bioRxiv preprint doi: https://doi.org/10.1101/2021.09.22.461331; this version posted October 27, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Nano3P-seq: transcriptome-wide analysis of gene expression and tail dynamics using end-capture nanopore sequencing

Oguzhan Begik[1,2,3], Huanle Liu[1], Anna Delgado-Tejedor[1,5], Cassandra Kontur[6], Antonio J. Giraldez[6], Jean-Denis Beaudoin[7], John S. Mattick[2,4] and Eva Maria Novoa[1,5*]

[1] *Centre for Genomic Regulation (CRG), The Barcelona Institute of Science and Technology, Dr. Aiguader 88, Barcelona 08003, Spain*

[2] *Garvan Institute of Medical Research, 384 Victoria St, Darlinghurst, 2010, NSW, Australia.*

[3] *Faculty of Medicine, UNSW Sydney, Sydney NSW 2052, Australia.*

[4] *School of Biotechnology and Biomolecular Sciences, UNSW Sydney, Sydney NSW 2052, Australia.*

[5] *Universitat Pompeu Fabra, Barcelona, Spain.*

[6] *Department of Genetics, Yale University, USA*

[7] *Department of Genetics and Genome Sciences, Institute for Systems Genomics, University of Connecticut Health Center, Farmington, CT, USA.*

* Correspondence to: Eva Maria Novoa (eva.novoa@crg.eu)

bioRxiv preprint doi: https://doi.org/10.1101/2021.09.22.461331; this version posted October 27, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

ABSTRACT

RNA polyadenylation plays a central role in RNA maturation, fate and stability. In response to developmental cues, polyA tail lengths can vary, affecting the translatability and stability of mRNAs. Here we develop Nano3P-seq, a novel method that relies on nanopore sequencing to simultaneously quantify RNA abundance, tail composition and tail length dynamics at per-read resolution. By employing a template switching-based sequencing protocol, Nano3P-seq can sequence any given RNA molecule from its 3'end, regardless of its polyadenylation status, without the need of PCR amplification or ligation of RNA adapters. We demonstrate that Nano3P-seq captures a wide diversity of RNA biotypes, providing quantitative estimates of RNA abundance and tail lengths in mRNAs, lncRNAs, sn/snoRNAs, scaRNAs and rRNAs. We find that, in addition to mRNAs and lncRNAs, polyA tails can be identified in 16S mitochondrial rRNA in both mouse and zebrafish. Moreover, we show that mRNA tail lengths are dynamically regulated during vertebrate embryogenesis at the isoform-specific level, correlating with mRNA decay. Overall, Nano3P-seq is a simple and robust method to accurately estimate transcript levels and tail lengths in full-length individual reads, with minimal library preparation biases, both in the coding and non-coding transcriptome.

bioRxiv preprint doi: https://doi.org/10.1101/2021.09.22.461331; this version posted October 27, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

INTRODUCTION

RNA molecules are subject to multiple co- and post-transcriptional modifications, shaping them to their final mature form [1]. Polyadenylation of RNA is one such modification, which is known to affect the stability and translatability of the RNA molecule [2–4], and play an essential role in determining the fate of RNA molecules in a wide range of biological processes [5,6].

One context where polyadenylation has been shown to play a major role in determining RNA fate and decay is vertebrate embryogenesis [6]. Indeed, vertebrate embryos undergo a major cellular reprogramming in the first hours post-fertilization, known as the Maternal-to-Zygotic Transition (MZT) [7]. During the MZT, maternally inherited RNAs and proteins are responsible for the activation of the zygotic genome, and are later replaced by the zygotic program [8,9]. Because the MZT begins in a transcriptionally silent embryo, it relies heavily on post-transcriptional regulatory mechanisms [7], including the modulation of the polyadenylation status of the RNA molecules [6,10]. Therefore, characterizing the dynamics of RNA polyadenylation is key to understanding how these modifications regulate the fate and function of RNA molecules.

In the last few years, several transcriptome-wide methods to study the dynamics of polyadenylated (polyA) tails using next-generation sequencing (NGS) have become available, such as PAL-seq or TAIL-seq [10,11]. While these methods have been successfully employed to characterize the dynamics of polyA tail lengths in various contexts, they have several important caveats: (i) they provide a limited perspective on isoform-tail relationships due to the short read length nature of NGS-based technologies; (ii) they do not provide single molecule resolution; (iii) they are severely affected by PCR amplification biases; and (iv) they can only measure tail lengths that are shorter than the read length.

To overcome these limitations, the direct RNA sequencing (dRNA-seq) platform offered by Oxford Nanopore Technologies has been proposed as a means to study both the transcriptome and polyA tail lengths simultaneously [12,13]. To sequence native RNAs using dRNA-seq, polyA-tailed RNA molecules are ligated to a 3' adapter that contains an oligo(dT) overhang (**Figure 1A**). Consequently, dRNAseq libraries will capture the full-length polyA tail; however, ligation will only occur on RNA molecules that anneal to the oligo(dT) overhang, thus capturing exclusively polyadenylated transcripts with tail lengths greater than 10nt. An alternative approach to study the transcriptome using nanopore technologies is direct cDNA sequencing (dcDNA-seq), but this approach is unable to sequence the polyA(-) transcriptome, in addition to being unable to capture the polyA tail length information (**Figure 1A**). Overall, both dRNA-seq and dcDNA-seq nanopore library preparation protocols are limited to the sequencing of polyadenylated transcripts, and thus cannot provide a comprehensive view of both polyadenylated and deadenylated RNA molecules, in addition to being unable to capture RNA molecules with other types of RNA tails (e.g., polyuridine).

bioRxiv preprint doi: https://doi.org/10.1101/2021.09.22.461331; this version posted October 27, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Here, we present a novel method that employs nanopore sequencing to simultaneously obtain per-isoform transcriptome abundance and tail lengths in full-length individual reads, with minimal library preparation steps, which we term **Nano**pore **3 P**rime end-capture **seq**uencing (Nano3P-Seq) (**Figure 1A**). Notably, Nano3P-seq uses template switching to initiate the reverse transcription, and therefore, does not require 3' end adapter ligation steps, PCR amplification, nor second strand cDNA synthesis. We demonstrate that Nano3P-Seq can capture any type of RNA molecule regardless of its 3' sequence, including polyA-tailed and non-tailed RNAs. Moreover, we show that Nano3P-seq can accurately quantify RNA abundances in both the coding and non-coding transcriptome, and can be used to estimate tail lengths in individual RNA molecules, and is highly reproducible across biological replicates.

RESULTS

Nano3P-Seq captures both polyadenylated and non-polyadenylated RNA molecules in a quantitative and reproducible manner

Because nanopore sequencing is typically limited to the analysis of polyA(+) RNA molecules (**Figure 1A**), previous efforts have opted to perform an *in vitro* polyadenylation reaction of the total RNA [14] to capture non-polyadenylated RNAs in the sequencing run. While this option allows capture of any given transcript present in the sample, it also leads to a loss of polyA tail length information. Therefore, we reasoned that by coupling template switching to cDNA nanopore sequencing we would simultaneously capture the polyA(+) and polyA(-) transcriptome, while retaining the polyA tail length information from each individual RNA molecule (**Figure 1A**).

In order to assess the ability of Nano3P-seq to sequence both polyA(+) and polyA(-) RNAs, we first sequenced two synthetic RNAs, one lacking a polyA tail, and a second that had been *in vitro* polyadenylated (see *Methods*) (**Figure S1A-C**). Our results show that Nano3P-seq is able to capture both polyadenylated and non-polyadenylated RNA molecules, as well as the diversity of polyA tail lengths in individual RNAs (**Figure S1C**). We then examined the performance of Nano3P-Seq *in vivo* samples, and sequenced total RNA samples from mouse brain, previously enriched in nuclear and mitochondrial content via subcellular fractionation to increase the content of non-coding RNAs [15] (see *Methods*). We confirmed that Nano3P-seq captured RNA biotypes that are typically polyadenylated (i.e. mRNAs, lincRNAs, processed transcripts) as well as non-adenylated (i.e. rRNA, miscRNA, snoRNA), the majority of them being rRNA and mRNAs (**Figure 1B,** see also **Figure S1D**). In addition, our results confirmed that polyA tail length information was retained in individual reads. Specifically, the majority of reads corresponding to mRNAs had polyA tails (**Figure 1C,** see also **Figure S1E,F**), whereas non-coding RNAs such as snoRNAs (**Figure S1F)** or snRNAs did not have polyA tails (**Figure S1G**), as could be expected.

bioRxiv preprint doi: https://doi.org/10.1101/2021.09.22.461331; this version posted October 27, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

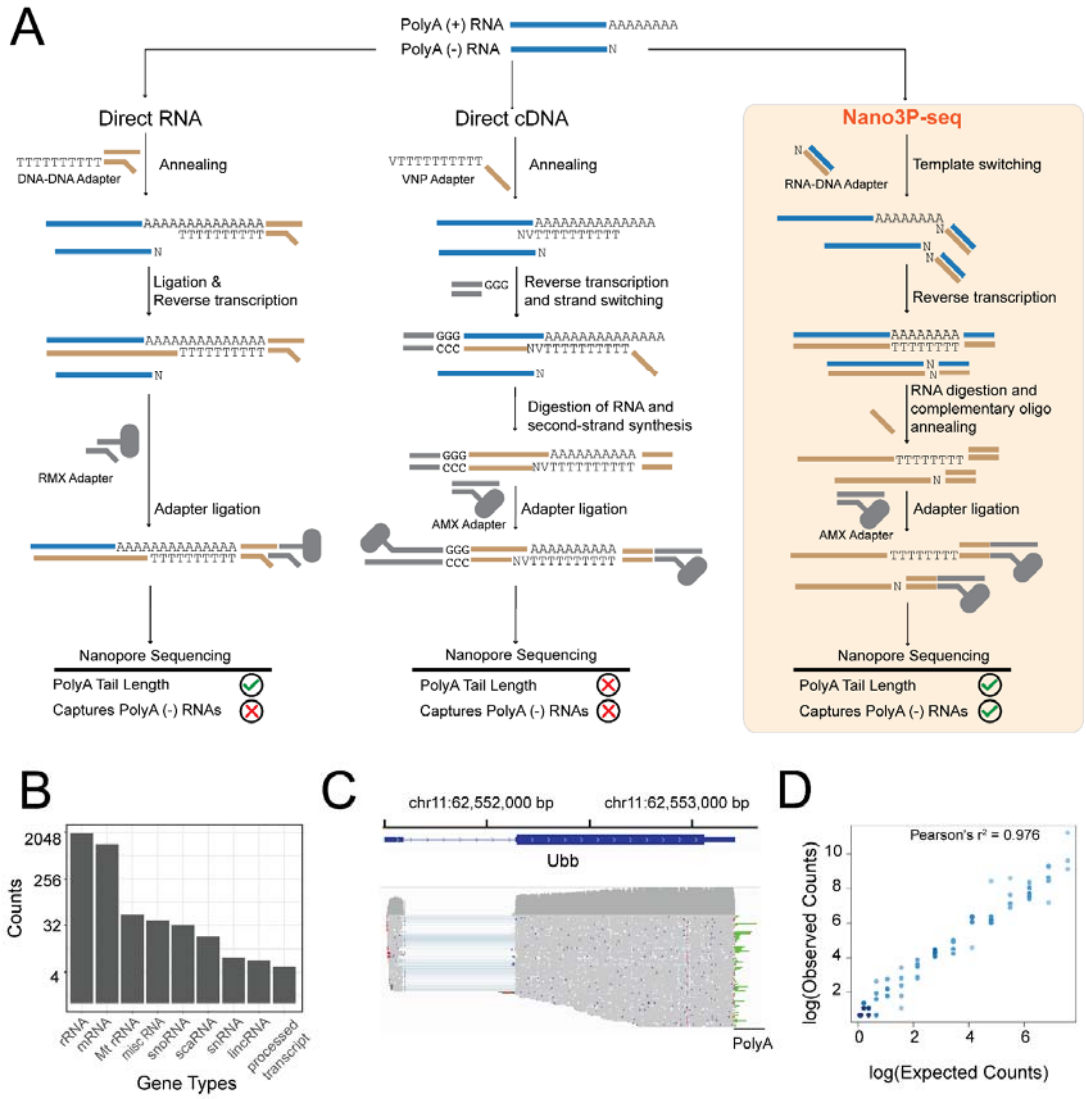

Figure 1. Nano3P-seq captures polyadenylated and non-polyadenylated RNAs, while retaining polyA tail length information. **(A)** Schematic overview comparing 3 different library preparation methods to study the transcriptome using nanopore sequencing: i) standard direct RNA nanopore sequencing (left panel), ii) standard direct cDNA nanopore sequencing (middle panel) and iii) Nano3P-seq (right panel). **(B)** Nano3P-seq captures a wide range of RNA biotypes in the mouse brain. **(C)** IGV snapshot of reads generated with Nano3P-seq, mapped to Ubb gene, illustrating the diversity of polyA tail lengths captured across different reads. The polyA tail region is shown in green. **(D)** Scatterplot of the expected and observed counts of sequins (Pearson's r^2 : 0.976). Each dot represents a sequin. See also Figure S1. Abbreviations: RMX, RNA adapter mix (provided with SQK-RNA002 direct RNA sequencing library preparation kit); AMX: adapter mix (provided with SQK-DCS109, direct cDNA sequencing library preparation kit).

bioRxiv preprint doi: https://doi.org/10.1101/2021.09.22.461331; this version posted October 27, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

To assess the accuracy and reproducibility of Nano3P-seq to quantify RNA abundances, we then examined the performance of Nano3P-seq in synthetic RNA mixes (sequins) [16] that had been spiked into the samples in independent flow cells. Our results showed that Nano3P-seq provided accurate estimates of RNA abundances (Pearson's r^2: 0.976) (**Figure 1D**), and that these quantifications were highly reproducible across replicables (Pearson's r^2: 0.995) (**Figure S1H**).

Nano3P-seq recapitulates the dynamics of coding and non-coding RNAs during vertebrate embryogenesis

Next, we examined the RNA dynamics that occurs during the MZT (**Figure 2A**) at single molecule resolution using Nano3P-seq. To this end, we isolated total RNAs from zebrafish embryos at 2, 4 and 6 hours post-fertilization (hpf) in biological duplicates, ribo-depleted the samples, and sequenced them using the Nano3P-seq protocol **(Figure S2A-B,** see also *Methods*). Quantification of the RNA abundances in both biological replicates showed that per-gene counts obtained using Nano3P-seq were highly reproducible across biological replicates (r^2= 0.931-0.956) (**Figure 2B)**.

A comparative analysis of RNA population dynamics across time points showed that Nano3P-seq recapitulated the transcriptomic switch that occurs during the MZT [7,8], with a drastic decay of maternal mRNAs (**Figure 2C**), in agreement with previous studies [17,18]. Notably, in addition to polyadenylated RNAs, Nano3P-seq captured a wide variety of RNA biotypes without polyA tail that are also present in early embryo stages, finding that the abundance of non-coding RNA populations, including misc RNAs, scaRNAs and snoRNAs, increased as the MZT progressed (**Figure 2D**). By contrast, much fewer non-coding RNA populations were globally captured when using direct RNA nanopore sequencing on the same samples (**Figure 2E**).

We noted, however, that mitochondrial rRNAs were not enriched in Nano3P-seq datasets relative to dRNAseq datasets (**Figure 2F**). Indeed, per-read analysis of mitochondrial rRNA reads revealed that a significant proportion of 16S mitochondrial rRNA contained a polyA tail, which explained the lack of enrichment of mitochondrial rRNAs in Nano3P-seq datasets, relative to dRNAseq datasets. In agreement with this observation, we found that polyA tailed 16S mitochondrial rRNAs were not only present in zebrafish, but also in mouse, suggesting that this feature is conserved across species, and not a sequencing artefact (**Figure S2C-E**), in agreement with previous reports [19].

bioRxiv preprint doi: https://doi.org/10.1101/2021.09.22.461331; this version posted October 27, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

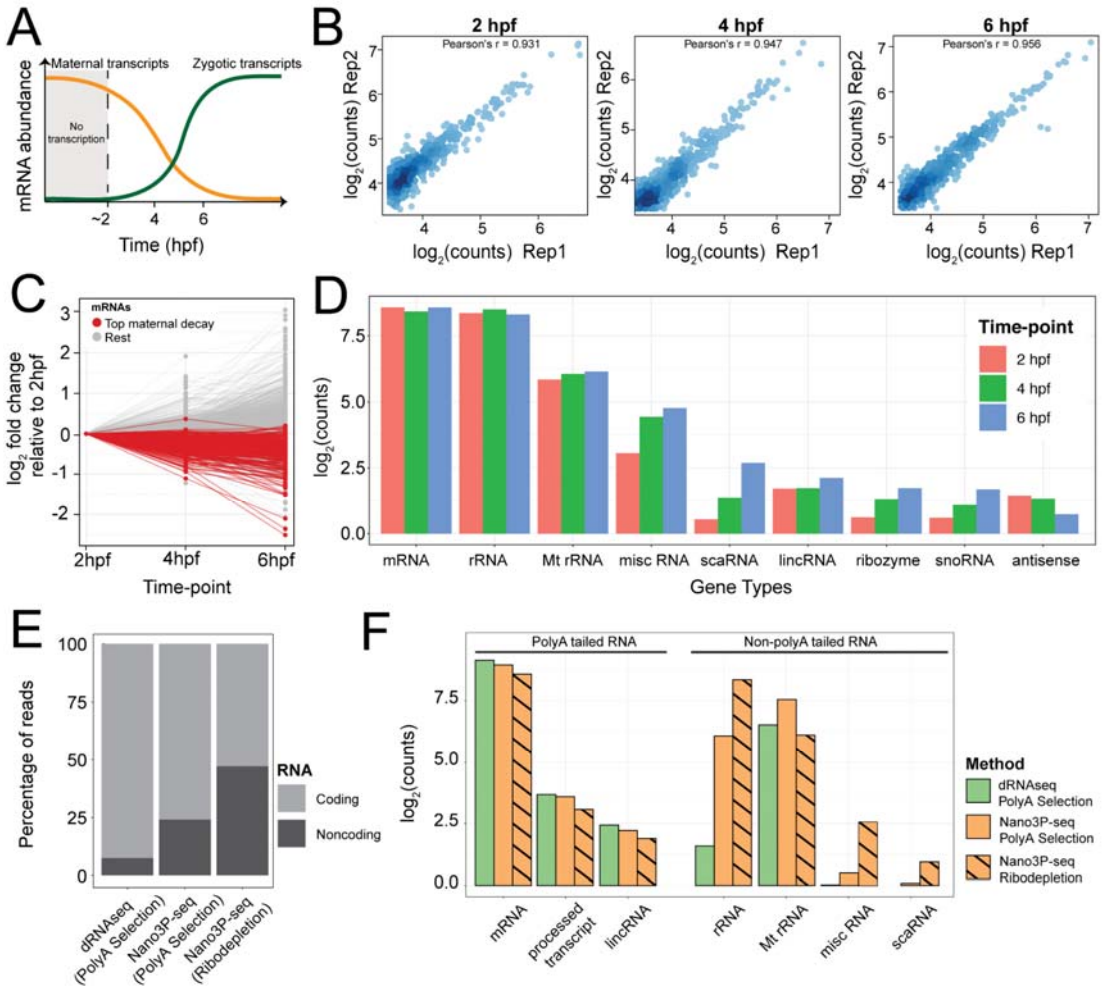

Figure 2. Nano3P-seq captures a wide diversity of coding and non-coding RNAs and their expression dynamics during the MZT. (A) Schematic overview of the transcriptional change that occurs during the maternal-to-zygotic transition (MZT) in zebrafish. **(B)** Scatterplots depicting the correlation of mRNA gene counts between biological replicates in three different time points during the MZT. **(C)** Changes in mRNA abundance during the MZT (t=2, 4 and 6 hpf), relative to 2 hours post-fertilization (hpf). Genes previously reported to have 'maternal decay mode' are depicted in red. **(D)** Barplots depicting the abundance of different RNA biotypes captured by Nano3P-seq during the MZT. **(E)** Relative proportion of coding and noncoding RNAs captured using direct RNA sequencing (on PolyA-selected samples), Nano3P-seq (on PolyA-selected samples) and Nano3P-seq (on Ribo-depleted samples). **(F)** RNA abundances of distinct biotypes captured using direct RNA sequencing (on PolyA -selected samples) (green), Nano3P-seq (on PolyA-selected samples) (orange) and Nano3P-seq (on Ribodepled samples) (dashed orange). See also Figure S2.

bioRxiv preprint doi: https://doi.org/10.1101/2021.09.22.461331; this version posted October 27, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

PolyA tail lengths can be accurately estimated using Nano3P-seq

We then examined whether Nano3P-seq could be used to accurately estimate polyA tail lengths. We should note that algorithms to detect polyA tails in native RNA nanopore sequencing reads are well established and benchmarked [12,20–22], but their applicability to cDNA reads, such as those from Nano3P-seq, remains unclear. To this end, we first examined whether the *tailfindR* polyA tail prediction software [21] would be able to capture the presence or absence of polyA tails on synthetic RNAs that were either polyadenylated or non-polyadenylated and had been sequenced using Nano3P-seq, finding that *tailfindR* was able to capture both polyadenylated and non-polyadenylated Nano3P-seq reads (**Figure 3A)**. Then, we assessed the accuracy of the polyA tail length predictions of *tailfindR* in Nano3P-Seq datasets that included a battery of synthetic RNAs (sequins) [16] with known polyA tail lengths. Our results showed that polyA tail length estimations in Nano3P-seq data were highly reproducible across replicates (r^2= 0.993, **Figure S3A,B**), and with an accuracy similar to that observed when performing polyA tail length estimations in sequins that had been sequenced using dRNAseq (**Figure 3B**). Moreover, the variance of tail length estimates across reads that belonged to the same transcript was smaller in Nano3P-seq datasets than in direct RNA sequencing datasets (**Figure S3C,D**). Finally, we performed a comparative analysis of mouse mRNA polyA tail lengths with those from yeast and zebrafish. We observed that mouse mRNAs showed longest mRNA tails among the 3 species with median polyA tail lengths of 106nt, whereas the shortest polyA tail lengths were observed in yeast, with median polyA tail length of 25nt (**Figure 3C**), in agreement with previous studies [10].

Charting polyA tail length dynamics *in vivo* with Nano3P-seq

We then wondered whether Nano3P-seq could be used to investigate the polyA tail length dynamics *in vivo*. We first examined the ability of Nano3P-seq to properly identify which RNA biotypes were polyadenylated in mouse brain total RNA samples, which had been previously enriched in nuclear/mitochondrial content to increase the proportion of ncRNAs. We found that polyA tails were mainly predicted on mRNAs, but also in lincRNAs and processed transcripts, which are also known to be polyadenylated [23,24] (**Figure 3D,** see also **Table S1**).

We next analyzed the polyA tail length dynamics across developmental stages of zebrafish mRNAs during the MZT (t=2, 4 and 6 hpf). PolyA tail length estimates were highly reproducible across independent biological replicates sequenced in independent flowcells, and for all 3 time points studied (r^2= 0.899-0.970) (**Figure 3E**). We observed an overall increase in mean mRNA polyA tail lengths during the MZT (**Figure 3F**, see also **Figure S3E,F**), in agreement with previous reports (**Figure S4A-B**). All mRNAs examined were found to be polyadenylated, with the exception of histone mRNAs, which had a median polyA tail length of zero (**Figure S3G,** see also **Table S2**), in agreement with previous studies reporting their non-polyadenylated status [25]. Moreover, these findings show that Nano3P-seq is able to capture RNA molecules with structured 3'ends, such as those found in histones [26].

bioRxiv preprint doi: https://doi.org/10.1101/2021.09.22.461331; this version posted October 27, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

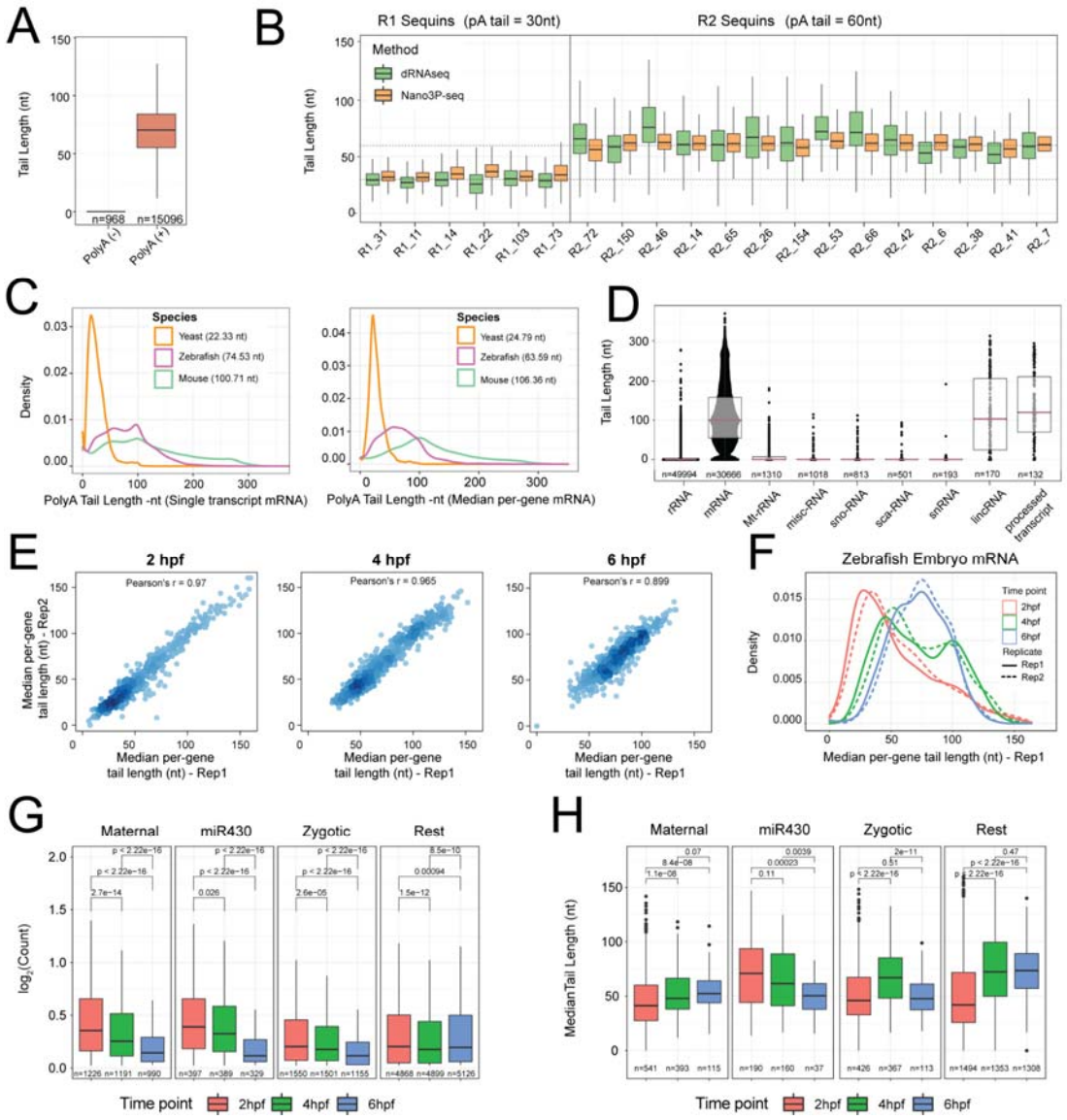

Figure 3. Nano3P-seq can be used to accurately estimate polyA tail lengths in individual molecules. (A) polyA tail length estimates of non-polyadenylated and polyadenylated synthetic RNAs sequenced with Nano3P-seq. **(B)** Comparison of per-read polyA tail length estimates of R1 and R2 sequins, which contain 30 nt and 60 nt polyA tail lengths, respectively, calculated using dRNAseq (green) and Nano3P-seq (orange). **(C)** PolyA tail length distribution of yeast, zebrafish and mouse mRNAs represented as single transcript (left panel) and per-gene median (right panel).**(D)** PolyA tail length estimation of different gene types from nuclear/mitochondrial enriched mouse brain total RNA. Each dot represents a read. **(E)** Replicability of per-gene polyA tail length distributions of zebrafish embryonic mRNAs between two biological replicates, for the three different time points analyzed (2, 4 and 6 hpf) . **(F)** Median per-gene polyA tail length distribution of zebrafish embryonic mRNAs at 2, 4 and 6 hpf. The first biological replicate is represented by a continuous line, whereas the second biological replicate is represented by a dashed line. **(G)** Comparative analysis of the mRNA abundances (shown as log_2 counts) of zebrafish mRNAs that have been binned according to their previously annotated decay mode (maternal, miR-430, zygotic and rest) during embryogenesis (t= 2, 4 and 6 hpf). Statistical comparison of means was performed using Kruskal-Wallis test. **(H)** Median tail length estimations of zebrafish mRNAs that have been binned according to their decay mode (maternal, miR-430, zygotic and rest) at 2, 4 and 6 hpf. Statistical analyses were performed using Kruskal-Wallis test. See also Figures S3 and S4.

bioRxiv preprint doi: https://doi.org/10.1101/2021.09.22.461331; this version posted October 27, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Finally, we examined the correlation between polyA tail length dynamics and mRNA decay. To this end, mRNA transcripts were binned depending on their decay mode (maternal decay, zygotic activation-dependent decay, miR-430-dependent decay and no decay), as previously described [27]. We observed that the 3 groups of mRNAs that are known to decay (maternal, zygotic and miR-430) showed a significant decrease in their mRNA abundances (**Figure 3G**), as could be expected. However, the patterns of polyA tail length dynamics heavily varied depending on the decay mode of the transcript (**Figure 3H**). Specifically, we observed that transcripts that decayed in a mir430-dependent manner showed a significant shortening of their polyA tail lengths during the MZT, in agreement with previous studies [17,28]. By contrast, in mRNAs with zygotic genome activation-dependent decay mode this shortening only occurred after 4 hpf, and maternal mRNAs did not present a shortening in their polyA tail lengths, but rather showed a consistent increase in their tail lengths throughout the MZT. These observations were also consistent with the reanalysis of the PAL-seq data (**Figure S4C-D**). These results show that not all decay modes are associated with shortening of transcript polyA tail lengths, and demonstrate the applicability of Nano3P-seq to identify polyadenylated RNA populations, study their RNA abundance and estimate their polyA tail length dynamics, at both the global level and the level of individual transcripts. Moreover, these results highlight the potential of Nano3P-seq to provide mechanistic insights on different gene regulatory programs.

Nano3P-seq captures isoform-specific differences in polyA tail dynamics during the MZT

A major feature that sets apart nanopore sequencing from next-generation sequencing is its ability to produce long reads, allowing to study polyadenylation dynamics at the isoform level. Thus, we examined whether Nano3P-seq could identify differentially polyadenylated transcript isoforms during the MZT.

To perform isoform-specific polyA tail dynamics analyses, individual reads were first assigned to their corresponding isoform based on the latest genome annotations (see *Methods)*. Only those reads mapping to genes encoding for at least 2 annotated isoforms and with mapping coverage greater than 10 reads per isoform were kept for further analyses. We first compared isoform-specific polyA tail lengths across isoforms encoded by the same gene, finding that 5.2% of analyzed genes presented significant differences in their polyA tail lengths across isoforms (**Table S3**), and that these differences were often conserved across the different time points analyzed (**Figure 4A**). However, in other cases, the behaviour of polyA tails across isoforms was markedly distinct as the MZT progressed (**Figure 4B,C,** see also **Figure S5**), showing that the regulation of polyA tail dynamics occurs at the level of individual isoforms, and that these changes are likely missed in per-gene analyses if the dynamically regulated isoform is not the most abundant.

We then examined whether polyA tail lengths significantly diverged across time points at per-isoform level, finding that 11.7% of analyzed transcripts significantly varied their polyA tail lengths during the MZT (**Figure 4D,E,** see also **Table S4**). These results show that polyA tail length dynamics is not only

bioRxiv preprint doi: https://doi.org/10.1101/2021.09.22.461331; this version posted October 27, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

dependent on the gene and embryogenesis stage, but is also specific to individual transcript isoforms. Moreover, it demonstrates that Nano3P-seq can provide transcriptome-wide measurements of the polyadenylation status of diverse biological samples with both single read and single isoform resolution.

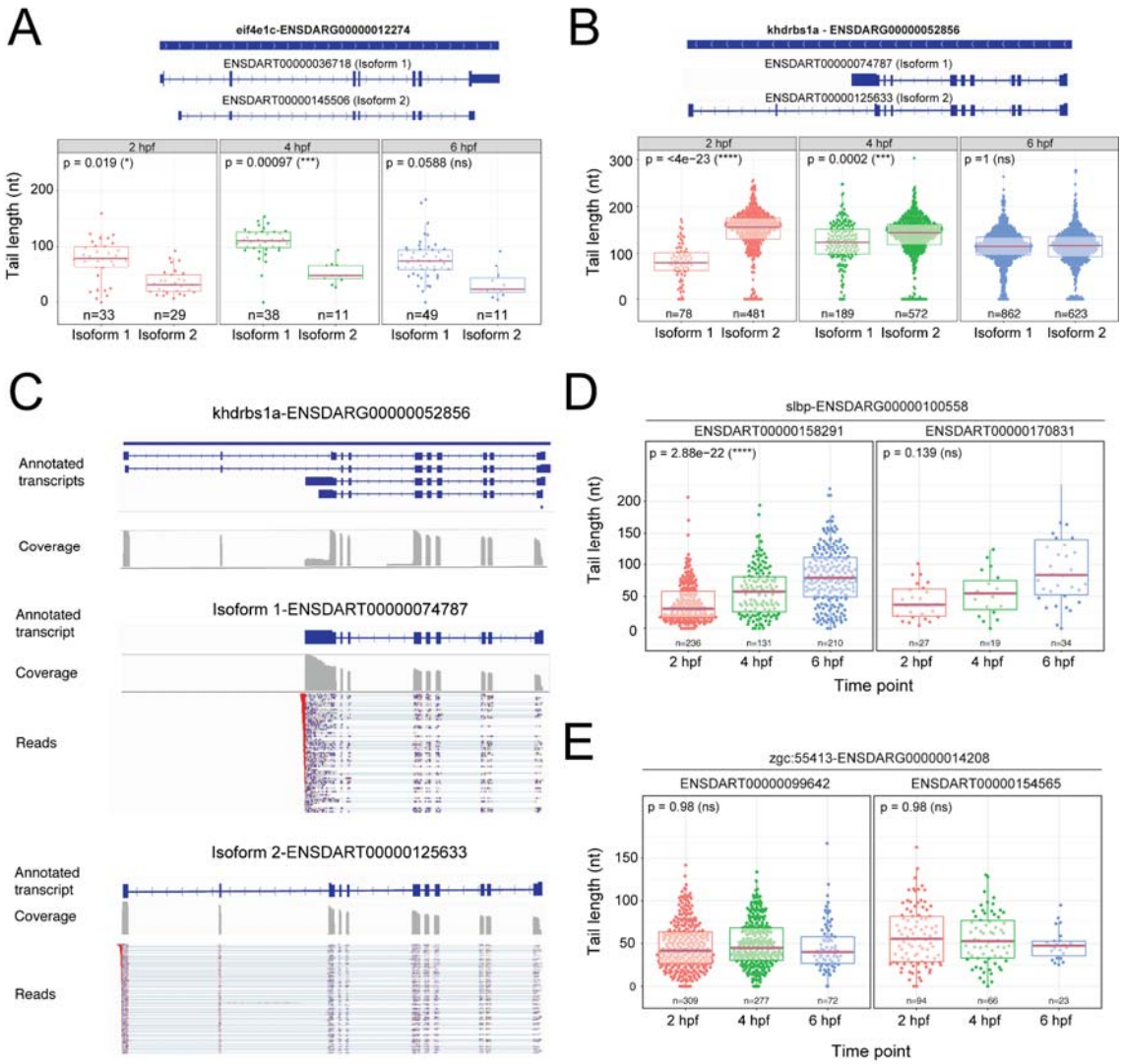

Figure 4. Isoform-specific polyA tail dynamics can be captured using Nano3P-seq. (A,B) Comparison of polyA tail length distributions of reads mapping to distinct isoforms of *eif4e1c* (A) and *khdrbs1a* (B), measured at 3 time points during the zebrafish MZT. Only isoforms with more than 10 reads are shown. Each dot represents the estimated polyA tail length from an individual read. The number of reads included in the analysis is shown below each boxplot. P-values have been computed using Kruskal-Wallis test and corrected for multiple testing using Benjamini-Hochberg. **(C)** Detailed analysis of isoform-specific polyadenylation patterns of the *khdrbs1a* gene, from zebrafish embryos at 2 hpf using Nano3P-seq. All 4 annotated transcripts are shown at the top of the panel, from which only two are detected at 2 hpf. Polyadenylated tails of individual reads are shown in red. **(D,E)** Comparison of polyA tail length distributions at the isoform level across developmental stages from zebrafish embryos during the MZT. Each dot represents an individual read. P-values were computed using Kruskal-Wallis

bioRxiv preprint doi: https://doi.org/10.1101/2021.09.22.461331; this version posted October 27, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

test, and corrected for multiple testing using Benjamini-Hochberg (p>0.05:ns, p≤0.05:*, p≤0.01:**, p≤0.001:***). See also Figure S5.

Detection of isoform-specific RNA modifications using Nano3P-Seq

RNA molecules are decorated with chemical modifications, which have been shown to be essential for the stability, maturation, fate and function of the RNAs [29–32]. Some of these modifications occur in base positions that are involved in the Watson-Crick base-pairing, causing a disruption during the reverse transcription, and consequently, can be seen as increased 'errors' and drop-off rates in RNA-seq datasets [33–36].

We examined the mismatch frequencies of pre-rRNAs and mature rRNAs in mouse and yeast using Nano3P-seq. Pairwise comparison of mismatch frequencies observed in reads mapped to pre-rRNAs relative to those mapped to mature rRNAs revealed that the correlation between the two samples was very high, with the exception of one nucleotide position, which corresponded to the m^1acp^3Y-modified position (**Figure 5A**). This observation was consistent both in mouse and yeast, and was accompanied by a marked drop-off in coverage just before the m^1acp^3Y modified site (**Figure 5B**). Therefore, our results demonstrate that m^1acp^3Y is a rRNA modification that is acquired in late rRNA maturation stages, as it is not present in pre-rRNA molecules, which is in agreement with previous observations [37]. Altogether, our results demonstrate that Nano3P-seq can identify isoform-specific and/or maturation-dependent RNA modifications in the form of altered mismatch frequencies and/or reverse transcription drop-offs.

Analysis of tail composition using Nano3P-seq

Recent works using TAIL-seq have reported a number of terminal modifications in poly(A) tails, such as polyuridine stretches, to play a role in mRNA decay [11,18,38,39]. However, TAIL-seq cannot be used to detect tail modifications within the vast majority of tail nucleotides, as Illumina sequencing quality strongly deteriorates in homopolymeric stretches. On the other hand, PacBio-based approaches such as FLAM-seq can sequence through the entire tails, yet they cannot unambiguously identify 3′ terminal modifications [40].

Here we examined whether Nano3P-seq would identify polyuridylated 3' tail ends. To this end, for each mapped read, we extracted the nucleotide sequence that is found between the nanopore adapter and the last nucleotide mapped to the zebrafish genome, for each developmental stage studied (t=2, 4 and 6hpf). We observed that, in addition to polyadenylated transcripts, Nano3P-seq was able to capture polyuridylated molecules (**Figure 5C,D**), with global U nucleotide content of tails increasing in later developmental stages (**Figure S7A**). We then examined whether polyuridylated samples would be present in mRNAs with shorter RNA tails. We found that tail length significantly decreased with increasing number of U nucleotides in the 3' tail ends both at 2 and 4hpf (**Figure 5C**), suggesting that polyuridylation is playing a role in the decay of molecules during the MZT, in agreement with previous works[18]. By contrast, at 6hpf, polyuridylated RNAs did not globally show shortened tail lengths compared to their A-ended counterparts, suggesting that polyuridylation might

bioRxiv preprint doi: https://doi.org/10.1101/2021.09.22.461331; this version posted October 27, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

be playing a different biological role in later developmental stages, once the maternal transcriptome has been cleared (**Figure 5C, see also Figure S7B**).

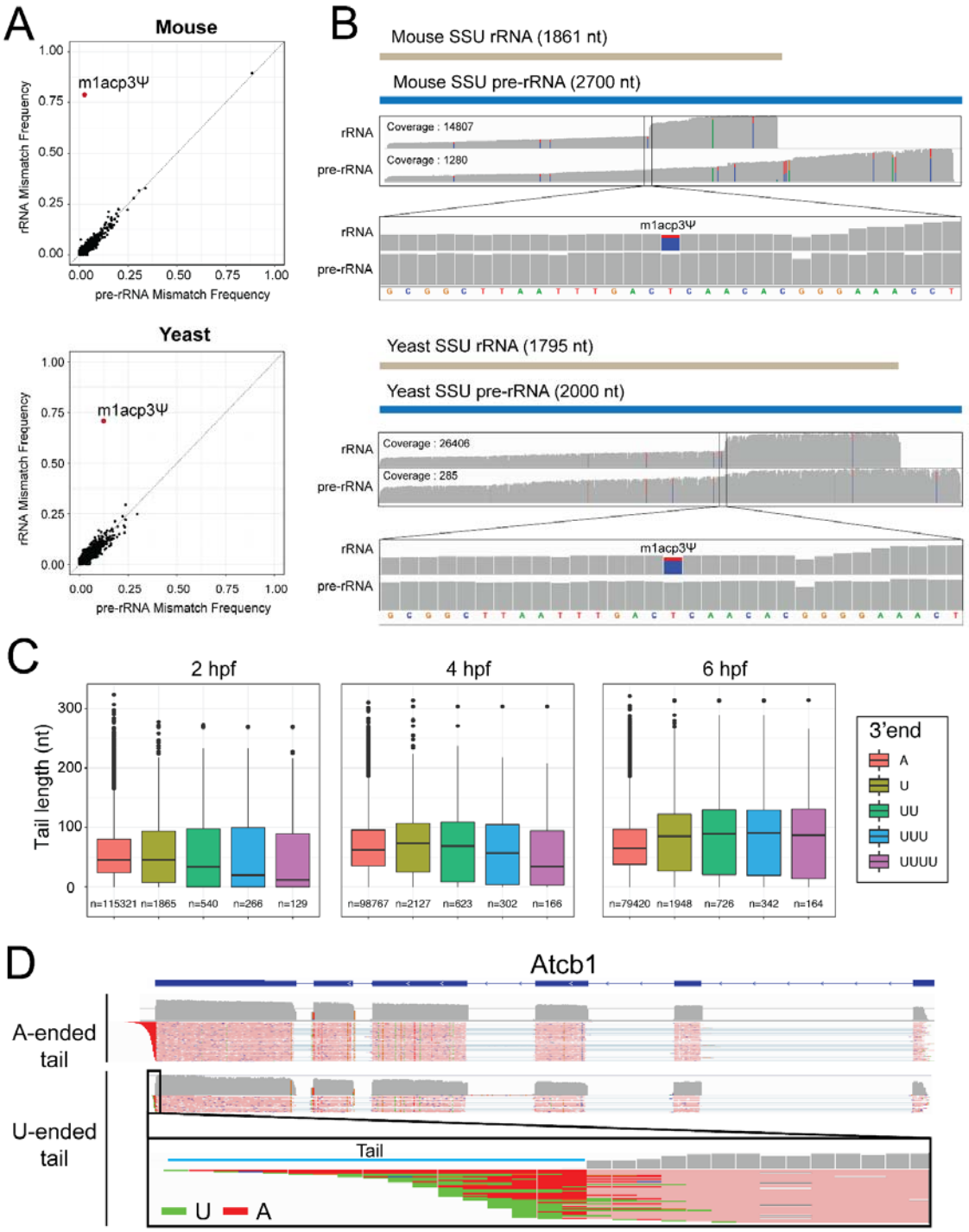

Figure 5. Nano3P-seq identifies differential RNA modified sites in pre-rRNAs and mature rRNAs. (A) Comparison of the per-site mismatch frequencies observed in reads mapping to mouse (upper panel) and yeast (lower panel) 18s pre-rRNA, relative to 18s rRNA, showing that the unique outlier identified is m^1acp^3Y. **(B)** IGV coverage tracks of reads mapping to mouse (upper panel) and yeast (lower panel) 18s rRNA (upper track) and 18s pre-rRNA (lower track), including a zoom on the position that is known to be modified with m^1acp^3Y.

bioRxiv preprint doi: https://doi.org/10.1101/2021.09.22.461331; this version posted October 27, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Positions with mismatch frequency lower than 0.1 are shown in gray. **(C)** Tail length estimation of RNAs binned based on their 3' tail end composition: A-ended (red), U-ended (olive), UU-ended (green), UUU-ended (blue) and UUUU-ended (purple). The number of molecules included in each bin is depicted under each boxplot. Data has been computed for zebrafish ribodepleted samples extracted at t=2hpf (left panel), 4hpf (middle panel) and 6hpf (right panel). **(D)** IGV snapshots of the reads mapping to *actb1* gene in zebrafish samples (t=6hpf). Reads ending with A nucleotide are grouped as "A-ended tail", whereas reads ending with U nucleotide are grouped as "U-ended tail". See also Figure S7.

DISCUSSION

In the last few years, a variety of NGS-based high-throughput methods have been developed to characterize the 3' ends of RNA molecules at a transcriptome-wide scale, including methods to reveal polyA tail sites (e.g., 3P-seq [6], PAS-seq [41], PAT-seq [42]) and to estimate polyA tail lengths (e.g., PAL-seq [10], TAIL-seq [11], mTAIL-seq [43]). A major limitation of NGS-based methods, however, is their inability to assign a given polyA tail length to a specific transcript isoform, thereby losing the isoform-specific tail length information. In addition, NGS-based methods cannot measure tail lengths greater than the read length, thus biasing our view of polyA tail dynamics to those transcripts that display shorter tail lengths.

More recently, novel methods to estimate polyA tail lengths using Pacific BioSciences (PacBio) long-read sequencing technologies have been developed, such as FLAM-seq [40] and PAIso-seq [44]. Compared to NGS, these methods are able to capture isoform-tail relationships; however, they are still affected by PCR amplification biases and ligation biases, in addition to producing relatively modest outputs in terms of number of reads [45–47]. Moreover, PacBio typically requires expensive sequencing instruments that are not widely available. On the other hand, direct RNA nanopore sequencing [48] has also been proposed as an alternative long-read sequencing technology to study polyA tail lengths [12,13]; however, this approach is unable to capture deadenylated RNAs, molecules with non-canonical tailings (e.g. polyuridine), or molecules with polyA tails shorter than 10 nucleotides (**Figure 1A**), thus biasing the view of the transcriptome towards polyadenylated molecules. Therefore, Nano3P-seq addresses these limitations by offering a simple, robust and cost-effective solution to study the coding and non-coding transcriptome simultaneously regardless of the presence or absence of polyA tail or 3' tail composition, without PCR nor ligation biases, and with single-read and single-isoform resolution. Moreover, the use of TGIRT as reverse transcriptase in the Nano3P-seq protocol not only maximizes the production of full-length cDNAs, but also ensures the inclusion of RNA molecules that are highly structured and/or modified, which would often not be captured -or their representation would be significantly biased- using standard viral reverse transcriptases [49,50].

Nano3P-seq provides quantitative measurements of RNA abundances (**Figure 1D**) as well as captures diverse RNA biotypes regardless of their tail end composition (**Figure 2D**). We show that it can be applied to diverse species with a distinct range of polyA tail lengths (**Figure 3B,C**), and can be used to study the dynamics of polyadenylation (Figure **3F-H** and **4**). Specifically, we demonstrate how

bioRxiv preprint doi: https://doi.org/10.1101/2021.09.22.461331; this version posted October 27, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Nano3P-seq provides per-read resolution transcriptome-wide maps of RNA abundance and polyadenylation dynamics during the zebrafish MZT. Our results show that transcripts targeted by mir-430 decay in a deadenylation-dependent manner, whereas those targeted by the maternal and zygotic decay programs have distinct polyA tail length dynamics during the MZT (**Figure 3G,H**). Moreover, we identified isoform-specific regulation of polyadenylation, demonstrating that analyses at per-gene level are insufficient to capture the dynamics of polyadenylation during the zebrafish MZT (**Figure 4A,B**, see also **Figure S5**). Overall, we demonstrate that Nano3P-seq can successfully identify polyadenylation changes across time points, across mRNA decay programs, and across isoforms, providing mechanistic insights on different gene regulatory programs.

Using Nano3P-seq, we compared the zebrafish transcriptome from both ribo-depleted and polyA+ selected transcriptomes during the zebrafish MZT. Because the vast majority of cellular RNA is composed of ribosomal RNA (rRNA), transcriptomic studies typically remove a significant portion of rRNA molecules to sequence a wider diversity of RNA biotypes. This can be achieved by i) ribo-depletion of the sample using biotinylated oligos that are complementary to rRNAs, or ii) via selective enrichment of polyA+ transcripts using oligo(dT) beads. We should note, however, that while these two approaches are often used interchangeably, its effect in the transcriptome composition is not equal. Nano3P-seq allows us to compare the effects of these two approaches both on the transcriptome composition and polyA tail length distribution. In terms of its effects in transcriptome composition, we find that ribo-depletion captures a larger variety of RNA biotypes compared to polyA+ selection, including several non-polyA tailed RNA biotypes, as expected (**Figure 2F**). However, we did not observe a significant difference in the distribution of mRNA polyA tail lengths between the two methods (**Figure S6**), suggesting that oligo(dT) enrichments do not significantly bias the mRNA populations by preferentially enriching for those having longer polyA tails.

In addition, we performed a comparative analysis of zebrafish polyA tail lengths in libraries sequenced using either Nano3P-seq or direct RNA sequencing (dRNA-seq). We show that Nano3P-seq captures RNA molecules regardless of their tail ends, resulting in the capture of diverse RNA biotypes (**Figure 1B** and **S1C-G**) including deadenylated mRNA molecules (**Figure S3E-F** and **S8A**) as well as polyuridylated molecules (**Figure 5C-D** and **S7**). By contrast, dRNA-seq could only capture longer polyadenylated transcripts as it relies on the presence of polyA tail lengths greater than 10 nucleotides. Indeed, when comparing the distribution of per-read polyA tail length estimations of mRNAs, we observed that Nano3P-seq captured mRNAs with predicted zero tail lengths, whereas dRNA-seq only captured reads with longer tails (**Figure S8**).

Overall, our work demonstrates that Nano3P-seq can simultaneously capture both non-polyA tailed and polyA tailed transcriptome, making it possible to accurately quantify the RNA abundances and polyA tail lengths at a per-read and per-isoform level, while minimizing the amount of biases introduced during the library preparation. These features set Nano3P-seq as a potent, and low-cost

bioRxiv preprint doi: https://doi.org/10.1101/2021.09.22.461331; this version posted October 27, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

method that can provide mechanistic insights on the regulation of RNA molecules and improve our understanding of mRNA tailing processes and post-transcriptional control.

MATERIAL AND METHODS

In vitro transcription of RNAs

The synthetic 'curlcake' sequences [51] (Curlcake 1, 2244 bp and Curlcake 2, 2459 bp) were *in vitro* transcribed using Ampliscribe™ T7-Flash™ Transcription Kit (Lucigen-ASF3507). Curlcake 2 was polyadenylated using *E. coli* polyA Polymerase (NEB-M0276S). polyA-tailed RNAs were purified using RNAClean XP beads. The quality of the *in vitro* transcribed (IVT) products as well as the addition of polyA tail to the synthetic constructs was assessed using Agilent 4200 Tapestation **(Figure S1A**). Concentration was determined using Qubit Fluorometric Quantitation. Purity of the IVT products was measured with NanoDrop 2000 Spectrophotometer.

Yeast culturing and total RNA extraction

Saccharomyces cerevisiae (strain BY4741) was grown at 30ºC in standard YPD medium (1% yeast extract, 2% Bacto Peptone and 2% dextrose). Cells were then quickly transferred into 50 mL pre-chilled falcon tubes, and centrifuged for 5 minutes at 3,000 g in a 4ºC pre-chilled centrifuge. Supernatant was discarded, and cells were flash frozen. Flash frozen pellets were resuspended in 700 µL Trizol with 350 µL acid washed and autoclaved glass beads (425-600 µm, Sigma G8772). The cells were disrupted using a vortex on top speed for 7 cycles of 15 seconds (the samples were chilled on ice for 30 seconds between cycles). Afterwards, the samples were incubated at room temperature for 5 minutes and 200 µL chloroform was added. After briefly vortexing the suspension, the samples were incubated for 5 minutes at room temperature. Then they were centrifuged at 14,000 g for 15 minutes at 4ºC and the upper aqueous phase was transferred to a new tube. RNA was precipitated with 2X volume Molecular Grade Absolute ethanol and 0.1X volume Sodium Acetate. The samples were then incubated for 1 hour at -20ºC and centrifuged at 14,000 g for 15 minutes at 4ºC. The pellet was then washed with 70% ethanol and resuspended with nuclease-free water after air drying for 5 minutes on the benchtop. Purity of the total RNA was measured with the NanoDrop 2000 Spectrophotometer. Total RNA was then treated with Turbo DNase (Thermo, #AM2238) with a subsequent RNAClean XP bead cleanup.

RNA isolation from mouse brain

In order to isolate nuclear/mitochondrial-enriched RNA from the mouse (*Mus musculus*) brain, we followed previously published protocols [52] with minor changes. A quarter of a C57BL6/J mouse brain was used for this protocol, and all samples and reagents were kept on ice during the protocol. Brain tissue was mined with a razor blade into smaller pieces. Cold Nuclei EZ Lysis Buffer (0.01 M Tris-Cl,pH7.5, 0.06M KCl, 0.001M EDTA, 1X Protease Inhibitor, 0.5% NP40) was added to the tissue in 1.5 mL eppendorf tube. The sample was homogenized using a dounce, and the homogenate was

bioRxiv preprint doi: https://doi.org/10.1101/2021.09.22.461331; this version posted October 27, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

transferred into a 2mL eppendorf tube. 1 mL of cold Nuclei EZ Lysis Buffer was added and mixed, followed by 4 minutes incubation on ice. During the incubation, the sample was gently mixed a couple of times using a pipette. Homogenate was filtered using a 70 um strainer mesh, and the flowthrough was collected in a polystyrene round-bottom FACS tube and subsequently transferred into a new 2 mL tube. The sample was centrifuged at 500g for 5 minutes at 4°C and the supernatant was removed. The nuclei/mitochondria enriched sample was resuspended in another 1.5 mL EZ Lysis buffer and incubated for 5 minutes on ice. The sample was centrifuged at 500 g for 5 minutes 4°C and the supernatant was discarded (cytoplasm). 500 uL Nuclei Wash and Resuspension Buffer (1M PBS, 1%BSA, RNase Inhibitor) was added to the sample and incubated for 5 minutes without resuspending to allow buffer interchange. After incubation, 1 mL of Nuclei Wash and Resuspension Buffer was added and the sample was resuspended. The sample was centrifuged at 500g for 5 minutes at 4°C. The supernatant was removed and only ~50 ul was left. Using 1.4 mL Nuclei Wash and Resuspension Buffer, the sample was resuspended and transferred to a 1.5 mL eppendorf tube. The last washing step was repeated and the pellet was resuspended in 300 uL Nuclei Wash and Resuspension Buffer. RNA was extracted using Trizol.

Zebrafish breeding

Wild-type zebrafish (*Danio rerio*) embryos were obtained through natural mating of the TU-AB strain of mixed ages (5–18 months). Mating pairs were randomly chosen from a pool of 60 males and 60 females allocated for each day of the month. Embryos and adult fish were maintained at 28 °C.

Zebrafish total RNA extraction and polyA selection

For RNA samples, 25 embryos per developmental stage and per replicate were collected and flash frozen in liquid nitrogen. Frozen embryos were thawed and lysed in 1 mL TRIzol (Life Technologies) and total RNA was extracted using the manufacturer's protocol. Total RNA concentration was calculated by nanodrop.

For polyA-selected RNA samples, polyadenylated RNAs were isolated with oligo (dT) magnetic beads (New England BioLabs) according to the manufacturer's protocol and eluted in 30 µL prior to nanodrop quantification.

Zebrafish total RNA ribodepletion

Ribodepletion was performed on zebrafish total RNA using riboPOOL oligos (siTOOLs, cat #055) following the manufacturer's protocol. Briefly, 5 ug total RNA in 14 uL was mixed with 1 uL resuspended riboPOOL oligos, 5 uL hybridization buffer and 0.5 uL SUPERase•In RNase Inhibitor (Thermo Fisher, AM2694). The mix was incubated for 10 minutes at 68°C, followed by a slow cool down (3°C/min) to 37°C for hybridization. In the meantime, Dynabeads MyOne Streptavidin C1 (Thermo Fisher, 65001) beads were resuspended by carefully vortexing at medium speed. 80 uL of bead resuspension (10 mg/mL) was transferred into a tube, which then was placed on a magnetic rack. After aspirating the supernatant, 100 uL of bead resuspension buffer was added to the sample and beads were resuspended in this buffer by agitating the tube. Sample was placed on a magnet

bioRxiv preprint doi: https://doi.org/10.1101/2021.09.22.461331; this version posted October 27, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

and the supernatant was aspirated. This step was performed twice. Beads were then resuspended in 100 uL of bead wash buffer and placed on magnet in order to aspirate the supernatant. Beads were then resuspended in a 160 uL depletion buffer. This suspension was then divided into two tubes of 80 uL, which will be used consecutively. 20 uL of hybridized riboPOOL and total RNA was briefly centrifuged to spin down droplets and it was pipetted into the tube containing 80 uL of beads in depletion buffer. The tube containing the mix was agitated to resuspend the solution well. Then the mix was incubated at 37°C for 15 minutes, followed by a 50°C incubation for 5 minutes. Immediately before use, the second tube containing 80 ul of beads was placed on a magnetic rack and the supernatant was aspirated. After the incubation at 50°C, the first depletion reaction was placed on a magnet and the supernatant was transferred into the tube containing the other set of beads. The mix was incubated again at 37°C for 15 minutes, followed by a 50°C incubation for 5 minutes. After briefly spinning down the droplets, the mix was placed on a magnet for 2 minutes. The supernatant was transferred into a different tube and cleaned up using RNA Clean & Concentrator-5 (Zymo, R1013).

Nano3P-Seq library preparation

The protocol is based on the direct cDNA Sequencing ONT protocol (DCB_9091_v109_revC_04Feb2019), with several modifications to be able to perform TGIRT template switching. Before starting the library preparation, 1 µL of 100 µM R_RNA (Oligo: 5' rGrArArGrArUrArGrArGrCrGrArCrArGrGrCrArArGrUrGrArUrCrGrGrArArG/3SpC3/ 3') and 1 µL of 100 µM D_DNA (5' /5Phos/CTTCCGATCACTTGCCTGTCGCTCTATCTTCN 3') were mixed with 1 µL 0.1 M Tris pH 7.5, 1 µL 0.5 M NaCl, 0.5 ul RNAse Inhibitor Murine (NEB, M0314S) and 5.5 ul RNase-free water. The mix was incubated at 94°C for 1 minute and the temperature was ramped down to 25°C (-0.1°C/s) in order to pre-anneal the oligos. Then, 100 ng RNA was mixed with 1 µL pre-annealed R_RNA+D_DNA oligo, 1uL 100 mM DTT, 4 µL 5X TGIRT Buffer (2.25 M NaCl, 25 mM MgCl2, 100 mM Tris-HCl, pH 7.5), 1 µL RNasin® Ribonuclease Inhibitor (Promega, N2511), 1 µL TGIRT (InGex) and nuclease-free water up to 19 µL. The reverse-transcription mix was initially incubated at RT for 30 minutes before adding 1 µL 10 mM dNTP mix. Then the mix was incubated at 60°C for 60 minutes and inactivated by heating at 75°C for 15 minutes before moving to ice. RNAse Cocktail (Thermo Scientific, AM2286) was added to the mix in order to digest the RNA, and the mix was incubated at 37°C for 10 minutes. The reaction was then cleaned up using 0.8X AMPure XP Beads (Agencourt, A63881). In order to be able to ligate the sequencing adapters to the first cDNA strand, 1 µL 100 µM CompA_DNA (5' GAAGATAGAGCGACAGGCAAGTGATCGGAAGA 3') was annealed to the 15 µL cDNA in a tube with 2.25 µL 0.1 M Tris pH 7.5, 2.25 µL 0.5 M NaCl and 2 µL nuclease-free water. The mix was incubated at 94°C for 1 minute and the temperature was ramped down to 25 °C (-0.1°C/s) in order to anneal the complementary to the first strand cDNA. Then, 22.5 µL first strand cDNA was mixed with 5 µL Adapter Mix (AMX), 22.5 µL Rnase-free water and 50 µL Blunt/TA Ligase Mix (NEB, M0367S) and incubated in room temperature for 10 minutes. The reaction was cleaned up using 0.8X AMPure XP beads, using ABB Buffer for washing. The sample was then eluted in Elution Buffer (EB) and mixed with Sequencing Buffer (SQB) and Loading Beads (LB) prior to loading onto a primed R9.4.1 flowcell. Libraries were run on either Flongle or MinION flowcells with MinKNOW acquisition

bioRxiv preprint doi: https://doi.org/10.1101/2021.09.22.461331; this version posted October 27, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

software version v.3.5.5. A detailed step-by-step Nano3P-seq protocol is provided as an additional file (**File S1**).

Annealing based direct cDNA-Sequencing library preparation with TGIRT

Some adjustments were made to the original Direct cDNA-Sequencing ONT protocol (SQK-DCS109), in order to be able to use TGIRT (InGex) as reverse transcription enzyme for nanopore sequencing, as this enzyme does not produce CCC overhang that is typically exploited by the direct cDNA sequencing library preparation protocol (**Figure 1A**). Briefly, 1 µL of 100 µM Reverse transcription primer VNP (5' /5Phos/ACTTGCCTGTCGCTCTATCTTCTTTTTTTTTTTTTTTTTTTTVN 3') and 1 µL of 100 µM of *in-house* designed complementary oligo (CompA: 5' GAAGATAGAGCGACAGGCAAGTA 3') were mixed with 1 µL 0.2 M Tris pH 7.5, 1 µL 1 M NaCl and 16ul RNase-free water. The mix was incubated at 94°C for 1 minute and the temperature was ramped down to 25°C (-0.1°C/s) in order to pre-anneal the oligos. Then, 100 ng polyA-tailed RNA was mixed with 1 µL pre-annealed VNP+CompA,, 1uL 100 mM DTT, 4 µL 5X TGIRT Buffer (2.25 M NaCl, 25 mM MgCl2, 100 mM Tris-HCl, pH 7.5) , 1 µL RNasin® Ribonuclease Inhibitor (Promega, N2511), 1 µL TGIRT and nuclease-free water up to 19 µL. The reverse-transcription mix was initially incubated at RT for 30 minutes before adding 1 µL 10 mM dNTP mix. Then the mix was incubated at 60°C for 60 minutes and inactivated by heating at 75°C for 15 minutes before moving on to ice. Furthermore, RNAse Cocktail (Thermo Scientific, AM2286) was added to the mix in order to digest the RNA and the mix was incubated at 37°C for 10 minutes. Then the reaction was cleaned up using 0.8X AMPure XP Beads (Agencourt, A63881). In order to be able to ligate the sequencing adapters the the first strand, 1 µL 100 µM CompA was again annealed to the 15 µL cDNA in a tube with 2.25 µL 0.1 M Tris pH 7.5, 2.25 µL 0.5 M NaCl and 2 µL nuclease-free water. The mix was incubated at 94°C for 1 minute and the temperature was ramped down to 25 °C (-0.1°C/s) in order to anneal the complementary to the first strand cDNA. Furthermore, 22.5 µL first strand cDNA was mixed with 5 µL Adapter Mix (AMX), 22.5 uL Rnase-free water and 50 µL Blunt/TA Ligase Mix (NEB, M0367S) and incubated in room temperature for 10 minutes. The reaction was cleaned up using 0.8X AMPure XP beads, using ABB Buffer for washing. The sample was then eluted in Elution Buffer (EB) and mixed with Sequencing Buffer (SQB) and Loading Beads (LB) prior to loading onto a primed R9.4.1 flowcell and ran on a MinION sequencer with MinKNOW acquisition software version v.3.5.5.

Analysis of dRNA datasets

Barcoded direct RNA sequencing (dRNA-seq) run was demultiplexed using *DeePlexiCon* [53]. Reads with demultiplexing confidence scores greater than 0.95 were kept for downstream analyses. For sequins, reads were base-called using stand-alone Guppy version 3.0.3 with default parameters and then the base-called reads were mapped to sequin sequences [16] with minimap2 with -ax splice -k14 -uf --MD parameters [54]. For zebrafish dRNA-seq samples, reads were base-called with Guppy version 4.0. Base-called reads were first mapped to maternal and somatic zebrafish ribosomal RNA sequences taken from [55] and then to the genome (GRCz11) with minimap2 [54] with -ax splice -k14 -uf -

bioRxiv preprint doi: https://doi.org/10.1101/2021.09.22.461331; this version posted October 27, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

-MD parameters. Mapped reads were intersected with ENSEMBL version 103 annotation (Danio_rerio.GRCz11.103.2.gtf) using bedtools intersect option [56].

Analysis of Nano3P-seq datasets

All the Nano3P-seq runs were basecalled and demultiplexed using stand-alone Guppy version 3.6.1 with default parameters. All runs were mapped using minimap2 [54]) with the following parameters: minimap2 -ax splice -k14 -uf --MD. For the synthetic constructs (curlcakes), base-called reads were mapped to Curlcake 1 and 2 sequences [51], and mapped reads were then intersected with the annotations of Curlcake 1 and 2 sequences to filter out the incomplete reads using bedtools. For yeast total RNA, we mapped the base-called reads to the *S. cerevisiae* genome (SacCer3), supplemented with ribosomal RNA; mapped reads were then intersected with SacCer64 annotation to filter out incomplete reads. For nuclear/mitochondrial enriched mouse brain RNA spiked in with sequins [16], we mapped the base-called reads to genome (GRCm38), supplemented with ribosomal RNA and sequin chromosome (chrIS). Mapped reads were then intersected with ENSEMBL version 102 annotation (Mus_musculus.GRCm38.102.gtf) and sequin annotation (RNAsequins.v2.2.gtf) in order to filter the incomplete reads. For zebrafish RNA, we first mapped the base-called reads to ribosomal RNAs and then to the genome (GRCz11). Mapped read starts were then intersected with ENSEMBL version 103 annotation (Danio_rerio.GRCz11.103.2.gtf) exon ends, in order to filter the incomplete reads. For assignment of reads to isoforms, IsoQuant package was used (https://github.com/ablab/IsoQuant) with Danio_rerio.GRCz11.103.2.gtf annotation. A complete step-by-step command line of the bioinformatic analysis done on Nano3P-seq datasets can be found in the GitHub repository https://github.com/novoalab/Nano3P_Seq.

Estimation of polyA tail lengths

For direct RNA sequencing reads, polyA tail length estimation was performed using *NanoTail*, a module from *Master of Pores* [57], a nextflow workflow for the analysis of direct RNA datasets, which uses internally Nanopolish v0.11.1 [22]. In *NanoTail*, all reads stored in the fastq files are first indexed with *nanopolish index* using default parameters, and the function *nanopolish polya* is used to perform polyA tail length estimations.

For Nano3P-seq reads, polyA tail length estimation was performed using *tailfindR* [21] with default parameters. We observed a consistent bias of 15nt in all tailfindR predictions benchmarked with known polyA tail lengths, possibly caused by the fact that tailfindR algorithms expect a double stranded cDNA, whereas Nano3P-seq polyA tail regions are single stranded cDNA regions. Therefore, all subsequent measurements were post-processed to adjust for this bias by subtracting 15nt to the predicted tail length. All code used to estimate polyA tail lengths and post-process Nano3P-seq data can be found at https://github.com/novoalab/Nano3P_Seq.

Analysis of tail composition

bioRxiv preprint doi: https://doi.org/10.1101/2021.09.22.461331; this version posted October 27, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Base-called reads mapped to the zebrafish genome were first trimmed using the Porechop tool (https://github.com/rrwick/Porechop) in order to remove the adapter sequences. Because Porechop sometimes failed at removal of the adapter sequences, only reads containing more than 80% A/U bases in their tail composition were kept for downstream analyses, thus ensuring that untrimmed reads are not included in downstream analyses (see **Figure S9**). Finally, an in-house python script was used to extract the tail FASTA sequences of the tail regions from trimmed reads, which has been made available in GitHub (https://github.com/novoalab/Nano3P_Seq/soft_clipped_content.py).

Animal Ethics

Fish lines were maintained according to the International Association for Assessment and Accreditation of Laboratory Animal Care research guidelines, and protocols were approved by the Yale University Institutional Animal Care and Use Committee (IACUC).

DATA AVAILABILITY

Base-called FAST5 reads from Nano3P-seq and dRNA-seq libraries have been made publicly available in ENA, under accession codes PRJEB46978. polyA tail lengths using PAL-Seq were obtained from GEO with the accession code GSE52809 [10]. All sequencing runs included in this work are listed in **Table S5**.

CODE AVAILABILITY

All scripts and code used in this work have been made available in GitHub: https://github.com/novoalab/Nano3P_Seq.

ACKNOWLEDGEMENTS

We thank all the members of the Novoa lab for their valuable insights and discussion. We thank Dr. Tim Mercer for providing us with the sequins used as spike-ins in Nano3P-seq zebrafish runs. OB is supported by a UNSW International PhD fellowship and Australian Government Research Training Program Scholarship. AD-T is supported by an FPI-SO fellowship (PRE2019-088498). This work was supported by the Australian Research Council (DP180103571 to EMN) and the Spanish Ministry of Economy, Industry and Competitiveness (MEIC) (PGC2018-098152-A-100 to EMN). We acknowledge the support of the MEIC to the EMBL partnership, Centro de Excelencia Severo Ochoa and CERCA Programme/Generalitat de Catalunya.

AUTHOR CONTRIBUTIONS

OB performed all the wet lab experiments. OB analyzed the data, together with EMN. HL contributed code for the analysis of polyA tail lengths and nucleotide content. AD-T processed and analyzed the direct RNA sequencing zebrafish data. CK, AJG and JDB provided the zebrafish total and polyA-selected RNA samples used in this study. EMN conceived the project. EMN supervised the work, with the assistance of JSM. OB and EMN wrote the paper, with contributions from all authors.

bioRxiv preprint doi: https://doi.org/10.1101/2021.09.22.461331; this version posted October 27, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

DECLARATIONS OF INTERESTS

EMN has received travel and accommodation expenses to speak at Oxford Nanopore Technologies conferences. All authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

REFERENCES

1. Licatalosi, D. D. & Darnell, R. B. RNA processing and its regulation: global insights into biological networks. *Nat. Rev. Genet.* **11**, 75–87 (2010).
2. Berkovits, B. D. & Mayr, C. Alternative 3' UTRs act as scaffolds to regulate membrane protein localization. *Nature* **522**, 363–367 (2015).
3. Dreyfus, M. & Régnier, P. The Poly(A) Tail of mRNAs. *Cell* vol. 111 611–613 (2002).
4. Colgan, D. F. & Manley, J. L. Mechanism and regulation of mRNA polyadenylation. *Genes Dev.* **11**, 2755–2766 (1997).
5. Tian, B., Hu, J., Zhang, H. & Lutz, C. S. A large-scale analysis of mRNA polyadenylation of human and mouse genes. *Nucleic Acids Res.* **33**, 201–212 (2005).
6. Ulitsky, I. *et al.* Extensive alternative polyadenylation during zebrafish development. *Genome Res.* **22**, 2054–2066 (2012).
7. Yartseva, V. & Giraldez, A. J. The Maternal-to-Zygotic Transition During Vertebrate Development: A Model for Reprogramming. *Curr. Top. Dev. Biol.* **113**, 191–232 (2015).
8. Lee, M. T., Bonneau, A. R. & Giraldez, A. J. Zygotic genome activation during the maternal-to-zygotic transition. *Annu. Rev. Cell Dev. Biol.* **30**, 581–613 (2014).
9. Walser, C. B. & Lipshitz, H. D. Transcript clearance during the maternal-to-zygotic transition. *Curr. Opin. Genet. Dev.* **21**, 431–443 (2011).
10. Subtelny, A. O., Eichhorn, S. W., Chen, G. R., Sive, H. & Bartel, D. P. Poly(A)-tail profiling reveals an embryonic switch in translational control. *Nature* **508**, 66–71 (2014).
11. Chang, H., Lim, J., Ha, M. & Kim, V. N. TAIL-seq: genome-wide determination of poly(A) tail length and 3' end modifications. *Mol. Cell* **53**, 1044–1052 (2014).
12. Workman, R. E. *et al.* Nanopore native RNA sequencing of a human poly(A) transcriptome. *Nat. Methods* **16**, 1297–1305 (2019).

bioRxiv preprint doi: https://doi.org/10.1101/2021.09.22.461331; this version posted October 27, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

13. Tudek, A. *et al.* Global view on the metabolism of RNA poly(A) tails in yeast Saccharomyces cerevisiae. *Nat. Commun.* **12**, 4951 (2021).
14. Begik, O. *et al.* Quantitative profiling of pseudouridylation dynamics in native RNAs with nanopore sequencing. *Nat. Biotechnol.* (2021) doi:10.1038/s41587-021-00915-6.
15. Levy, B. W., Johnson, C. B. & McCarthy, B. J. Diversity of sequences in total and polyadenylated nuclear RNA from Drosophila cells. *Nucleic Acids Research* vol. 3 1777–1790 (1976).
16. Hardwick, S. A. *et al.* Spliced synthetic genes as internal controls in RNA sequencing experiments. *Nat. Methods* **13**, 792–798 (2016).
17. Giraldez, A. J. *et al.* Zebrafish MiR-430 promotes deadenylation and clearance of maternal mRNAs. *Science* **312**, 75–79 (2006).
18. Chang, H. *et al.* Terminal Uridylyltransferases Execute Programmed Clearance of Maternal Transcriptome in Vertebrate Embryos. *Mol. Cell* **70**, 72–82.e7 (2018).
19. Bratic, A. *et al.* Mitochondrial Polyadenylation Is a One-Step Process Required for mRNA Integrity and tRNA Maturation. *PLoS Genet.* **12**, e1006028 (2016).
20. Kim, D. *et al.* The Architecture of SARS-CoV-2 Transcriptome. *Cell* **181**, 914–921.e10 (2020).
21. Krause, M. *et al.* tailfindr: alignment-free poly(A) length measurement for Oxford Nanopore RNA and DNA sequencing. *RNA* **25**, 1229–1241 (2019).
22. Loman, N. J., Quick, J. & Simpson, J. T. A complete bacterial genome assembled de novo using only nanopore sequencing data. *Nat. Methods* **12**, 733–735 (2015).
23. Carninci, P. *et al.* The transcriptional landscape of the mammalian genome. *Science* **309**, 1559–1563 (2005).
24. Guttman, M. *et al.* Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals. *Nature* **458**, 223–227 (2009).
25. Marzluff, W. F., Wagner, E. J. & Duronio, R. J. Metabolism and regulation of canonical histone mRNAs: life without a poly(A) tail. *Nature Reviews Genetics* vol. 9 843–854 (2008).

bioRxiv preprint doi: https://doi.org/10.1101/2021.09.22.461331; this version posted October 27, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

26. Battle, D. J. & Doudna, J. A. The stem-loop binding protein forms a highly stable and specific complex with the 3' stem-loop of histone mRNAs. *RNA* **7**, 123–132 (2001).

27. Vejnar, C. E., Messih, M. A., Takacs, C. M. & Yartseva, V. Genome wide analysis of 3′ UTR sequence elements and proteins regulating mRNA stability during maternal-to-zygotic transition in zebrafish. *Genome* (2019).

28. Vejnar, C. E. *et al.* Genome wide analysis of 3' UTR sequence elements and proteins regulating mRNA stability during maternal-to-zygotic transition in zebrafish. *Genome Res.* **29**, 1100–1114 (2019).

29. Roundtree, I. A., Evans, M. E., Pan, T. & He, C. Dynamic RNA Modifications in Gene Expression Regulation. *Cell* **169**, 1187–1200 (2017).

30. Begik, O. *et al.* Integrative analyses of the RNA modification machinery reveal tissue- and cancer-specific signatures. *Genome Biol.* **21**, 97 (2020).

31. Boo, S. H. & Kim, Y. K. The emerging role of RNA modifications in the regulation of mRNA stability. *Exp. Mol. Med.* **52**, 400–408 (2020).

32. Zhao, B. S. *et al.* m6A-dependent maternal mRNA clearance facilitates zebrafish maternal-to-zygotic transition. *Nature* **542**, 475–478 (2017).

33. Ryvkin, P. *et al.* HAMR: high-throughput annotation of modified ribonucleotides. *RNA* **19**, 1684–1692 (2013).

34. Khoddami, V. *et al.* Transcriptome-wide profiling of multiple RNA modifications simultaneously at single-base resolution. *Proc. Natl. Acad. Sci. U. S. A.* **116**, 6784–6789 (2019).

35. Motorin, Y., Muller, S., Behm□Ansmant, I. & Branlant, C. Identification of Modified Residues in RNAs by Reverse Transcription□Based Methods. in *Methods in Enzymology* vol. 425 21–53 (Academic Press, 2007).

36. Werner, S. *et al.* Machine learning of reverse transcription signatures of variegated polymerases allows mapping and discrimination of methylated purines in limited transcriptomes. *Nucleic Acids Res.* **48**, 3734–3746 (2020).

37. Boccaletto, P. & Bagiński, B. MODOMICS: An Operational Guide to the Use of the RNA

bioRxiv preprint doi: https://doi.org/10.1101/2021.09.22.461331; this version posted October 27, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Modification Pathways Database. *Methods Mol. Biol.* **2284**, 481–505 (2021).

38. Lim, J. *et al.* Uridylation by TUT4 and TUT7 marks mRNA for degradation. *Cell* **159**, 1365–1376 (2014).
39. Morgan, M. *et al.* mRNA 3′ uridylation and poly(A) tail length sculpt the mammalian maternal transcriptome. *Nature* vol. 548 347–351 (2017).
40. Legnini, I., Alles, J., Karaiskos, N., Ayoub, S. & Rajewsky, N. FLAM-seq: full-length mRNA sequencing reveals principles of poly(A) tail length control. *Nat. Methods* **16**, 879–886 (2019).
41. Shepard, P. J. *et al.* Complex and dynamic landscape of RNA polyadenylation revealed by PAS-Seq. *RNA* vol. 17 761–772 (2011).
42. Harrison, P. F. *et al.* PAT-seq: a method to study the integration of 3'-UTR dynamics with gene expression in the eukaryotic transcriptome. *RNA* **21**, 1502–1510 (2015).
43. Lim, J., Lee, M., Son, A., Chang, H. & Kim, V. N. mTAIL-seq reveals dynamic poly(A) tail regulation in oocyte-to-embryo development. *Genes Dev.* **30**, 1671–1682 (2016).
44. Liu, Y., Nie, H., Liu, H. & Lu, F. Poly(A) inclusive RNA isoform sequencing (PAIso-seq) reveals wide-spread non-adenosine residues within RNA poly(A) tails. *Nat. Commun.* **10**, 5292 (2019).
45. Oikonomopoulos, S. *et al.* Methodologies for Transcript Profiling Using Long-Read Technologies. *Front. Genet.* **11**, 606 (2020).
46. Ramsköld, D. *et al.* Full-length mRNA-Seq from single-cell levels of RNA and individual circulating tumor cells. *Nat. Biotechnol.* **30**, 777–782 (2012).
47. Au, K. F. *et al.* Characterization of the human ESC transcriptome by hybrid sequencing. *Proc. Natl. Acad. Sci. U. S. A.* **110**, E4821–30 (2013).
48. Ozsolak, F. *et al.* Direct RNA sequencing. *Nature* **461**, 814–818 (2009).
49. Xu, H., Yao, J., Wu, D. C. & Lambowitz, A. M. Improved TGIRT-seq methods for comprehensive transcriptome profiling with decreased adapter dimer formation and bias correction. *Sci. Rep.* **9**, 7953 (2019).
50. Mohr, S. *et al.* Thermostable group II intron reverse transcriptase fusion proteins and

bioRxiv preprint doi: https://doi.org/10.1101/2021.09.22.461331; this version posted October 27, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

their use in cDNA synthesis and next-generation RNA sequencing. *RNA* **19**, 958–970 (2013).

51. Liu, H. *et al.* Accurate detection of m6A RNA modifications in native RNA sequences. *Nat. Commun.* **10**, 4079 (2019).
52. Martelotto, L. 'Frankenstein' protocol for nuclei isolation from fresh and frozen tissue for snRNAseq v2 (protocols.io.3fkgjkw). *protocols.io* (2019) doi:10.17504/protocols.io.3fkgjkw.
53. Smith, M. A. *et al.* Barcoding and demultiplexing Oxford Nanopore native RNA sequencing reads with deep residual learning. *bioRxiv* 864322 (2019) doi:10.1101/864322.
54. Li, H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094–3100 (2018).
55. Locati, M. D. *et al.* Expression of distinct maternal and somatic 5.8S, 18S, and 28S rRNA types during zebrafish development. *RNA* **23**, 1188–1199 (2017).
56. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842 (2010).
57. Cozzuto, L. *et al.* MasterOfPores: A Workflow for the Analysis of Oxford Nanopore Direct RNA Sequencing Datasets. *Front. Genet.* **11**, 211 (2020).