further observations I found NOG2, which contains an intronic snoRNA snR191 gene that has similar features see in IMD4 (Fig 6C). In the chromatin-associated sample, I saw that there are 3' ends that span across the gene body with many complete transcripts of snR191 and a few snR191 intermediates that span across the first exon. Also shown is the NOG2 gene found in the RiboMinus-treated Total RNA sample. For space, most of the snR191 reads are omitted in the figure. Similarly to IMD4 in figure 6B, I see many reads that span across the gene body and almost all of the reads end near the 3' end at the poly(A) site. When comparing the genome browser alignments of both these runs, I can conclude that at least some of the IMD4 of the chromatin-associated RNA is nascent.

To determine if there are any population changes in the chromatin-associated RNA, I compared the mRNA and snoRNA populations of the chromatin-associated RNA to the RiboMinus-treated Total RNA (Fig 6D). Log2 scores of the proportional difference in RPM of the genes found in both samples were plotted on histogram bar graphs, with the log2 scores on the x-axis and the number of genes for that category in the y-axis. In the chromatin-associated sample I found that there was a general decrease with mRNAs, with most of the mRNAs decreasing to ~5-fold but some were increasing to ~3-fold. For snoRNAs, I saw a general increase in transcripts in the chromatin-associated samples. Up to ~7-fold increase for most of the snoRNA genes, but some genes had up to ~4-fold decrease. The poly(I) method was able to detect these differences that are expected in chromatin-associated samples, which indicates that it has the potential to be used to study an RNA sample that contains little poly(A) RNA.

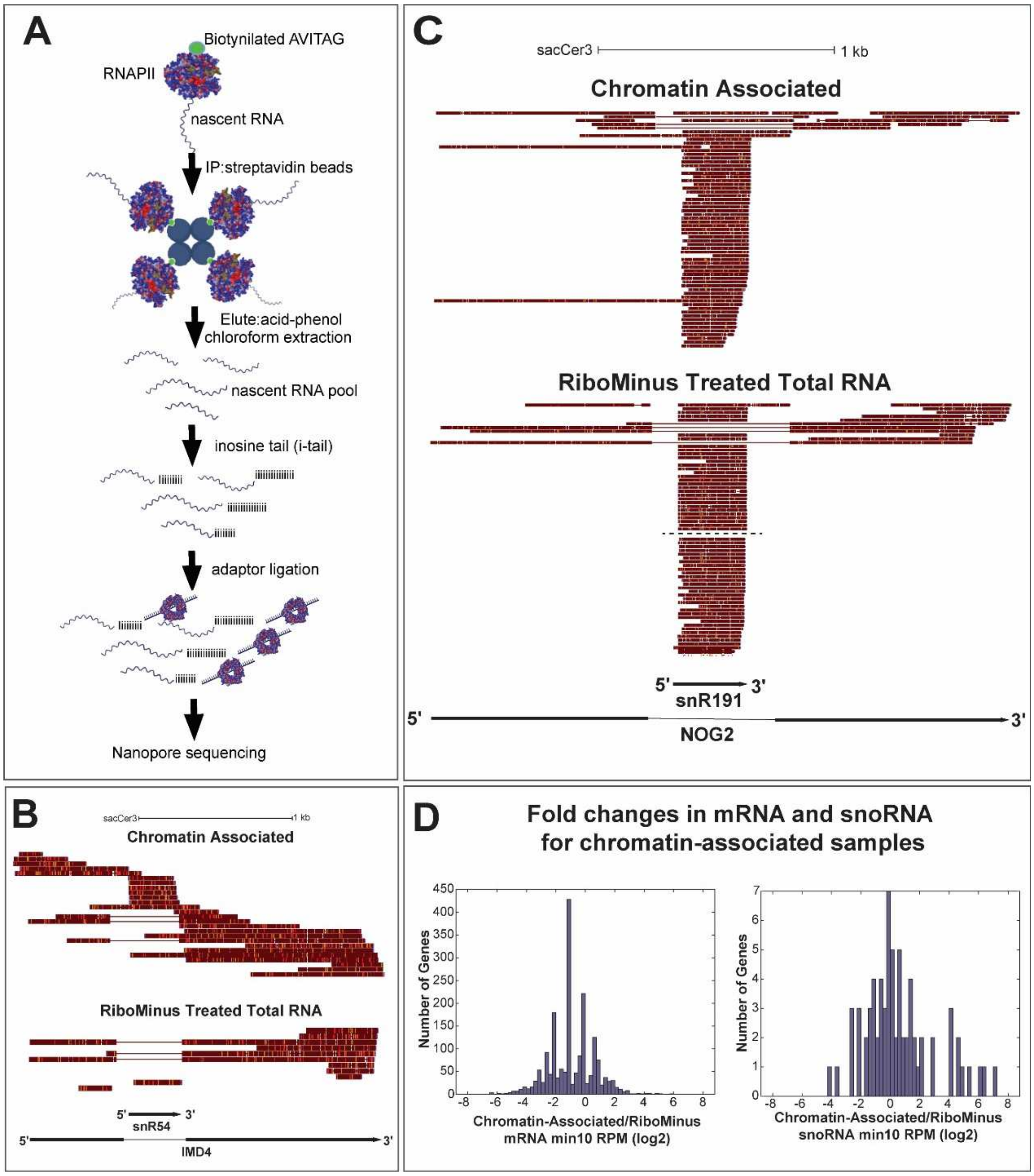

A
Biotynilated AVITAG
RNAPII
nascent RNA
IP:streptavidin beads
Elute:acid-phenol
chloroform extraction
nascent RNA pool
inosine tail (i-tail)
adaptor ligation
Nanopore sequencing
B
sacCer3
1 kb
Chromatin Associated
RiboMinus Treated Total RNA
5'
3'
snR54
5'
3'
IMD4
C
sacCer3
1 kb
Chromatin Associated
RiboMinus Treated Total RNA
5'
3'
snR191
5'
3'
NOG2
D
Fold changes in mRNA and snoRNA
for chromatin-associated samples
Number of Genes
450
400
350
300
250
200
150
100
50
0
-8
-6
-4
-2
0
2
4
6
8
Chromatin-Associated/RiboMinus
mRNA min10 RPM (log2)
Number of Genes
7
6
5
4
3
2
1
0
-8
-6
-4
-2
0
2
4
6
8
Chromatin-Associated/RiboMinus
snoRNA min10 RPM (log2)

Figure 6: Initial examination of nascent transcript structure using I-tailing of chromatin associated RNA. (Panel **A**) Purification of chromatin associated RNA using recombinant AVITAG RNA Polymerase II and streptavidin bead capture (Panel **B**) Representative aligned reads of chromatin associated RNA in IMD4 (Panel **C**) Representative aligned reads of chromatin associated RNA in NOG2 (Panel **D**) Histogram of log2 scores of mRNA and snoRNAs in the chromatin-associated sample in comparison to the RiboMinus Total-RNA treated sample.

Discussion

Previous studies have investigated the ability for Cid-1 and PAP to extend a short oligo of A15 using canonical nucleotides (15, 16), and crystal structures (25-27) for modeling their activities. Cid-1 is a non-canonical poly(A) polymerase, but later was re-classified as a poly(U) polymerase after the discovery of its high poly(U) polymerase activity (28, 29). Canonical poly(A) polymerase for yeast has been used for 3' end labeling with modified nucleotides (16), but our accidental discovery of Cid-1's unknown abilities for incorporating inosine tails onto RNAs has allowed us to exploit its abilities to easily sequence biological samples of mixed RNA species. I confirmed past findings of both Cid-1 and PAP's ability to extend on a short A primer (15, 16), and found that Cid-1 inosine tails long model mRNA with a +50 size (Fig 1C, Fig 2 A), but has trouble extending on shorter mRNAs (Suppl Fig 1A) and non-polyadenylated RNA (Fig 1C, Suppl Fig 1A).

I found that Cid-1 extended inosine tails onto long mRNAs with a discrete size of 50 nucleotides at a fast reaction rate and long inosine tails on complex molecules (Fig 2). Studies of Cid-1 structure required a truncated sequence of the protein for stability, and mostly its binding site with bound nucleotides have been observed (23, 24). However, this leaves behind the binding site for the 3' end of the RNA to be observed. How do substrates bind to that site, and how does it affect the activity of the polymerase? Cid-1 has a peculiar activity of inefficiently adding certain nucleotides onto A24 (Fig 1A) and to GLuc200(A+) and GLuc200(A-), which may indicate that Cid-1 requires the RNA to be of a particular length for the 3' end to be bound to the Cid-1 binding site.

Using our Cid-1 findings, I designed a sequencing protocol to accommodate capture of 3' inosine tails (Fig 3A). I was able to observe that poly(I) creates a distinct signal and its transition to the poly(A) that can be differentiated from poly(A) tails and the complex RNA in control samples (Fig 3B). When observing the raw current traces from biological samples, I found that they have the same characteristics as the control (Fig 3C).

I currently do not have a tool that can classify reads of poly(A), poly(I), and poly(A)poly(I) samples, and estimate the length of the poly(A) and the poly(I) homopolymers. Our control samples of poly(I)-ligated GLuc200 would make a great training data set that can be used to train the parameters for a new poly(I) classifier to contain all of the features stated above. Poly(U)-tails do exist in vivo, but I have not sequenced RNA that contains those samples. I do not know what the inosine homopolymer looks like compared to poly(U), and ensuring that the poly(I) can be used on samples containing poly(U) can be beneficial and can be part of a classifier if built.

Poly(I) sequencing was found to be comparable to poly(A) sequencing for non-rRNA genes and when there is a high enough amount of reads in the samples (Fig 4). I found that Cid-1 generally adds a 50 nucleotide poly(I)-tail on an mRNA (Fig 1C), but Nanopolish underestimates this as a ~25-30 nucleotide extension (Fig

4E and 4F). Nanopolish and other homopolymer tail length programs are subject to error in calling homopolymers since they do this by transit time rather than by discriminating changes in sequence signal. Calibration will be very important in order to be able to estimate homopolymer lengths accurately by nanopore methods.

Cid-1's ability to incorporate inosine tails on non-polyadenylated RNAs (Fig 1C and 2D-F) can help expand the scope of direct RNA sequencing, as both polyadenylated and non-adenylated RNAs can now be sequenced in the same sample (Fig5E-F). However, the reads being an accurate representation of the sample population means that ribosomal RNAs make up the majority of the reads in total RNA. In our attempts to find the best ribosomal depletion method for the poly(I) method to better study the transcriptome (Fig 5B), I found that RiboMinus had the best depletion of the large ribosomal subunits, but had an unusual enrichment of 5.8S from the manufacture's lack of 5.8S oligo in the kit. I found that the developed 5.8S 3' oligo blocker was efficient at blocking those reads. For future developments, incorporating 5.8S 3' oligo blockers with the RiboMinus kit may be a way to further deplete the ribosomal RNAs.

However, RiboMinus-treatment alone was able to detect a high abundance of snoRNAs where only a small amount was found in the poly(A) samples (Fig 5D). Most likely these RNAs are not poly(A) tailed or some contain very short poly(A) tails from the slight enrichment of these samples in the poly(I)-tailed poly(A) sample if they were too short to hybridize to the poly(T) adapter in standard ONT poly(A) sequencing. This shows potential for poly(I) sequencing to be used to study RNA samples that do not contain only poly(A) RNA.

In our chromatin-associated samples, I may have sequenced nascent RNA (Fig 6B-C). The low number of reads and the high contamination of mRNAs indicated by the complete transcripts made chromatin-associated RNA difficult to study. Even though no conclusive results can be obtained about the nascent RNA, I found that I am able to detect a suppression of mRNA and an enrichment of snoRNA between the chromatin-associated RNA and the RiboMinus-treated RNA samples (Fig 6D) which is expected. Along with the enrichment of snoRNAs found in the RiboMinus sample (Fig 5D), this shows the potential for poly(I) sequencing to be used for studying populations of non-poly(A) RNA.

The crucial barrier when using the poly(I) method for sequencing is the low throughput for all samples. I found that I generally get 10-30% the number of reads compared to standard ONT poly(A) sequencing (Appendix Table 1). This may be due to the length of the poly(I) tails. Cid-1 is seen to tail some RNAs with extremely long poly(I) tails up to >1000nt and sometimes adds extremely short poly(I) tails of only a few nucleotides (Figure 1,2,5). The MinKnow software used to run the sequencing program may consider the long poly(I) tails as stalling events, as it contains code that considers a stalling event as an unchanged state in the pore for more than 5 seconds. If a stalling event is found, the MinKnow software reverses the voltage across the pore and the RNA is ejected. At least in DNA, it is found that incomplete reads that are ejected are less likely to be sequenced again as the motor protein is no longer available after its migration across the molecule (30). Since ONT Direct RNA sequencing uses the same sequencing principle as DNA, it is logical to assume that the motor protein RNA polymerase may work the same way. These long poly(I) tails may be considered as stalling events and ejected out of the pore, resulting in low reads. The inefficiency of Cid-1 poly(I) tailing on some RNAs may create a problem during the library preparation, as the adapter split used for poly(I) sequencing is 10 nucleotides in length. The short poly(I) tails may cause inefficient hybridization of the adapter resulting in in-complete ligation. Un-adapted RNA caused by inefficient ligation or un-poly(I) tailed RNA cannot be sequenced and may cause clogs in the nanopore. Further development may be necessary to increase the efficiency of tailing and decreasing the length of the modified tails.

Poly(I) sequencing has been demonstrated to show potential in being used to sequence any RNA sample with poly(A) and non-poly(A) RNAs. The ability to sequence RNAs with inosine homopolymers can be expanded to other modified nucleotides. If each signal can be distinguished from each other and natural RNA, this can be used potentially as a barcoding technique for sequencing multiple samples within the same sequencing run, and to help generate training data to detect RNA modifications.

Materials and Methods:

Total RNA extraction from Yeast

Total RNA was extracted as described previously (30). Briefly, *S. cerevisiae* BY4741 cells were grown to an A_{600} of 0.5 in YEPD, and 10mL pellets were resuspended in 440ul of 50mM Sodium Acetate pH 5.2, 10mM EDTA, 1% SDS and 400ul of phenol:chloroform: isoamyl alcohol. After vortexing, the cells were incubated at 65C for 10 minutes with intermittent 5-10 second vortexing every minute. After a 5minute incubation on ice, the samples were added to pre-spun 2mL Phase Lock Gel Heavy tubes (Eppendorf #955154045). After a 5-minute centrifugation at max speed, 400ul of chloroform was added, shaken, then centrifuged again. Another 400ul of chloroform was added, shaken, and centrifuged at full speed before transferring the top aqueous phase into a new 2mL microcentrifuge tube. The samples were brought up to 2mL volume with 0.3M sodium acetate pH 5.2 and 70% ethanol. After inversion and centrifugation, the pellets were rinsed with 70% ethanol and centrifuged before drying the pellets with a speed-vac then resuspended in nuclease-free water.

Tissue Culture and Total RNA isolation of GM12878

GM12878 Total RNA was gifted from the Nanopore group and cells were culture and isolated as described previously (7).

Poly(A) RNA Selection from Total RNA

Poly(A) RNA of BY4741 and GM12878 was selected using NEXTflex Poly(A) beads (BIOO Scientific Cat#NOVA-512980) according to the manufacturer's instructions. Briefly, RNA was heated to 65C for 2 minutes in 50% Binding Buffer at >200ul, and chilled on ice before resuspending the prepared magnetic beads. The beads were rotated at room temperature for 5 minutes before pelleting and removal of the supernatant. Washing buffer was added, pelleted, and removed for a total of two washes. The beads were resuspended with 50ul elution buffer, incubate at 80C for 2 minutes and immediately pelleted before transfer of the supernatant to 100ul of binding buffer. Eluate was heated to 65C for 2 minutes and placed on ice before washing the beads with 200ul of washing buffer twice. The beads were resuspended with eluate then rotated for 5 minutes in room temperature. The beads were pelleted, then washed with washing buffer twice before adding 17ul of elution buffer, heated to 80C for 2 minutes, then placed on the magnet for transfer of eluate to a fresh tube.

Ribosomal Depletion with RiboMinus Transcriptome kit

Ribosomal depletion of total RNA from BY4741 yeast and GM12878 cell line using RiboMinus Transcriptome kits and Concentration Modules (Invitrogen #K155001, K155003) were prepared by bead capture according to the manufacturer's instructions. Briefly, total RNA, probe, and hybridization buffer were combined and incubated at 37C for 5 minutes, iced, and then added to prepared magnetic beads. The beads and sample were incubated at 37C for 15 minutes, then pelleted on a magnetic rack before transferring the supernatant to a different tube containing binding buffer and ethanol. The sample was bound to the concentration spin column then washed with washing buffer twice, then eluted with nuclease-free water.

Ribosomal Depletion with Terminator™ 5′ - Phosphate-Dependent Exonuclease

Ribosomal depletion of yeast BY4741 Total RNA using Terminator™ 5′ - Phosphate-Dependent Exonuclease (Epicentre) was achieved by incubating RNA (1x Terminator Reaction Buffer A, 3ug of BY4741 Total RNA and 1unit of Terminator Exonuclease) in 20µl volume. The reaction was incubated at 30°C for 1 hour, then terminated with the addition of 1µl 100mM EDTA. The ribosomal depleted RNA was purified using 25:24:1 phenol:chloroform:isoamyl alcohol followed by ethanol precipitation.

Ribosomal blocking during polyinosine poly(U) polymerase tailing

To block the addition of ~50 inosine homopolymers on the 3' end of ribosomal RNAs, denaturation of RNA before I-tailing with poly(U) polymerase was adjusted accordingly: RNA in 0.1mM EDTA and 0.5pmol – 4pmol of 3'overhang or 3'stemloop oligo pools were added to a final volume of 2.95uL, denatured at 95°C for 2 minutes, 55°C for 2 minutes then placed on ice for 2 minutes before proceeding to the poly(I) tailing reaction.

Polynucleotide tailing with Poly(U) Polymerase

To add a homopolymer to the 3'end of RNA using poly(U) polymerase, RNA in 0.1mM EDTA to a final volume of 2.95uL is denatured at 95°C for 2 minutes then placed on ice for 2 minutes. The RNA is added to a reaction containing 4mM NTP, 50mM NaCl, 13.5mM $MgCl_2$, 1mM DTT, BSA 500ug/ml, pH 7.9, and 1ul of NEB poly(U) polymerase in a final volume of 7.5ul and incubated at 37°C for 1 hour. For library preparation 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 minutes. 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 11ul of water and incubated in 10 minutes at room temperature, pelleted, then eluate was transferred to a new tube.

Polynucleotide tailing with PolyA Polymerase, Yeast

To add a homopolymer to the 3' ends of RNAs using Poly(A) Polymerase, Yeast (ThermoScientific 74225Y/Z) a reaction containing (1x Poly(A) Polymerase reaction buffer, 200fmol RNA, 0.5mM NTP) was incubated at 37°C for 30 minutes, then two volumes of Gel Loading Buffer II (Invitrogen: AM8546G) was immediately added to stop the reaction. Reaction products were separated on 15%, 8%, 6% 8M urea denaturing PAGE.

vCID1 expression in *E. coli*:

Plasmid 10H-tev-vCID was made by standard cloning techniques using synthetic DNA. It carries a beta-lactamase gene and a colE1 origin of DNA replication in E. coli. E. coli BL21(DE3) pLysS cells are transformed with the plasmid and transformants are selected on LB (Luria Broth) agar plates with 100 µg/ml ampicillin at 37°C overnight. A single colony is used to inoculate 50mL of LB supplemented with 100 µg/ml ampicillin, shaking at 300rpm overnight at 37°C. Overnight cultures are diluted to 1L with LB supplemented with 25 µg/ml ampicillin and grown at 37°C shaking at 300rpm to OD600 = 0.6. Cells are induced to express vCID by adding IPTG to 1mM and shaking at 300rpm for 16-18 hours at 18°C. Cells are harvested by centrifugation at 5000rpm for 10 minutes, and cell pellets are resuspended in 10mL of 50mM Tris-HCl, 1mM EDTA pH 8.0 and centrifuged again at 4°C at 5000rpm, for 30 minutes. Supernatant is decanted, and washed pellets are stored at -80°C.

vCID1 Purification

The buffers used for purification contain 50mM NaH2PO4, 300mM NaCl, 100mM KCl, 1mM DTT, 10% glycerol and various concentrations (10-500mM) imidazole. Frozen pellets are resuspended in 14mL of 10mM imidazole buffer with 1 mg/mL lysozyme, 0.5mM PMSF, then incubated on ice for 5 minutes. Cells are lysed with glass bead vortexing for 3 minutes in 15 second intervals. Cells are incubated on ice for 15 seconds between vortexing. Lysates are clarified by centrifugation at 14000xg for 1 hour at 4°C. Resulting supernatant is incubated for 30 minutes with 2.5mL of cobalt-chelate resin (pre-equilibrated in 10 mM imidazole buffer) with gentle shaking at 4°C, and then is poured into a column. The column is washed 2x with 12mL of 20mM imidazole buffer, and 1x with 5mL of 50mM imidazole buffer. Eluates are collected using 100mM, 150mM, 200mM, 250mM, 400mM, and 500mM imidazole buffers at 5mL each in succession. Each eluted sample is concentrated in a protein concentrator tube (Millipore) until the total volume is <500μL. The samples are resuspended in 10mL storage buffer (10mM Tris-HCl pH 7.5, 100mM NaCl, 1mM DTT, 0.1mM EDTA, 50% glycerol v/v), and concentrated again until the sample volume was under 500μL.

GLuc200I30 and GLuc200A44I30 sample and library preparation

To prepare GLuc200I30 and GLuc200A44I30 samples, 15pmol of GLuc200 RNA with 30pmol of the appropriate bottom splint adapter (C25TTTTTTTTTT (IDT) for GLuc200A44 or C25CCT AAG AGC AAG AAG AAG (IDT) for GLuc200), 1.4 nmols of a synthetic 5’p-15mer inosine homopolymer (Stanford PAN facility) in 10mM tris pH 8.0, 1mM EDTA, and 50mM NaCl in 6µl reaction volume was heated to 55°C and slow cooled to 16°C in 25 minutes. 1ul 10X T4 ligation reaction buffer (NEB B0202S) and 2,000 units T4 DNA ligase (NEB M0202T) were added to each reaction and brought to 10ul 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 minutes before loading into a 10% acrylamide gel and ran for 3.5 hours at 28 watts.

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 in 850µl of 1x TAE buffer using a D-tubeTM Dialyzer Midi (Millipore Sigma 71507) for at least 90 minutes at 130 volts. The electro-eluted samples were precipitated with 85µl 0.3M NaOAc and 850µl isopropanol at -20°C overnight. The next day the samples were centrifuged at 4000 x g for 30 minutes, decanted the supernatant, then the pellets were washed with 70% ethanol twice with subsequent centrifugations at 16000 x g for 15 minutes. The pellets were air dried for 15 minutes and resuspended in 10µl water with yields between 25-100ng. The libraries for each ligation product were prepared following ONT's Direct-RNA Nanopore sequencing library preparation with ~50ng of RNA and optional reverse transcription step skipped.

in vitro **transcription template generation**

MYL6 and MYL6 no polyA templates were generated using linearized pUC13MYL6 plasmid. pUC13-MYL6 was digested with BbsI (NEB: R0539S), or BsmI (NEB: R0134S) for MYL6 or MYL6 no polyA respectively. Gluc200 and Gluc200A44 templates were generated using PCR amplification of GLuc at the first 200nt residues at the 5'end from of pCMV-GLuc 2 Control Plasmid (NEB). The sequence was targeted using a forward primer containing a T7 promotor region, and a reverse primer that terminates the PCR product at the 3' end of the truncated GLuc sequence or the addition of 40nt 3' polyA tail for GLuc200 and GLuc200A44 respectively. PCR amplification was obtained using Platinum® Taq DNA Polymerase High Fidelity (Invitrogen) using 1x High Fidelity PCR buffer, 0.01ug/ul of plasmid, 0.4mM dNTP Mix, 0.4uM Forward Primer, 0.4uM Reverse primer, 2mM MgSO4, 1unit Platinum® Taq DNA Polymerase High Fidelity. PCR cycles were: 94C 30s, / 94C 10s, 58C 15s, 65C 45s / 65C 10m, 4C hold. DNA In vitro transcription templates were purified using NucleoSpin® Gel and PCR Clean-up (Macherey-Nagel) following the manufacturer's instructions.

T7 *in vitro* transcription

Templates were in vitro transcribed using the MEGAscript™ T7 Transcription Kit (Invitrogen). Reaction products were separated, and gel excised using 6% 8M urea polyacrylamide gel electrophoresis. Gel slices were rotated at 4°C overnight in RNA Elution Buffer (0.3M NaOAc pH 5.2, 0.2% SDS, 1mM EDTA, 10µg/mL proteinase K). Eluted product was purified using 25:24:1 phenol:chloroform:isoamyl alcohol and ethanol precipitation.

pCp labeling

T7 transcripts and their I-tailed counterparts were labeled with pCp [5'-32P] Cytidine 3', 5' bis(phosphate) 3000 Ci/mmol, 10mCi/ml. Used NEB T4 RNA Ligase 1 (NEB Cat no), 1X reaction buffer, 0.15mM ATP, 10% DMSO, 20 µCi pCp, 6pmol of RNA, and 333 units RNA Ligase 1 (NEB M0204S). Incubated at 16°C for ~16-18 hours. The products were purified by extracting with an equal volume of 25:24:1

phenol:chloroform:isoamyl alcohol (PCA) and 0.3mM NaOAc pH 5.2 was added before ethanol precipitation.

RNAseA digestion

RNA samples in 100mM NaCl, 10mM EDTA, 0.025ug/ul RNAse A digestion were incubated at 37°C for 15 minutes, then immediately placed in equal amounts of 25:24:1 phenol:chloroform: Isoamyl and 0.3mM NaOAc pH 5.2 and proceeded with cleanup before ethanol precipitation.

CKY2647 cell harvest

CKY2647 cells were grown to an A600 of 0.5-0.8 in YEPD in 100mL cultures. Cells were harvested at 1100 x *g* centrifugation for 5 minutes at 4°C. Pellets were washed with 40ml ice-cold PBS twice, then transferred to 1.7ml Eppendorf tubes and washed with cold PBS with centrifugation at 1100x g for 5 minutes at 4°C. Supernatant was removed and pellets were snap frozen in liquid nitrogen before storing in -80C.

CKY2647 Chromatin Associated RNA Purification

A bead column was assembled accordingly: A hole was pierced in the bottom of a 15mL centrifuge tube with a 22-gauge needle and placed inside a 50mL falcon tube lid with a pre-cut hole. At the 12mL marker on the 15mL centrifuge tube, parafilm was warped around as a stabilizer then the 50mL falcon tube lid containing the 15mL centrifuge was screwed back onto the bottom of the 50mL falcon tube lid.

Working in a 4°C room, the CKY2647 pellets stored in -80C were thawed on ice for 5 minutes, then resuspended in 1ml of Buffer 1 (20mM HEPES pH 8.0, 60mM KCl, 15mM NaCl, 5mM MgCl2, 1mM CaCl2, 0.8% Triton-X100, 0.25mM sucrose, and freshly added 0.5mM spermine and 2.5mM spermidine). For lysing, the sample was added to a 2ml Eppendorf tube containing 1mL 0.5mm zirconia beads, then vortexed for 1 minute with 1 minute pulses for six cycles using a pre-warmed Turbomix attachment on a Vortex Genie 2 (Scientific Industries Inc SKU: SI-0564) that was pre-warmed by running at max speed for 1 minute preceding the vortexing. The sample was transferred to the assembled bead column. For transfer of remaining lysed cells still stuck in the 2ml tube, 1ml of Buffer 1 was added then the tube was inverted before transferring to the assembled bead column. This process was repeated three times. The assembled bead column was centrifuged at 400 x g for 6 minutes at 4°C. While avoiding the pellet, the supernatant at the bottom of the 50ml falcon tube was transferred into two 1.7ml Eppendorf tubes at 750ml each then the pellet was discarded. The samples were pelleted by centrifuged at 2000 x g for 15 minutes at 4°C. The chromatin pellets were resuspended with 800ml of Buffer 1, then the two pellets were combined into one 1.7ml Eppendorf tube before pelleting again by centrifugation at 2000 x g for 15 minutes at 4°C. Using a pipet, the pellet was aggressively resuspended in 800ul of Buffer 2 (20mM HEPES pH 7.6, 450mM NaCl, 7.5mM MgCl2, 20mM EDTA, 10% glycerol, 1% NP-40, 2M urea, 0.5M sucrose, and freshly added 1mM DTT and 0.2mM PMSF). The sample was vortexed for 5-10 seconds, then incubated on ice for 5 minutes. The sample was centrifuged at 2000 x g for 15 minutes at 4°C, then pellet was re-suspended in 800ul of Buffer 2. The sample was centrifuged at 2000 x g for 15 minutes at 4°C, then the pellet was re-suspended in 300ul of Buffer 3 (HEPES pH 8.0, 60mM KCl, 15mM NaCl, 5mM MgCl2, 1mM CaCl2) before adding to prepared streptavidin beads for RNPII capture. (Beads were prepared by: 150ul of streptavidin beads (NEB #S1420S) were wash with 700ul of bead buffer (0.5M NaCl. 20mM Tris pH 7.5, 1mM EDTA), then applied to a magnet. Supernatant was discarded then the beads were resuspended with 300ul of Buffer 3). The sample and beads were rotated in 4°C for 2 hours, then the beads were washed three times accordingly: application to a magnet, discarding supernatant, resuspending bead pellet in 500ul Buffer 3. For RNA purification, the beads were resuspended in 500ul of RNA Extraction Buffer (0.3M NaOAc pH 5.3, 1mM EDTA, 1% SDS), 100ul acid phenol, and 100ul chloroform, then centrifuged at max speed for 5 minutes. 480ul of the aqueous phase was combined with 1.2mL of 100% ethanol in a fresh 1.7mL Eppendorf tube. The samples were incubated at -80°C for 1 hour or overnight, then was spun down at max speed for 20 minutes in room temperature. The supernatant was discarded then the pellet was washed with 600ul of 70% ethanol by spinning down at max speed in a centrifuge for 15 minutes in room temperature. For digestion of contaminating DNA, the pellet was resuspended in 100ul of DNAse solution (1x Turbo DNAse Buffer, 10units TURBO DNAse (Invitrogen AM2238)) and incubated at 37°C for 30 minutes. The sample was purified using a RNA Clean and Concentrator-5 kit (Zymo R1013) following the manufacturer's instructions and eluted in 35ul of RNAse-free water.

Library Preparation

Purified RNA (500-775 ng) was prepared for nanopore direct RNA sequencing as follows: PolyA enriched samples were prepared using the ONT SQK-RNA001/SQKRNA002 kit following the manufacturer's instructions. I-tailed RNA was prepared using the ONT SQK-RNA002 kit using a custom poly(C) adapter solution in place of the RTA adapter provided by the kit. Superscript IV (Thermo Fisher) was used for the optional reverse transcription for PolyA and polyI-tailed libraries. To assemble the custom poly(C) adapter, the splint was hybridized in 100pmols top oligo (5′ - pGGCTTCTTCTTGCTCTTAGGTAGTAGGTTC - 3′), and 100pmol bottom oligo (5′ - CCTAAGAGCAAGAAGAAGCCCCCCCCCCCC - 3′), 50mM Tris pH 8, 100mM NaCl, 0.1mM EDTA in 10µl volume and incubating at 75°C, then immediately slow cooling at a ramp rate of 0.1°C/sec to 23°C. The hybridized adapter was diluted to 100µl with 90µl water, and 1µl of adapter solution was used for polyItailed libraries. RNA sequencing on the MinION was performed using ONT R9.4 flow cells and the standard MinKNOW protocol script RNA001 or RNA002 recommended by ONT with one exception: collection of bulk phase raw files for the first 2 hours of sequencing then standard sequencing for ~48 hours.

Base-calling

We used Guppy Base-calling Software from Oxford Nanopore Technologies, Limited. Version 3.0.3+7e7b7d0 with the configuration file "rna_r9.4.1_70bps_hac.cfg" (31) for base-calling direct RNA. NanoFilt version 2.5.0

(32) was used for classification of passed reads. Reads classified as “pass” had phredscore threshold of ≥ 7 and “failed” if < 7. A custom script was used for fastq file “U” to “T” conversion.

Mapping

We used minimap2 (33) to map passed RNA reads determined by NanoFilt version 2.5.0 to (genomes for yeast: sacCer3, or human: GRCh38). For yeast alignments the mapping parameters were -ax splice -uf -k10 -G2000, for human GM12878 alignments the mapping parameters were -ax splice -uf -k14.

Appendix

Table 1: Sequencing Runs

Sample	Poly(A) Prep	Sample	Seq Method	library prep input	# reads	# pass reads
Poly(A) Control #1	Prep 1	Poly(A)	ONT	750ng	1,507,013	1,408,009
Poly (A) Control #2	Prep 2	Poly(A)	ONT	600ng	4,544,817	4,154,482
Poly(A) Control #3	Prep 2	Poly(A)	ONT	600ng	2,976,161	2,755,245
Poly(A) Poly(I) #1	Prep 1	Poly(A)	Poly(I)	750ng	496,472	435,203
Poly(A) Poly(I) #2	Prep 1	Poly(A)	Poly(I)	750ng	206,170	170,135
Poly(A) Poly(I) #3	Prep 1	Poly(A)	Poly(I)	706ng	347,347	302,453
Poly(A) Poly(I) #4	Prep 3	Poly(A)	Poly(I)	600ng	769,807	666,455
Poly(A) Poly(I) #5	Prep 3	Poly(A)	Poly(I)	386ng	34,427	26,291
Poly(A) Poly(I) #6	Prep 3	Poly(A)	Poly(I)	407ng	346,754	297,307
Poly(A) Poly(I) #7	Prep 4	Poly(A)	Poly(I)	600ng	367,679	319,259
Total RNA		Total RNA	Poly(I)	750ug	128,919	57,276
RiboMinus		RiboMinus	Poly(I)	600ng	280,590	79,587
Terminator		Terminator	Poly(I)	696ng	13,746	7650
3' oligo blockers		3' oligo blockers	Poly(I)	750ng	13,243	4496
Chromatin-Associated		Chromatin-Associated	Poly(I)	470ng	320,969	258,896

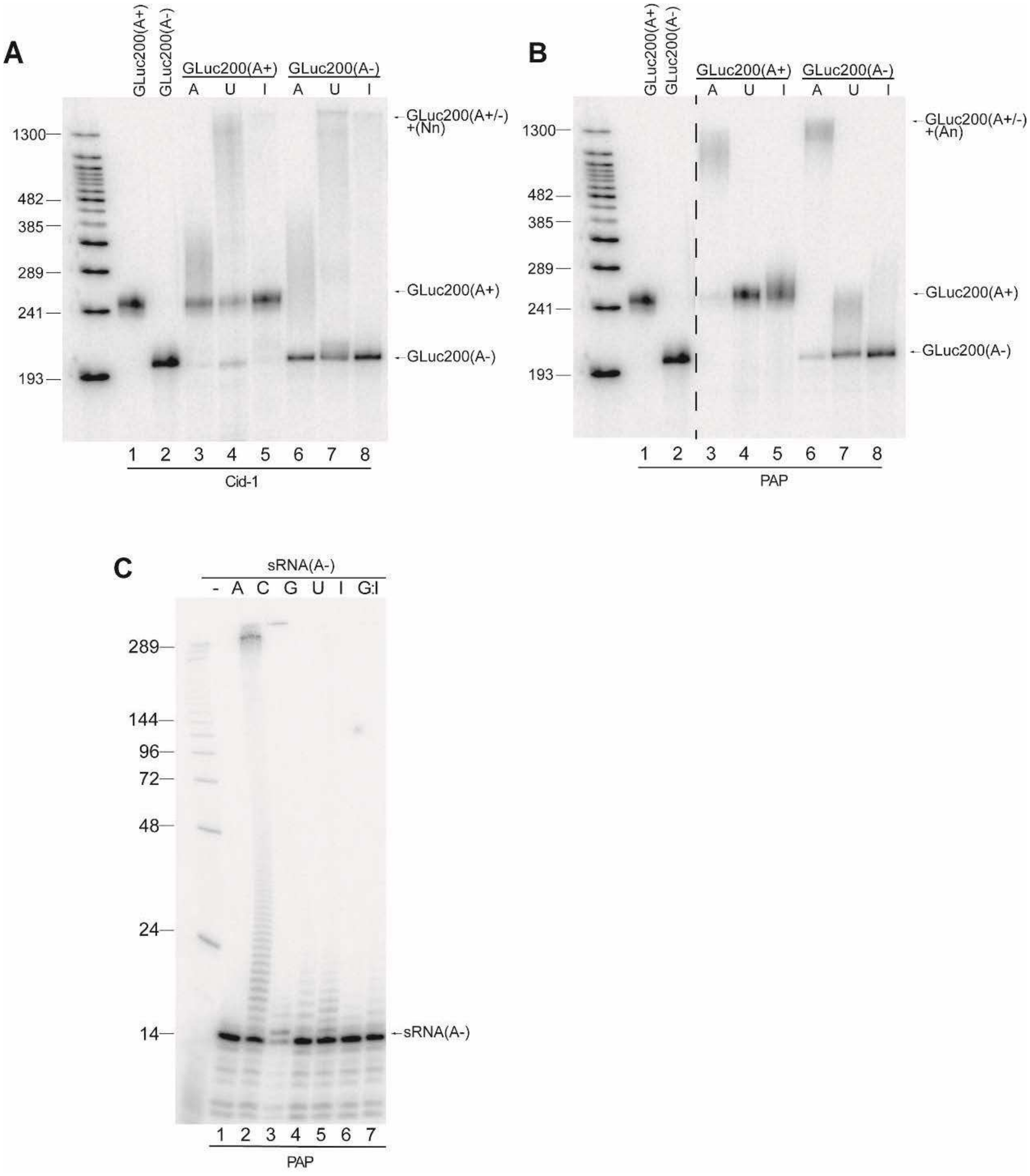
Table 2: Example alignments found in Poly(A) Control, Poly(A)Poly(I), and Total RNA.

BY4741 Total RNA	ORFs identified	rRNA alignments	tRNA alignments	snoRNA alignments
Poly(A) Control #1	5726 (85.3%)	483 (0.04%)	141 (0.01%)	1271 (0.10%)
Poly(A) Poly(I) #1	5518 (82.2%)	26028 (6.20%)	41 (0.01%)	399 (0.10%)
Total RNA	2293 (34.16%)	39197 (73.00%)	0	240 (0.45%)

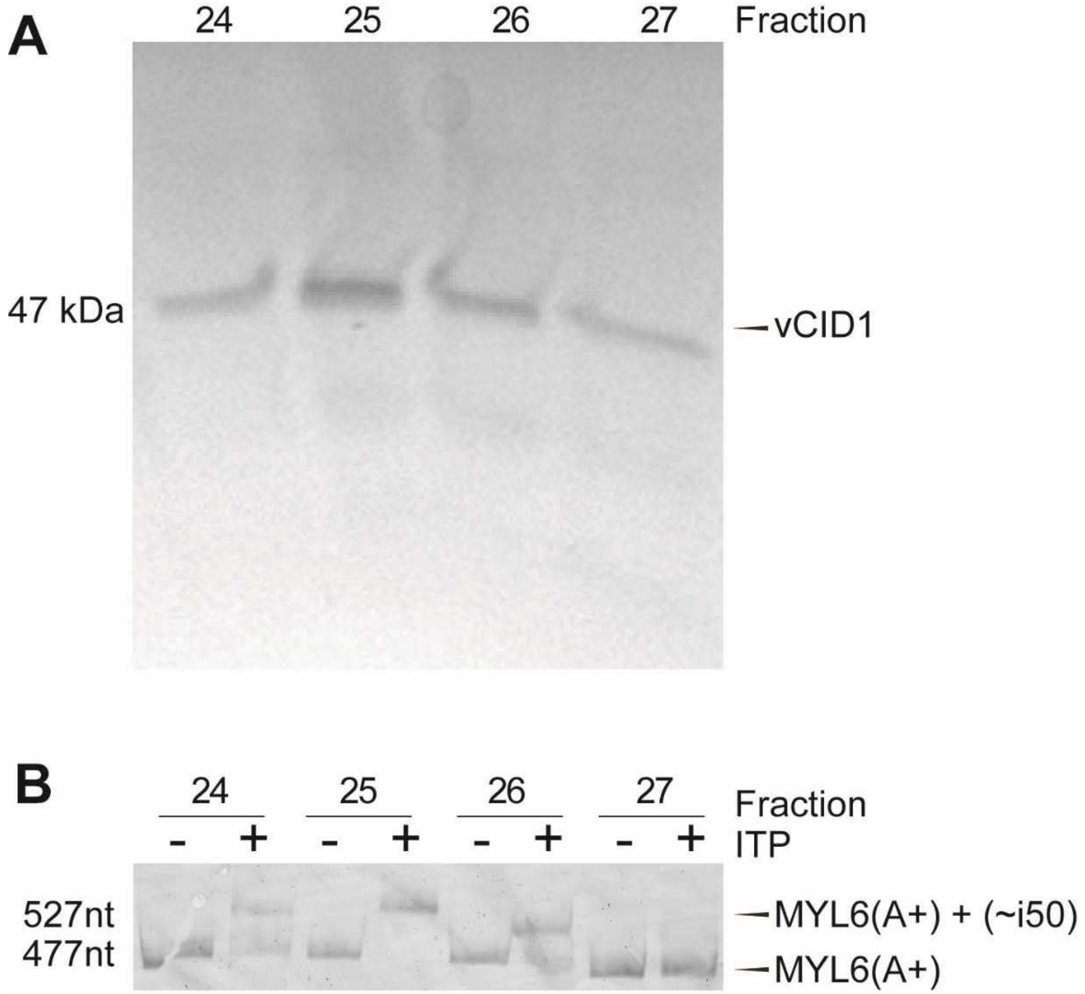
Table 3: Percentage of rRNA found in Total RNA and rRNA depleted samples

rRNA	Terminator (4496)	3' Oligo Blockers (7650)	RiboMinus (69083)	Total RNA (55635)
RDN18-2	532 / 12%	2604 / 34%	2240 / 3%	16098 / 29%
RDN25-2	269 / 6%	724 / 9%	2773 / 4%	9161 / 16%
RDN58-2	16 / 0.3%	19 / 0.24%	28328 / 41%	740 / 1.3%
RDN5-2	4 / 0.09%	1 / 0.01%	57 / 0.08%	23 / 0.04%

Supplement


A
GLuc200(A+)
GLuc200(A-)
GLuc200(A+)
A U I
GLuc200(A-)
A U I
1300
482
385
289
241
193
GLuc200(A+/-) +(Nn)
GLuc200(A+)
GLuc200(A-)
1 2 3 4 5 6 7 8
Cid-1
B
GLuc200(A+)
GLuc200(A-)
GLuc200(A+)
A U I
GLuc200(A-)
A U I
1300
482
385
289
241
193
GLuc200(A+/-) +(An)
GLuc200(A+)
GLuc200(A-)
1 2 3 4 5 6 7 8
PAP
C
sRNA(A-)
- A C G U I G:I
289
144
96
72
48
24
14
sRNA(A-)
1 2 3 4 5 6 7
PAP

Supplemental Figure 1: A test of commercial preparations of two enzymes for their ability to add nucleotide homopolymers to the 3’ end of short RNAs. Incorporation of AMP, UMP, and IMP onto the 3' ends of GLuc200(A+) and GLuc200(A–) using (Panel **A**) Cid-1 Poly(U) Polymerase from New England Biolabs and (Panel **B**) Poly(A) Polymerase from ThermoFisher. Incorporation of rNMPs onto the 3' end of ^{32}P-5' end-labeled 14 nucleotide complex RNA using (Panel **C**) Poly(A) Polymerase from ThermoFisher.


A
24
25
26
27
Fraction
47 kDa
vCID1
B
24
25
26
27
Fraction
- + - + - + - +
ITP
527nt
477nt
MYL6(A+) + (~i50)
MYL6(A+)

Supplemental Figure 2: vCID1 has the same activity as Cid-1 from NEB. (Panel **A**) SDS-PAGE analysis of the protein composition in each FPLC fraction (Panel **B**) Activity of inosine incorporation to MYL6(A+) of each FPLC fraction.

Bibliography

1. Rohan Lowe NS, Mark Bleackley, Stephen Dolan, Thomas Shafee. Transcriptomics technologies. PLos Computational Biology 2017; 13:e1005457.

2. Milos FOaPM. RNA sequencing: advances, challenges and opportunities. Nature Review 2010; 12:87-98.

3. Rory Stark MGaJH. RNA sequencing: the teenage years. Nature Review 2019; 20:631-56.

4. Julie Cocquet AC, Guanglan Zhang, Reiner A. Veitia Reverse transcriptase template switching and false alternative transcripts. Genomics 2006; 88:127-31. 5. Daniel R Garalde EAS, Daniel Jachimowicz, Botond Sipos, Joseph H Lloyd, Mark Bruce, Nadia Pantic, Tigist Admassu, Phillip James, Anthony Warland, Michael Jordan, Jonah Ciccone, Sabrina Serra, Andrew J Heron & Daniel J Turner. highly parallel direct RNA sequencing on an array of nanopores. Nature Methods 2018; 15.

6. Franka J. Rang WPKaJdR. From squiggle to basepair: computational approaches for improving nanopore sequencing read accuracy. Genome Biology 2018; 19.

7. Andrew M. Smith MJ, Logan Mulroney, Daniel R. Garalde, Mark Akeson. Reading canonical and modified nucleobases in 16S ribosomal RNA using nanopore native RNA sequencing. PLos One 2019.

8. Rachael E. Workman ADT, Paul S. Tang, Miten Jain, John R. Tyson,, Roham Razaghi PCZ, Timothy Gilpatrick, Alexander Payne, Joshua Quick,, Norah Sadowski NH, Jaqueline Goes de Jesus, Karen L. Jones, Cameron M. Soulette,, Terrance P. Snutch NL, Benedict Paten, Matthew Loose, Jared T. Simpson,, Hugh E. Olsen ANB, Mark Akeson, and Winston TimpNanopore native RNA sequencing of a human poly(A) transcriptome. Nature Methods 2019; 16:1297–305.

9. Matthew T Parker KK, Anna V Sherwood,, Nicholas J Schurch KM, Peter D Gould, Anthony JW Hall,, Geoffrey J Barton GGS. Nanopore direct RNA sequencing maps the complexity of Arabidopsis mRNA processing and m6A modification. eLife 2020; 9.

10. Ramya Rangan INZ, Rhiju Das. RNA genome conservation and secondary structure in SARS-CoV-2 and SARS-related viruses. bioRxiv 2020.

11. Adrian Viehweger SK, Kevin Lamkiewicz, Ramakanth Madhugiri, John Ziebuhr, Martin Hölzer, Manja Marz. Direct RNA nanopore sequencing of full-length coronavirus genomes provides novel insights into structural variants and enables modification analysis. Genome Research 2019; 29:1545-54.

12. Nathan P. Roach NS, Amelia F. Alessi, Winston Timp, James Taylor, Kim aJK. The full-length transcriptome of C. elegans using direct RNA sequencing. Genome Research 2020; 30:299–312

13. Runsheng Li XR, Qiutao Ding, Yu Bi, Dongying Xie, and Zhongying Zhao. Direct full-length RNA sequencing reveals unexpected transcriptome complexity during Caenorhabditis elegans development. Genome Research 2020; 30:287-98.

14. Manyun Yang AC, Xiaobo Liu, Yaguang Luo, Daniel Sun,, Shaohua Li TG, Luo Sun, Hayden Dillow, Jack Lepine, Mingqun Xu2 and Boce Zhang. Direct Metatranscriptome RNA-seq and Multiplex RT-PCR Amplicon Sequencing on Nanopore MinION –Promising Strategies for Multiplex Identification of Viable Pathogens in Food. Frontiers in Microbiology 2020; 11.

15. Olivia S. Rissland AM, and Chris J. Norbury. Efficient RNA Polyuridylation by Noncanonical Poly(A) Polymerases. Molecular and Cellular Biology 2007; 27:3612-24.

16. Keller GMaW. Tailing and 39-end labeling of RNA with yeast poly(A) polymerase and various nucleotides. RNA 1998; 4:226-30.

17. Davis ABSaRW. The Poly(A) Binding Protein Is Required for Poly(A) Shortening and 60s Ribosomal Subunit-Dependent Translation Initiation. Cell 1989; 58.

18. Nicholas J Loman JQJTS. A complete bacterial genome assembled de novo using only nanopore sequencing data. nature Methods 2015; 12:733–5.

19. John L. Woolford Jr. aSJB. Ribosome Biogenesis in the Yeast Saccharomyces cerevisiae. Genetics 2013; 195:643-81.

20. Warner JR. The economics of ribosome biosynthesis in yeast. Cell 1999; 24:437-40.

21. J. Michael Cherry ELH, Craig Amundsen, Rama Balakrishnan, Gail Binkley, Esther T. Chan KRC, Maria C. Costanzo, Selina S. Dwight,, Stacia R. Engel DGF, Jodi E. Hirschman, Benjamin C. Hitz, Kalpana Karra,, Cynthia J. Krieger SRM, Rob S. Nash, Julie Park, Marek S. Skrzypek,, Matt Simison SWaEDW. Saccharomyces Genome Database: the genomics resource of budding yeast. Nucleic Acids Research 2012; 40:D700–D5.

22. Nikolov DB, Burley SK. RNA polymerase II transcription initiation: A structural view. Proceedings of the National Academy of Sciences 1997; 94:15-22. 23. Neugebauer KM. Nascent RNA and the Coordination of Splicing with Transcription. Cold Spring Harbor Perspectives in Biology 2019.

24. Daiss JLR, NY, US), Nishitani, Kohei (Rochester, NY, US), Schwarz, Edward M. (Rochester, NY, US), Kates, Stephen L. (Pittsford, NY, US). DIAGNOSTIC DEVICE AND METHOD FOR DETECTION OF STAPHYLOCOCCUS INFECTION. United States: UNIVERSITY OF ROCHESTER (Rochester, NY, US), 2017.

25. Joel Bard AMZ, Steffen Helmling, Thomas N. Earnest, Claire L. Moore,

Andrew Bohm. Structure of Yeast Poly(A) Polymerase Alone and in Complex with 3*-dATP. Science 2000; 289.

26. Paola Munoz-Tello CG, and Stephane Thore. Functional Implications from the Cid1 Poly(U) Polymerase Crystal Structure. Cell 2012; 20:977-86.

27. Paola Munoz-Tello CGaST. A critical switch in the enzymatic properties of the Cid1 protein deciphered from its product-bound crystal structure. Nucleic Acids Research 2013; 42:3372-80.

28. Norbury ALSaCJ. The Cid1 family of non-canonical poly(A) polymerases. Yeast 2006; 23:991-1000.

29. Olivia S. Rissland CJN. The Cid1 poly(U) polymerase. Biochimica et Biophysica Acta 2008; 1779:286–94.

30. Matthew Loose SMMS. Real-time selective sequencing using nanopore technology. Nature Methods 2016.