Characterising the RNA modification and polyadenylation landscape at single molecule resolution using third-generation sequencing technologies

Author:
Begik, Oguzhan

Publication Date:
2021

DOI:
https://doi.org/10.26190/unsworks/1973

Characterising the RNA modification and polyadenylation landscape at single molecule resolution using third-generation sequencing technologies

Oguzhan Begik

A thesis in fulfilment of the requirements for the degree of Doctor of Philosophy

St Vincent's Clinical School, Faculty of Medicine UNSW Sydney

Garvan Institute of Medical Research

Centre for Genomic Regulation

Supervisors : John S Mattick and Eva Maria Novoa

December 2021

UNSW SYDNEY | Australia's Global University

Thesis/Dissertation Sheet

Surname/Family Name	:	**Begik**
Given Name/s	:	**Oguzhan**
Abbreviation for degree as give in the University calendar	:	**PhD**
Faculty	:	**Medicine**
School	:	**St Vincent's Clinical School**
Thesis Title	:	**Characterising the RNA modification and polyadenylation landscape at single molecule resolution using third-generation sequencing technologies**

Abstract 350 words maximum: (PLEASE TYPE)

RNA modifications, collectively referred to as the 'epitranscriptome', are not mere decorations of RNA molecules, but can be dynamically regulated upon environmental queues and changes in cellular conditions. This dynamic behaviour is achieved through the RNA modification machinery, which comprises "writer", "reader" and "eraser" proteins that modify, recognize and remove the modification, respectively.

Chapter 2 presents a comprehensive analysis of the RNA modification machinery (readers, writers and erasers) across species, tissues and cancer types, revealing gene duplications during eukaryotic evolution, changes in substrate specificity and tissue- and cancer-specific expression patterns.

Chapters 3 and 4 present the exploration and development of novel methods to map and analyze RNA modifications transcriptome-wide. Nanopore direct-RNA sequencing technology was used to provide RNA modification maps in full-length native RNA molecules. Firstly, it is shown that RNA modifications can be detected in the form of base-calling 'errors', thus allowing us to train Support Vector Machine models that can distinguish m6A-modified from unmodified sites, both *in vitro* and *in vivo*. Secondly, it is demonstrated that distinct RNA modification types have unique base-calling 'error' signatures, allowing us to exploit these signatures to distinguish different RNA modification types. It is found that pseudouridine has one of the most distinct signatures, appearing in the form of C-to-U mismatches. Finally, this information was used to predict novel pseudouridine sites on ncRNAs and mRNAs transcriptome-wide, as well as to obtain quantitative measurements of the stoichiometry of modified sites.

Chapter 5 presents the development of a novel nanopore-based method, which is termed 'Nano3P-seq', to simultaneously quantify RNA abundance and tail length dynamics in individual molecules in both the coding and non-coding transcriptome, from cDNA reads. It is demonstrated that Nano3P-seq offers a simple approach to study the coding and non-coding transcriptome at single molecule resolution regardless of the tail ends.

Together, this work provides a comprehensive framework for the study of RNA modifications and polyA tail dynamics using third generation sequencing technologies, opening novel avenues for future works that aim to characterize their dynamics and biological roles both in health and in disease.

Thesis submission for the degree of Doctor of Philosophy

Thesis Title and Abstract | Declarations | **Inclusion of Publications Statement** | Corrected Thesis and Responses

UNSW is supportive of candidates publishing their research results during their candidature as detailed in the UNSW Thesis Examination Procedure.

Publications can be used in the candidate's thesis in lieu of a Chapter provided:

- The candidate contributed **greater than 50%** of the content in the publication and are the "primary author", i.e. they were responsible primarily for the planning, execution and preparation of the work for publication.
- The candidate has obtained approval to include the publication in their thesis in lieu of a Chapter from their Supervisor and Postgraduate Coordinator.
- The publication is not subject to any obligations or contractual agreements with a third party that would constrain its inclusion in the thesis.

☑ The candidate has declared that **some of the work described in their thesis has been published and has been documented in the relevant Chapters with acknowledgement**.

A short statement on where this work appears in the thesis and how this work is acknowledged within chapter/s:

> The results from “Integrative analyses of the RNA modification machinery reveal tissue- and cancer-specific signatures.” paper in Genome Biology is contained in Chapter 2. The results from “Accurate detection of m6A RNA modifications in native RNA sequences” paper in Nature Communications is contained in Chapter 3. The results from “Quantitative profiling of pseudouridylation dynamics in native RNAs with nanopore sequencing.” paper in Nature Biotechnology is contained in Chapter 4. The results from “Nano3P-seq: transcriptome-wide analysis of gene expression and tail dynamics using end-capture nanopore sequencing” paper submitted for publication (BioRxiv) is contained in Chapter 5. Acknowledgement of the work of the other authors of this paper has been made at the beginning of the chapter.

Candidate's Declaration

I declare that I have complied with the Thesis Examination Procedure.

ORIGINALITY STATEMENT

'I hereby declare that this submission is my own work and to the best of my knowledge it contains no materials previously published or written by another person, or substantial proportions of material which have been accepted for the award of any other degree or diploma at UNSW or any other educational institution, except where due acknowledgement is made in the thesis. Any contribution made to the research by others, with whom I have worked at UNSW or elsewhere, is explicitly acknowledged in the thesis. I also declare that the intellectual content of this thesis is the product of my own work, except to the extent that assistance from others in the project's design and conception or in style, presentation and linguistic expression is acknowledged.'

Signed ....

01 October 2021

Date .....

COPYRIGHT STATEMENT

Signed ....

Date 01 October 2021

.....

AUTHENTICITY STATEMENT

'I certify that the Library deposit digital copy is a direct equivalent of the final officially approved version of my thesis.'

Signed ....

Date 01 October 2021

.....

Publications, presentations and awards during the course of this thesis

Publications

1. Liu H *, **Beğik O***, Lucas MC, Ramirez JM, Mason C, Wiener D, Schwartz S, Mattick JS, Smith M, Novoa EM. (2019) Accurate detection of m6A RNA modifications in native RNA sequences. *Nature Communications,* article 4079. https://doi.org/10.1038/s41467-019-11713-9 (**Results presented in Chapter 3**)
2. Smith MA*, Ersavas T*, Ferguson JM*, Liu H, Lucas MC, **Beğik O,** Bojarski L, Barton K, Novoa EM. (2020) Barcoding and demultiplexing Oxford Nanopore native RNA sequencing reads with deep residual learning. *Genome Research* 30:1345–1353. https://genome.cshlp.org/content/30/9/1345.full.pdf **(Appendix)**
3. **Beğik O**, Lucas MC, Liu H, Ramirez JM, Mattick JS, Novoa EM (2020) Integrative analyses of the RNA modification machinery reveal tissue- and cancer-specific signatures. *Genome Biology* 21, article 97. https://doi.org/10.1186/s13059-020-02009-z (**Results presented in Chapter 2**)
4. Liu H*, **Beğik O***, Novoa EM (2021) EpiNano: Detection of m^6A RNA Modifications Using Oxford Nanopore Direct RNA Sequencing. *Methods in Molecular Biology*, 2298:31-52. https://link.springer.com/protocol/10.1007%2F978-1-0716-1374-0_3 **(Appendix)**
5. **Beğik O***, Lucas MC*, Pryszcz LP, Ramirez J, Medina R, Milenkovic I, Cruciani S, Liu H, Vieira HGS, Sas-Chen A, Mattick JS, Schwartz S, Novoa EM (2021) Quantitative profiling of pseudouridylation dynamics in native RNAs with nanopore sequencing. *Nature Biotechnology,* epub in advance of print. https://doi.org/10.1038/s41587-021-00915-6 (**Results presented in Chapter 4**)
6. **Beğik O**, Liu H, Delgado-Tejedor A, Kontur C, Giraldez AJ, Beaudoin JD, Mattick JS, Novoa EM (2021) Nano3P-seq: transcriptome-wide analysis of gene expression and tail dynamics using end-capture nanopore sequencing. *Submitted for publication.* https://doi.org/10.1101/2021.09.22.46133 (**Results presented in Chapter 5**)

Key: * Authors contributed equally

Oral Presentations

1. **RNA Society Meeting 2021** (RNA 2021), *Quantitative profiling of pseudouridylation dynamics in native RNAs with nanopore sequencing*, Virtual Conference, June 2021
2. **Cambridge RNA Club**, Quantitative profiling of pseudouridylation dynamics in native RNAs with nanopore sequencing, Virtual Talk, April 2021
3. **RNA Institute Mini-Symposium**, Quantitative profiling of pseudouridylation dynamics in native RNAs with nanopore sequencing, Virtual Symposium, March 2021
4. **XIII. PhD Symposium of Centre for Genomic Regulation**, Identifying the molecular fingerprints of rRNA modifications using paired direct RNA and cDNA sequencing, Barcelona, November 2019

Conference Presentations

1. **Beğik O**, Liu H, Lucas MC, Ramirez JM, Mason C, Wiener D, Schwartz S, Mattick JS, Smith M, Novoa EM. "Accurate detection of m6A RNA modifications in native RNA sequences" RNA Editing Gordon Research Conference, Lucca, Italy, March 2019
2. **Begik O**, Vieira H, Mattick J.S, Novoa EM. "Deciphering m3C Modifications on mRNA", XII. PhD Symposium of Centre for Genomic Regulation, Barcelona, Spain, November 2018
3. **Begik O,** Mattick J.S, Novoa EM. **"**Deciphering m^3C Modifications on mRNA", EMBL Post-Graduate Symposium, Garvan Institute of Medical Research, Sydney, Australia, November 2017

Awards

1. Poster Prize Winner at GRC RNA Editing Conference March 2019
2. Boehringer Ingelheim (Travel Grant) for EMBL Course, January 2019
3. University International Postgraduate Award (UIPA) at UNSW, Semester 2 2017

Acknowledgements

I would like to begin my acknowledgement by quoting Isaac Newton: “If I have seen further, it is by standing on the shoulders of giants”. I would like to thank all women and men of science who pursued their curiosity and dedicated their lives to science. Without them, the world we know would not exist.

For around 10 years of my scientific career, I have met many amazing scientists. I would like to thank all of these people who guided me in my journey and helped me become the scientist I am now.

I would also like to express how honoured and grateful I am to be co-supervised by one of my scientific idols, Professor John Mattick. I still remember how excited I felt when I found out about my acceptance to the lab. It has been exceptional to be a part of the team.

I would like to express my gratitude to my co-supervisor Dr Eva Novoa for seeing the potential in me and becoming my mentor and friend. Thanks to her, I got to do my PhD in two different continents! It has been a really great journey and she is the reason why.

I am also most grateful to Dr Guillaume Filion, who was my supervisor when I was an undergraduate student. His mentorship ever since then has been extraordinary and every time I interacted with him, I learned something new.

I would like to thank all my labmates for the wonderful working environment and countless scientific discussions. I have learnt so much from everyone.

A very special thanks from the bottom of my heart goes to very special people in my life; Harun Cingöz, Moritz Bauer and Júlia Urgel i Solas for always being there for me.

En büyük teşekkür ise aileme… Bu hayattaki en değerli varlığım sizsiniz. Beni her zaman destekleyen, her zaman bana inanan bir ailem olduğu için dünyanın en şanslı insanıyım. Uzakta olsanız dahi hep aklımdasınız. Sizi çok seviyorum.

“Stands at the sea, wonders at wondering: I, a universe of atoms, an atom in the universe.”

Richard P. Feynman

List of Abbreviations

2′O MTase	M7G-specific 2′O methyltransferase
A-to-I	Adenosine-to-inosine
Ac4c	N_4-acetylcytosine
ACC	Adrenocortical carcinoma
ACF	Apobec-1 complementation factor
ADAD	Adenosine deaminase domain containing
ADAR	Adenosine deaminases acting on RNA
ADAT	Adenosine deaminases that act on tRNAs
AID	Activation induced cytidine deaminase
ALKBH	AlkB homolog
apoB	Apolipoprotein B
APOBEC	Apolipoprotein B mRNA editing enzyme, catalytic polypeptide
BCDIN3D	BCDIN3 domain containing RNA methyltransferase
BLCA	Bladder urothelial carcinoma
BRCA	Breast invasive carcinoma
BS	Branch site
CESC	Cervical squamous cell carcinoma and endocervical adenocarcinoma
CFI	Cleavage factor I
cDNA	Complementary DNA
Cm	2’-O-methylcytidine
COAD	Colon adenocarcinoma

CPSF	Cleavage and polyadenylation specificity factor
CstF	Cleavage and stimulation factor
CTD	C-terminal domain
DDX4	DEAD-Box helicase 4
DKC1	Dyskerin pseudouridine synthase 1
DNMT2	DNA methyltransferase 2
dRNAseq	Direct RNA sequencing
DS	Dysregulation score
dsRNA	Double-stranded RNA
eIF4G	Eukaryotic translation initiation factor 4 G
ESCA	Esophageal carcinoma
FBL	Fibrillarin
FBLL1	Fibrillarin-like 1
FTO	Fat mass and obesity-associated protein
GBM	Glioblastoma multiforme
GMP	Guanosine monophosphate
GTEx	Genotype tissue expression
HENMT1	HEN methyltransferase 1
hm^5C	5-Hydroxymethylcytosine
HMM	Hidden markov model
HNSC	Head and neck squamous cell carcinoma
HPA	Human protein atlas
i(6)A	N_6-isopentenyladenosine
IGF2BP	Insulin-like growth factor 2 mRBP

IHC	Immunohistochemistry
KICH	Kidney chromophobe
KIRC	Kidney renal clear cell carcinoma
KIRP	Kidney renal papillary cell carcinoma
LAGE3	L antigen family member 3
LAML	Acute myeloid leukemia
LCMS/MS	Liquid chromatography-tandem mass spectrometry
LGG	Brain lower grade glioma
LIHC	Liver hepatocellular carcinoma
lncRNA	Long non-coding RNA
LUAD	Lung adenocarcinoma
LUSC	Lung squamous cell carcinoma
m^1A	N_1-methyladenosine
m^1acp^3Y	1-methyl-3-(3-amino-3-carboxypropyl)pseudouridine
m^1G	N_1-methylguanosine
m^3C	N_3-methylcytosine
m^3U	N_3-methyluridine
m^5C	5-methylcytosine
m^6A	N_6-methyladenosine
m^7G	N_7-methylguanosine
MepCE	Methylphosphate capping enzyme
MeRIP-seq	Methylated RNA immunoprecipitation sequencing
METTL	Methyltransferase-like
miRNA	Micro RNA

mRNA	Messenger RNA
MZT	Maternal-to-Zygotic transition
Nano3P-seq	Nanopore 3 Prime end-capture sequencing
NAT10	N-Acetyltransferase 10
ncRNA	Non-coding RNA
NGS	Next-generation sequencing
Nm	2'-O-methylation
NSUN	NOP2/Sun RNA methyltransferase
ONT	Oxford Nanopore Technologies
OV	Ovarian serous cystadenocarcinoma
PAAD	Pancreatic adenocarcinoma
PABP	Poly(A) binding protein
PacBio	Pacific Biosciences
PAP	Poly(A) polymerase
PCA	Principal component analysis
PCPG	Pheochromocytoma and paraganglioma
PCR	Polymerase chain reaction
PPT	Polypyrimidine tract
PRAD	Prostate adenocarcinoma
pre-mRNA	Precursor mRNA
pri-miRNA	Primary miRNA
PUS	Pseudouridine synthases
qRT-PCR	Quantitative real time PCR
RBPs	RNA binding proteins

READ	Rectum adenocarcinoma
RMPs	RNA modification-related proteins
RMWs	RNA modification writers
RNMT	RNA guanine-7 methyltransferase
rRNA	Ribosomal RNA
RT	Reverse transcription
SAM	S-adenosylmethionine
SARC	Sarcoma
SBS	Sequencing-by-synthesis
SD	Standard deviation
SKCM	Skin cutaneous melanoma
snoRNA	Small nucleolar RNA
snRNA	Small nuclear RNA
STAD	Stomach adenocarcinoma
t6A	N_6-threonylcarbamoyladenosine
TE	Transposable element
TGCT	Testicular germ cells tumor
THCA	Thyroid carcinoma
TMA	Tumor microarray
TPM	Transcripts per kilobase million
TRDMT1	TRNA aspartic acid methyltransferase 1
TRIT1	TRNA isopentenyltransferase 1
TRMT	TRNA methyltransferase
tRNA	Transfer RNA

TS	Tissue specificity
UCEC	Uterine corpus endometrial carcinoma
UCS	Uterine carcinosarcoma
UTR	Untranslated region
YTHD	YTH domain
ZCCHC4	CCHC domain-contaning protein 4
Ψ	Pseudouridine

List of Figures

List of Tables

Table of Contents

1. Introduction

1.1. Processing and regulation of RNA

RNA is the centerpiece of life at the cellular level. It is much more than an intermediate molecule between DNA and proteins. Starting from its production process called transcription, RNA is subject to numerous regulatory steps to ensure the correct repertoire, quantity, and quality of gene expression. Moreover, there is considerable diversification of the components involved in transcriptional and post-transcriptional regulation as organismal complexity increases. Transcription is carried out by a single RNA polymerase enzyme in prokaryotes, but by three RNA polymerase complexes in plants and animals, which transcribe different classes of RNAs and recognise different promoter types.

1.1.1. Transcriptional machinery

In eukaryotes, RNA polymerase I (Pol I) transcribes ribosomal RNA (rRNA) genes, RNA polymerase II (Pol II) transcribes messenger RNA (mRNA), long non-coding RNA (lncRNA), micro RNA (miRNA), small nuclear RNA (snRNA) and small nucleolar RNA (snoRNA) genes, and RNA polymerase III (Pol III) transcribes transfer RNA (tRNA) and 5S rRNA genes (**Table 1.1**) [1,2]. Eukaryotic RNA polymerases also require proteins called “transcription factors” that regulate transcription [3–5].

Type	**Transcription Product**	**Location**
RNA Polymerase I	rRNA (45S precursor > 28S, 18S, 5.8S)	Nucleolus
RNA Polymerase II	mRNA, lncRNA, miRNA, snRNA, snoRNA	Nucleoplasm
RNA Polymerase III	tRNA, 5S rRNA	Nucleoplasm

Table 1.1 - Types of RNA polymerases, their products, and localisation

RNA polymerase I mediated ribosomal RNA transcription takes place in the nucleolus, where there are many copies of the rDNA gene. The transcription produces 45S pre-rRNA, which is post-transcriptionally cleaved into 3 subsidiary species (28S,

18S, and 5.8S) and modified by Fib/Dkc with the help of C/D box and H/ACA box snoRNAs [6].

Unlike other rRNAs, 5S rRNA is transcribed by Pol III, alongside tRNAs and other small RNAs [7]. The genes transcribed by Pol III are mostly defined as "housekeeping genes" and their transcription is tightly connected to the cell cycle and growth regulation. It, therefore, interacts with fewer regulatory proteins than Pol II [8]

Pol II produces mRNAs and a range of regulatory and infrastructural RNAs and is highly regulated in response to developmental and environmental signals [2]. Co-transcriptional processing of the RNA polymerase II transcripts ensures the maturation of these RNAs. The C-terminal domain (CTD) of RNA polymerase II is phosphorylated during elongation, leading to the recruitment of the proteins that catalyse 5' capping, splicing, and 3' processing, [9]. As polymerase II transits the polyadenylation site, usually indicated by a conserved AATAAA motif, enzymes responsible for 3' processing and polyadenylation catalyses the cleavage of the precursor mRNA (pre-mRNA) and polyadenylation from the 3' end **(Figure 1.1)** [10].

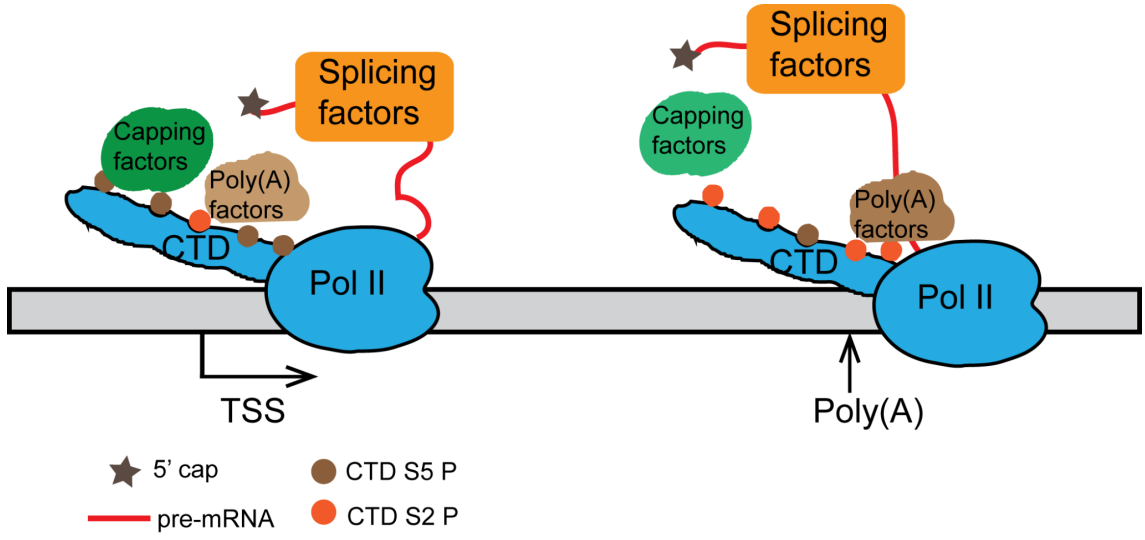


Figure 1.1 - The mRNA factory model: Coupling of pre-mRNA processing factors with the transcription machinery.

Phosphorylation on different positions of the CTD leads to the recruitment of capping and poly(A) factors. Phosphorylated Ser5 (S5 P) residues are enriched on the 5' end of the genes, whereas Phosphorylated Ser2 (S2 P) residues are enriched on the 3' end of the genes. Figure adapted from Saldi et al 2016 [11].

1.1.2. Post-transcriptional regulation of RNA

Post-transcriptional regulation begins with the Gppp-cap formation and subsequent methylation of the cap leading to 7-methylguanosine at the 5' end of the RNA. It is followed by the intron splicing and 3'end cleavage and polyadenylation, which results in a mature and functioning RNA [12]. Furthermore, RNA modifications occur during and after transcription, affecting processes such as splicing, localization, decay, and translation [13–19].

5' end capping occurs once the first 25-30 nucleotides of the RNA have been transcribed [20]. There are three enzymes involved in this process to produce cap 0 (Figure 1.2), whereas another enzyme produces cap1 [21,22]. Firstly, RNA triphosphatase enzyme removes one phosphate from the triphosphorylated RNA molecule on the 5' end, leaving a diphosphate group. Secondly, RNA guanylyltransferase adds guanosine monophosphate (GMP) to the 5'end with diphosphate, which makes up the G cap. Finally, guanine-N7 methyltransferase adds a methyl group to the G cap at the seven position using S-adenosyl-methionine (SAM) as a source of methyl, which forms the cap 0 structure [21]. Furthermore, the ribose methylation of the nucleotide at the +1 position by m7G-specific 2′O methyltransferase (2′O MTase) leads to the formation of cap1 in higher eukaryotes **(Figure 1.2)** [22]. Cap structures play important roles in many cellular processes such as export of RNA from nucleus to cytoplasm, circularization of RNA via interaction of eukaryotic translation initiation factor 4 G (eIF4G) and poly(A) binding protein (PABP1), translation initiation, and creation of self RNA signature [22].

During transcription, mRNA is synthesised as a large precursor (pre-mRNA), containing coding parts (exons) and non-coding parts (introns). The introns are removed from the pre-mRNA and exons are brought together, to form mature mRNAs. This process also enables alternative splicing, which takes place in the majority of the transcripts, leading to an enormous diversity of coding and regulatory RNA isoforms [23,24]. Splicing is carried by the spliceosome, which is composed of small nuclear guide RNAs (snRNAs) and almost 100 proteins [25]. The spliceosome machinery recognises specific sites located at or near intron-exon junctions: the 5' splice site (5' SS), 3' splice site (3' SS), and branch site (BS). The branch site is usually located upstream of a pyrimidine-rich region called polypyrimidine tract (PPT) [26]. Intron sequences usually start with a GU sequence at their 5' end, and end with an AG sequence at its 3' end [27]. Splicing is initiated with the nucleophilic attack from the 2'-OH group of the adenosine located at the branch site to the phosphodiester bond of the 5'SS. This exposes the 3'-OH end of the 5'exon and forms a lariat structure from the

interaction of branch site and 5'SS. Next, the 3'-OH end of the 5' exon attacks the 3'SS, which leads to the ligation of two exons and complete removal of the intron (**Figure 1.3**) [26].

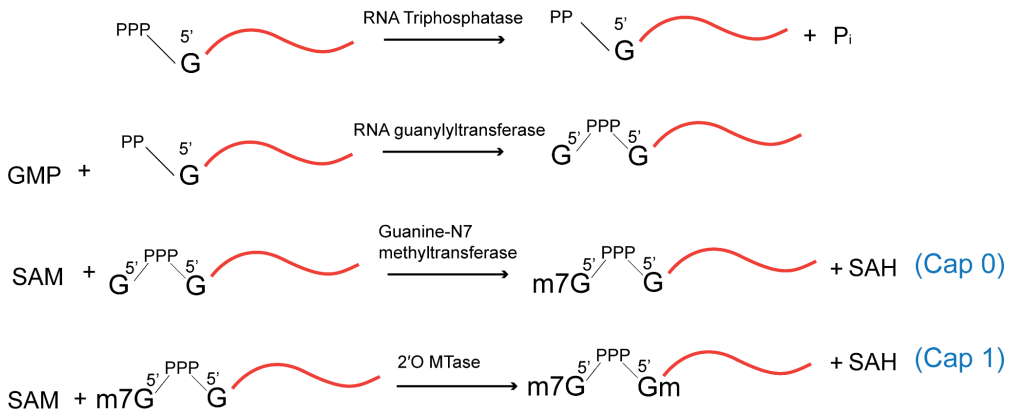


Figure 1.2 - Formation of the 5' cap.

Many enzymes are involved in the catalytic process of producing the 5' cap, leading to the formation of cap 0 and in higher eukaryotes, cap 1. Figure adapted from Ramanathan et al, 2016 [22].

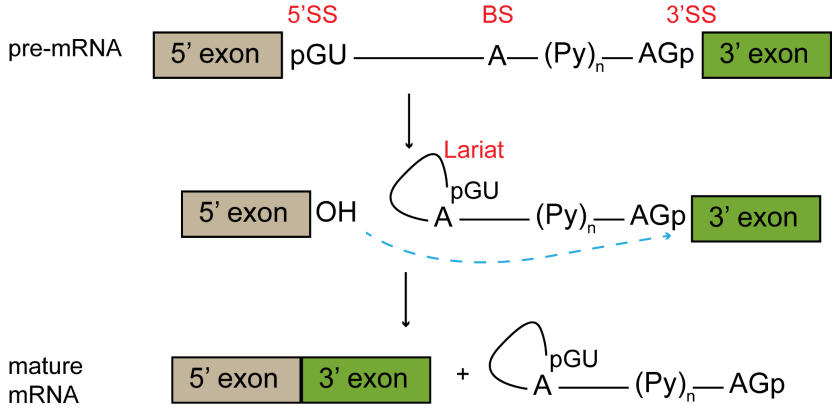


Figure 1.3 - RNA splicing overview.

Precursor mRNA (pre-mRNA) is processed via splicing that leads to the formation of mature mRNA. Cis-regulatory sequences such as 5' splice site (5' SS), 3' splice site (3' SS), and branch site (BS) are located at or near intron-exon junctions. Splicing consists of two main steps; nucleophilic attack from A base of the BS to the 5'SS and formation of lariat structure, and attack of 5'exon to the 3'SS and formation of mature mRNA. Figure adapted from [28]

All mRNAs except histone mRNAs undergo 3' end processing that leads to the endonucleolytic cleavage of the nascent transcript and the addition of a poly(A) tail [29]. The polyadenylation signal (AAUAAA), which determines the cleavage site, is located approximately 10-30 nucleotides upstream of the cleavage site. The polyadenylation signal is recognised by cleavage and polyadenylation specificity factor (CPSF), which is composed of at least five subunits (CPSF 30, CPSF 73, CPSF 100, CPSF 160, and hFip1) and catalyses the 3' cleavage and recruitment of the poly(A) polymerase (PAP) to the cleaved site [30,31]. In addition to the polyadenylation signal, there are two other cis-elements that are important for the 3' end processing and polyadenylation, namely the upstream UGUA-containing sequences (USE) and downstream GU- and G- rich sequences (DSE) [32–35]. USE is recognised by the cleavage factor I (CFI), whereas the DSE is recognised by the cleavage and stimulation factor (CstF) complexes [36]. Subsequent to the recognition of the cis-regulatory elements by the protein complexes, CPSF 73 catalyses the endonucleolytic cleavage [37] and the exposed 3' end of the RNA template is polyadenylated by a nuclear poly(A) polymerase.

The length of the poly(A) tail is dependent on the interaction between CPSF complex, PAP, and nuclear poly(A)-binding proteins PABPN1 and PABP2 [38]. Poly(A) tails play an important role in the stability and translation of the mRNA by interacting with the poly(A)-binding proteins in the nucleus and cytoplasm [39,40]. Cytoplasmic poly(A)-binding proteins (PABPC) interact with poly(A) tail to protect mRNA from degradation and promote translation [41]. Subsequent to the deadenylation of mRNA, PABPC is dissociated, exposing mRNA for degradation and translational inactivation. For example, during maternal-to-zygotic transition (MZT) in zebrafish and frog embryos, a group of mRNAs are marked for decay by deadenylation [42,43]. Once deadenylated, however, mRNA can still be polyadenylated in cytoplasm by noncanonical PAPs in order to be reactivated for translation in response to a cellular signal [44,45].

1.1.3. Chemical RNA Modifications

Chemical alterations after their synthesis are common to all biological molecules including DNA, RNA and protein [46]. RNA modifications, collectively referred to as the "epitranscriptome" [47], have been known to exist for the last 60 years [47,48]. There are over 170 types of RNA modifications described to date, which alter the chemistry of all four bases, as well as the ribose moiety [47] of RNAs. Eukaryotic rRNA molecules contain about 200 modifications per molecule on average,

whereas eukaryotic tRNAs contain about 13 modifications per molecule. The function of RNA molecules heavily depends on the modifications decorating them. Moreover, modifications can alter the fate of the RNA molecules by affecting molecular processes such as their splicing pattern, stability, translation efficiency, and subcellular localization [13–19]. Furthermore, at least some RNA modifications are known to be reversible, which suggests that these modifications are dynamically regulated (**see section 1.1.3.1**). Consequently, RNA modifications have a great impact on biological processes such as development [49], inheritance [50], and cell fate [51]. Disruption of modifications has been associated with as many as 100 human diseases [52–56]. However, a major drawback to understanding the role and dynamics of modifications in specific RNAs is the paucity of sequencing-based detection techniques, which are only available for a handful of modifications (**Figure 1.4**).

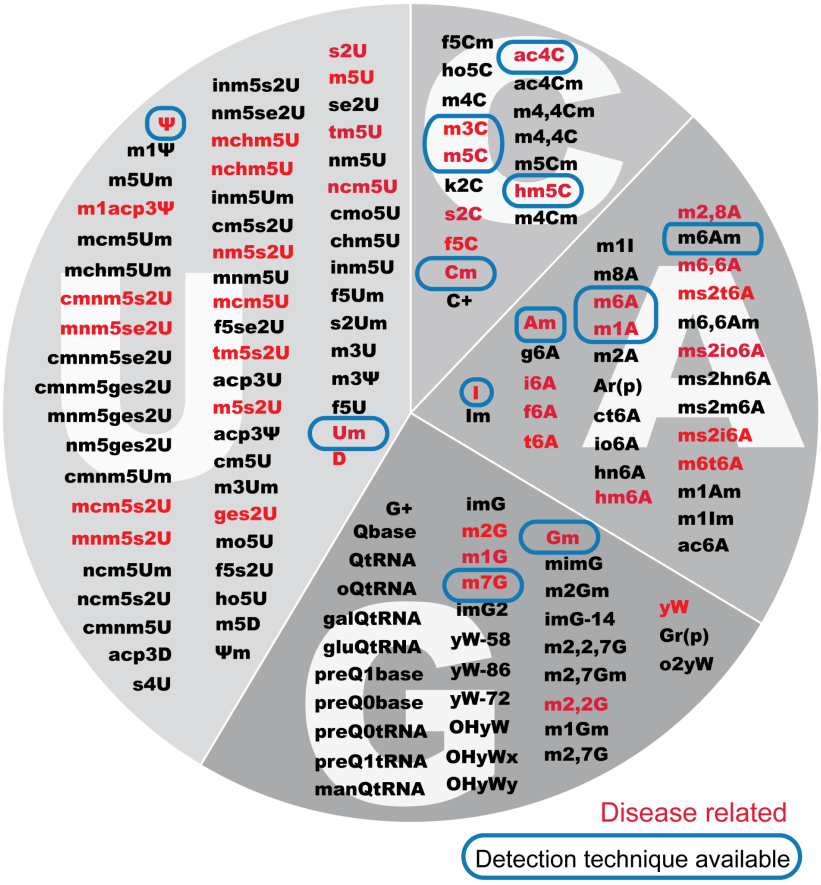


Figure 1.4 - Diversity of RNA modifications.

Pie chart shows the RNA modifications discovered so far. RNA modifications associated with any disease are highlighted in red. RNA modifications with an established detection technique are circled in blue. Adapted from Jonkhout et al 2017 [52].

1.1.3.1. RNA Modification Machinery

Recently, it has been shown that RNA modifications, which had been thought to be static alterations, are actually dynamic. This conclusion was drawn from the observation that N_6-methyladenosine (m^6A) can be removed by a protein called fat mass and obesity-associated protein (FTO) [57]. Subsequently, it was shown that it is actually the alkB homolog 5 (ALKBH5), not FTO, that can remove the m^6A mark *in-vivo* (**Figure 1.5**) [58]. Enzymes that are responsible for adding the chemical modification are termed as "writers", whereas the ones responsible for removing the modification are termed as "erasers".

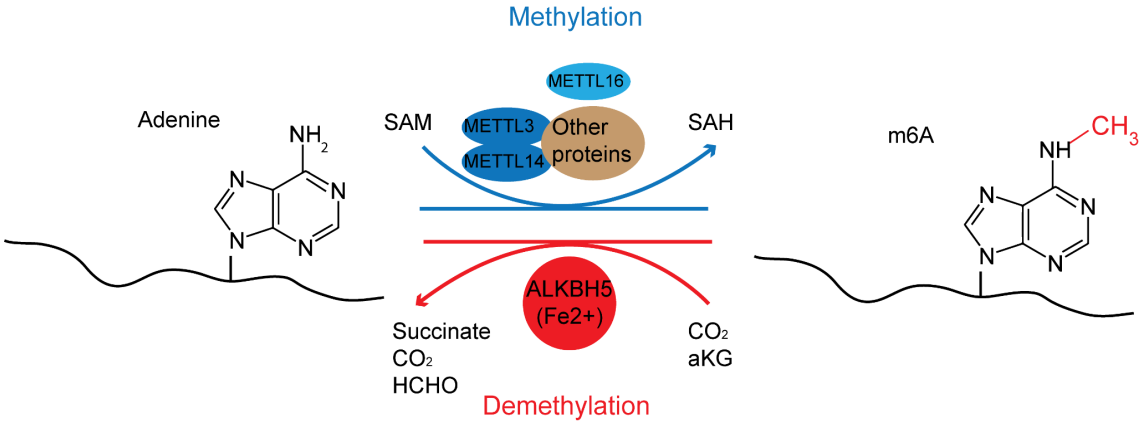


Figure 1.5 - Mechanism of reversible m^6A methylation.

Writers for m^6A modification, namely METTL3/METTL14 and METTL16 catalyses the modification, whereas ALKBH5 enzyme removes the modification. Figure adapted from Qin et al, 2020 [59].

Methylation is one of the most common types of RNA modifications and RNA methyltransferase enzymes catalyse the methylation of RNA by using SAM as a source of methyl group [60]. METTL3 is the first methyltransferase-like (METTL) gene family member discovered in 1994. METTL3, interacting with other proteins, catalyses the transfer of a methyl group to the N6 position of the adenine base (m^6A) in mRNAs and ncRNAs [61]. A number of RNA methyltransferase writers are now known, including METTL, RNA Guanine-7 Methyltransferase (RNMT), NOP2/Sun RNA Methyltransferase (NSUN), Methylphosphate Capping Enzyme (MEPCE), tRNA Methyltransferase (TRMT), and BCDIN3 Domain Containing RNA Methyltransferase (BCDIN3D). There are also non-methyltransferase writes such as N-acetyltransferase 10 (NAT10), tRNA isopentenyltransferase 1 (TRIT1), and Dyskerin Pseudouridine Synthase 1 (DKC1), which catalyse addition of N_4-acetylcytosine (Ac4c), N_6-isopentenyladenosine(i(6)A), and pseudouridine (Ψ), respectively [60].

In addition to FTO and ALKBH5, several other erasers have been identified. Studies have shown that ALKBH1 can demethylate RNAs possessing N_3-methylcytosine (m^3C) and N_1-methyladenosine (m^1A) [62,63]. Despite the fact that both ALKBH2 and ALKBH3 can erase m^1A and m^3C, ALKBH2 is more active on both single and double-stranded DNAs, whereas ALKBH3 preferentially demethylates single-stranded RNAs and DNAs [64,65].

Proteins that can recognise RNA modifications and interact with them to create action are termed as RNA “readers". The best-studied family of RNA readers are the YTH domain proteins (YTHD), which are readers for m^6A modifications. YTHDC1, for example, interacts with m^6A-modified molecules and regulates splicing and translation efficiency [15,66,67]. Moreover, YTHDC2 mediates the interaction of m^6A-modified RNA and CCR4-NOT complex, which then leads to faster decay of the mRNA [68]. Another family of m^6A readers is the insulin-like growth factor 2 mRBP (IGF2BP) family, which are associated with increased half-life and translation of m^6A methylated mRNAs [69]. In addition to m^6A reader proteins, there is also 5-methylcytosine (m^5C) “reader” called ALYRED that interacts with m^5C modified mRNAs and aids in their export from nucleus to cytosol [70].

Due to their ability to manipulate and interact with RNA modifications, these RNA modification-related proteins (RMPs) have been studied for their involvement in various biological processes [66,71–74]. Despite our knowledge about a group of RNA modification-related proteins that were studied in specific phenotypes with their loss of function, the literature is lacking a systematic analysis of these proteins [52].

1.1.3.2. Main Types of mRNA Modifications

m^6A is the most common type of modification on mRNAs [75]. In addition to mRNAs, they are also found in lncRNAs [76], primary miRNAs (pri-miRNA) [77] and rRNAs [78,79]. Enzymes responsible for the addition of m^6A to mRNA, lncRNA and miRNA are identified as METTL3-METTL14 complexes [80], as well as stand-alone METTL16 enzymes [81]. The catalysis of m^6A addition to rRNAs are performed by tRNA methyltransferase 112 (TRMT112)-METTL5 complex for 18S rRNA [82] and CCHC domain-containing protein 4 (ZCCHC4) for 28S rRNA [83]. m^6A is involved in molecular processes such as RNA splicing, stability, localization, translation and structure [84–86]

It should be noted that m6A had been mapped in PuGm6ACU sequences already in the 1980s [87,88], and that in 2012 Dominissini et al proposed the high-throughput approach for deep sequencing-based m6A mapping [89]. This study

showed that the m^{6}A sites were conserved throughout evolution and they were dynamically regulated. They also showed that m^{6}A modification usually occurs in certain motifs containing RRACH (R: A or G, H:A, C, or U) (**Figure 1.6**) [89]. Furthermore, many more studies aimed to map the m^{6}A using different approaches [90–92], and many aimed to quantify the m^{6}A modification amount in a modified site [93–95]. However, a high-throughput and full quantitative measurement of modification fractions for each modified site have not been established.

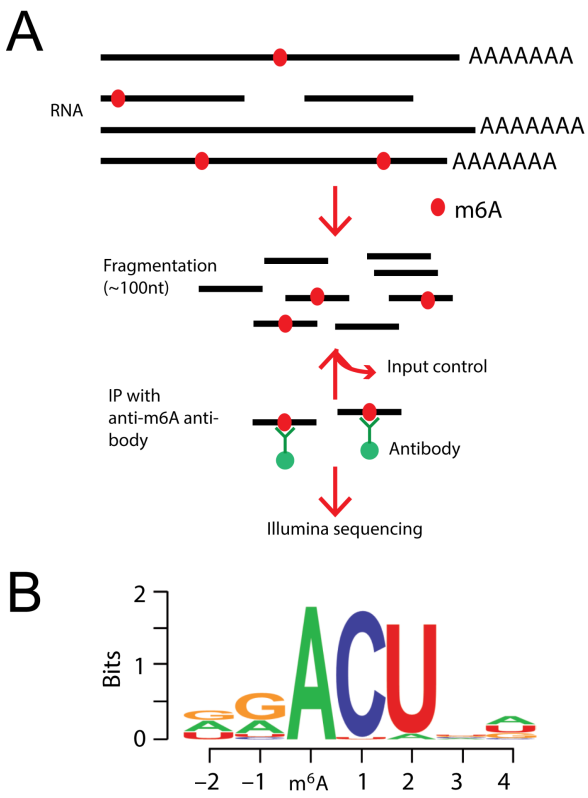


Figure 1.6 - Overview of the m^{6}A-seq and findings.

m^{6}A-seq method relies on the fragmentation of the input RNA followed by immunoprecipitation (IP) of the m^{6}A containing RNA fragments. Both immunoprecipitated and input are then sequenced using next-generation sequencing. m^{6}A-seq revealed that m^{6}A modifications are enriched within the RRACH (R: A or G, H:A,C, or U) motif. Figure adapted from Dominissini et al, 2012 [89]

m^{5}C is a widespread modification in many types of RNAs, including tRNAs, mRNAs, rRNAs and enhancer RNAs (eRNAs) [96]. There are many functional implications of m^{5}C modification, including RNA structure, stability, translation accuracy

and translational readthrough of stop codons [97–99]. Enzymes responsible for catalyzing m^5C modification on RNAs include NSUN family proteins (NSUN1 to 7), and DNMT2 (TRMDT1) [96].

Methods to map m^5C modification include antibody pulldown and bisulphite treatment that converts unmethylated cytosine to uracil (U) and leaves m^5C as it is [100–104]. Surprisingly, there have been many discrepancies between different studies in terms of the m^5C location on the RNAs, which can be explained by the limitations in m^5C antibody specificity and/or the difference in the efficiency of bisulfiphe treatments. A recent study with an improved pipeline pointed out that there are only a few hundred of m^5C sites on human and mouse RNAs, which were found within a sequence that was similar to the m^5C motif on tRNAs [104].

m^1A is present in high stoichiometry in tRNAs and rRNAs [105–108]. TRMT10 and TRM6-TRM61 complexes are responsible for catalysing m^1A modification on tRNA and mitochondrial rRNA [106,108], whereas RRP8 is responsible for catalysing the modification on rRNA [105]. Although it has been reported that both TRMT10 and TRM6-TRM61 complex also catalyse m^1A modification on mRNAs, there is a controversy about the reliability of the discovered positions, such that one study reports a very low number and stoichiometry of sites [109], whereas the other one reports high number and stoichiometry of sites [110]. The author of the earlier study later published another analysis, confirming their results again and claiming that the latter study reported many false positives [111]. Methylation on the N1 position of adenosine on RNA disrupts the Watson-Crick base-pairing and alters the overall charge of the molecule drastically [112]. Due to these dramatic changes in the molecule, m^1A is known to affect RNA-protein interactions, as well as secondary structure [113,114].

Since m^1A disrupts the Watson-Crick base pairing and hence affects reverse transcription (RT), mutations are introduced in the RT product. This property of the modification is used with the combination of antibody pull down and next generation sequencing, to map m^1A modifications on mRNAs. [115–117]. Using this approach, m^1A localization was shown to be enriched near the 5' end of the RNA [112]. Surprisingly, a recent study showed that this enrichment was due to the cross-reactivity of the m1A antibody to the m7G cap [118].

Ψ is the most common modification in cellular RNA and it was also one of the first modifications to be discovered [119]. It is highly present in rRNA and tRNA, as well as many ncRNAs and mRNAs [120–123]. Uridine (U) is converted into Ψ by base-specific isomerization, which is catalysed by pseudouridine synthases (PUSs). Pseudouridylation occurs either via RNA-dependent or RNA-independent pathway. The

RNA-dependent pathway includes dyskerin protein in humans that interact with H/ACA snoRNPs, which comprises H/ACA snoRNAs and their associated proteins (**Figure 1.7**) [124]. This complex mainly modifies rRNAs via base-pairing of the snoRNAs with the target region. The RNA-independent pathway includes PUS proteins that target specific sites on tRNAs, rRNAs, ncRNAs and mRNAs [121–123,125]. The addition of pseudouridine into RNA alters RNA secondary structure, as well as strengths the sugar-phosphate backbone, base pairing and base stacking [120]. Furthermore, certain positions in ncRNAs and mRNAs were shown to be stress-responsive [120,121].

Two main methods used to map pseudouridine transcriptome-wide, namely Ψ-seq and PSI-seq, rely on using CMC to modify pseudouridine positions, which in turn leads to reverse transcription termination. Both of these methods identified hundreds of Ψ sites in human and yeast mRNAs[121,122,126]. An alternative approach is the RBS-seq, which is a variation of the bisulphite sequencing. RBS-seq enables detection of Ψ at single-base resolution transcriptome-wide [127].

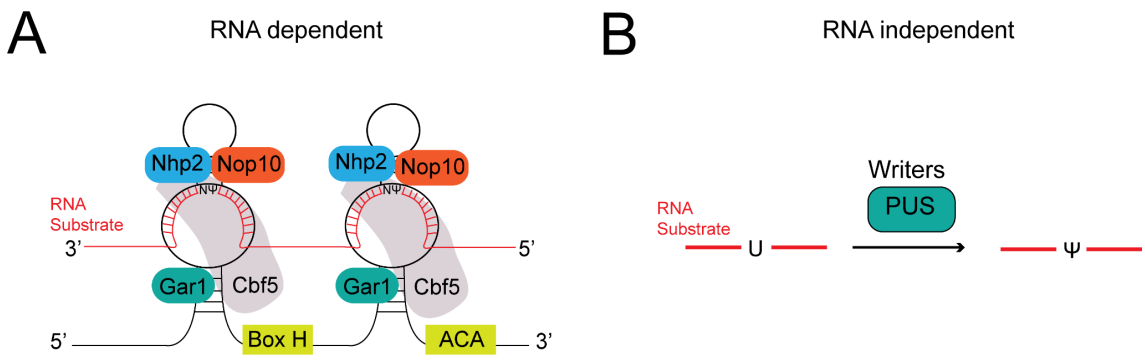


Figure 1.7 - Pseudouridylation pathways.

(A) The RNA-dependent pseudouridylation is guided by H/ACA snoRNAs and their associated proteins, catalysed by Cbf5 (dyskerin in humans). **(B)** The RNA-independent pseudouridylation is catalysed by PUS enzymes without guide snoRNAs. Figure adapted from Zhao et al, 2018 [128].

Internal 2'-O-methylation (Nm) is one of the most common types of RNA modification where the 2'hydroxyl (-OH) of the ribose sugar is methylated. Methylation of the ribose sugar can occur on any nucleotide and Nm is present in tRNAs, rRNAs, ncRNAs and mRNAs [129], with rRNA having more than 100 Nm sites [130]. Moreover, snRNAs in the spliceosomal complex contain Nm modifications that are essential for

the assembly and function of the complex [131,132]. Nm modification can be catalysed either by fibrillarin (FBL) which interacts with box C/D snoRNAs [133,134] or stand-alone enzymes [135,136]. Since the 2'-OH group of the RNA is a key player in the RNA structure formation, its methylation could lead to drastic changes in RNA-protein interactions and RNA secondary structures [137,138]. Furthermore, it alters the hydrophobicity of the molecule and protects the RNA from nuclease degradation [139,140].

Three deep sequencing-based approaches have been proposed to map ribose methylations. RiboMethSeq relies on the ability of ribose methylation to be resistant to cleavage at alkaline conditions [141–143]. This method was used to map ribose methylation dynamics in rRNA, tRNA and snRNAs [141–143]. 2′-OMe-Seq method, on the other hand, takes advantage of the low dNTP conditions, which leads to reverse-transcription terminations at the ribose methylated positions. With this method, 12 previously unannotated rRNA modifications were discovered [144]. Finally, two different studies are relying on the ability of ribose methylation to be resistant to periodate ($IO4^-$) cleavage, namely [145–147] RibOxi-Seq and Nm-Seq. RibOxi-Seq was used to map ribose methylations in rRNA, whereas Nm-Seq was used to map ribose methylations in rRNA and mRNA [145–147].

1.1.3.3. RNA editing

RNA is also 'edited' in eukaryotes by deamination of cytosine and adenosine to form uracil and inosine, respectively, which contributes to genetic diversity and plasticity by leading to changes in RNA sequence compared to its DNA template [148]. Impairment of RNA editing has been reported to be involved in cancer and neurodegenerative disorders [71,149,150].

The first occurrence of RNA editing was observed in the mRNA of human apolipoprotein B (apoB), where a C>U editing introduces a new stop codon which then leads to a smaller version of apoB protein (apoB48) in intestine (**Figure 1.8**) [151]. Later on, it was shown that the enzyme responsible for this editing is a deaminase protein called apolipoprotein B mRNA editing enzyme, catalytic polypeptide 1 (APOBEC1), alongside with its cofactor apobec-1 complementation factor (ACF) [152].A>I editing was first observed in Xenopus, where RNA duplexes and double-stranded RNAs (dsRNAs) were targeted for adenosine deamination [153,154].

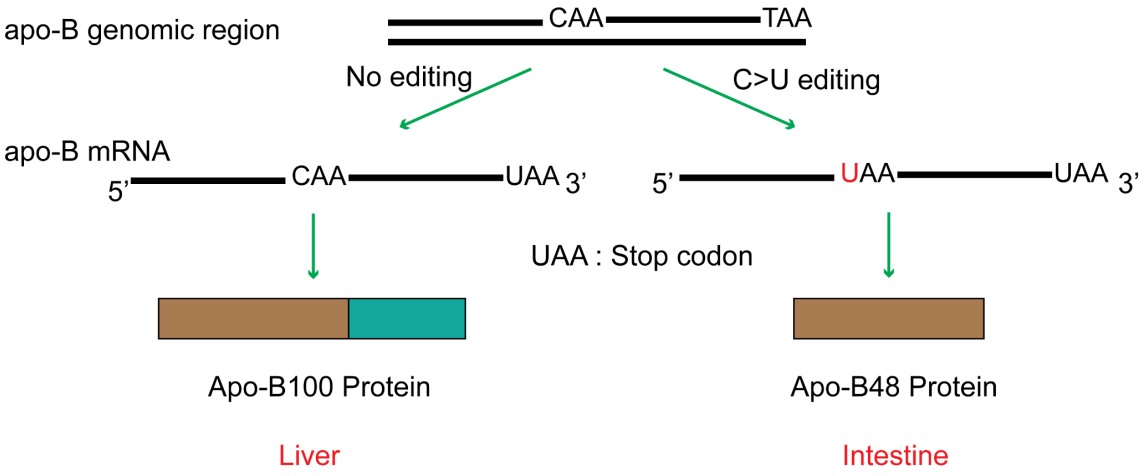


Figure 1.8 - C>U editing of the apo-B gene.

In the absence of the C>U editing in the liver, apo-B mRNA produces the Apo-B100 protein. However, in the presence of the C>U editing in the intestine, CAA sequence is converted into UAA, which leads to an early stop codon. As a consequence, a smaller protein, Apo-B48 is produced in the intestine.

In addition to APOBEC1, which arose in the amniotes, many other proteins are belonging to the activation induced cytidine deaminase/apolipoprotein B editing complex (AID/APOBEC) family, all of which are vertebrate-specific. The first member to arise was AID, which is involved in somatic rearrangement and hypermutation of immunoglobulin domain in adaptive immunity [155]. Other members of this family are APOBEC2 in the vertebrates, APOBEC3 in placental mammals and APOBEC4 in tetrapods [156]. Interestingly, a great expansion of the APOBEC3 in primates has led to the emergence of 7 APOBEC3 orthologs [157]. The first gene that appears in the vertebrates from the family is the AID, which is involved in adaptive immunity [155]. APOBECs edit cytosines in RNA and DNA [158], although only three of them are known to catalyse RNA editing; APOBEC1, APOBEC3A [159] and APOBEC3H [160].

A>I editing of coding and noncoding sequences in RNA is catalysed by a family of proteins called adenosine deaminases acting on RNA (ADARs) [161,162]. Although the catalytic activity of ADAR (ADAR1) and ADARB1 (ADAR2) has been widely studied and proven, the catalytic activity ADARB2 (ADAR3) remains a question [163]. Additionally, the expression patterns of ADAR1 and ADAR2 proteins show that they are ubiquitously expressed in all tissues, whereas ADAR3 is specifically expressed in the brain [164]. ADAR2 targets GluR-B mRNA at a specific coding position, which leads to

the conversion of glutamine at the 607th amino acid position to arginine. This conversion diminishes the permeability of the AMPA receptor to calcium [165] and the absence of Adar2 in mice causes lethal seizures after birth [166].

Despite uncertainty about its catalytic activity, ADAR3 has been shown to play a role in cognitive processes such as learning and memory in mammals [167].

A>I editing can be detected in the form of A-to-G mismatch in RNA sequencing data since inosine shows guanosine-like base-pairing properties. To distinguish RNA editing from single-nucleotide polymorphism (SNP), whole-genome sequencing (WGS) data needs to be used as reference [168–170]. Most A>I editing takes place in long duplex RNAs located in the noncoding regions of the mRNAs - untranslated regions (UTRs) and introns - which indicates involvement of RNA editing in the regulation of gene expression, splicing and binding of regulatory elements to the UTRs [171].

Adenosine editing has expanded in vertebrate, mammalian and primate evolution, and in humans occurs largely in Alu elements, which invaded the primate lineage in three waves and occupy over 10% of the genome, with over 1 million copies [171,172]. The presence of inosine in RNA molecules has been reported to suppress the innate immune response [173–175].

There are two more ADAR-like vertebrate-specific proteins, ADAD1 (or testis nuclear RNA-binding protein, TENR) and ADAD2 (TENR-like) proteins, whose targets are still unknown [176,177]. Finally, tRNAs are edited by adenosine deaminases that act on tRNAs (ADATs) which are evolutionarily conserved in both prokaryotes and eukaryotes [178]. There are three ADAT enzymes (ADAT1-3) in humans and many other eukaryotes [179].

1.1.3.4. Biological functions of RNA modifications

As the most common modification on mRNA, and one of the few whose positions can be assayed transcriptome-wide (using modification-specific antibodies, see below), m^6A is one of the most extensively studied RNA modifications. m^6A has been found to participate in numerous molecular processes including transcription [180], splicing [67,74], mRNA export [181], mRNA structure [182], mRNA stability [183], translation efficiency [15] and miRNA biogenesis [77]. Alterations in molecular processes due to the presence of m^6A modifications is achieved by the interaction between m^6A and reader proteins.

The best-characterised family of the m^6A reader proteins is the YTHD protein family, which was highlighted before. The archetypal representative of this family, YTHDF2, is involved in two mRNA degradation pathways via interacting with m^6A sites

and other proteins [68]. First, m^6A-containing mRNA is targeted by the deadenylase complex (CCR4–NOT complex) via its interaction with YTHDF2 [184–187]. Deadenylation of the mRNA exposes its 3' end, which is then targeted by the exosome complex or DIS3-like proteins for 3'-5' exoribonucleolytic cleavage [188,189]. Second, YTHDF2 recruits heat-responsive protein 12 (HRSP12), which is an adaptor between YTHDF2 and RNase P/MRP complex [185]. RNase P/MRP complex then acts as an endoribonuclease to cleave the mRNA containing m^6A modification [185].

Other YTHD proteins (e.g. YTHDF1, YTHDF2, YTHDF3, and YTHDC2) are also involved in the degradation of m^6A-containing mRNA [190–193]. In addition, other proteins regulate the stability of RNA, including the IGF2BP family of proteins [69], human antigen R (HuR) [194], proline-rich coiled-coil 2 A (PRRC2A) [195], Ras-GTPase-activating protein SH3 domain-binding protein (G3BP1) [196], and fragile X mental retardation protein (FMRP) [197].

The role of m^6A in biological processes has also been widely studied. Depletion of methyltransferases Mettl3 or Mettl14 in mice revealed that m^6A is essential for embryo development and differentiation [51,198,199]. In zebrafish, many maternal mRNAs are degraded by a Ythdf2-mediated pathway and the depletion of Ythdf2 causes a developmental delay [200]. In flies, m^6A is essential for neuronal functions and sex determination [49,201].

Demethylation of m^6A is also essential for biological processes, including spermatogenesis [202] and adipogenesis [183,203]. Finally, targeted inhibition of erasers FTO and/or ALKBH5 by small molecule inhibitors leads to reduced cancer progression [13], which shows the importance of m^6A methylation in cancer.

m^5C is another well-studied RNA modification, most of our understanding of which comes from the abundant RNAs such as tRNAs and rRNAs that are modified by m^5C in multiple locations. In tRNAs, m^5C modification has been suggested to affect tRNA structure, stability and codon-anticodon interactions [204]. In rRNAs, on the other hand, loss of m^5C modification causes disruptions in the rRNA folding, as well as translational read-through of premature stop codons [99]. Furthermore, m^5C modifications were reported to be highly enriched in the 3'UTR of mRNAs or near translation initiation codon, revealed by bisulfiphe-sequencing analysis [70,101,205,206].

Despite their known locations, the functional consequences of the m^5C modifications on mRNA are still uncertain. One interesting study suggests that a nuclear export factor ALYREF interacts with the m^5C containing RNAs and the depletion of this protein leads to nuclear retention of these RNAs [70]. Consistent with

its critical role in RNA metabolism that is highlighted above, m^5C modification has been associated with various biological processes. Most of our understanding of the role of m^5C in these processes comes from the loss-of-function studies of the methyltransferases. Mutations in the NSUN2 gene have been associated with autosomal-recessive intellectual disability [207], Dubowitz-like syndrome [208], disrupted neurogenesis, and impaired brain development [209]. Furthermore, mutations in the NSUN3 gene have been associated with mitochondrial deficiency, leading to developmental disorders in humans [210] and impaired differentiation of embryonic stem cells in mice [211]. Finally, NSUN7, which is predominantly expressed in testis, has been shown to play an important role in male fertility [212].

Ψ is a product of C-C glycosidic isomerization of a uridine base. The difference in its chemical properties from uridine leads to a stronger base pairing with adenine and a firmer phosphodiester backbone [213]. This alteration in the chemical properties of the RNA molecule can affect the secondary/tertiary structure, protein interactions, anticodon recognition, and misreading in the modified RNA. In tRNAs, for example, Ψ modifications are essential for the tertiary of tRNA and these conformational changes due to the modification affect the role of the modified tRNA in translation [214–216]. On the other hand, the presence of Ψ modification in the stop codons was shown to convert them into missense codons [217]. Although the effect of Ψ modification on mRNA stability remains controversial [121,218], it has been established that Ψ modification of synthetic mRNAs increases their evasion of the innate immune system and enhances their translation, the basis of the new generation of mRNA vaccines, pioneered against Covid19 [219–221]. Moreover, studies on PUS enzymes have shown their association with multiple human diseases including Celiac disease [222], X-linked ichthyosis [223], Crohn's disease [222], and mitochondrial myopathy and sideroblastic anemia [224].

Nm is another common type of RNA modification on RNA molecules, which differs from the other methylations by its presence in the ribose sugar, rather than the base. Nm alters the fate of the modified RNA molecule by increasing its hydrophobicity, protecting it against nucleolytic cleavage, altering its interactions with protein and other RNAs, and stabilizing its helical structures [139,140,225–227]. For instance, Nm has been shown to stabilise A-form duplexes and stabilise the RNA-RNA base pairs on modified RNAs [228–231]. Moreover, Nm modification has been shown to distort RNA-protein interactions[138], as well as RNA tertiary structures [137].

Being an abundant modification, Nm is involved in many biological processes. Loss-of-function of the pseudouridine writes enzyme fibrillarin resulted in disruption of

the translation machinery [232], whereas overexpression of the same enzyme led to enhanced translation, leading to increased proliferation of breast cancer cells [233]. Nm modification is also important for spliceosome assembly and function due to its presence in snRNAs [131,132]. For example, depletion of Nm modification of U6 snRNA leads to alterations in splicing, resulting in impaired spermatogenesis in mice [234].

1.2. Mapping RNA modifications

Due to their abundance, most of our understanding of RNA modifications came from tRNA and rRNA modifications detected by thin-layer chromatography (TLC) and high-performance liquid chromatography (HPLC), methods that are outdated now [235,236]. These methods historically detected only abundant RNA modifications since they have low sensitivity. Although the catalogue of the modifications on abundant RNAs was established decades ago, it remains a challenge to map RNA modifications in less abundant RNAs.

1.2.1. Liquid Chromatography Mass Spectrometry (LC-MS/MS)

Liquid chromatography-tandem mass spectrometry (LCMS/MS) is an accurate method that provides both qualitative and quantitative information on the RNA modifications and can be used as an orthogonal method to “validate” predicted sites [237]. It relies on differences in biophysical properties between modified and canonical nucleosides, such as molecular mass and lipophilicity [238]. To be analysed by LC-MS/MS, RNA molecules first need to be digested with enzymes such as nuclease P1, phosphodiesterase, and alkaline phosphatase, which leads to the formation of nucleosides [239]. The main RNA modifications that can be analysed by LC-MS/MS are the abundant types of RNA modifications that occur mostly on tRNAs and rRNAs [237]; therefore a transcriptome-wide mapping of RNA modifications using LC-MS/MS is far from reality. Finally, an exceptional advantage of LC-MS/MS is its ability to do untargeted identification of the RNA modifications, which can be used for a novel type of modification discovery [240].

1.2.2. Next-generation sequencing

Most transcriptome-wide modification mapping has been achieved using NGS-based methods. Since RNA modifications are not abundant in mRNA, earlier methods relied on enriching modification sites by immunoprecipitation using modification-specific antibodies (**Figure 1.9**) [89]. These methods were called methylated RNA

immunoprecipitation followed by sequencing (MeRIP-seq) and initially used for m^6A modification (named m^6A-seq). These early studies showed that m^6A modifications are more common on mRNA than previously thought before and are dynamically regulated [89,90].

Most NGS techniques are based on sequencing-by-synthesis (SBS) technology, such that, reverse-transcription of the RNA into cDNA is followed by a second strand is synthesis with fluorophore tagged DNA molecules. Some naturally occurring RNA modifications (m^1A, m^3C, m1acp3Y,m^3U, m^1G, m22G) are known to disrupt Watson-Crick base pairing and therefore cause reverse transcription errors [117,241]. When encountered with these modifications, reverse transcriptase either falls from the template, leading to RT drop-off or introduces a mismatch. This error signature has been used to detect modifications for over a decade [115,116,242–244].

As outlined before, however, only a few RNA modifications cause reverse transcription errors. Therefore, treatment of the RNA template with chemicals that selectively react with the modified residues has been widely used to expand the repertoire of RNA modifications that can be detected. RNA modifications that can be detected by treating with a chemical include Ψ via CMCT treatment [121,122,126], m^5C via bisulfite treatment [245], and ac4C via sodium cyanoborohydride treatment [246] (**Figure 1.9**). Additionally, Nm can be detected in the reverse transcription product by either alkaline treatment, which exposes the Nm sites in the form of fragments protected from the alkaline hydrolysis or reverse transcription in the presence of low dNTP concentration, which leads to pausing of the reaction [117,247].

NGS-based methods have provided valuable information about RNA modifications since their discovery. Without NGS coupled mapping, we would not have been able to appreciate the diversity, plasticity, and abundance of RNA modifications. However, NGS-based methods have important limitations. First, specific chemical reagents or antibodies are only available for a handful of modifications [248],[117]. In addition, an antibody specific to a modification may cross-react with other modifications, and current studies suggest a high amount of false positives in reported sites [111,118]. A recent example is the demonstration of an m^1A antibody cross-reacting with the 5'cap, resulting in a high amount of false-positive m^1A sites near the 5'UTR [118].

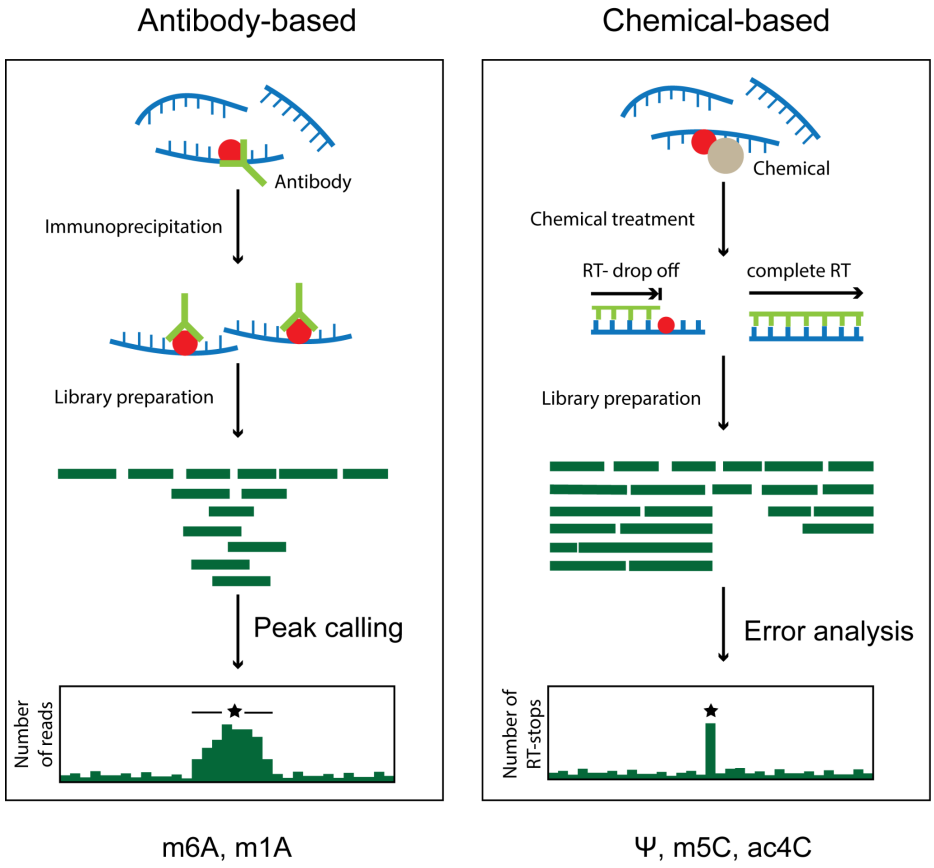


Figure 1.9 - Main NGS-based methods to map RNA modifications.

Antibody-based enrichment, coupled to NGS sequencing, leads to occurrence of "peaks" around the modified positions (left panel). Chemical-based method relies on the RT-drop off that is introduced by the modified site after interacting with the chemical (right panel). Figure adapted from Jonkhout et al, 2017 [52].

There are other limitations of the NGS technology, as the library preparation involves fragmentation, multiple ligation steps, and PCR amplification. Fragmentation of the RNA leads to the loss of isoform-specific information and the inability to detect co-occurring distant modifications in the same transcripts [249]. Moreover, PCR amplification and multiple ligations introduce biases in the sequencing data [250]. Finally, quantitative measurement of the modified sites, namely their stoichiometry, is often not possible with NGS-based methods [249].

1.2.3. Third-generation sequencing

The third-generation sequencing technology platforms developed by Oxford Nanopore Technologies (ONT) [251] and Pacific Biosciences (PacBio) [47] have been proposed to be able to overcome the limitations in detecting RNA modifications in native RNA sequences. PacBio-based RNA modification detection techniques use the information from the kinetic changes of reverse transcriptases in the presence of a modified site [252]. Nanopore sequencing relies on the measurement of disruption in the ionic current when a nucleic acid template is passing through the protein nanopore, which is embedded on the membrane of the flow cell. Since it is a long-read sequencing technology and it can sequence direct RNAs without any manipulation, nanopore sequencing provides a unique opportunity to study RNA modifications directly in a single nucleotide and single molecule resolution in a quantitative manner [251].

A pioneering study using direct ONT RNA sequencing showed that modified sites exhibit different current signals than unmodified sites in the same position in synthetic RNAs (**Figure 1.10**) [251]. Shortly afterwards, another study on direct RNA sequencing of the bacterial 16s rRNA illustrated that change in current intensity is also observed in *in-vivo* samples [253]. These findings triggered a number of subsequent studies to detect RNA modifications using nanopore sequencing [254–260].

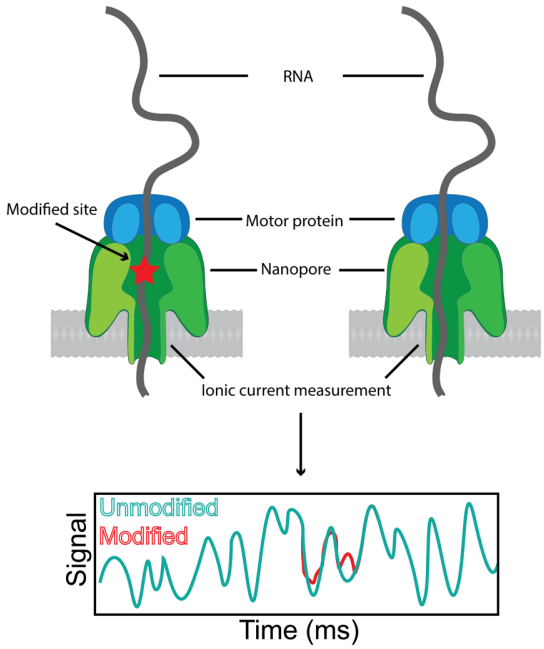


Figure 1.10 - RNA modification detection using ONT direct RNA sequencing.

As RNA goes through a nanopore with the help of a motor protein, disruption in the ionic current is converted to a signal. This signal is altered in the presence of a modified site.

1.3. Detection of Poly(A) tail lengths

Until recently, the measurement of poly(A) tail lengths had mostly focused on individual mRNAs. The view on poly(A) tail length dynamics has vastly changed in the last few years, thanks to the new technologies. In this section, I will briefly canvass the traditional methods, as well as the recently established methods to measure poly(A) tail lengths.

1.3.1. Low throughput methods

The poly(A) tail lengths of many individual genes have been determined using low throughput methods such as Northern Blot, 3' RACE, PCR, and Sanger sequencing-based methods. Northern Blot-based method assesses the poly(A) tail length by comparing full-length mRNAs with mRNAs lacking poly(A) tails as a result of RNase H treatment upon oligo(dT) annealing [261]. Due to its limitations including the intensive labor, requirement of a large amount of input, difficulty in assessing the difference in poly(A) tail length in long mRNAs, non-specific digestion of internal poly(A) stretches within the mRNA body, this method provided only a limited amount of information [262]. 3' RACE uses an oligo-dT adapter primer that uses the poly(A) tail to prime the reverse transcription, and therefore the poly(A) tail length information is kept in the synthesised product, which can be further amplified using polymerase chain reaction (PCR) [263].

Another PCR-based method relies on the enzymatic addition of G or I nucleotides to the 3'end of an RNA. PCR reaction then uses two primers, one of which is designed to anneal to the part where poly(A) ends and G/I tail starts, and another one is designed to anneal to the 150-200 bases upstream of the polyadenylation site. An alternative to the G/I tailing comes from using a linker RNA to ligate to the 3'end of the RNA, instead of a G/I tail and designing a primer to anneal to the linker, which then can be used for amplification [42].

1.3.2. Next-generation sequencing-based methods

Despite the extensive use of NGS technology in almost all aspects of RNA biology, its use in measuring poly(A) tail dynamics fell behind. This was mostly because NGS methods do not favor long homopolymer sequences and library preparation methods selectively exclude the poly(A) tails [264]. Recently, several NGS-

based methods have been made available with adjustments to the widely used protocols to characterise the poly(A) tail length dynamics. One such method is PAL-seq [42], which measures poly(A) tail lengths by incorporating fluorescent tags on biotinylated deoxyuridine triphosphate (dUTP) and then uses signal intensity to quantify poly(A) tail length. While PAL-seq can provide accurate estimates of poly(A) tail lengths, it is technically complex, can present efficiency-related issues during the biotin-dUTP extension step, and can only capture 3' ends of polyadenylated molecules.

An alternative NGS-based method to quantify poly(A) tail lengths is TAIL-seq [265], which relies on the use of RNAse T1 to obtain short fragments, which are then ligated to 3' and 5' RNA adapters, subjected to cDNA synthesis, PCR amplification, and finally sequenced using Illumina platforms. Although these methods have provided information about poly(A) tail lengths in various organisms transcriptome-wide, they have several limitations. First, due to the way the library is prepared, a given poly(A) tail length cannot be assigned to a specific isoform, which leads to the loss of isoform-specific tail length information. Second, NGS-based methods are unable to measure the length of tails if they are longer than the read length, which limits the analyses to shorter tail lengths.

1.3.2. Long-read sequencing-based methods

With the development of long-read sequencing technologies, interest in using this technology to estimate poly(A) tail lengths have increased. The first methods include FLAM-seq [266] and PAIso-seq [267], which are based on PacBio technologies and can identify the tail-isoform relationship. However, these methods still include PCR amplification, multiple ligations, and/or G/I tailing, which introduce biases. Additionally, PacBio technology involves expensive sequencing instruments and only produces a limited amount of reads [268–270]. Another long-read sequencing technology, direct RNA sequencing by ONT [271], has also been established as an alternative way to estimate poly(A) tail lengths [272]. However, this approach cannot capture RNAs that are deadenylated, contain non-canonical tail composition (e.g. polyuridine), or contain poly(A) tails shorter than 10 nucleotides.

1.4. Thesis Objectives

This thesis aimed to develop and apply new approaches to characterise RNA modification and polyadenylation at single-molecule resolution using ONT nanopore sequencing.

First (chapter 2), the thesis aimed to characterise the evolution and expression of RNA modification-related proteins (RMPs), which were previously poorly annotated and largely uncharacterised. The results reveal surprising heterogeneity in the expression patterns of RMPs across mammalian tissues and that RMPs are dysregulated in multiple cancer types.

Second (chapter 3) we set out to develop a proof-of-principle method for the detection of RNA modifications, specifically m^6A, by exploiting systematic errors and low base-calling qualities in direct RNA sequencing data. By using this information, an algorithm was trained with both m^6A-modified and unmodified synthetic sequences which then was used to predict m^6A modifications both *in-vitro* and in yeast mRNA.

Third (chapter 4) we set out to expand the repertoire of the RNA modifications that can be detected with direct RNA sequencing by characterising specifically the distinct signature of the Ψ-modified sites and de novo prediction of the Ψ modifications in mRNAs, ncRNAs, and rRNAs. In doing so we uncovered a novel Ψ modification in yeast mitochondrial rRNA, which we validated using orthogonal methods. The pseudouridylation dynamics across different environmental stresses were also explored. NanoRMS software, which can estimate per-site modification stoichiometries by using the current signal information, was developed.

The fourth (chapter 5) part of this thesis introduces a method called Nano3P-seq that is designed to capture any given RNA molecule from its 3'end, regardless of its polyadenylation status, without the need of PCR amplification or ligation of RNA adapters. Nano3P-seq was able to estimate abundances and tail lengths of various RNA biotypes, and their dynamically regulated tail lengths during vertebrate embryogenesis at the isoform-specific level, correlating with mRNA decay.

2. Integrative analyses of the RNA modification machinery reveal tissue- and cancer-specific signatures

This chapter contains material described in the publication published in Genome Biology (Begik et al., 2020) [256].

I generated all the data, performed all the analyses, and drafted the manuscript and figures with the help of other authors in the publication.

Illustrations in Figures 2.8, 2.12 were drawn by Morghan C Lucas.

Immunohistochemistry images in Figures 2.8, 2.9, 2.17,2.18, 2.19 were edited by Morghan C Lucas.

TMA staining scores in Figures 2.17, 2.18 were calculated by Morghan C Lucas.

Immunofluorescence experiments in Figure 2.12 were performed by Morghan C Lucas.

Huanle Liu and Jose Miguel Ramirez contributed to developing custom scripts for the expression analysis.

This work was supervised by Eva Maria Novoa and John S. Mattick.

2.1. Introduction

As canvassed in the general introduction, technological advancements have revolutionised our understanding of RNA modifications, which can occur by removal (by deamination, often called 'RNA editing') or by the addition of chemical side groups on the ribose or base moieties. Insights into the physiological roles of specific RNA modification-related proteins (RMPs) have mostly come from naturally occurring phenotypes or diseases associated with their loss of function [52–56]. However, prior to the present study, a systematic annotation and characterization of RMPs across human tissues, cell types, and disease states were lacking.

This chapter presents the compilation and analysis of the evolutionary history of 90 RNA modification writers and the gene expression patterns of 146 human RMPs (**Table S2.1**) from 32 tissues, 10 species and 13,358 tumor-normal samples. The analyses revealed that many RMPs display restricted gene expression patterns and/or are dysregulated in specific types of cancer. Specifically, a vast proportion of RNA modification 'writers' were found to have undergone duplications (84%), typically accompanied by a change in their RNA target specificity and/or tissue expression patterns (82%). The most frequent change in tissue specificity is the acquisition of restricted testis-specific expression, suggesting that a significant portion of the human RNA modification machinery is likely devoted to sperm formation and maturation. 27% of human RMPs were also found to be significantly dysregulated in cancers, with the expression of several dysregulated RMPs being correlated with cancer prognosis. Overall, this work reveals an unanticipated heterogeneity of RMP expression across both normal and malignant cell types, and points towards several less-characterised RMPs, such as HENMT1 or LAGE3, as promising drug targets for anti-tumor therapies.

2.2. Comprehensive annotation and evolutionary analysis of RNA modification writers

To reveal the evolutionary history of the RNA modification machinery, I first compiled and manually curated a list of human RMPs (**Table S2.1**, see also *Methods*). Due to the wide chemical variety of RNA modifications, the evolutionary analysis was restricted to the catalytic domain of three major RNA modification 'writer' (RMW) classes: i) methyltransferases, ii) pseudouridylases and iii) deaminases. For each annotated RMW [52,273,274], Pfam domains of the catalytic domain were extracted and used as input for HMM-based searches against the human proteome. This

resulted in a total of 90 human RMWs, doubling the amount of annotated human RMWs in other resources [273]. To determine the evolutionary history and identify duplication events that occurred in each family, ortholog proteins from representative species were retrieved (see *Methods*), and phylogenetic trees were built to identify the number of duplications occurring within each family. Overall, this analysis identified 46 duplication events (**Figure 2.1A**), which have mainly occurred in the base of Eukaryota, Metazoa and Vertebrata (**Figure 2.1B).**

Duplications are often accompanied by changes in substrate specificity (**Figure 2.1C,D**), at least in those RMWs where the substrate specificity has been reported. One such case is the family of m^3C RNA methyltransferases, where the ancestral protein METTL2 modifies both $tRNA^{Arg}$ and $tRNA^{Thr}$, whereas its paralog enzymes, METTL6 and METTL8, methylate $tRNA^{Ser}$ and mRNA, respectively [275] (**Figure 2.1C**). It is important to note that it is still questionable whether METTL8 is capable of methylating mRNAs, since there is a recent paper indicating that it methylates mitochondrial tRNAs [276]. Similarly, the N_1-methylguanosine (m^1G) methyltransferases TRMT10A and TRMT10B modify tRNAs in position m^1G9 [277], whereas its paralog TRMT10C has been reported to place N1-methyladenosine (m^1A) in mitochondrial tRNAs and mRNAs [109], in addition to m^1G in tRNAs (**Figure 2.1D**).

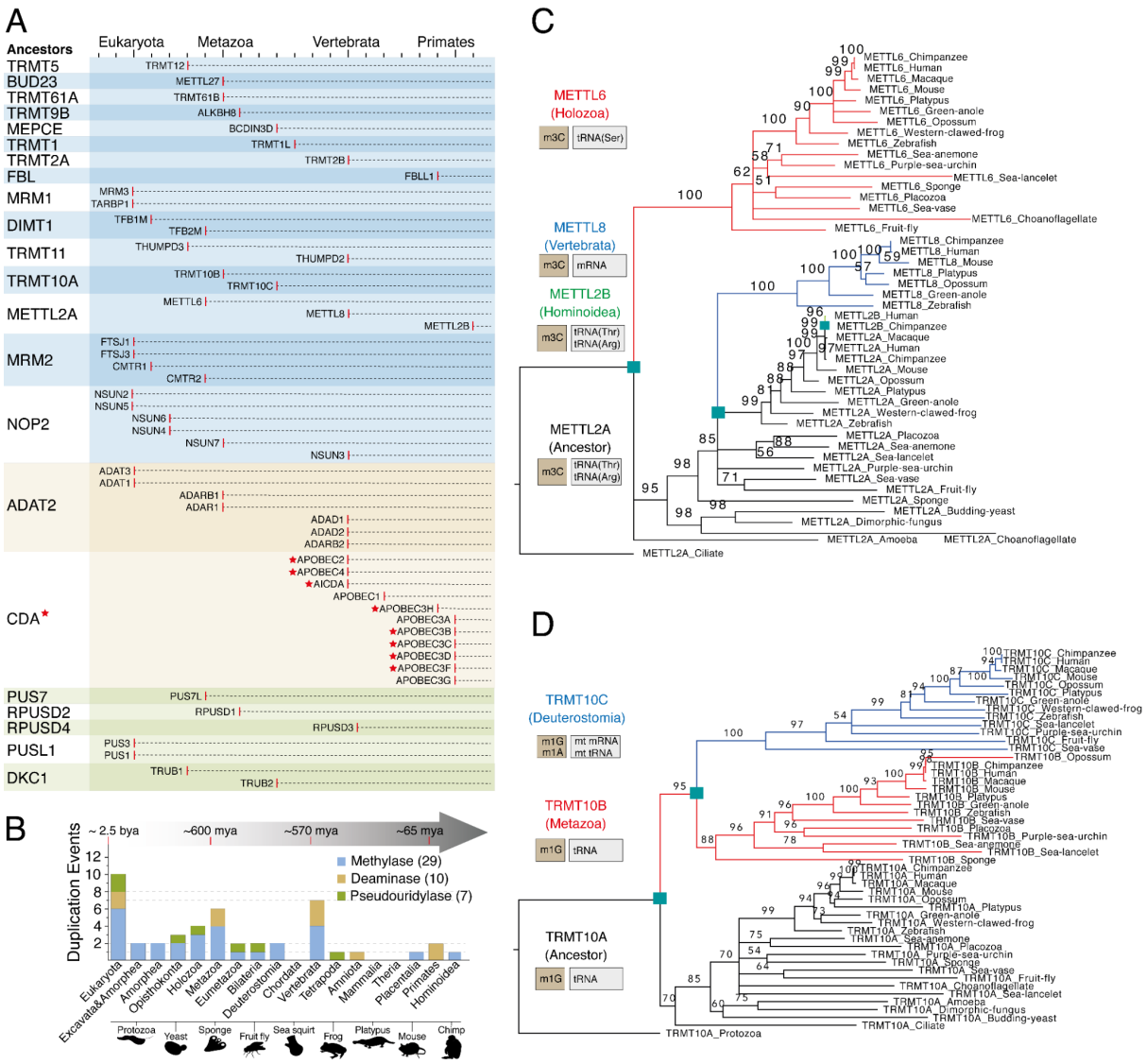


Figure 2.1 - Evolutionary analysis of RNA modification "writers".

(A) Detailed overview of the evolutionary history of RMW duplications during eukaryotic evolution. Red stars indicate that proteins do not target RNAs but they are in the same family with an RNA writer protein. Red lines indicate the evolutionary group in which the enzyme has appeared. **(B)** Histogram of RMW duplication events throughout eukaryotic evolution. Duplication events were inferred using multiple sequence alignments, coupled to maximum likelihood tree generation, for each family.**(C)**, **(D)** Maximum likelihood phylogenetic trees of methyltransferase family METTL2A/2B/6/8 **(C)** and TRMT10A/B/C **(D)**. Cyan squares indicate the node where the duplication occurred. Numbers shown on the branches indicate bootstrapping values

2.3. Heterogeneity of expression patterns among duplicated RMPs is conserved across species

We then wondered whether duplicated RMPs might have acquired distinct tissue expression patterns than the ancestral gene. To test this, I examined the heterogeneity of RMP expression patterns across tissues in human and mice, using publicly available RNASeq datasets [278–280] (**Figure 2.2A,** see also **Figure 2.3A** for gene-labeled heatmaps). For each gene and tissue, I computed 'tissue specificity (TS) scores' [281], which is defined as the deviation of gene expression levels in a given tissue, relative to the average expression across all tissues (see *Methods*). The results show that testis is the most distinctive tissue in terms of RMP gene expression patterns, both in human and mouse (**Figure 2.2A,B)**. This was due to several RMPs being quasi-exclusively expressed in testis (e.g. ADAD1, ADAD2), but also to several RMPs whose expression levels have significantly increased this tissue (e.g. FBLL1, HENMT1, NSUN7). In contrast, other tissues such as the colon displayed none or few tissue-enriched RMPs (**Figure 2.2B,** see also **Figure 2.3B**). Moreover, RMP tissue expression patterns are largely conserved in both mouse and human (**Figure 2.2C**).

To validate the tissue-specific RMP expression patterns, I performed quantitative Real-Time PCR (qRT-PCR) in four mouse tissues (brain, liver, lung and testis), finding similar expression patterns to those observed in the RNAseq datasets (**Figure 2.4**). I then examined whether tissue-specific RMP expression patterns would also be observed at the proteomic level, finding, in agreement with the transcriptomic analysis, that testis showed the most distinctive RMP protein expression levels and patterns among the 17 tissues analysed [282], whereas other tissues, such as the colon, displayed none or few tissue-enriched RMPs (**Figure 2.5**).

The analysis was extended to additional amniote species, finding that testis was also the main outlier in terms of RMP expression patterns in all species analysed, supporting the notion that testis-specific RMP functionalities are evolutionarily conserved (**Figure 2.2D**, see also **Figure 2.6**). Overall, 89% of RMP duplication events were often followed by a change in tissue specificity (32.6%), target specificity (17.4%) or both (39.1%) (**Figure 2.2E**, and **Figure 2.7**), with a major over-representation of acquisition of testis-specific gene expression.

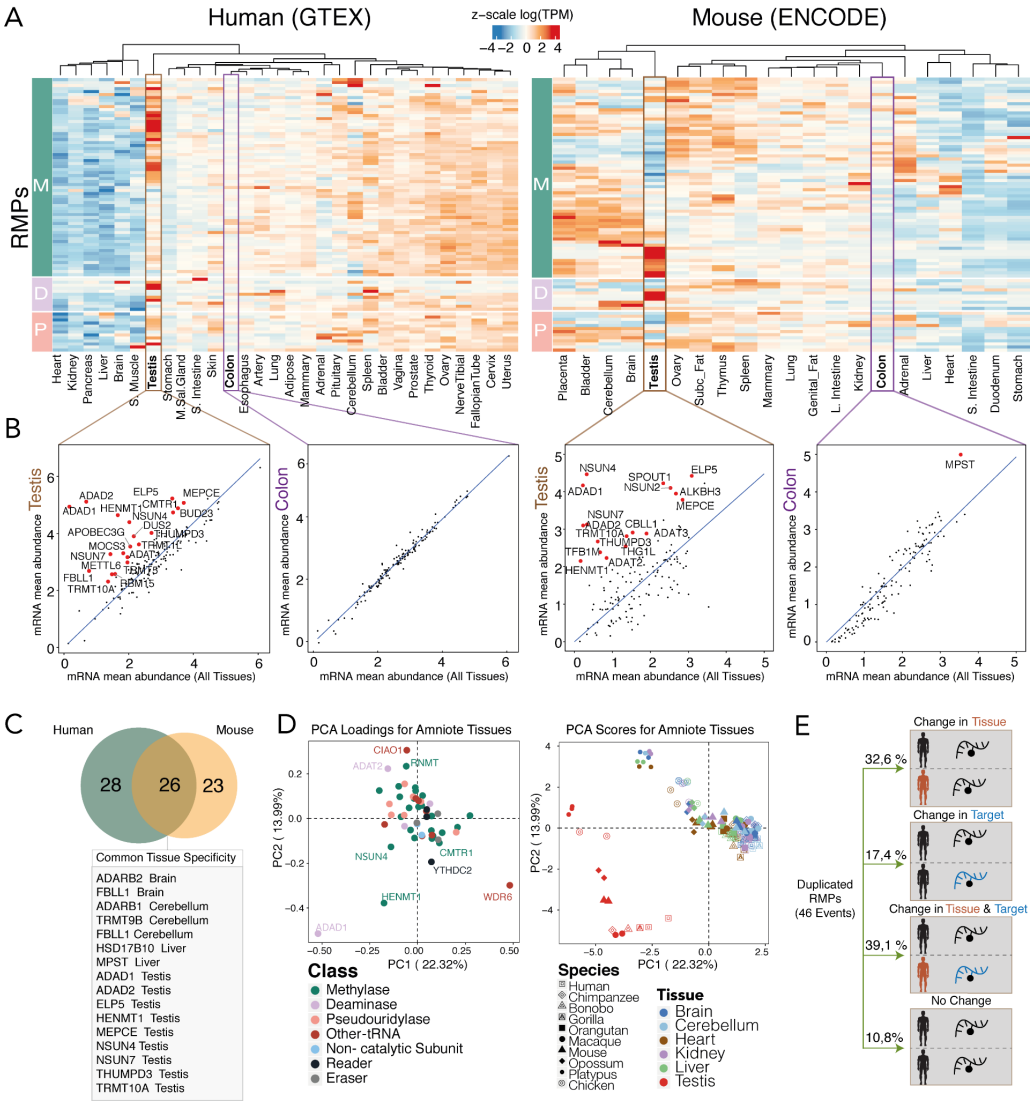


Figure 2.2 - Analysis of RMP tissue specificity expression in different species

(A) Heatmap of z-scaled log(TPM) values of catalytic RNA writer proteins (M: methyltransferases; D: deaminases; P: pseudouridylases) throughout human and mouse tissues. In both, testis has the most distinct RMP expression pattern in which many genes show very high expression, whereas other tissues such as colon show moderate expression level of RMPs. **(B)** Scatter plots depicting tissue-specificity analysis, computed by representing the RMP mRNA expression values in a given tissue (y axis) relative to the mean mRNA abundance in all tissues (x axis). Testis has a significant number of tissue-specific genes in both human and mouse, while colon shows no tissue-specific genes in human and only one in mouse. Tissue-specific genes are labeled in red. **(C)** Venn diagram of the conservation of tissue specificity between human and mouse. **(D)** PCA of amniote tissues based on the log(RPKM) mRNA expression of their RMPs. The loadings plot (left) shows the contribution of each RMP to the clustering of amniote tissues. The scores plot (right) shows the clustering of

each tissue, where testis tissue (in red) is the main contributor to the variance of the data, and is found apart from the rest of the amniote tissues for every given species. **(E)** Schematic representation of the fate of the 46 RMW duplication events shown in Figure 1, showing that 89% of them suffered a change in their tissue and/or target specificity.

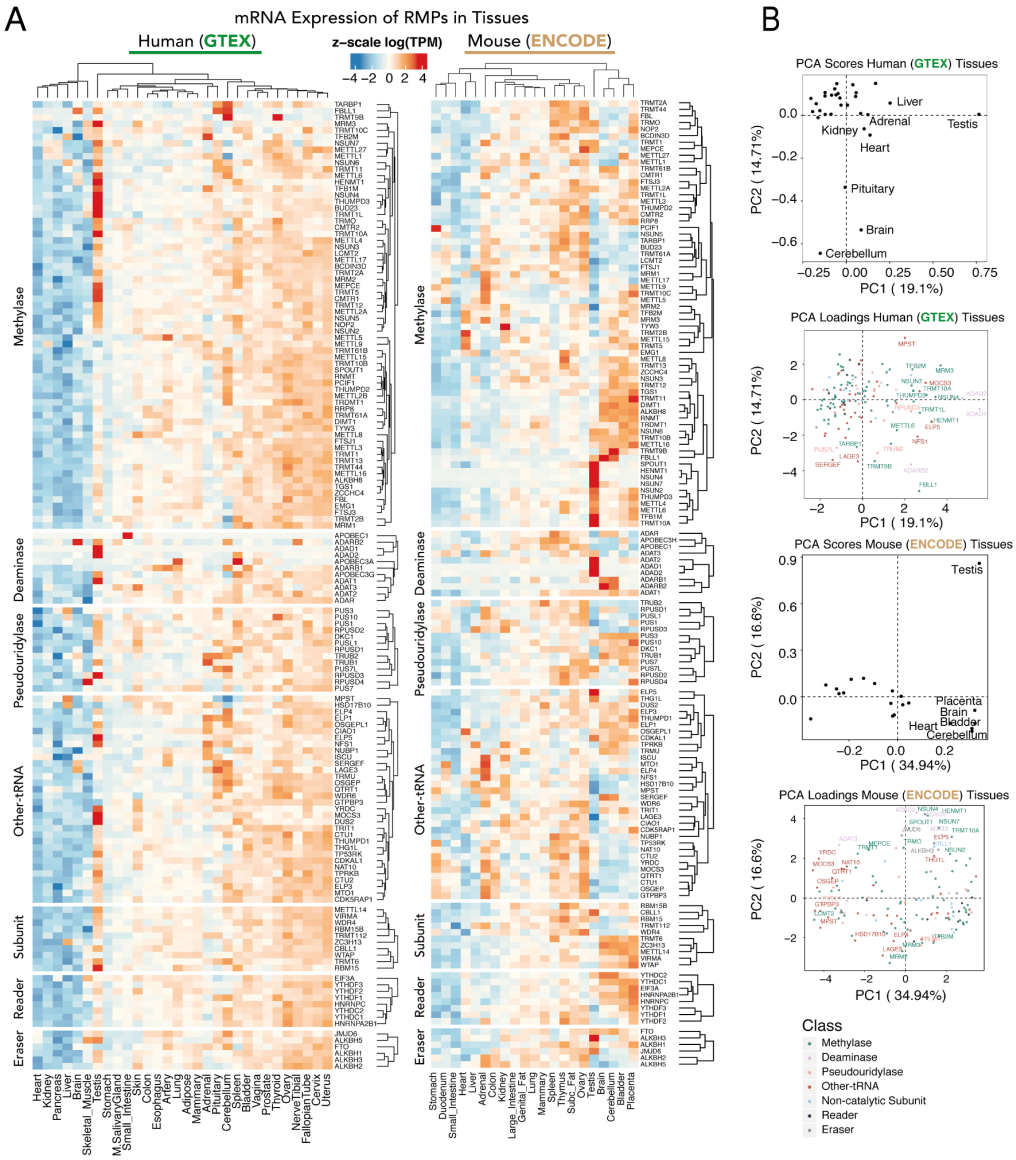


Figure 2.3 - Tissue-specificity of RMPs in various tissues in human and mouse

(A) Heatmap of z-scaled log2 TPM values of RMPs from human and mouse tissues. RMPs have been subdivided into 7 different classes depending on their annotated function: i) methylases, ii) deaminases, iii) pseudouridylases, iv) other writer activity, v) non-catalytic subunit, vi) readers and vii) erasers. RMPs have been individually clustered within each class. **(B)** PCA of human and mouse tissues based on the

expression of their RMPs across tissues. Only the first two principal components are shown. Variance explained by each principal component are shown in each axis. In the loadings plots, RMPs are colored following the same classification used in panel A.

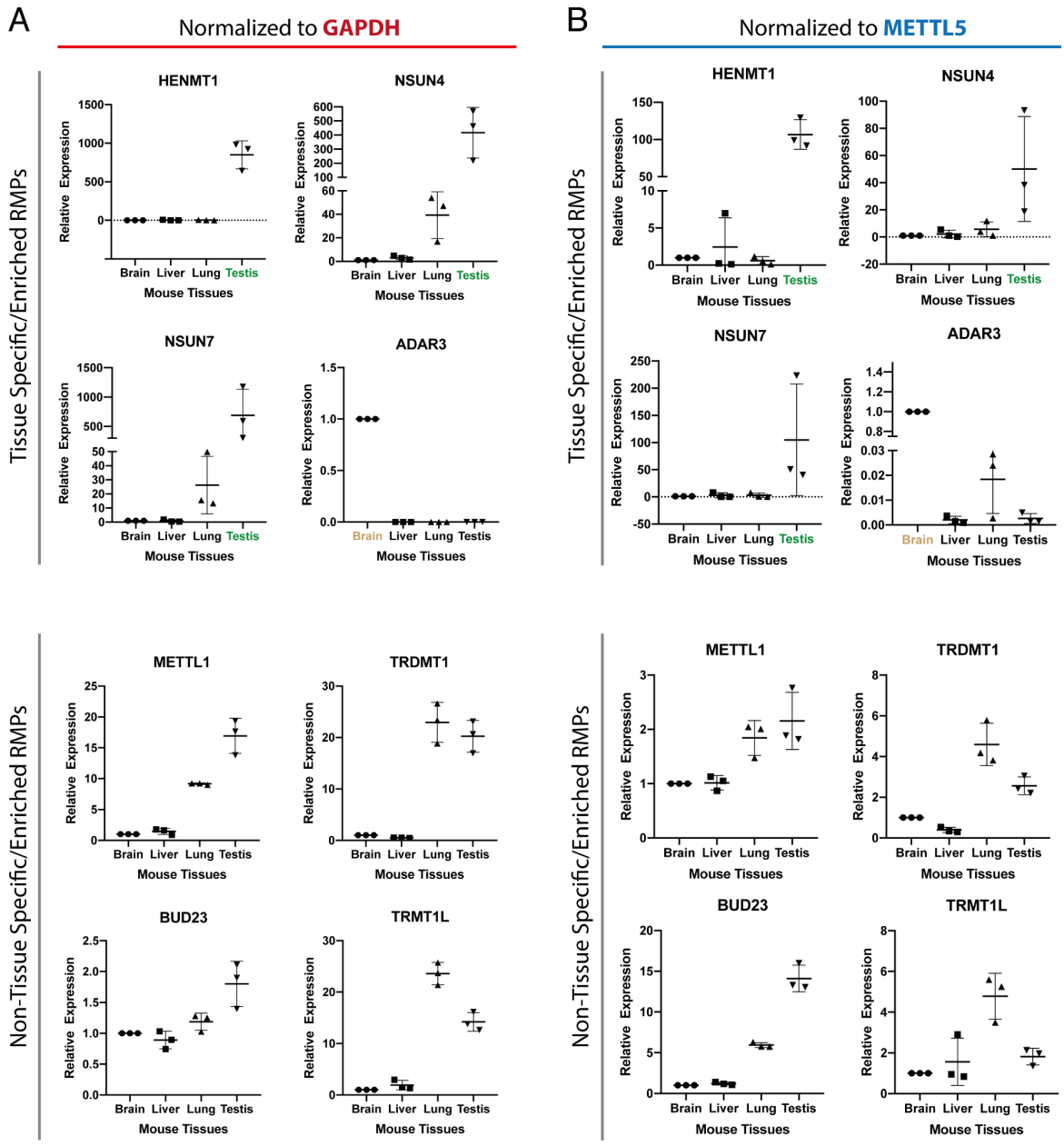


Figure 2.4 - Gene expression analysis of RMPs in mouse tissues

Quantitative real-time PCR of 8 RMPs expressed in four mouse tissues (brain, liver, lung and testis) normalised to either GAPDH **(A)** or METTL5 **(B)**. RMPs have been grouped into two categories, based on whether they are tissue-specific/enriched or non-tissue specific/enriched, as per RNAseq analysis. All the tissues are normalised to the brain tissue of their corresponding biological replicate.

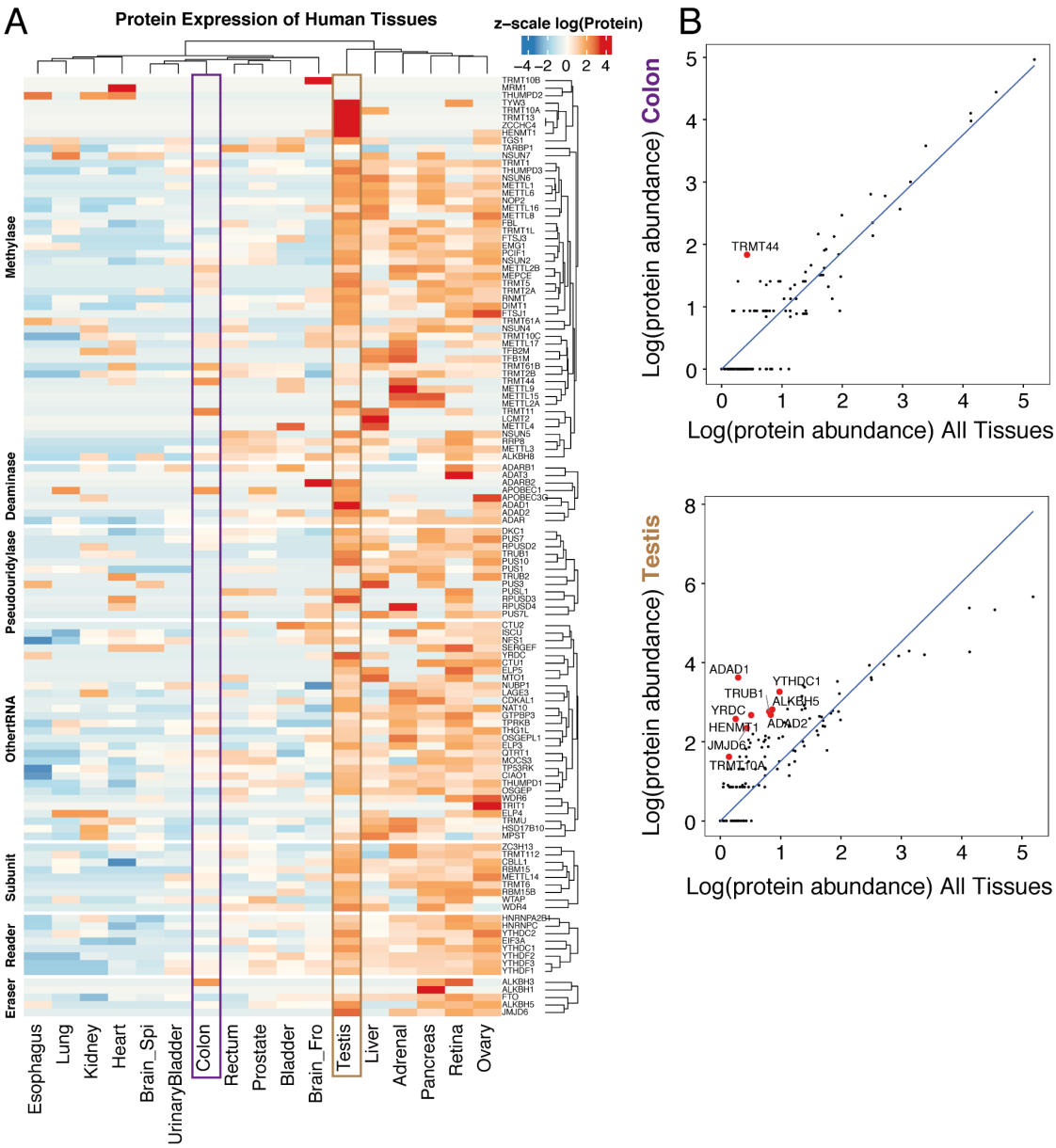


Figure 2.5 - Protein levels of RMPs in human tissues

(A) Heatmap of z-scaled log2 protein levels of RMPs in human tissues. RMPs have been subdivided into 7 classes depending on their annotated function: i) methylases, ii) deaminases, iii) pseudouridylases, iv) other writer activity, v) non-catalytic subunit, vi) readers and vii) erasers. RMPs have been individually clustered within each class. **(B)** Scatter plots depicting tissue-specificity analysis based on protein levels, which have been computed by representing the RMP mRNA expression values in a given tissue (y axis) relative to the mean mRNA abundance in all tissues (x axis). Scatter plots show that testis has a significant number of tissue-specific genes in human, whereas colon shows only one tissue-specific gene in human. Tissue-specific genes are labeled in red.

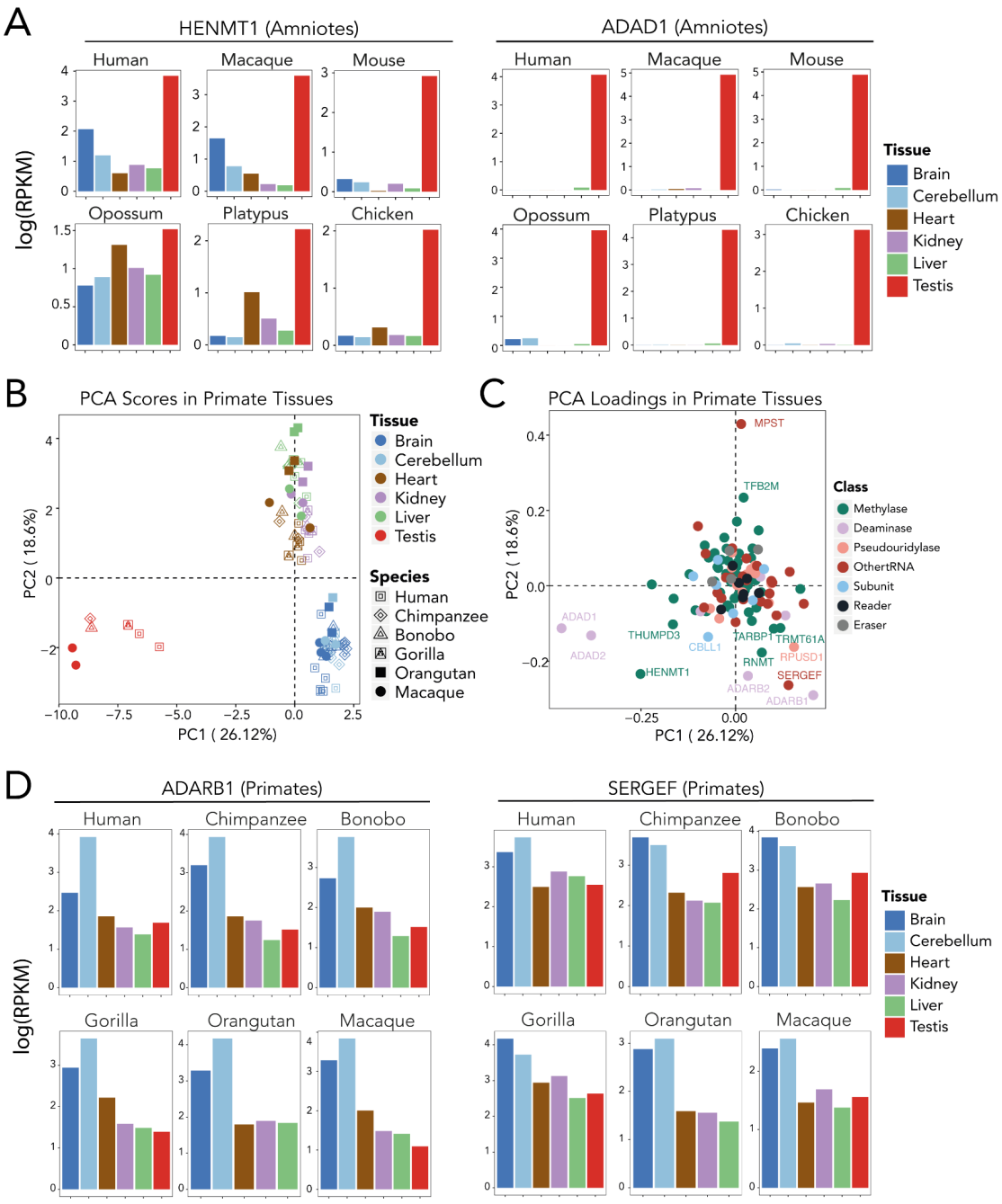


Figure 2.6 - Evolutionarily conserved tissue-specific expression patterns of RMPs

(A) Barplots depicting HENMT1 and ADAD1 expression values in different amniote species and tissues, showing conserved testis-specific expression of these enzymes. **(B, C)** Principal Component Analysis (PCA) of RMP expression values in primates, using as input the log(RPKM) expression of RMPs. Both scores (B) and loadings (C) are shown for the first two principal components. Variance explained by each PC are shown in each axis**. (D)** Barplots depicting ADARB1 and SERGEF expression values in different primate species and tissues, showing conserved brain- and cerebellum-specific expression of these enzymes.

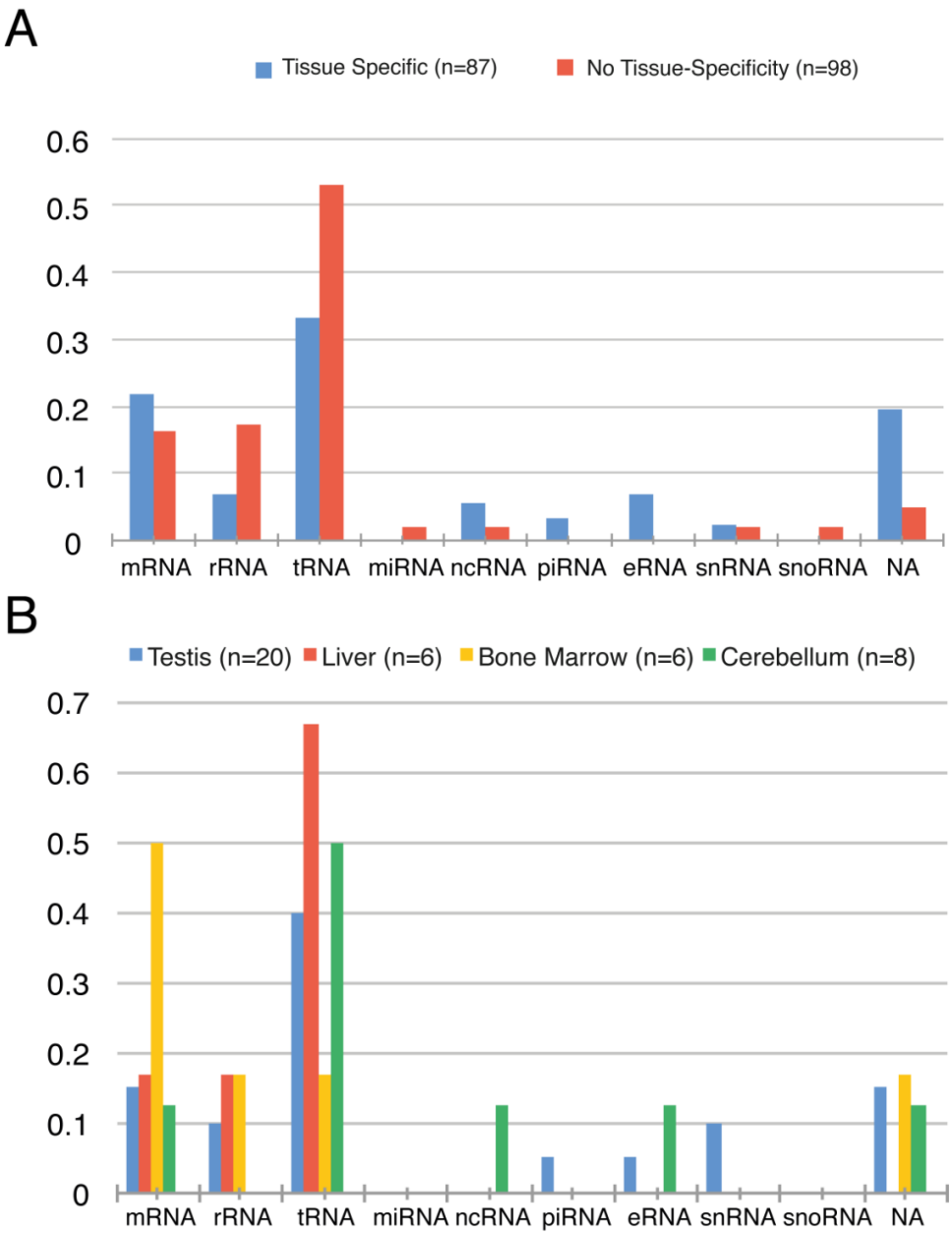


Figure 2.7 - Analysis of target specificity of tissue specific and non-tissue specific genes

(A) Non-tissue specific RMPs mainly target tRNAs, rRNAs and mRNAs. Tissue specific RMPs also target these RNAs, however the proportion of tRNA and rRNA targets is lower. Instead, tissue-specific RMPs target a higher proportion of small non-coding RNAs, including piRNA and eRNA. **(B)** RNA target specificity of tissue-specific RMPs. Only tissues that had 5 or more tissue-specific RMPs were included in the analysis (4/32). The final set of tissues that met this criterion were: testis (n=20), liver (n=6), bone marrow (n=6) and cerebellum (n=8). Liver and bone marrow mainly target tRNAs, rRNAs and mRNAs. In contrast, cerebellum and testis display a larger proportion of tissue-specific RMPs that target distinct families of small non-coding RNAs.

2.4. Testis-specific RMPs are mainly expressed during meiotic stages of spermatogenesis

The process of sperm formation, termed spermatogenesis (**Figure 2.8A**), is a highly-specialised differentiation process in which transcriptional, post-transcriptional and translational regulation are highly orchestrated [283–286]. RNA modifications can influence pre-mRNA splicing, mRNA export, turnover, and translation, which are controlled in the male germline to ensure coordinated gene expression [287]. Recent works have shown that m^6A depletion in mice dysregulates the translation of transcripts that are required for spermatogonial proliferation and differentiation [288]. Moreover, m^5C modifications are essential for the transmission of diet-induced epigenetic information across generations in the epididymis [50]. However, whether additional RNA modifications may be involved in such orchestration is largely unknown.

To identify at which stage of sperm formation and maturation testis-specific RMPs are involved, I gathered publicly available single-cell RNA sequencing data from mouse testis [289] (**Figure 2.8B**, see also **Fig. 2.9A** for gene-labeled heatmap). I first classified RMPs based on their gene expression patterns (see *Methods*), identifying four main expression patterns: (i) high expression only during mitotic stages (spermatogonia); (ii) high expression in both mitotic and meiotic stages (spermatogonia, spermatocytes, spermatids), although decreased in the latter; (iii) low expression throughout spermatogenesis; and (iv) high expression only during meiotic stages (spermatocytes and spermatids) (**Figure 2.8B,C,** see also **Fig. 2.9B,C**).

The results show that the majority of RMPs, including those involved in placing, reading and removing m^6A (VIRMA, YTHDC2, YTHDF2, ALKBH5, METTL14, METTL3) are highly expressed in spermatogonial cells, whereas their expression rapidly drops as the spermatogenic process begins (**Figure 2.8B,C**, see also **Figure 2.9C**). Interestingly, this is not the case for all RMPs, such as m^5C methyltransferase NSUN7, which is not expressed in the early stages of spermatogenesis, but whose expression levels are drastically increased in spermatocytes and spermatids (**Figure 2.9A,C**). Similarly, the testis-specific adenosine deaminase ADAD1 is not expressed in the early stages of spermatogenesis, but its expression levels are greatly increased in meiotic stages. Depletion of NSUN7 or ADAD1 is known to cause infertility [212,290], suggesting that RMPs that are selectively expressed in meiotic stages of spermatogenesis are essential for proper sperm formation and/or maturation. However, the molecular mechanisms behind these infertility phenotypes are largely uncharacterised. Similar expression patterns were observed when analyzing other

publicly available single-cell mouse spermatogenesis RNAseq datasets [291,292] **(Figure 2.9, 2.10** and **2.11**).

I then investigated whether specific RMPs also showed increased expression patterns in the epididymis, relative to other tissues (**Figure 2.9D**). The analysis identified two RMPs as epididymis-enriched: (i) TRDMT1 -also known as DNMT2-, an m^5C methyltransferase modifying position 38 in specific tRNAs [293], and (ii) METTL1, an N_7-methylguanosine (m^7G) tRNA methyltransferase, which has been shown to act on tRNAs [294]. Interestingly, TRDMT1 has been shown to play a major role in the transmission of paternal epigenetic information across generations [50]; however, whether METTL1 is involved in the transmission of such information is yet to be determined.

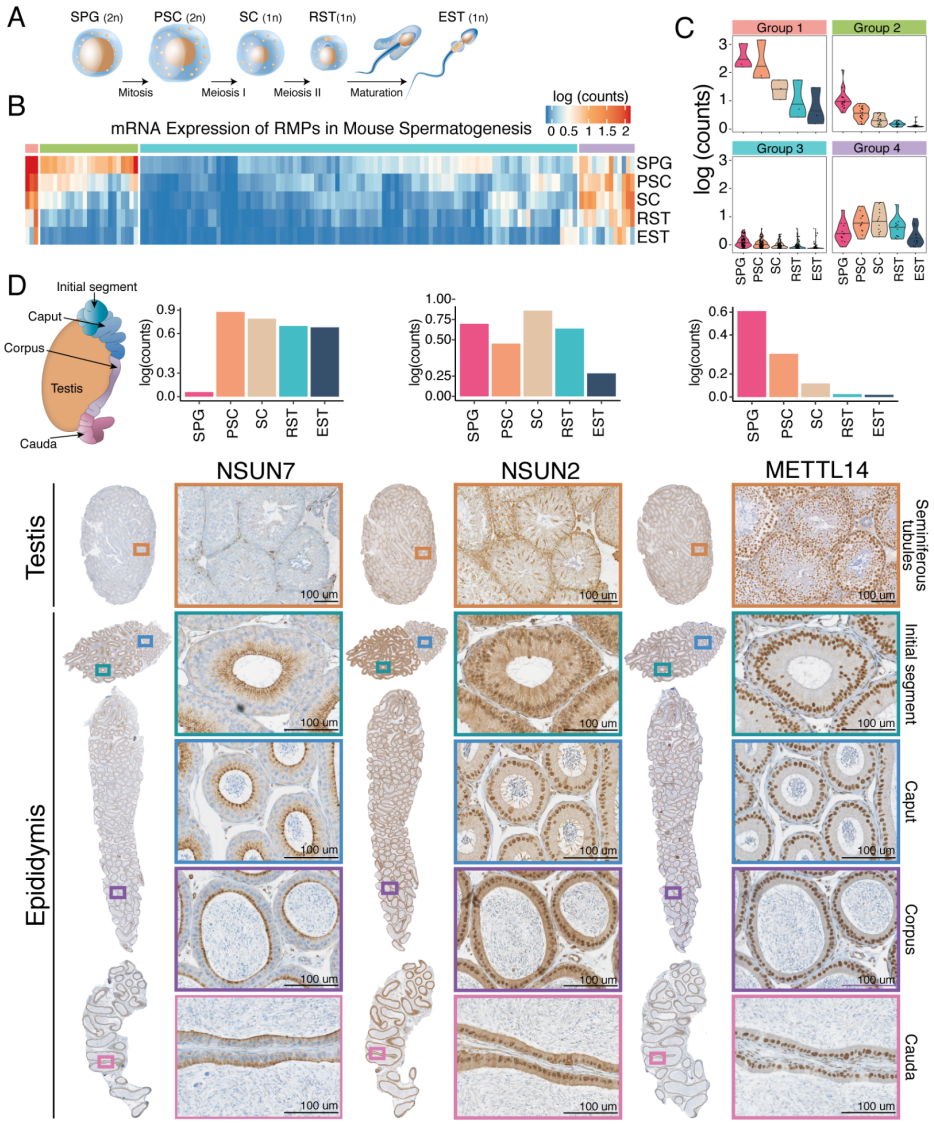


Figure 2.8 - Analysis of RMP gene expression during spermatogenesis.

(A) Schematic representation of the four main phases of spermatogenesis: i) mitotic division of spermatogonia (SPG) into primary spermatocytes (PSC) ii) meiotic division of PSCs into secondary spermatocytes (SC), iii) meiotic division SCs into round spermatids (RST) and iv) spermiogenesis, in which round spermatids (RST) mature into elongated spermatids (EST). **(B)** Heatmap of RMP expression levels in mouse testis. RMPs were clustered into 4 groups based on k-means analysis of their normalised average mRNA expression values. **(C)** Violin plots of the expression patterns of each of the 4 identified clusters **(D)** RNA median expression barplot and immunohistochemistry of NSUN7, NSUN2 and METTL14, depicting distinct protein expression levels along the different sections of the testis and epididymis, as well as different subcellular localizations. Brown color indicates a specific staining of the antibody whereas blue represents hematoxylin counterstain.

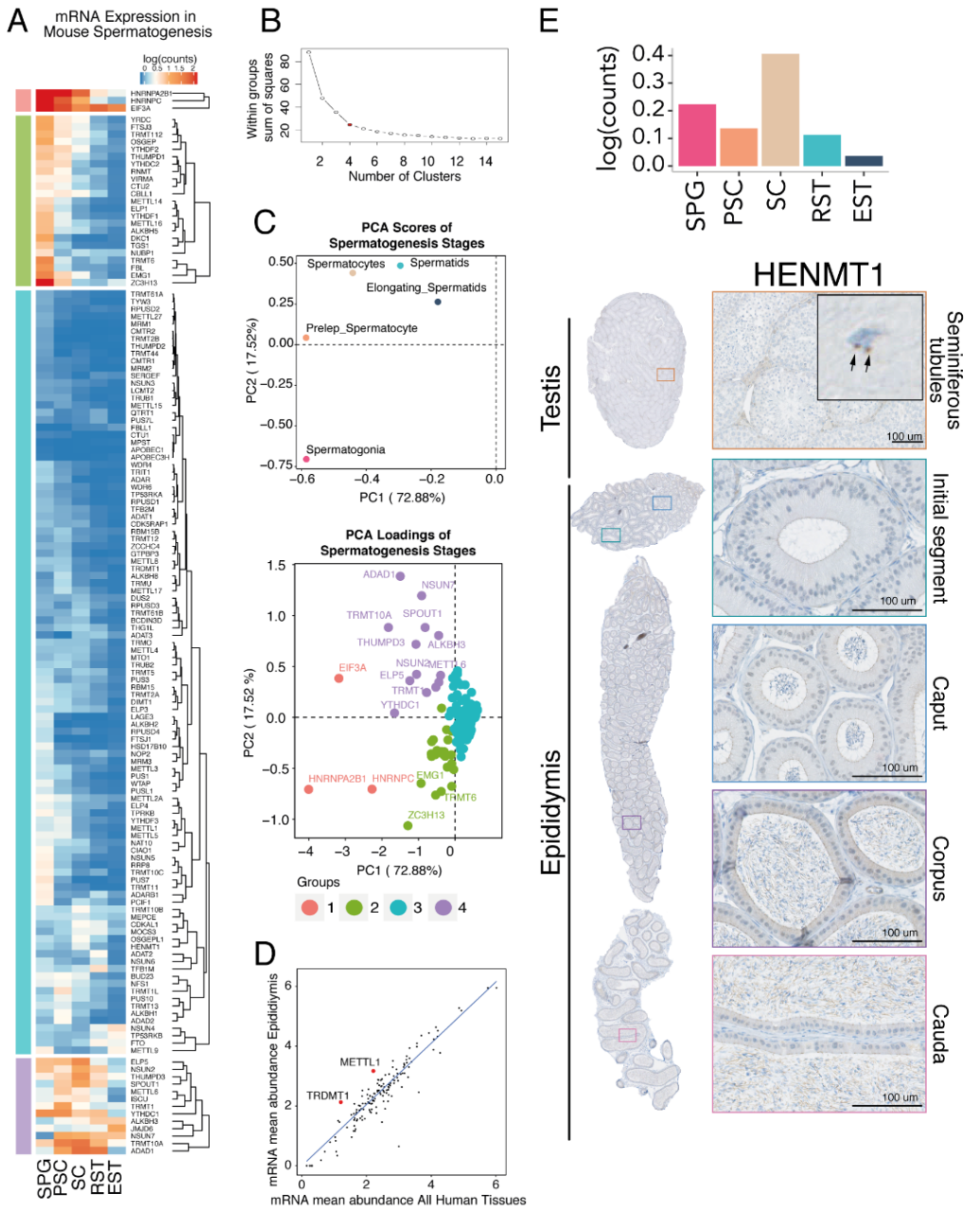


Figure 2.9 - Gene expression patterns of RMPS in spermatogenesis

(A) Heatmap of RMP expression patterns during spermatogenesis, for each of the 4 clusters identified using k-means. **(B)** Within groups, sum of squares was used to determine the optimal number of clusters in spermatogenesis RMP analysis, which is referred to as 'Scree's test'. Based on this test, the optimal number of clusters is 4, which corresponds to the elbow in the curve. **(C)** Principal Component Analysis (PCA) of spermatogenesis RMP expression values. Genes have been colored according to their corresponding cluster. **(D)** Scatter plot depicting tissue-specificity analysis of epididymis tissue, using as input the HPA dataset. Tissue-specific genes are labeled in red. **(E)** In the upper plot, mRNA expression values of HENMT1 for each

spermatogenesis maturation stage are shown. In the lower plot, immunostaining of mouse testis and epididymis using HENMT1 antibody is shown. Brown color indicates a specific staining whereas blue shows hematoxylin counterstain. Arrows show subcellular localization of HENMT1. SPG : spermatogonia PCS : primary spermatocytes, SC: secondary spermatocytes , RS: round spermatids, ES: elongating spermatids.

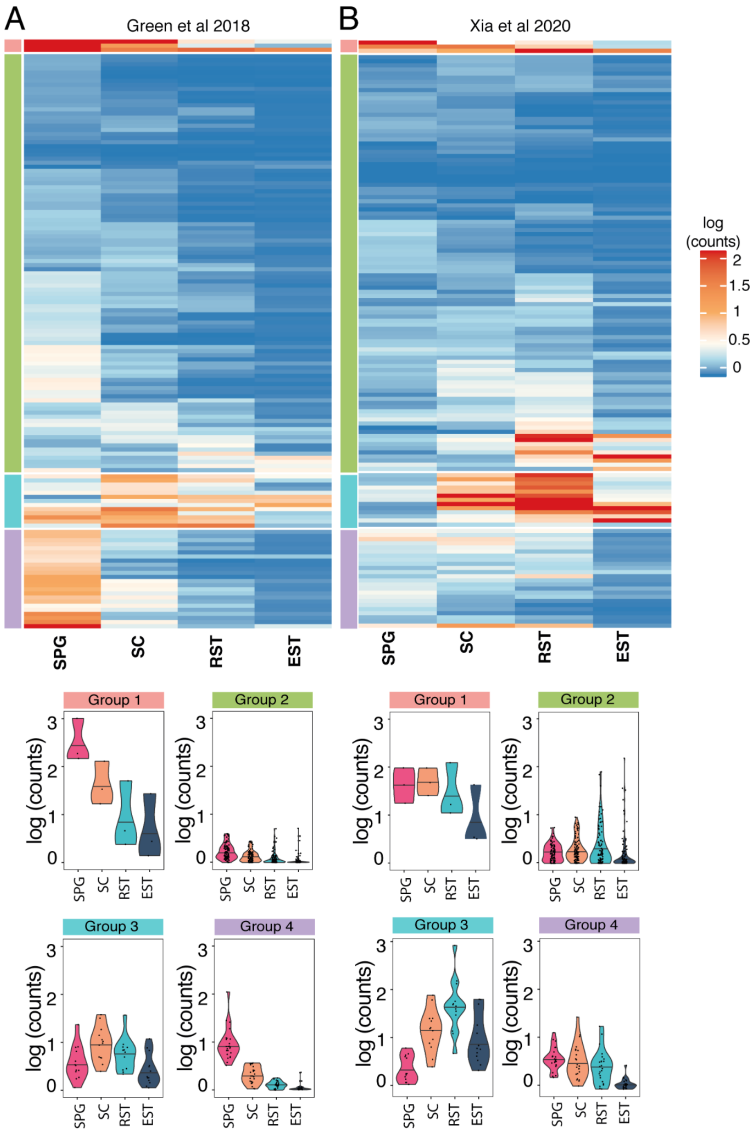


Figure 2.10 - Comparison of RMP expression changes during spermatogenesis using published single-cell RNA sequencing datasets.

Heatmaps and violin plots of log counts shows the expression patterns of different gene clusters based on their expressional behavior during spermatogenesis from **(A)** Green et al., 2018 and **(B)** Xia et al., 2020. SPG : spermatogonia PCS : primary spermatocytes, SC: secondary spermatocytes , RS: round spermatids, ES: elongating spermatids.

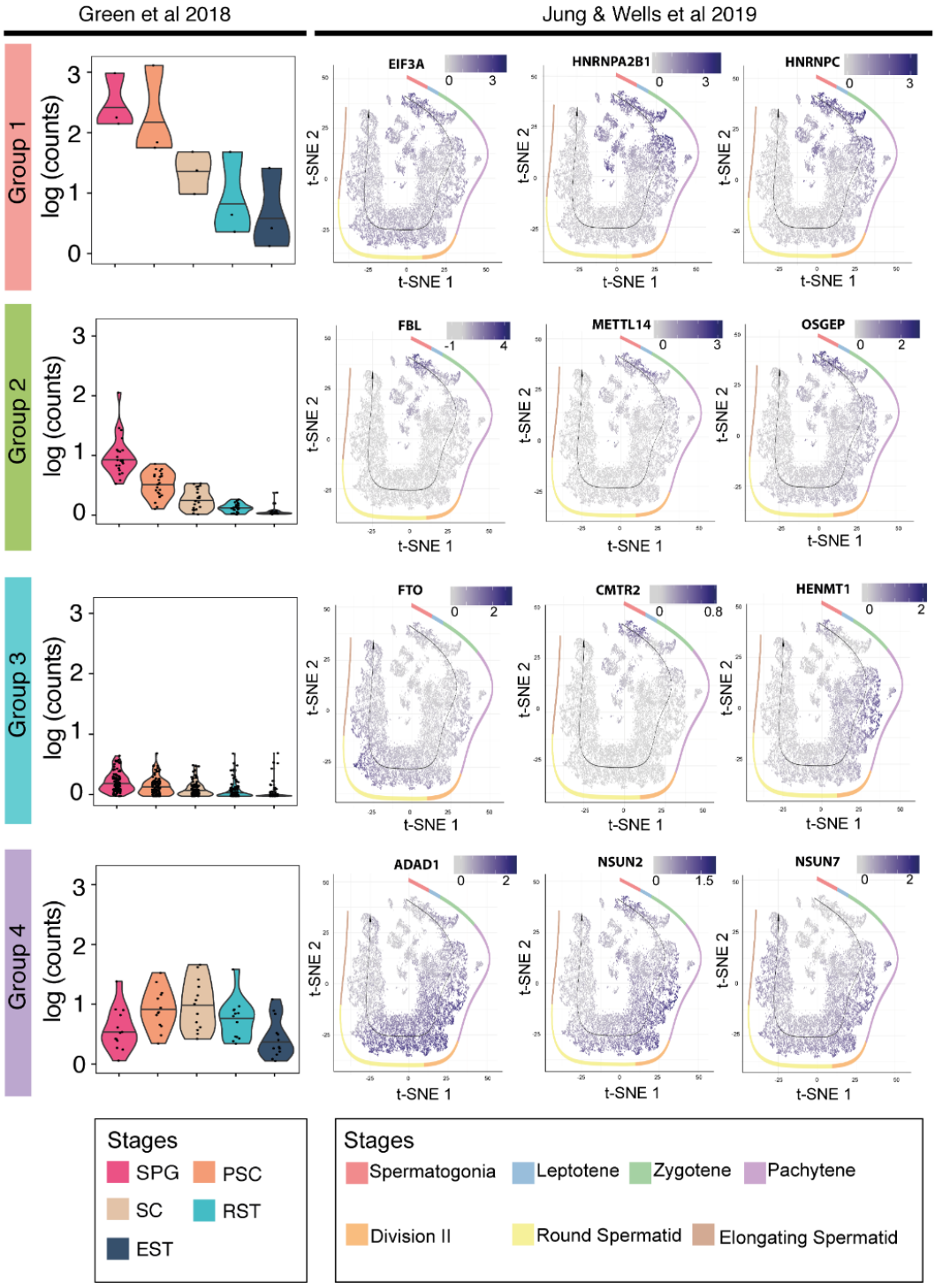


Figure 2.11 - Comparison of RMP expression patterns during spermatogenesis, using the data published by Green et al., 2018 and Jung & Wells et al., 2019

Gene expression profiles from randomly selected genes – three genes per group identified using the initial analysis of single-cell RNA sequencing data (Green et al., 2018) – were extracted using the interactive website from Jung and Wells et al. 2019. SPG : spermatogonia PCS : primary spermatocytes, SC: secondary spermatocytes , RS: round spermatids, ES: elongating spermatids.

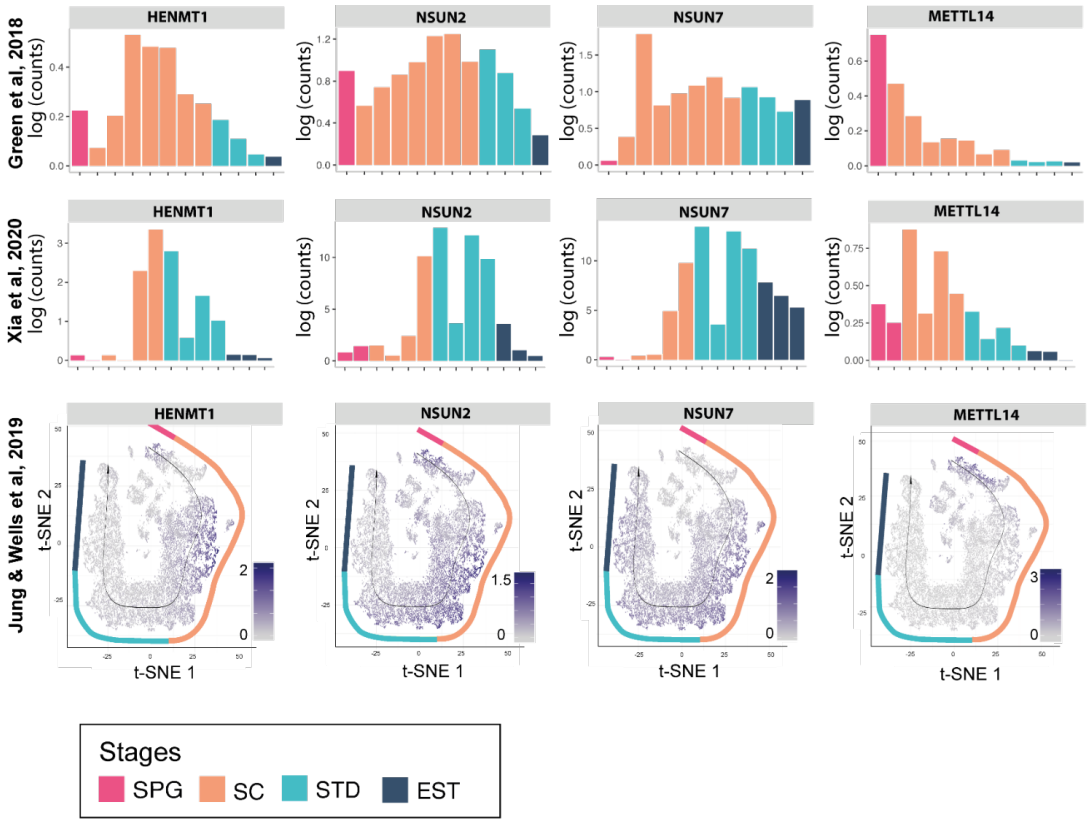


Figure 2.12 - RNA expression patterns of a group of RMPs in different datasets

Comparison of four RMP (HENMT1, NSUN2, NSUN7 and METTL14) expression changes during spermatogenesis using published single-cell RNA sequencing datasets. Barplots of log counts shows the expression patterns of different genes during spermatogenesis. SPG : spermatogonia PCS : primary spermatocytes, SC: secondary spermatocytes , RS: round spermatids, ES: elongating spermatids.

2.5. Immunohistochemistry reveals heterogeneity in RMP expression patterns along the epididymis

It is well established that mRNA levels do not always correlate well with protein levels [295]. Thus, to assess whether our findings would hold at the protein level, immunohistochemistry was performed in both testis and epididymis to characterise the expression patterns of 4 RMPs at the protein level: (i) NSUN7, a putative m^5C methyltransferase that has been shown to affect sperm motility [212,296]; (ii) NSUN2, an m^5C tRNA methyltransferase involved in sperm differentiation [297]; (iii) METTL14, a component of the m^6A methyltransferase complex, which has been shown to be dynamically regulated during spermatogenesis [288], and (iv) HENMT1, a piRNA 2'-O-

methyltransferase responsible for transposon silencing during spermatogenesis [298] (**Figure 2.8D**, see also **Figure 2.9E**).

NSUN7 is most highly expressed in spermatocytes, as well as in the initial segment and caput regions of the epididymis, in agreement with its role in the acquisition of sperm motility [212,296,299,300] (**Figure 2.8D,** left panels). Intriguingly, NSUN7 displayed vesicular-like localization in the epithelial cells of epididymal ducts, with significant accumulation in the apical surface. It is yet to be determined how NSUN7 depletion causes defects in sperm motility, as well as which are the targets of NSUN7 in testis and epididymal tissues. On the other hand, NSUN2 displayed high expression levels in spermatocytes and spermatids (**Figure 2.8D,** middle panels). It is also observed that NSUN2 is highly expressed in the initial segment of the epididymis, with decreased expression in the remaining epididymal sections. To identify the subcellular localization of NSUN7 and NSUN2, immunofluorescence assays were performed in mice testis, co-staining with either fibrillarin (FBL, nucleolar marker) or DEAD-Box Helicase 4 (DDX4, chromatoid body marker [301]) (**Figure 2.13**). NSUN2 was mainly expressed in the adluminal compartment in the later stages of spermatogenic maturation in seminiferous tubules. Surprisingly, the expression of NSUN2 and DDX4 was quasi mutually exclusive, DDX4 being expressed in earlier stages of spermatogenesis, and NSUN2 being expressed in later stages. Colocalization of NSUN2 with DDX4 was not observed, in contrast to previous reports [297].

METTL14 is also highly expressed in early spermatogenesis and down-regulated during the subsequent stages at the mRNA level (**Figure 2.8D,** right panels), in agreement with the dynamic regulation of m^6A levels during spermatogenesis [288]. This result was corroborated at the protein level using IHC, where METTL14-positive early spermatogenic cells are found in the periphery of the seminiferous tubules, while round spermatids and elongated spermatids, located in the very interior of the seminiferous tubules, and spermatozoa, found in the lumen of the seminiferous tubules and epididymis (see **Figure 2.13**), were negative. Finally, HENMT1 was mostly highly expressed in spermatogonia and secondary spermatocytes at the RNA level, however IHC of HENMT1 did not show stage- or cell-specific staining (**Figure 2.9E**). Overall, these analyses showed that RMPs are dynamically expressed during spermatogenesis and during sperm maturation, and that, for the four genes investigated, protein expression patterns were largely in agreement with mRNA expression.

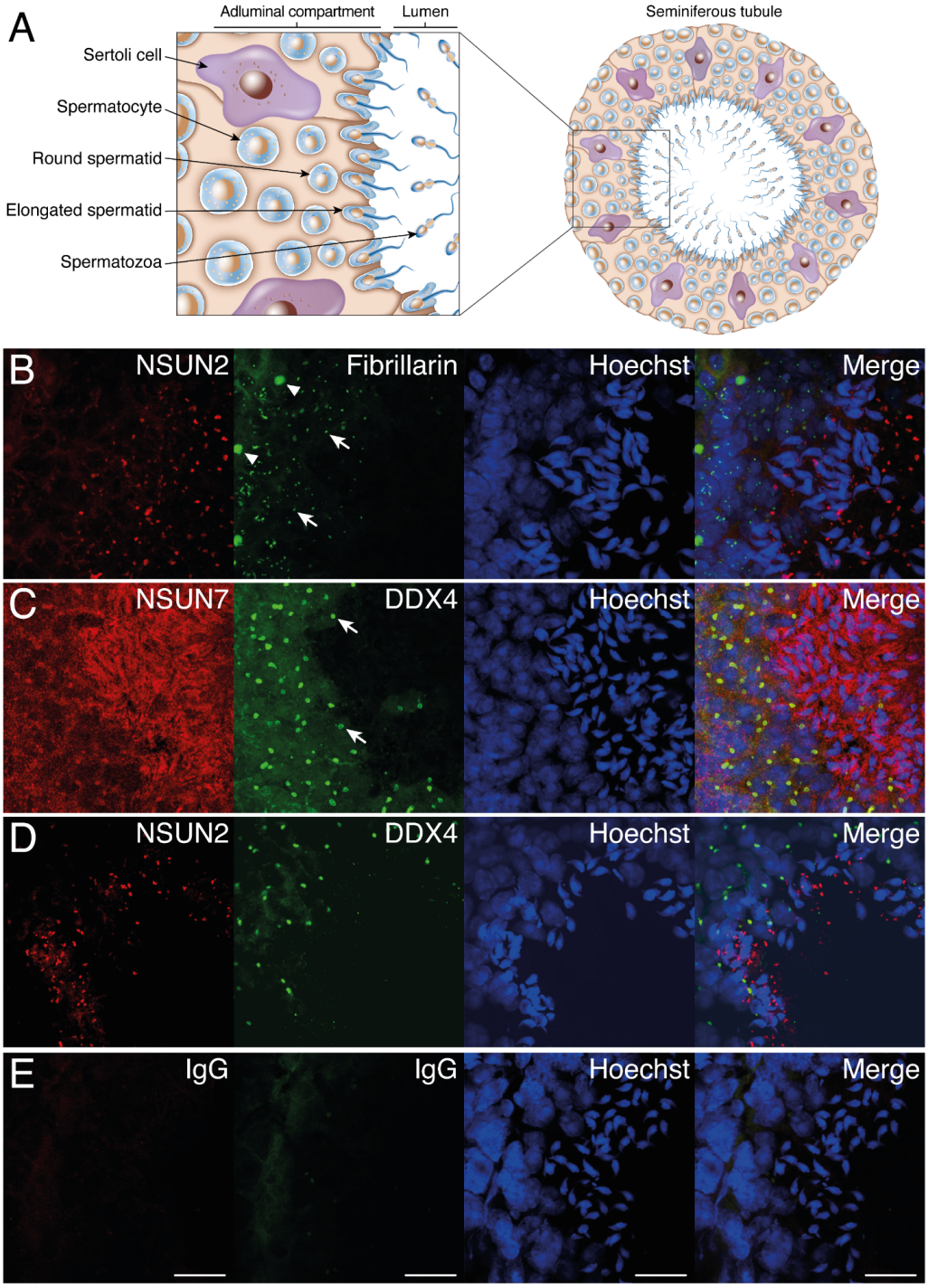


Figure 2.13 - Immunofluorescence of NSUN2 and NSUN7 RMPs in mouse testis

(A) Schematic of the area and orientation of the seminiferous tubules in the confocal images. **(B)** Localization of NSUN2 (red) and nucleolus marker Fibrillarin (green), with arrows indicating nucleoli and arrow heads the nucleoli of Sertoli cells. **(C)** Localization of NSUN7 (red) and chromatoid body marker DDX4 (green), with arrows indicating chromatoid body structures. **(D)** Localization of NSUN2 (red) and DDX4 (green). **(E)** IgG isotype controls. Nuclei were counter-stained with Hoechst 33342 (blue) and a merge of all channels is shown in the far-right column. Scale bar = 25 μm.

2.6. Analysis of RMP expression in tumor-normal paired human samples reveals heterogeneity in RMP dysregulation across cancer types

Due to their ability to modulate RNA metabolism and influence protein synthesis rates, RNA modifications have recently emerged as important regulators of cancer [302–304]. Several studies have shown that modulation of the RNA modification machinery can decrease the expression of specific oncogenes [72,305]. For example, in the case of glioblastoma, treatment with an FTO inhibitor was shown to decrease the expression levels of certain oncogenes [303]. Similarly, tRNA modifying enzymes NSUN2 and METTL1 can affect chemotherapy sensitivity by changing the methylation states of certain tRNAs [306]. Thus, understanding which epitranscriptomic players are dysregulated in each tumor type is essential to guide the research for future anticancer therapies targeting this regulatory layer.

To this end, I performed an integrative analysis of RMPs gene expression across 13,358 tumor-normal paired human samples gathered from publicly available datasets [307], which included 28 different cancer types (**Table 2.1**). Firstly, I compared the expression patterns between paired tumor-normal samples by measuring the log2 fold changes of median gene expression between tumor and normal paired samples, for each RMP and cancer type (**Figure 2.14** and *Methods*).

I found that certain cancer types, such as pancreatic adenocarcinoma (PAAD) and acute myeloid leukemia (LAML) showed significant dysregulation of a vast proportion of RMPs (**Figure 2.14**). Surprised by this result, we wondered whether these global up/down-regulation patterns could in fact be an artefact generated by the use of external datasets. Indeed, certain TCGA cancer types do not have real ‘matched’ tumor-normal data readily available, and often employ data from other publicly available datasets (e.g. GTEx) as ‘normal’ human tissue (**Table 2.1**).

Type	Cancer Type Description	TCGA (Tumor)	TCGA (Normal)	GTEX (Normal)	Normal Tissue	Normal Total
ACC	Adrenocortical carcinoma	77		127	Adrenal Gland	127
BLCA	Bladder Urothelial Carcinoma	407	19	9	Bladder	28
BRCA	Breast invasive carcinoma	1099	113	179	Breast	292
CESC	Cervical squamous cell carcinoma and endocervical adenocarcinoma	306	3	10	Cervix Uteri	13
COAD	Colon adenocarcinoma	290	41	304	Colon	345
ESCA	Esophageal carcinoma	182	13	271	Esophagus (Mucosa)	284
GBM	Glioblastoma multiforme	166	5	206	Brain (Cortex, Frontal Cortex)	211
HNSC	Head and Neck squamous cell carcinoma	520	44		-	44
KICH	Kidney Chromophobe	66	25	27	Kidney	52
KIRC	Kidney renal clear cell carcinoma	531	72	27	Kidney	99
KIRP	Kidney renal papillary cell carcinoma	289	32	27	Kidney	59
LAML	Acute Myeloid Leukemia	173		70	Bone Marrow (K562 Cells)	70
LGG	Brain Lower Grade Glioma	523		206	Brain (Cortex, Frontal Cortex)	206
LIHC	Liver hepatocellular carcinoma	371	50	110	Liver	160
LUAD	Lung adenocarcinoma	515	59	287	Lung	346
LUSC	Lung squamous cell carcinoma	498	50	287	Lung	337
OV	Ovarian serous cystadenocarcinoma	426		88	Ovary	88
PAAD	Pancreatic adenocarcinoma	179	4	165	Pancreas	169
PCPG	Pheochromocytoma and Paraganglioma	182	3		-	3
PRAD	Prostate adenocarcinoma	496	52	100	Prostate	152
READ	Rectum adenocarcinoma	93	10	304	Colon	314
SARC	Sarcoma	262	2		-	2
SKCM	Skin Cutaneous Melanoma	469	1	557	Skin	558
STAD	Stomach adenocarcinoma	414	36	173	Stomach	209
TGCT	Testicular Germ Cell Tumors	154		165	Testis	165
THCA	Thyroid carcinoma	512	59	278	Thyroid	337
UCEC	Uterine Corpus Endometrial Carcinoma	181	23	78	Uterus	101
UCS	Uterine Carcinosarcoma	57		78	Uterus	78

Table 2.1 - Number of samples analysed for each cancer type, both in Normal and Tumor tissues

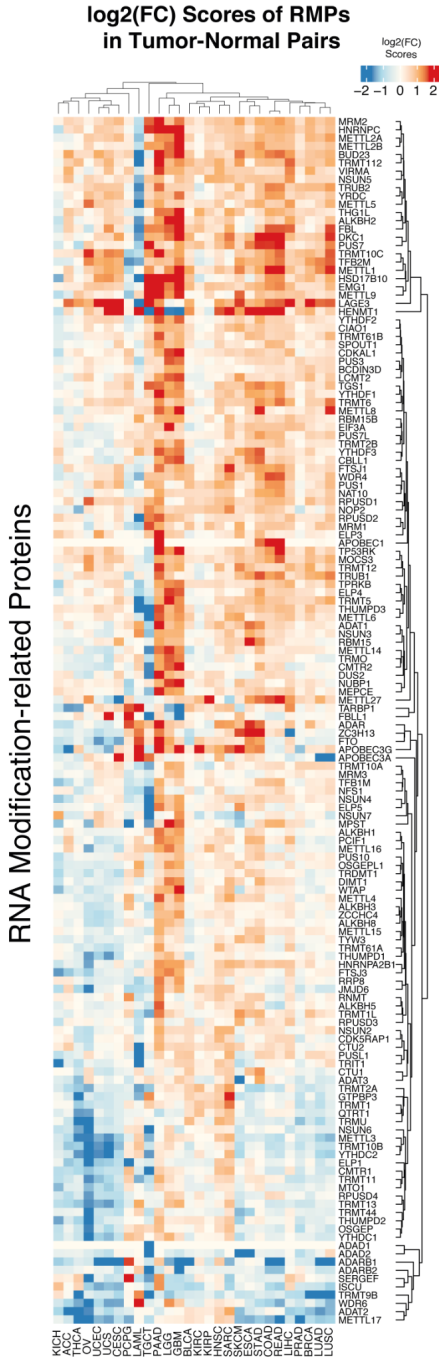


Figure 2.14 - Heatmap of the RMP expression changes between tumor and normal samples, across 28 cancer types.

RMP expression is measured as log2 fold change (log2FC), using the mean differences of all patients. Positive (red) values indicate up-regulation in the tumor, whereas negative (blue) values indicate down-regulation.

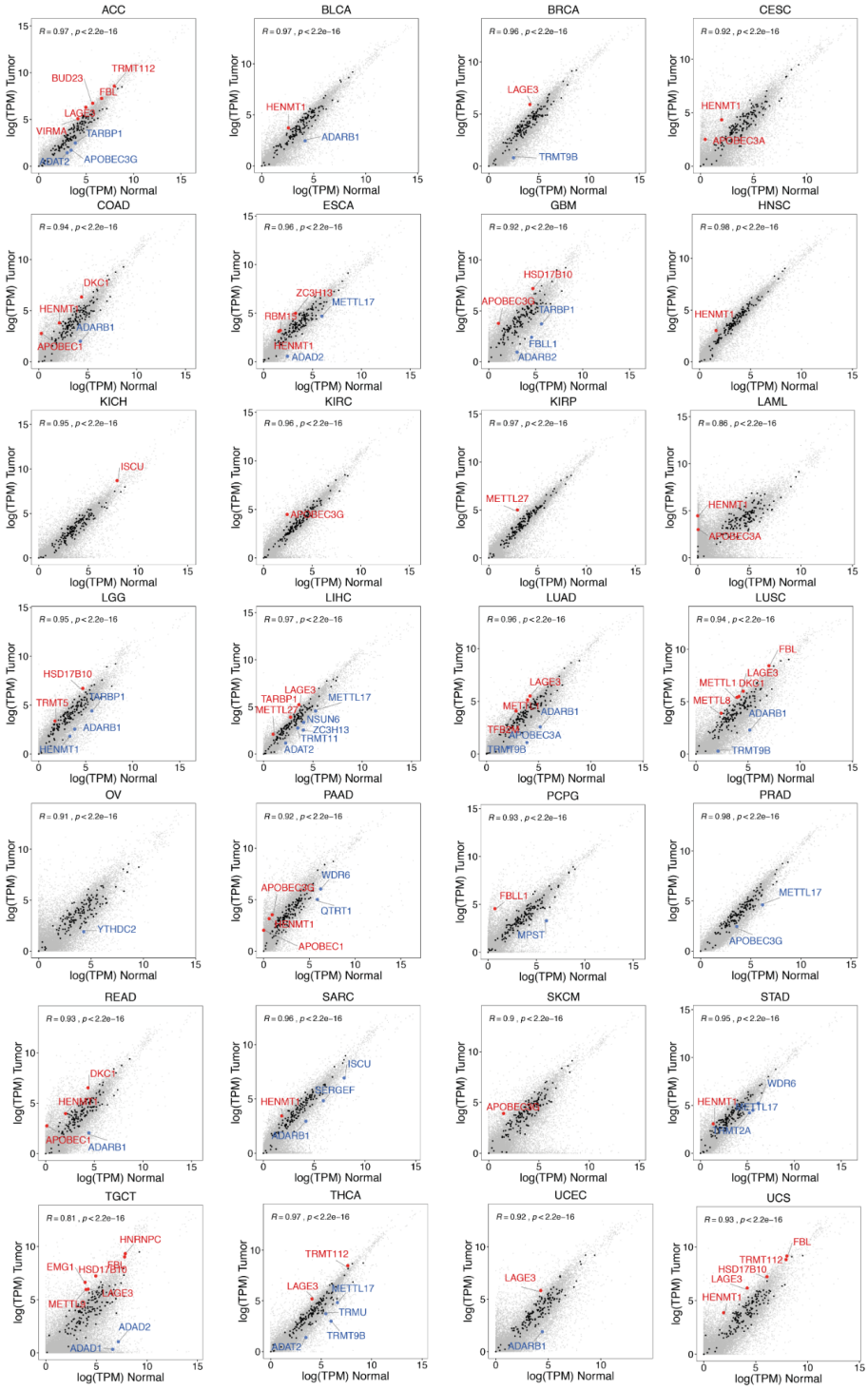


Figure 2.15 - Scatterplots showing expression levels of RMPs in matched tumor-normal samples for all 28 cancer types analysed

Values represent median log(TPM) across all patients. RMPs are shown in black, unless they are significantly up-regulated (red) or down-regulated (blue). Non-RMP genes are shown in grey. Pearson correlation values are shown for each cancer type. See Table 2.1 for abbreviations used for each cancer type.

To address this issue, I extracted the gene expression levels of all genes - not just RMPs - for each cancer type, finding that certain cancer types that employ GTEx data as source of 'normal' human tissues, such as LAML, display low Pearson correlation values between matched tumor-normal samples (r^2=0.86), compared to those observed in other cancer types such as prostate adenocarcinoma (PRAD) (r^2=0.98) **(Figure 2.15**). Thus, to identify which RMPs were significantly dysregulated in each cancer type, I computed 'dysregulation scores' [281], which take into account the global variance of the tumor-normal paired data, for each cancer type (**Figure 2.16A**). An RMP was considered as dysregulated in a given cancer type if its dysregulation score was higher than 2.5 standard deviations (SD) relative to the linear fit to the gene expression in the matched normal tissue (see *Methods*). Using this strategy, a total of 40 RMPs were identified to be dysregulated in at least one cancer type (**Table 2.2**). Moreover, the 'global' up/down-regulation patterns found using log2 fold change comparisons were not observed (**Figure 2.16B**), suggesting that these results were in fact artefacts caused by the lack of proper 'matched' normal tissues for certain cancer types.

2.7. Dysregulation score analyses of tumor-normal paired human samples identify LAGE3 and HENMT1 as top-ranked dysregulated RMPs

We then asked whether specific RMP genes were recurrently up- or down-regulated in multiple cancer types, as these could constitute promising drug targets that could be used to treat diverse cancer types. I identified 11 RMPs that were up-regulated in two or more cancer types, as well as 8 RMPs which were consistently down-regulated in at least 2 cancer types (**Figure 2.16C,** see also **Table 2.2**). I found that the most frequently up-regulated RMP was HENMT1 (**Figure 2.16D**), a piRNA 2'-O-methyltransferase which is highly expressed in gonadal cells, involved in transposable element (TE) mutagenesis protection [298,308,309]. Whether the global up-regulation of HENMT1 in cancer samples might be contributing to increased TE mutagenesis is currently unknown.

The second most frequently up-regulated RMP was the L antigen family member 3 (LAGE3), a component of the complex responsible for the formation of N_6-threonylcarbamoyladenosine (t^6A) in position 37 of tRNAs (**Figure 2.16D**). Interestingly, this modification is found in the anticodon stem-loop of many tRNAs decoding ANN

codons [310], and has been shown to affect both translation accuracy as well as efficiency [311]. Up-regulation of HENMT1 and LAGE3 expression levels was consistently observed in tumors from distinct stages, with the highest expression in stages III and IV (**Figure 2.17**).

Short Name	Up-regulated RMPs	Down-regulated RMPs
ACC	BUD23, FBL, LAGE3, TRMT112, VIRMA	ADAT2, APOBEC3G, TARBP1
BLCA	HENMT1	ADARB1
BRCA	LAGE3	TRMT9B
CESC	APOBEC3A, HENMT1	-
COAD	APOBEC1, DKC1, HENMT1	ADARB1
ESCA	HENMT1, RBM15, ZC3H13	ADAD2, METTL17
GBM	APOBEC3G, HSD17B10	ADARB2, FBLL1, TARBP1
HNSC	HENMT1	-
KICH	ISCU	-
KIRC	APOBEC3G	-
KIRP	METTL27	-
LAML	APOBEC3A, HENMT1	-
LGG	HSD17B10, TRMT5	ADARB1, HENMT1, TARBP1
LIHC	LAGE3, METTL27, TARBP1	ADAT2, METTL17, NSUN6, TRMT11, ZC3H13
LUAD	LAGE3, METTL1, TFB2M	ADARB1, APOBEC3A,TRMT9B
LUSC	DKC1, FBL, LAGE3, METTL1, METTL8	ADARB1, TRMT9B
OV	-	YTHDC2
PAAD	APOBEC1, APOBEC3G, HENMT1	QTRT1, WDR6
PCPG	FBLL1	MPST
PRAD	-	APOBEC3G, METTL17
READ	APOBEC1, DKC1, HENMT1	ADARB1
SARC	HENMT1	ADARB1, ISCU, SERGEF
SKCM	APOBEC3G	-
STAD	HENMT1	METTL17,TRMT2A, WDR6
TGCT	EMG1, FBL, HNRNPC, HSD17B10, LAGE3, METTL9	ADAD1, ADAD2
THCA	LAGE3, TRMT112	ADAT2, METTL17, TRMT9B, TRMU
UCEC	LAGE3	ADARB1
UCS	FBL, HENMT1, HSD17B10, LAGE3, TRMT112	

Table 2.2 - **List of significantly dysregulated RMPs identified using dysregulation score-based analysis**

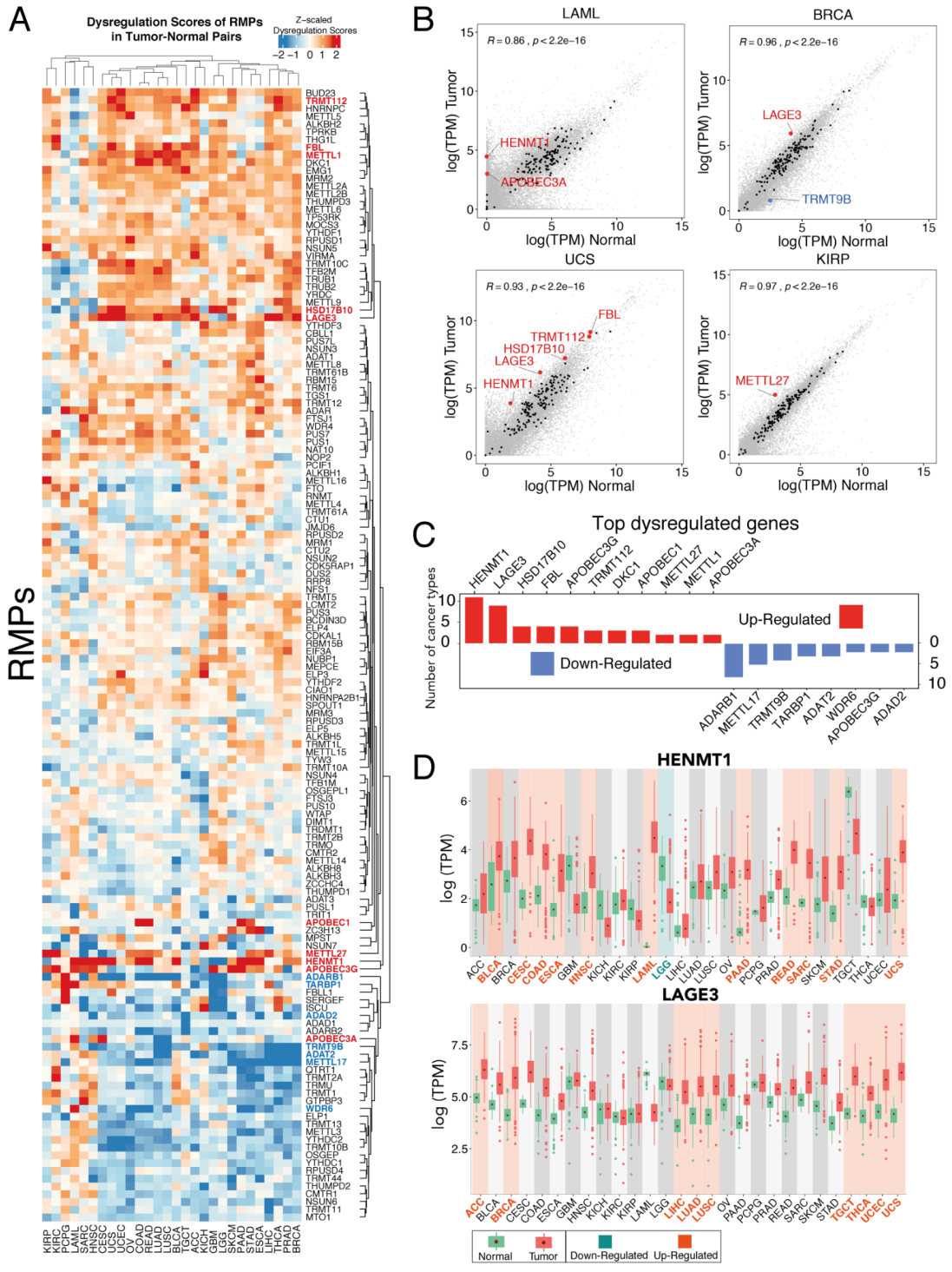


Figure 2.16 - Expression analysis of RMPs in human tumor-normal paired samples.

(A) Heatmap of z-scaled dysregulation scores of RMPs in tumor-normal paired samples, across 28 cancer types. Positive (red) values indicate up-regulation in tumor samples, whereas negative (blue) values indicate down-regulation. Genes labeled as

red (up-regulated) and blue (down-regulated) represent top significantly dysregulated genes, which are also individually listed in panel C. **(B)** Scatter plot comparing RMP expression levels of matched tumor-normal samples, for the following cancer types: LAML (Acute Myeloid Leukemia) and UCS (Uterine Carcinosarcoma) BRCA (Breast invasive carcinoma) and KIRP (Kidney renal papillary cell carcinoma). Values represent median log(TPM) across all patients. Black data points indicate the expression of RMPs, where dysregulated genes are highlighted in red (up-regulated) or blue (down-regulated). Non-RMP genes are depicted in grey. **(C)** Barplot illustrates the number of cancer types in which significantly dysregulated genes are highlighted in red (up-regulated) or blue (down-regulated). Only RMPs that are dysregulated in more than 2 cancer types are shown. For the full list of dysregulated RMPs, see Table 2.2. **(D)** Boxplots of log(TPM) mRNA expression values of HENMT1 (upper panel) and LAGE3 (bottom panel) across all 28 cancer types analysed in this work. Green box plots represent normal samples, whereas red box plots represent tumor samples. Tumor-normal pairs highlighted in cyan represent cancer types in which the RMP is significantly down-regulated, whereas those highlighted in orange represent those cancer types in which the RMP is up-regulated. Error bars represent standard deviation of mRNA expression levels across patients. Each data point represents a different patient sample. Abbreviations: ACC (Adrenocortical carcinoma), BLCA (Bladder Urothelial Carcinoma), BRCA (Breast invasive carcinoma), CESC (Cervical squamous cell carcinoma and endocervical adenocarcinoma), COAD (Colon adenocarcinoma), ESCA (Esophageal carcinoma), GBM (Glioblastoma multiforme), HNSC (Head and Neck squamous cell carcinoma), KICH (Kidney Chromophobe), KIRC (Kidney renal clear cell carcinoma), KIRP (Kidney renal papillary cell carcinoma), LAML (Acute Myeloid Leukemia), LGG (Brain Lower Grade Glioma), LIHC (Liver hepatocellular carcinoma), LUAD (Lung adenocarcinoma), LUSC (Lung squamous cell carcinoma), OV (Ovarian serous cystadenocarcinoma), PAAD (Pancreatic adenocarcinoma), PCPG (Pheochromocytoma and Paraganglioma), PRAD (Prostate adenocarcinoma), READ (Rectum adenocarcinoma), SARC (Sarcoma), SKCM (Skin Cutaneous Melanoma), STAD (Stomach adenocarcinoma), TGCT (Testicular Germ Cell Tumors), THCA (Thyroid carcinoma), UCEC (Uterine Corpus Endometrial Carcinoma), UCS (Uterine Carcinosarcoma).

We then examined whether LAGE3 and HENMT1 would be upregulated in patient-derived samples at the protein level. To this end, Tissue Microarrays (TMAs) were employed in combination with immunohistochemistry, analyzing a total of 72 samples (cores) from both tumor and normal tissues, for 12 different cancer types. The results show that both LAGE3 and HENMT1 are upregulated in specific tumor types at the protein level (**Figure 2.18A,B**), although the observed differences were not found to be statistically significant (**Figure 2.19**). Nevertheless, the results suggest that LAGE3 and HENMT1 have altered expression levels in specific cancer types also at the protein level.

Finally, we asked whether the expression levels of RMPs might be correlated with cancer prognosis. I identified 283 cases where RMP expression patterns are

significantly associated with patient survival (**Figure 2.18C**). For example, high NSUN5 expression levels in glioblastoma (GBM) are correlated with poor prognosis, in agreement with a recent study [312]. Similarly, this work revealed BUD23 expression to be correlated with cancer survival, in agreement with another recent study [313].

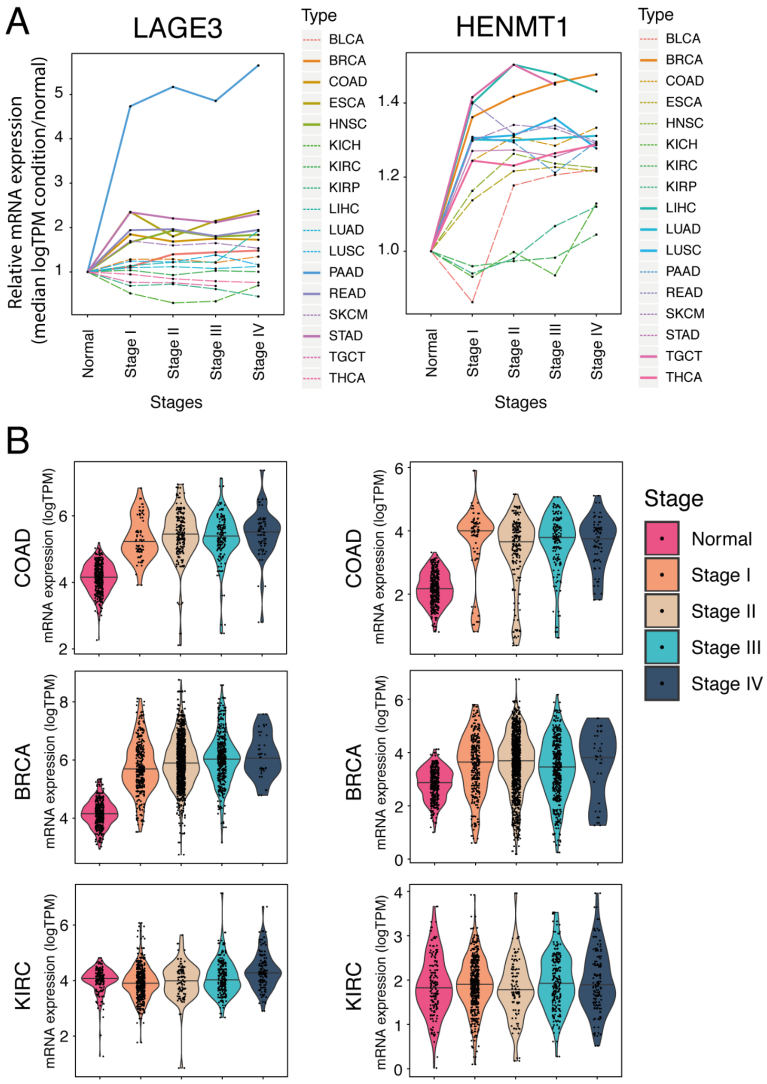


Figure 2.17 - Expression analysis of LAGE3 and HENMT1 across cancer types and stages

(A) Analysis of LAGE3 (left) and HENMT1 (right) mRNA expression levels across different cancer types and stages, relative to normal tissue expression levels. Median log(TPM) values from each cancer stage were normalised to the median log(TPM) of the normal tissue. Dashed lines depict cancer types in which LAGE3 or HENMT1 is not dysregulated, whereas full lines are used for cancer types in which the gene is dysregulated. **(B)** Violin plots of mRNA expression levels of LAGE3 and HENMT1 in individual cancer types (COAD, BRCA and KIRC) and across stages. Each dot represents the mRNA expression levels of a different individual.

Surprisingly, FTO expression levels are not significantly correlated with patient survival in LAML, despite this cancer type being used to test FTO inhibitors [314]. By contrast, LAGE3 expression levels were significantly correlated with patient survival in LAML (**Figure 2.18D**). Among all the RMP-cancer pairs studied, NSUN7 was identified as the top-ranked RMP in terms of prediction of lower grade glioma (LGG) patient survival (p=8e^{-24}); although its biological role still remains uncharacterised. Future research will be needed to functionally dissect the role that NSUN7 plays in glioma, as well as to decipher why its expression levels are highly predictive of patient survival.

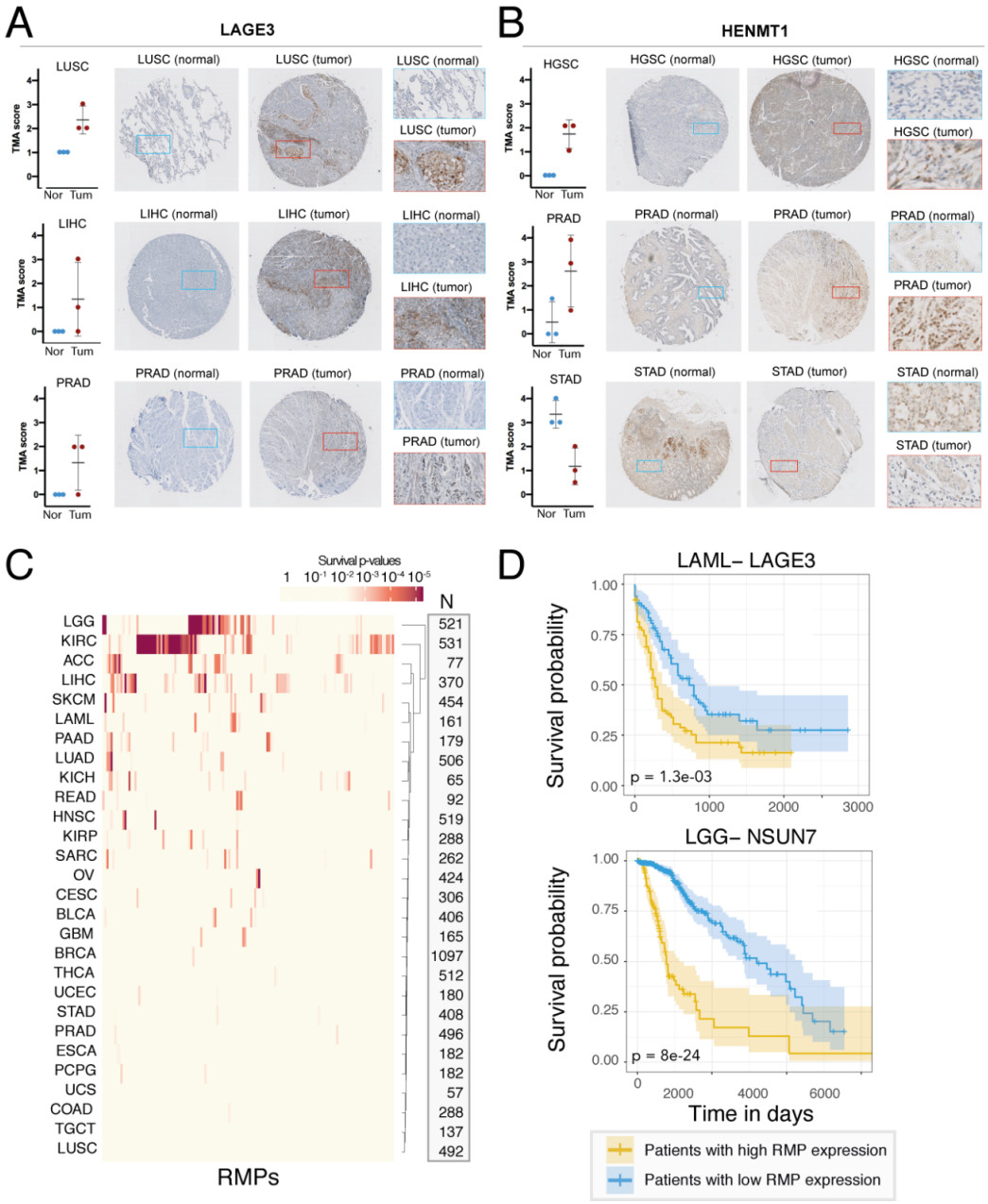


Figure 2.18. Immunohistochemical analysis and prognostic value of RMPs expression levels in different cancer types.

(A and B) Immunohistochemical analysis and images of normal and tumor LAGE-3 stained LUSC (Lung squamous cell carcinoma), LIHC (Liver hepatocellular carcinoma) and PRAD (Prostate adenocarcinoma) **(A)** and HENMT-1 stained HGSC (High-grade serous carcinoma), LUSC and STAD (Stomach adenocarcinoma) TMAs **(B)**. Representative cores and subsets are shown for each tissue and antibody, where the brown color indicates a specific staining of the antibody and blue represents the hematoxylin counterstain. Mean TMA score is plotted for each core, with three cores from different individuals per condition quantified. Two-sided Wilcoxon tests did not yield significant differences in any comparison, P-values of all tumor-normal comparisons for each cancer type and antibody are shown in Figure S13. **(C)** Heatmap of survival p-values of 146 RMPs across 28 cancer types. Survival p-values are calculated by comparing the prognosis of patients that express high (upper 50%) versus low (lower 50%) RMP levels. “N” column shows the number of patients included for the analysis of each cancer type. **(D)** Individual examples of survival plots where the expression levels of the RMP are predictive of cancer prognosis. P-values have been calculated by comparing the survival between patients expressing high levels (yellow, top 50%) versus low expression levels (blue, bottom 50%).

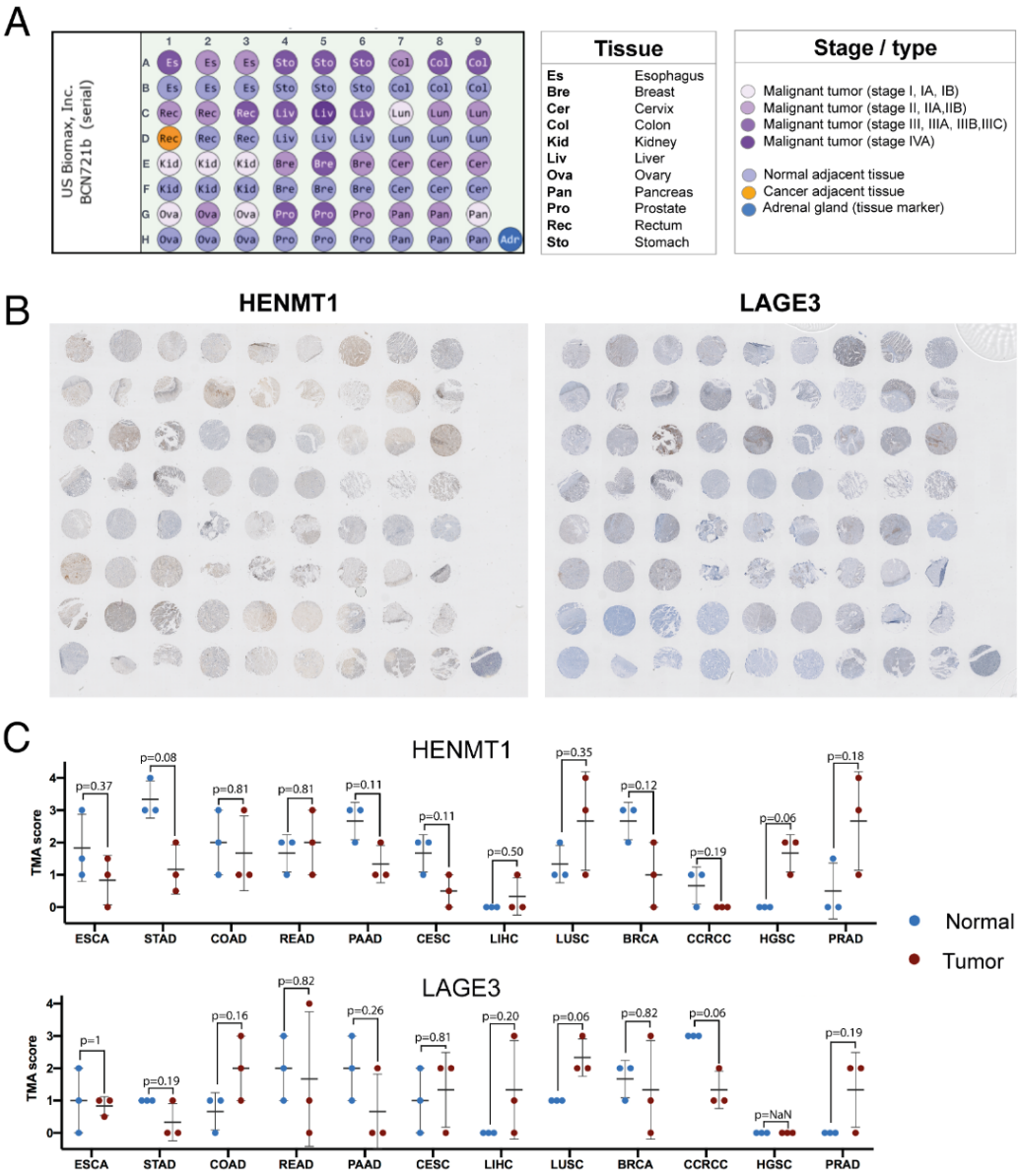


Figure 2.19 - Tissue microarray staining of HENMT1 and LAGE3

(A) Schematic representation of the layout of the Tissue microarray (TMA) slide used for immunohistochemical analysis. **(B)** Overall staining pattern of TMAs using HENMT1 and LAGE3 antibodies. Brown color indicates specific staining of the antibody, whereas blue represents the hematoxylin counterstain. **(C)** TMA scores (mean of two independent blinded scorers) obtained for each cancer type included in the slide. Mean TMA score for each core is depicted with a horizontal line, whereas scores given to each individual core are shown as dots. P-values were computed using two-sided Wilcoxon tests.

2.8. Materials and Methods

2.8.1. Compilation of human RNA modification-related proteins (RMPs)

An initial list of human methyltransferases, deaminases and pseudouridylases was obtained by merging the lists available in the MODOMICS database (http://modomics.genesilico.pl/) and from a recently published review [52]. These lists were further initially completed with candidate genes by the addition of annotated proteins on Uniprot [274]. For each of these proteins, hidden Markov model (HMM) profiles of the corresponding PFAM catalytic domains were retrieved (**Table 2.3**) by querying the PFAM database (https://pfam.xfam.org/). Each HMM profile was then used to query the human proteome using the hmmsearch function from HMMER software v.3.2.1 (http://hmmer.org/). Proteins above the default threshold were kept as candidate RMW proteins. Related information for each of these proteins (modification type, target RNA, localization) was extracted from Uniprot, as well as from relevant literature [274]. Additional tRNA writer proteins were gathered from a recent study matching tRNA modifications to their writers [315]. Readers, erasers and non-catalytic subunit proteins were obtained from annotated Uniprot genes as well as from published literature [316]**.** APOBEC3G and APOBEC3A were included in the analyses due to recent literature showing their deamination activity on RNA molecules *in vivo* in addition to acting on DNA [159,160].

2.8.2. Phylogenetic analysis

I first built a set of representative eukaryotic species, by choosing one species for each major phylogenetic clade for which complete proteomes were available. The final list of representative species consisted of 25 complete proteomes from UniProt [274], which included 23 eukaryal species, as well as 2 outgroups (1 bacteria and 1 archaea). For each proteome and RMW, I performed HMM-based searches, as described above. Candidate orthologs were manually curated to ensure that I did not miss any ortholog in the analysis. For each curated ortholog dataset, multiple sequence alignments were built using MAFFT with G-INS-1 method [317]. Alignment files were used to construct maximum-likelihood phylogenetic trees using IQ-Tree with bootstrapping (n=5000) [318]. Consensus trees were visualised using FigTree v 1.4.4 [319] and used to identify the duplication events.

PPFAM Domain	Search Number	PPFAM Domain	Search Number
A_deamin	1	Pox_MCEL	1
AdoMet_Mtase	1	PseudoU_synth_1	1
APOBEC_N	1	PseudoU_synth_2	1
Bin3	1	RrnaAD	1
dCMP_cyt_deam_1	1	SpoU_methylase	1
DNA_methylase	1	TRM	1
EMG1	1	TRM13	1
Fibrillarin	1	TrmO	1
FtsJ	1	tRNA_m1G_MT	1
Gcd10p	1	tRNA_U5-meth_tr	1
GCD14	1	TruB_N	1
LCM	1	TruD	1
MafB19-deam	1	TYW3	1
Met_10	1	UPF0020	1
Methyltr_RsmB-F	1	WD40	1
Methyltransf_4	1	zf-GRF	1
Methyltransf_10	1	WBS_methylT	1
Methyltransf_11	1	DREV	2
Methyltransf_15	1	Methyltransf_12	2
Methyltransf_8	1	Methyltransf_5	2
Methyltrn_RNA_3	1	MTS	2
MT-A70	1	Rsm22	2
PCIF1_WW	1		

Table 2.3 - PFAM domains used in phylogenetic analysis

2.8.3. Tissue specificity analysis

Human mRNA expression levels (TPM-Transcripts Per Kilobase Million) for each of the 146 human RMPs were downloaded from the Genotype Tissue Expression (GTEx) dataset [278], version v7, as well as from the Human Protein Atlas (HPA) [320]. Three GTEX tissues (whole blood, transformed lymphocytes and transformed fibroblasts) were discarded from downstream analyses, as these have been previously considered as outliers that can bias the analyses [278] or are not normal tissues of the human body. mRNA expression levels for adult mouse tissues (TPM, Transcripts Per Kilobase Million) were obtained from a study that is part of the ENCODE project [321]. For each dataset (HPA, GTEx, ENCODE), I log transformed the TPM values after the addition of a pseudocount. To determine which genes were tissue-specific, wecompared the expression levels of RMP in a given tissue to the median expression levels of RMPs across all tissues. I then calculated residuals (using *rlm* function), which is referred to as “tissue-specificity score” (TS), for each RMP to the regression line of each tissue. An RMP was considered tissue-specific if their TS was greater than 2.5 standard deviation (SD), as previously described [281], which, in a normal distribution of the standardised residuals, equals to the region outside of the 97.9 percentiles.

2.8.4. RNA Extraction from mice tissues and Quantitative Real-time PCR

Brain, liver, lung and testis tissues were collected from 20 week old C57BL/6J mice in triplicate. RNA was extracted from tissues using TRIzol™ Reagent (15596018, Thermo Fisher Scientific) and Chloroform (C2432, Vidra Foc) as per manufacturer's instructions, and precipitated with isopropanol (BP2618-500, Thermo Fisher Scientific) and Pellet Paint® Co-Precipitant (69049, Novagen). Samples were DNase treated with Turbo™ DNase (AM2238, Thermo Fisher Scientific) for 15 minutes at 37°C and cleaned-up using Agencourt RNAClean XP beads (A63987, Beckman Coulter) as per manufacturer's instructions. Quality of the extracted RNA was assessed using Nanodrop™ Spectrophotometer 2000. cDNA was synthesised using Superscript II™ (18064014, Thermo Fisher Scientific) following the manufacturer's instructions. Quantitative Real-Time PCR (qRT-PCR) was performed with Power SYBR™ Green PCR Mix (4367659, Thermo Fisher Scientific) using ViiA™ 7 Real-Time PCR System as per manufacturer's instructions. For each primer pair, three biological replicates with three technical replicate reactions were performed (total of 9 reactions per primer pair). METTL5, which is expressed stably among the four mouse tissues studied [322], was used for normalization purposes. Results were also analysed using GAPDH for

normalization purposes. qRT-PCR plots were built using GraphPad Prism 8. All oligonucleotides used for qRT-PCR can be found listed in **Table 2.4.**

Species	**Gene**	**Primer**	**Sequence 5' > 3'**
Mouse	ADAR3	Forward Primer	GTCTGGAGGGCTAAGCAGTC
Mouse	ADAR3	Reverse Primer	GCAAGGAAGGTTGACAGTATGC
Mouse	BUD23	Forward Primer	GCATCTCGTAGCCGGAGAC
Mouse	BUD23	Reverse Primer	CGTGAGTTGCGAACGTATTTCC
Mouse	NSUN7	Forward Primer	TGGACCCAACGAGTGAAAGG
Mouse	NSUN7	Reverse Primer	GTATTGGCGACTACATCCCCC
Mouse	RPUSD3	Forward Primer	CCCAGATGCCTTTGCACCT
Mouse	RPUSD3	Reverse Primer	GTCCGAGAGAAGTAAGGGGG
Mouse	TRDMT1	Forward Primer	TACCACCCAAGTTATTGCTGC
Mouse	TRDMT1	Reverse Primer	TCGTAAAGCACATGGACCTTC
Mouse	TRMT1L	Forward Primer	GATGCCCCTCTGATGCAGTTT
Mouse	TRMT1L	Reverse Primer	CGGACATCTCAACCCTGTCG
Mouse	METTL5	Forward Primer	AACTAGAGAGTCGCCTGCAAG
Mouse	METTL5	Reverse Primer	CTGCAACCGCTTTGTTTTCAA
Mouse	METTL1	Forward Primer	CAGACCACACACTGCGCTA
Mouse	METTL1	Reverse Primer	CATCCTTTGGATCATCATGGCTC
Mouse	HENMT1	Forward Primer	TGGCAGAAAGCATACCGTG
Mouse	HENMT1	Reverse Primer	ACCGTTGTTTGTATAATGGTGGT
Mouse	NSUN4	Forward Primer	TGGGATAGTGTGAGTGCTAAGC
Mouse	NSUN4	Reverse Primer	AAGCATCGAAGATTTGGGCTG
Mouse	GAPDH	Forward Primer	AGCCTCGTCCCGTAGACAAA
Mouse	GAPDH	Reverse Primer	AATCTCCACTTTGCCACTGC

Table 2.4 - Primers used for qPCR

2.8.5. RMP expression analysis across tissues in amniote species

mRNA expression levels of 12 amniote species (human, chimpanzee, bonobo, gorilla, orangutan, rhesus macaque, mouse, gray-short tailed opossum, platypus and chicken) were obtained from GSE30352 [323]. Normalised RPKM values of constitutive exons for both amniote and primate orthologs were used for downstream analyses. Heatmaps of the log transformed (with a pseudocount) and row (gene) z-scaled tissue-wide mRNA expression values were built using *complex heatmap* R package. PCA analysis was performed using *prcomp* function of R and plots of scores (amniote and primate tissues) and loadings (orthologous genes) were plotted for the first two principal components using *ggplot* R package.

2.8.6. Analysis of RMPs expression during spermatogenesis

Processed spermatogenesis data was extracted from GSE112393 [289]. Input data was used to perform k-means clustering of RMPs based on their expression profiles in different sperm cell populations. The optimal number of clusters was calculated by plotting the within groups sum of squares by number of clusters extracted using k-means function in R, following criteria used by Scree's test. Heatmaps were built using the *complex heatmap* R package. Violin plots were built using the *ggplot* R package. To assess the consistency of the results across diverse datasets, I analysed the RMP expression patterns from two additional mouse spermatogenesis studies [291,292]. For the first dataset, I used the same gene cluster groups and plotted the corresponding heatmap and violin plots using the ggplot R package (**Figure 2.10**). For the second dataset, I obtained the graphical representations for individual RMPs (**Figure 2.11**) from the interactive website accompanying the paper [292].

2.8.7. Immunohistochemistry

Testis and epididymis from 6-12 week old C57BL/6J mice were fixed overnight at 4ºC with neutral buffered formalin (HT501128-4L, Sigma-Aldrich) and embedded in paraffin. Paraffin-embedded tissue sections (3 µm in thickness) were air dried and further dried at 60ºC overnight. Immunohistochemistry was performed using The Discovery XT Ventana Platform (Roche). Antigen retrieval was performed with Discovery CC1 buffer (950-500, Roche). Primary antibodies rabbit polyclonal anti-NSUN2 (20854-1-AP, Proteintech), rabbit polyclonal anti-NSUN7 (PA5-54257, Thermo Fisher Scientific), rabbit polyclonal anti-HENMT1 (PA5-55866, Thermo Fisher Scientific), and rabbit polyclonal anti-METTL14 (HPA038002, HPA038002) were diluted

1:1000, 1:100, 1:150 and 1:2000 respectively with EnVision FLEX Antibody Diluent (K800621, Dako, Agilent) and incubated for 60 min. Secondary antibody OmniMap anti-rabbit HRP (760-4311) was incubated for 20 min. Detection of the labeling was performed using the ChromoMAP DAB (760-159, Roche). Sections were counterstained with hematoxylin (760-2021, Roche) and mounted with Dako Toluene-Free Mounting Medium (CS705, Agilent) using a Dako CoverStainer (Agilent). Specificity of staining was confirmed with a rabbit IgG, polyclonal Isotype Control (ab27478, Abcam). Brightfield images were acquired with a NanoZoomer-2.0 HT C9600 digital scanner (Hamamatsu) equipped with a 20X objective. All images were visualised with a gamma correction set at 1.8 in the image control panel of the NDP.view 2 U123888-01 software (Hamamatsu, Photonics, France). Mice samples were collected, prepared as paraffin blocks, sliced and stained at the IRB Histopathology Facility. Negative controls for each antibody were also included, which showed no staining (**Figure 2.20**). All IHC experiments were performed in biological triplicates.

2.8.8. Immunofluorescence

Testis and epididymis from 12 week old C57BL/6J mice were embedded in Tissue-Tek® O.C.T™ Compound (4583, Sakura) and 12 µm sagittal sections were mounted on SuperFrost™ microscope slides (12372098, Thermo Fisher Scientific). Tissue sections were defrosted, circled with a PAP pen (Z377821, Sigma-Aldrich), fixed in 4% PFA (28908, Thermo Fisher Scientific) for 10 minutes and permeabilised in 0.5% Triton-X 100 for 30 minutes (T8787, Sigma-Aldrich). Subsequently, sections were blocked in 5% BSA (A7906, Sigma-Aldrich) for 45 minutes at room temperature and incubated in primary antibody in 5% BSA overnight at 4°C. Primary antibodies were used at the following dilutions; 1:40 rabbit polyclonal anti-NSUN2 (20854-1-AP, Proteintech), 1:20 rabbit polyclonal anti-NSUN7 (PA5-55866, Thermo Fisher Scientific), 1:50 mouse monoclonal anti-DDX4 (ab27591, Abcam), 1:250 mouse monoclonal anti-Fibrillarin (38F3, Novus Biologicals) and 2 µg/mL IgG Isotype controls (G3A1 and 2791, Cell Signalling). Sections were incubated with 1:400 Alexa488-coupled anti-mouse (A-11001, Thermo Fisher Scientific) and Alexa555-coupled anti-rabbit (A-21429, Thermo Fisher Scientific) secondaries and counter-stained with 1:10,000 Hoechst 33342 (H3570, Life Technologies) for 2 hours at room temperature, then mounted with Fluoromount™ Aqueous Mounting Medium (F4680, Sigma Aldrich). Prepared slides were imaged on a Leica TCS SPE using a 63X NA1.4 oil objective. Three 1024x1024 representative regions of interest were imaged per testis (n=3) over a 3D stack (3-5 µm

depth with a z-step size of 1 µm), using a zoom factor of 2. All images were captured with a frame average of 4, with the exception of Hoechst which was imaged with a frame average of 2.

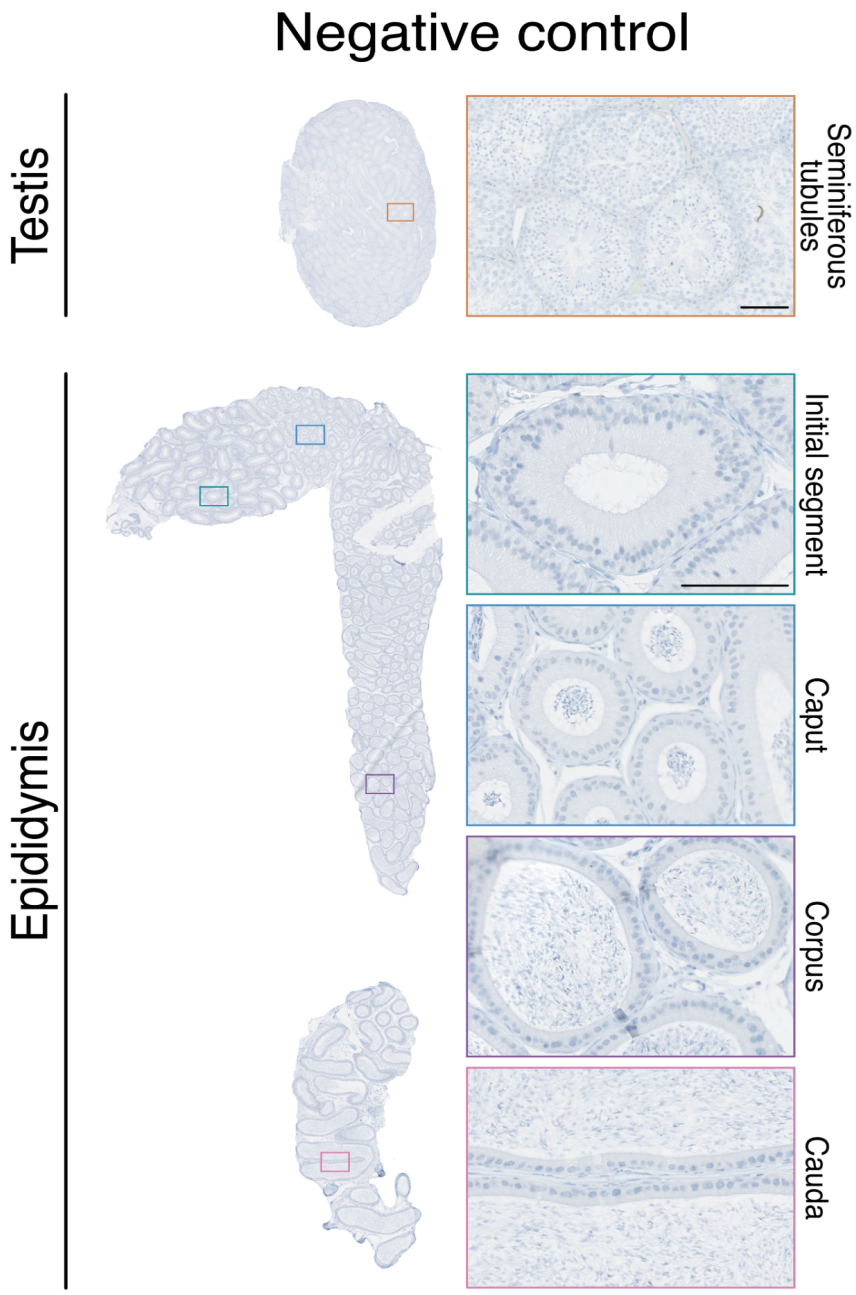


Figure 2.20. Immunohistochemical staining of mouse testis and epididymis using isotype control rabbit IgG antibody (negative control).

Brown color indicates antibody specific staining, whereas blue depicts hematoxylin counterstain.

2.8.9. Analysis of RMP expression in tumor-normal paired human datasets

TPM expression values were downloaded from the UCSC XENA Project, which contains the TCGA and GTEX RNA Seq data that is processed together to provide more reliable expression analysis with tumor and normal samples [307]. I discarded CHOL, THYM and DLBC tumor-normal tissue pairs due to lack of proper control of normal tissue (low number of patients) in these cancer types. Data was transformed into log2(TPM+1) for downstream analyses. For the log2(FC) analyses, I calculated the difference between median log2 expression levels between tumor and normal datasets, for each cancer type and RMP. For dysregulation analysis, I calculated the residuals (using *rlm* function in R) for all of the gene expression in a given tumor tissue and normal tissue pair, which has been previously termed as 'dysregulation score' (DS) [281]. I set the threshold of significance DS at 2.5 standard deviations (SD) as previously described [281], which, in a normal distribution of standardised residuals, equals to the region outside of the 97.9 percentiles. I then extracted the dysregulation scores of the RMPs and used it for further downstream analyses. For heatmap representations, dysregulation scores were scaled and centered, and the final heatmap was built using *complex heatmap* R package. Scatter plots of median log2 expression values for all genes in tumor-normal paired data were built using the *ggplot* R package, highlighting RMPs in black, significantly dysregulated RMPs in red (up-regulated in tumor) and blue (down-regulated in tumor), and non-RMP proteins were depicted in grey.

2.8.10. Survival Analyses

Survival phenotypes were downloaded from the XENA Platform, using the "TCGA TARGET GTEX" cohort [307]. In order to analyse the survival data, I first determined patients that have "high" (upper 50% relative to average expression) and "low" (lower 50% relative to average expression) expression of a specific gene, and matched these patients with their overall survival information. I then used the *survminer* R package to plot survival curves for each gene and every cancer type, as well as to extract the survival p-values. P-values were transformed by inversion and subsequent log-transformation with a pseudocount [log(1/p+1)]. Heatmap of the survival p-values was built using the complex *heatmap* R package. Transformed survival p-values were visualised using *ggplot*.

2.8.11. Tumor microarray immunohistochemistry and analysis

Multi organ tumor with adjacent normal tissue microarray slides with accompanying pathology grade, TNM (tumor, node and metastasis) classification and clinical stage information were purchased from US Biomax Inc (BCN721a). Each slide contains three malignant and three normal cores from 12 types of human organs (esophagus, stomach, colon, rectum, liver, lung, kidney, breast, cervix, ovary, prostate and pancreas), each core taken from different individuals. TMAs were stained at IRB Histopathology Facility. Primary antibodies rabbit polyclonal anti-HENMT1 (PA5-55866, Thermo Fisher Scientific) and rabbit polyclonal anti-LAGE3 (HPA036123, Sigma-Aldrich) were diluted to 1:50. Secondary antibody OmniMap anti-rabbit HRP (760-4311) was incubated for 20 min. Detection of the labelling was performed using the ChromoMAP DAB (760-159, Roche). For scoring of tissue microarrays, each core was given a score from 0 to 4 based on the proportion of positively stained cells: 0 represents <2% of cells staining positive, 1 represents 2-25%, 2 represents 26-50%, 3 represents 51-75% and 4 represents 76-100% staining of cells [324]. Two blinded independent people scored the stainings, following the guidelines described above. Both scorers were blind to both the antibody and tissue type. The scores from each scorer were averaged to obtain the final score per core. A two-sided Wilcoxon test was used to assess significance.

2.9. Discussion

Over the past decade, systematic efforts to detect and map RNA modifications have boosted the new field of epitranscriptomic research. Many proteins are involved in the writing, reading and erasing of RNA modifications but their roles in tumorigenesis and potential as therapeutic targets remain largely uncharacterised. To bridge this gap, here we have compiled a list of 146 human RNA modification-related proteins (RMPs) (**Table S2.1**), and have analysed the evolutionary history and gene expression patterns of 90 RMPs across 32 mammalian tissues, 10 species, 5 cell types and 13,358 tumor-normal paired cancer samples.

Through this analysis, a large number of duplication events were identified in multiple RNA modification families (**Figure 2.1**), often accompanied by the acquisition of restricted tissue expression patterns and/or change in its RNA target specificity. Therefore, RMP gene duplication is a strategy to acquire novel functions, and is typically achieved by altering the expression patterns and/or RNA target selectivity of the paralog proteins, in agreement with other analyses of gene evolution [325]. The majority of tissue-restricted RMPs are in fact testis-enriched (**Figure 2.2**), suggesting

that certain RMPs might play a pivotal role in sperm formation and maturation. Moreover, deletion of testis-enriched genes such as NSUN7, ADAD1 or HENMT1 leads to male sterility [212,290,296,298].

At the beginning of spermatid elongation, nuclear condensation starts, and consequently the transcriptional machinery is shut down. Therefore, to provide proteins for the following maturation steps of sperm assembly, mRNAs have to be premade in spermatocytes and round spermatids, before nuclear condensation happens, and translationally repressed until needed [283–286]. Chemical RNA modifications provide an ideal platform to achieve the fine regulation that is required upon transcriptional shutdown, determining which RNAs are expressed, repressed, or undergo decay [13]. In this regard, previous work has shown that METTL3/METTL14 mediated m^6A modification is dynamically regulated in spermatogenesis [288]. Similarly, piRNA molecules in germ line cells are tightly regulated by HENMT1, via 2'-O-methylation of their 3'ends [298]. Here we show that a vast proportion of RMPs are dynamically regulated during spermatogenesis as well as during sperm maturation in the epididymis, and as such, may be involved in the regulation and decay of specific transcripts that occur during sperm formation and maturation (**Figure 2.8**).

Recent studies have shown that specific RNA modifications are essential for the transmission of paternal diet-induced phenotypes intergenerationally [50]. I identified two RMPs (TRDMT1 and METTL1) whose expression is significantly enriched in epididymis (**Figure 2.5**), one of which (TRDMT1) was recently shown to be involved in the transmission of diet-induced paternal phenotypes across generations [50]. Whether METTL1 plays a role in intergenerational inheritance is yet to be deciphered; however, recent insights showing its role in miRNA maturation [326] suggest that this enzyme might be playing a role in miRNA-acquired inheritance of information.

In the last few years, several studies have placed RNA modifications in the forefront of cancer research [72,302,304,327,328], mostly focused on the machinery responsible for writing and erasing m^6A modifications. For many years, FTO was thought to be of special interest due to its association with obesity [329]. However, later studies proved this genome-wide association to be false [330], and that the single nucleotide variant present in the FTO intron was in fact associated with the activity of neighboring genes [330].

Nonetheless FTO kept receiving special attention due to its perceived activity as an eraser of m^6A [57], the most frequent type of RNA modifications present in mRNAs. However, this is now thought to be incorrect, as later studies showed that FTO is in fact an eraser of N6,2'O-methyladenosine (m^6Am), which is much less abundant in

mRNAs [58,331]. Similarly, FTO has been proposed to constitute a promising target for antitumor therapies [314,328,332]. While FTO has been shown to play an important role in leukaemia [332], it is possible that additinal RMPs such as HENMT1, which is drastically dysregulated in this cancer type, might constitute a better drug target to inhibit leukemogenesis (**Figure 2.16**).

Here I show that the expression of 40 RMPs is significantly altered in tumor samples, relative to their matched normal samples (**Table 2.2** and **Figure 2.16**). Moreover, I identify two enzymes, LAGE3 and HENMT1, as the top recurrently up-regulated RMPs across cancer types. Surprisingly, these proteins have so far received little attention in cancer research studies. LAGE3 mutations are known to cause multiple human diseases, including nephrotic syndrome and microcephaly [333]; however, its role in tumorigenesis and cancer progression is yet to be determined. The upregulation of LAGE3 and HENMT1 was validated using Tissue Microarrays (TMAs) across a battery of 12 cancer types (**Figure 2.18**). While several cancer types where LAGE3 and HENMT1 were consistently upregulated identified, the variability among cancer grades across the tumor cores, together with the low number of cores per tumor type (n=3) led to insufficient statistical power to identify significant expression changes. Future work will be needed to decipher the biological role of LAGE3 and HENMT1 in cancer, as well as its potential use as a target for diagnostic and prognostic purposes.

2.10. Supplementary Data

Table S2.1 - List of human RNA modification–related proteins (RMPs) used in this study

Symbol (HGNC)	Category	Modification Symbol	Target
ADAD1	Probable RNA Deaminase	A-I	NA
ADAD2	Probable RNA Deaminase	A-I	NA
ADAR	RNA Deaminase	A-I	mRNA,ncRNA
ADARB1	RNA Deaminase	A-I	mRNA,ncRNA
ADARB2	RNA Deaminase (Possibly inactive)	A-I	NA
ADAT1	RNA Deaminase	A-I	tRNA(37)
ADAT2	RNA Deaminase	A-I	tRNA(34)
ADAT3	RNA Deaminase	A-I	tRNA(34)
ALKBH1	RNA Methyl Eraser	m1A, m5C	tRNA, mt-tRNA
ALKBH2	RNA Methyl Eraser	m1A, m3C	NA
ALKBH3	RNA Methyl Eraser	m1A, m3C	mRNA, tRNA
ALKBH5	RNA Methyl Eraser	m6A	mRNA
ALKBH8	RNA Methylase	mchm5U	tRNA
APOBEC1	RNA Deaminase	C-U	mRNA, HPV viralRNA
APOBEC3A	RNA Deaminase	C-U	cellular RNA, ssDNA (HIV)
APOBEC3G	RNA Deaminase	C-U	mRNA, ssDNA (HIV), HIV viral RNA
BCDIN3D	RNA Methylase	mm(pN)	tRNA, miRNA
BUD23	RNA Methylase	m7G	rRNA
CBLL1	RNA Methylase non-catalytic subunit	m6A	mRNA
CDK5RAP1	Other tRNA Writer	ms2i6A	tRNA

CDKAL1	Other tRNA Writer	ms2t6A	tRNA
CIAO1	Other tRNA Writer	s2U, mcm5S2U	tRNA
CMTR1	RNA Methylase	2-O-M (Cap1)	mRNA, snRNA
CMTR2	RNA Methylase	2-O-M (Cap2)	mRNA, snoRNA
CTU1	Other tRNA Writer	s2U, mcm5S2U	tRNA
CTU2	Other tRNA Writer	s2U, mcm5S2U	tRNA
DIMT1	RNA Methylase	m6, 6A	rRNA(A1779-A1780)
DKC1	RNA Pseudouridylase	Ψ	rRNA
DUS2	Other tRNA Writer	D	tRNA
EIF3A	RNA Methyl Reader	m6A	mRNA
ELP1	Other tRNA Writer	cm5U, ncm5U, mcm5U, mcm5S2U	tRNA
ELP3	Other tRNA Writer	cm5U, ncm5U, mcm5U, mcm5S2U	tRNA
ELP4	Other tRNA Writer	cm5U, ncm5U, mcm5U, mcm5S2U	tRNA
ELP5	Other tRNA Writer	cm5U, ncm5U, mcm5U, mcm5S2U	tRNA
EMG1	RNA Methylase	m1Y	rRNA(18S rRNA- Y1248)
FBL	RNA Methylase	Xm	rRNA
FBLL1	Probable RNA Methylase	Xm Probable	NA
FTO	RNA Methyl Eraser	m6A(m) (Cap), m6A	mRNA,tRNA, snRNA
FTSJ1	RNA Methylase	Xm	tRNA(C32-G34)
FTSJ3	RNA Methylase	Xm	rRNA, HIV RNA
GTPBP3	Other tRNA Writer	tm5U	tRNA
HENMT1	RNA Methylase	Xm (3'end)	piRNA
HNRNPA2B1	RNA Methyl Reader	m6A	
HNRNPC	RNA Methyl Reader	m6A	

HSD17B10	Other tRNA Writer	m1G,m1A	tRNA
ISCU	Other tRNA Writer	s2U, mcm5S2U	tRNA
JMJD6	RNA Methyl Eraser		
LAGE3	Other tRNA Writer	t6A	tRNA
LCMT2	Probable RNA Methylase	NA	Probable tRNA
MEPCE	RNA Methylase	mm(pN)	snRNA
METTL1	RNA Methylase	m7G	tRNA(G46)
METTL14	RNA Methylase non-catalytic subunit	m6A	mRNA
METTL15	Probable RNA Methylase	m4C	16s rRNA in E.coli homolog
METTL16	RNA Methylase	m6A	ncRNA,mRNA, U6 snRNA
METTL17	Candidate RNA Methylase	Probable rRNA Methylase	rRNA in yeast homolog
METTL27	Candidate RNA Methylase	RNA Met ?	NA
METTL2A	RNA Methylase	m3C	tRNA(Thr-Arg)
METTL2B	RNA Methylase	m3C	tRNA(Thr-Arg)
METTL3	RNA Methylase	m6A	mRNA
METTL4	Probable RNA Methylaase	m6A (?)	NA
METTL5	Candidate RNA Methylase	Probable m2G	small rRNA in prokaryotic homolog
METTL6	RNA Methylase	m3C	tRNA(Ser)
METTL8	RNA Methylase	m3C	mRNA
METTL9	Candidate RNA Methylase	RNA Met ?	NA
MOCS3	Other tRNA Writer	s2U, mcm5S2U	tRNA
MPST	Other tRNA Writer	s2U, mcm5S2U	tRNA
MRM1	RNA Methylase	Gm	mt-rRNA (Gm1145)
MRM2	RNA Methylase	Um	mt-rRNA
MRM3	RNA Methylase	Gm	mt-rRNA(Gm1370)

MTO1	Other tRNA Writer	tm5U	tRNA
NAT10	Other tRNA Writer	ac4C	tRNA
NFS1	Other tRNA Writer	s2U, mcm5S2U	tRNA
NOP2	RNA Methylase	m5c	rRNA(28SRNA-c4447)
NSUN2	RNA Methylase	m5c	tRNA(Leu) (C34,48), tRNA(Gly) (C48,C49,C50)
NSUN3	RNA Methylase	m5c	mt-tRNA(C34)
NSUN4	RNA Methylase	m5c	mt-rRNA
NSUN5	RNA Methylase	m5c	rRNA(28SRNA-c3782)
NSUN6	Probable RNA Methylase	m5c	tRNA
NSUN7	Probable RNA Methylase	m5c	eRNA
NUBP1	Other tRNA Writer	s2U, mcm5S2U	tRNA
OSGEP	Other tRNA Writer	t6A	tRNA
OSGEPL1	Other tRNA Writer	t6A	tRNA
PCIF1	RNA Methylase	m6A(m) (Cap)	mRNA
PUS1	RNA Pseudouridylase	Ψ	tRNA\mttRNA(U27,U28,U30)
PUS10	RNA Pseudouridylase	Ψ	tRNA(U54,U55)
PUS3	RNA Pseudouridylase	Ψ	tRNA(U38-U39)
PUS7	RNA Pseudouridylase	Ψ	tRNA, tRFs, mRNA
PUS7L	Probable RNA Pseudouridylase	Ψ	NA
PUSL1	Probable RNA Pseudouridylase	Ψ	NA
QTRT1	Other tRNA Writer	Q	tRNA
RBM15	RNA Methylase non-catalytic subunit	m6A	mRNA
RBM15B	RNA Methylase non-catalytic subunit	m6A	mRNA
RNMT	RNA Methylase	m7G	mRNA

RPUSD1	Probable RNA Pseudouridylase	Ψ	NA
RPUSD2	Probable RNA Pseudouridylase	Ψ	NA
RPUSD3	RNA Pseudouridylase	Ψ	mt-mRNA
RPUSD4	RNA Pseudouridylase	Ψ	mt-rRNA(16s rRNA-U1397), mt-tRNA(U39)
RRP8	RNA Methylase	m1A	rRNA
SERGEF	Other tRNA Writer	s2U, mcm5S2U	tRNA
SPOUT1	Probable RNA Methylase	NA	miRNA
TARBP1	RNA Methylase	Gm	HIV RNA
TFB1M	RNA Methylase	m6, 6A	mt-rRNA
TFB2M	RNA Methylase	m6, 6A	mt-rRNA
TGS1	RNA Methylase	m2,2,7G	snoRNA, snRNA
THG1L	Other tRNA Writer	xG	tRNA
THUMPD1	Other tRNA Writer	ac4C	tRNA
THUMPD2	RNA Methylase	m2G	tRNA (Needs experimental validation)
THUMPD3	RNA Methylase	m2G	mt-tRNA (Needs experimental validation)
TP53RK	Other tRNA Writer	t6A	tRNA
TPRKB	Other tRNA Writer	t6A	tRNA
TRDMT1	RNA Methylase	m5C	tRNA(Asp)(C38)
TRIT1	Other tRNA Writer	i6A	tRNA
TRMO	RNA Methylase	m6t6A	tRNA (A37)(tRNA(Ser)(GCU))
TRMT1	RNA Methylase	m2,2G	tRNA, mt-tRNA
TRMT10A	RNA Methylase	m1G	tRNA
TRMT10B	RNA Methylase	m1G	tRNA
TRMT10C	RNA Methylase	m1G:m1A	mt-tRNA,mt-mRNA

TRMT11	RNA Methylase	m2G	tRNA(G10 in yeast homolog)
TRMT112	RNA Methylase non-catalytic subunit	m7G	rRNA
TRMT12	RNA Methylase	o2yW	tRNA (imG-14)
TRMT13	RNA Methylase	Cm, Am	tRNA (in rice homolog)
TRMT1L	Probable RNA Methylase	m2,2G	tRNA (?)
TRMT2A	RNA Methylase	m5u	tRNA(U54 in yeast homolog)
TRMT2B	RNA Methylase	m5u	tRNA(U54 in yeast homolog)
TRMT44	RNA Methylase	Um	Um44 in tRNA(Ser) in yeast homolog
TRMT5	RNA Methylase	m1G	tRNA(G37)
TRMT6	RNA Methylase non-catalytic subunit	m1A	tRNA,mRNA
TRMT61A	RNA Methylase	m1A	tRNA,mRNA
TRMT61B	RNA Methylase	m1A	mt-tRNA (58),mt-mRNA,mt-rRNA
TRMT9B	Probable RNA Methylase	mchm5U	tRNA
TRMU	Other tRNA Writer	tm5S2U	tRNA
TRUB1	RNA Pseudouridylase	Ψ	mRNA
TRUB2	RNA Pseudouridylase	Ψ	mt-mRNA
TYW3	RNA Methylase	yW	tRNA(Phe) in yeast homolog
VIRMA	RNA Methylase non-catalytic subunit	m6A	mRNA
WDR4	RNA Methylase non-catalytic subunit	m7G	tRNA(G46)
WDR6	Other tRNA Writer	Cm, Gm,f5Cm, hm5Cm	tRNA
WTAP	RNA Methylase non-catalytic subunit	m6A	mRNA
YRDC	Other tRNA Writer	t6A	tRNA
YTHDC1	RNA Methyl Reader	m6A	

YTHDC2	RNA Methyl Reader	m6A	
YTHDF1	RNA Methyl Reader	m6A	
YTHDF2	RNA Methyl Reader	m6A	
YTHDF3	RNA Methyl Reader	m6A	
ZC3H13	RNA Methylase non-catalytic subunit	m6A	mRNA
ZCCHC4	RNA Methylase	m6A	rRNA (28SrRNA-a4220)

3. Accurate detection of m6A RNA modifications in native RNA sequences

This chapter contains material described in the publication published in Nature Communications (Liu and Begik et al 2019) [255].

I performed the wetlab protocol optimisations and experiments with the assistance of other authors. I also contributed to drafting the manuscript. Following is the specific list of contributions that I made in this study:

- Optimised in-vitro transcription, capping, and polyadenylation experiments that are necessary to synthesise modified RNA molecules.
- Optimised direct RNA library preparation, which is necessary to produce data for the analyses.
- Prepared direct RNA libraries of *in-vitro* constructs and yeast mRNA, which is the source of all the data produced for this paper.

The co-first author of the paper Huanle Liu performed all the bioinformatics analyses.

Morghan C. Lucas contributed to the wetlab protocol optimisations.

Jose Miguel Ramirez performed base-caller comparison analyses.

David Wiener performed yeast culturing and mRNA purification.

Eva Maria Novoa supervised the project, with contributions of Schraga Schwartz, Christopher E. Mason, John S. Mattick, and Martin A. Smith.

3.1. Introduction

ONT direct RNA sequencing makes it possible to sequence native RNA molecules. RNA modifications in these native RNA molecules are known to cause disruptions in the pore current that can be detected upon comparison of raw current intensities – also known as 'squiggles'– [251,254]. However, current efforts have not yet yielded an efficient and accurate RNA modification detection algorithm, largely due to the challenges in the alignment and re-squiggling of RNA current intensities [334,335].

As an alternative strategy, in this chapter, we hypothesised that the current intensity changes caused by the presence of RNA modifications may lead to increased 'errors' and decreased qualities from the output of base-calling algorithms that do not model base modifications (**Figure 3.1A**). Indeed, here we found that base-calling 'errors' can accurately identify m^6A RNA modifications in native RNA sequences, and propose a novel algorithm, *EpiNano* (github.com/enovoa/EpiNano), which can be used to identify m^6A RNA modifications from RNA reads with an overall accuracy of ~90%. These results provide a proof of concept for the use of base-called features to identify RNA modifications using direct RNA sequencing, and open new avenues to explore additional RNA modifications in the future.

3.2. Optimisation of the wet-lab protocols

In order to produce *in-vitro* transcribed RNAs, I first transformed the E. coli bacteria with plasmids containing sequences to be *in-vitro* transcribed (see Methods). Next, I isolated the plasmids from the bacteria cells, digested them with the restriction enzymes and cleaned them up using phenol-chloroform extraction. I verified the digestion with agarose gel electrophoresis (**Figure 3.1**).

Linearised DNA was then used as a template for the *in-vitro* transcription (IVT), either in the presence of ATP or N^6-Methyladenosine-5'-Triphosphate(m^6ATP). This resulted in IVT products that are fully modified with m^6A on their Adenine positions. After cleanup, *in-vitro* transcribed RNA was run on TapeStation in order to ensure the full-length production (**Figure 3.2**). Then a 5'cap was added to the in-vitro transcribed RNAs to ensure their stability during manipulation. Furthermore, poly(A) tailing reaction ensured that the RNAs can be used for direct RNA sequencing library preparation. Poly(A) tail addition was confirmed on Tapestation **(Figure 3.2)**.

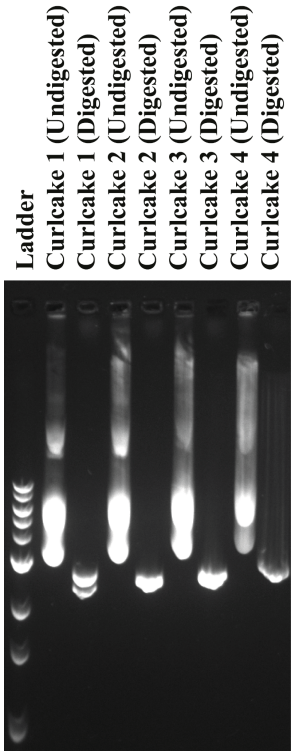


Figure 3.1 - Agarose gel electrophoresis of the plasmid DNA.
Curlcake 1-4 plasmids that contain the sequence to be in-vitro transcribed were digested and run on agarose gel after cleanup. Distinct bands in the digested lanes illustrate a complete digestion of the plasmid.

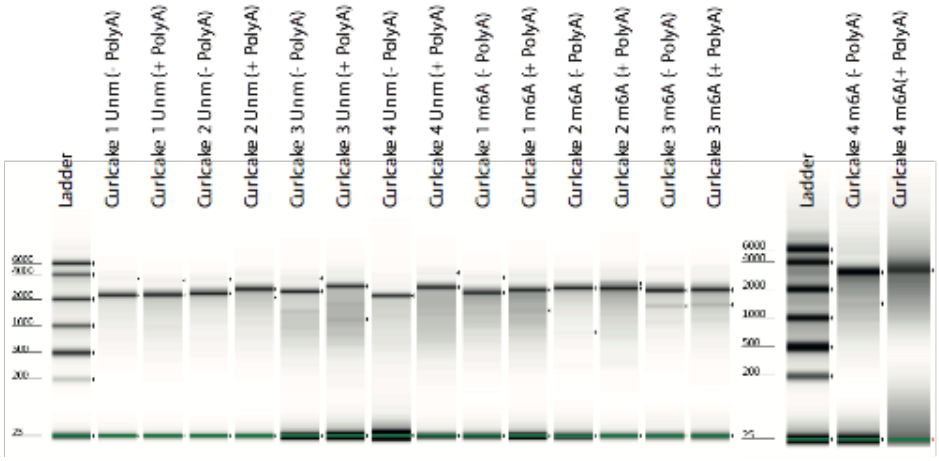


Figure 3.2 - TapeStation output of the quality and quantity of the m6A-modified and unmodified *in vitro* transcription products.

Curlcake 1-4 in-vitro transcribed RNAs before and after poly(A) tailing.

3.3. RNA modifications cause altered base-calling features in direct RNA sequencing reads

Previous work has shown that ONT raw current intensity signals, known as 'squiggles', can be subdivided into 'events', which correspond to consecutive 5-mer sequences shifted one nucleotide at a time (e.g., in the sequence AGACAAU, the corresponding 5-mer 'events' would be AGACA, GACAA, and ACAAU)[336–339]. Therefore, to systematically identify the current intensity changes caused by the presence of a given RNA modification, perturbations of the current intensity signals should be measured and analysed for each possible 5-mer (n=1024).

To this end, a set of synthetic sequences that comprised all possible 5-mers was designed (median occurrence of each 5-mer=10), while minimizing the RNA secondary structure (see Methods). I then employed direct RNA sequencing to characterise the differences of *in vitro*-transcribed constructs that incorporated either m^6A instead of adenine, or unmodified ribonucleotides ('unm') (**Figure 3.3A**). Comparison of the two datasets revealed that base-called m^6A-modified reads were significantly enriched in mismatches compared to their unmodified counterparts (**Figure 3.3B and 3.3C**), and that these 'errors' were mainly, but not exclusively, located in adenine positions. In addition to mismatch frequency, other metrics including per-base quality, insertion frequency, deletion frequency and current intensity, were significantly altered (**Figure 3.3C** and see also **Figure 3.5**). Moreover, these 'errors' were highly reproducible in independent biological replicates with respect to mismatch frequency, deletion frequency, per-base quality and current intensity (**Figures 3.3D-G**). By contrast, insertion frequencies were not reproducible across biological replicates, suggesting that this feature is likely unrelated to the presence of RNA modifications, and thus was not further considered in downstream analyses.

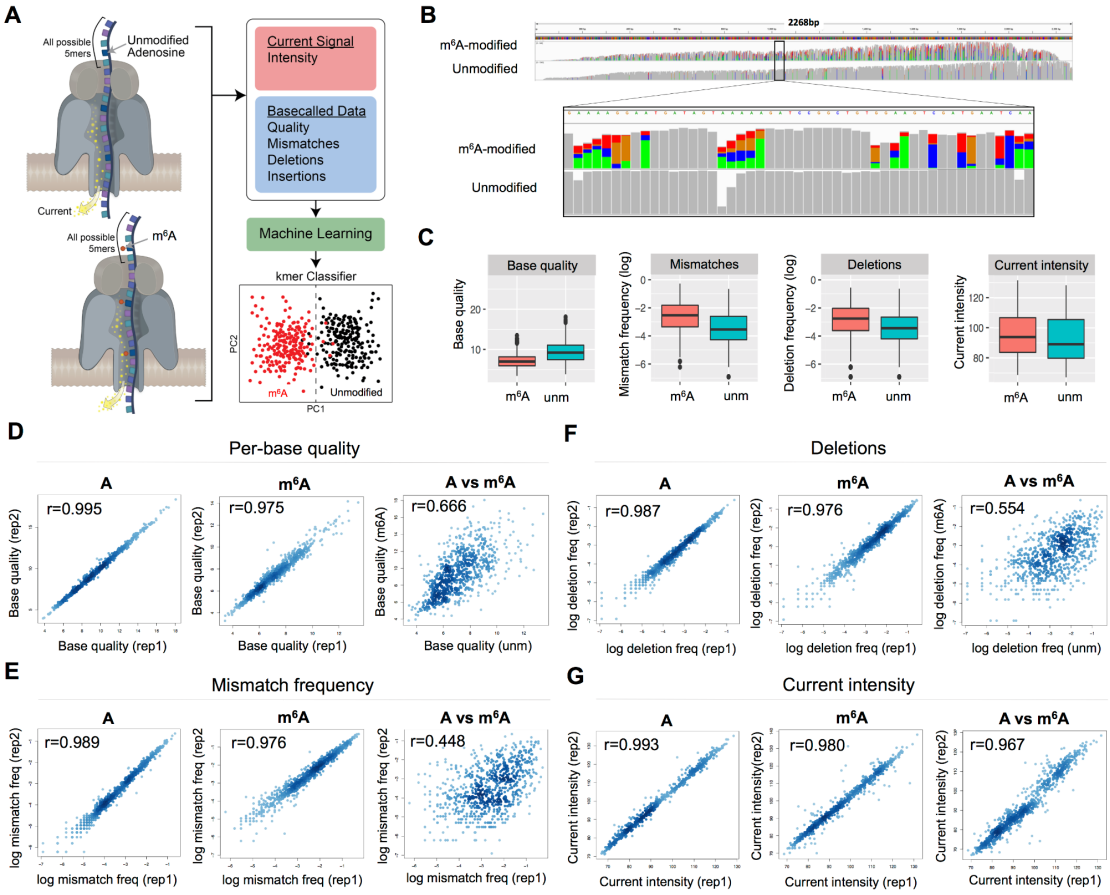


Figure 3.3 - Base-calling 'errors' can be used as a proxy to identify RNA modifications in direct RNA sequencing reads.

(**A**) Schematic overview of the strategy used in this work to train and test an m^{6}A RNA base-calling algorithm **(B)** IGV snapshot of one of the four transcripts used in this work. In the upper panel, *in-vitro* transcribed products containing m^{6}A have been mapped, whereas in the lower panel the unmodified counterpart is shown. Nucleotides with mismatch frequencies greater than 0.05 have been coloured. **(C)** Comparison of m^{6}A and A positions, at the level of per-base quality scores (left panel), mismatch frequencies (middle left panel), deletion frequency (middle right panel) and mean current intensity (right panel). All possible k-mers (computed as a sliding window along the transcripts) have been included for these comparisons (n=9,974) (**D, E, F, G**) Replicability of each individual feature - base quality (D), deletion frequency (E), mismatch frequency (F) and current intensity (G)- across biological replicates, for both unmodified ('A') and m^{6}A-modified ('m^{6}A') datasets. Comparison of unmodified and m^{6}A-modified ('A vs m^{6}A') is also shown for each feature. Correlation values shown correspond to Spearman's rho. Error bars indicate s.d.

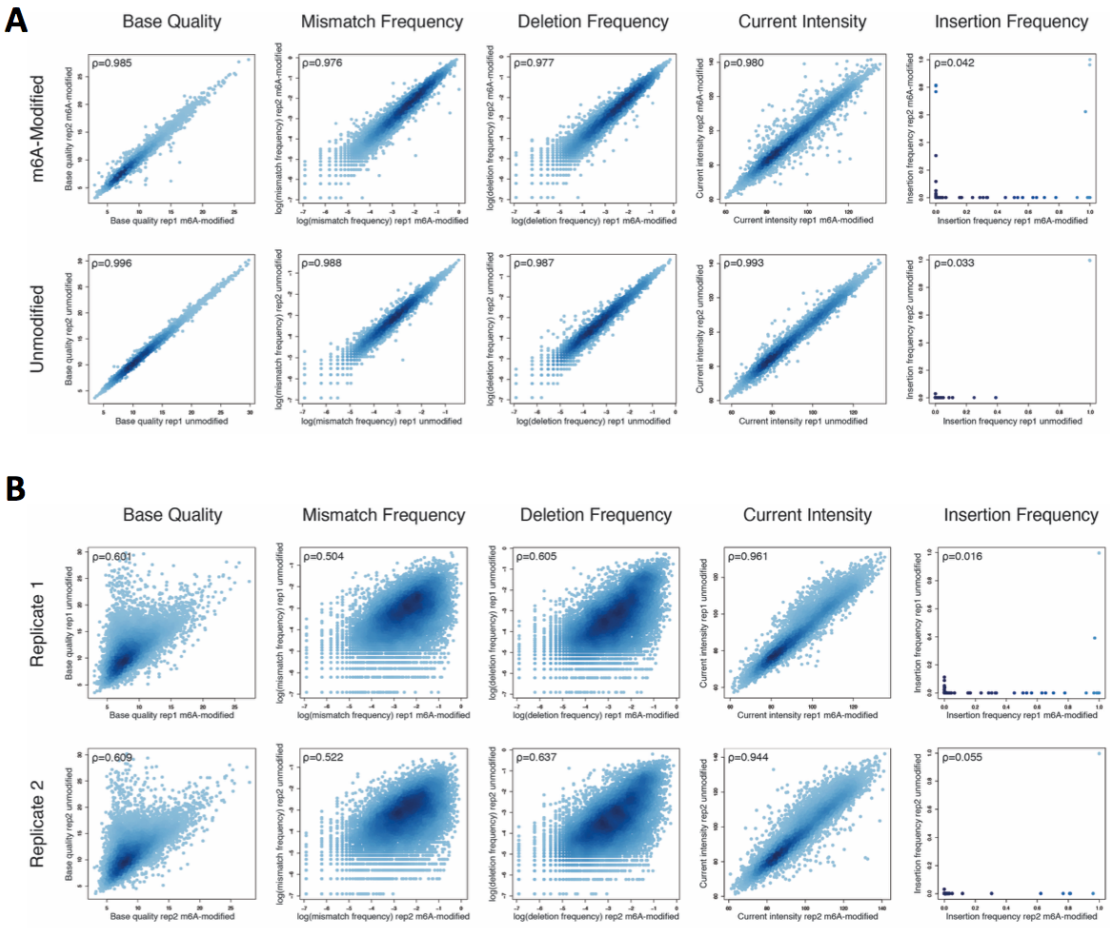


Figure 3.4 - **Replicability of the features extracted across replicates.**

(**A**) Comparison of features between m^{6}A-modified datasets (replicate 1 and replicate 2) and m^{6}A-unmodified datasets (replicate 1 and replicate 2). Each dot corresponds to a different nucleotide of the synthetic constructs (n=9978) (**B**) Comparison of features and across datasets, comparing m^{6}A-modified and unmodified datasets of replicate 1 (upper panels) and of replicate 2 (lower panels).

3.4. Base-calling 'errors' can accurately predict m^{6}A RNA modifications in direct RNA sequencing reads

These observed differences were then examined for their sufficiency to accurately classify a given site into "modified" or "unmodified". For this aim, the analysis focused on 5-mers that matched the known m^{6}A motif RRACH, as these would be the most relevant in which to detect m^{6}A modifications. To reveal whether

the features from m^{6}A-modified RRACH k-mers were distinct from unmodified RRACH k-mers, base-called features (base quality, mismatch frequency and deletion frequency) were compiled for each position of the k-mer (-2, -1, 0, +1, +2) (**Figure 3.5A,** see also **Figure 3.6**), and performed Principal Component Analysis (PCA) of the features, finding that the two populations (m^{6}A-modified and unmodified RRACH k-mers) were largely non-overlapping (**Figure 3.5B**). As a control, same analysis was performed in k-mers with identical sequence context, but centered in C, G, or U (instead of A), finding that no differences could be observed between these populations (**Figure 3.5C**), suggesting that the observed differences are m^{6}A-specific, and not dataset-specific.

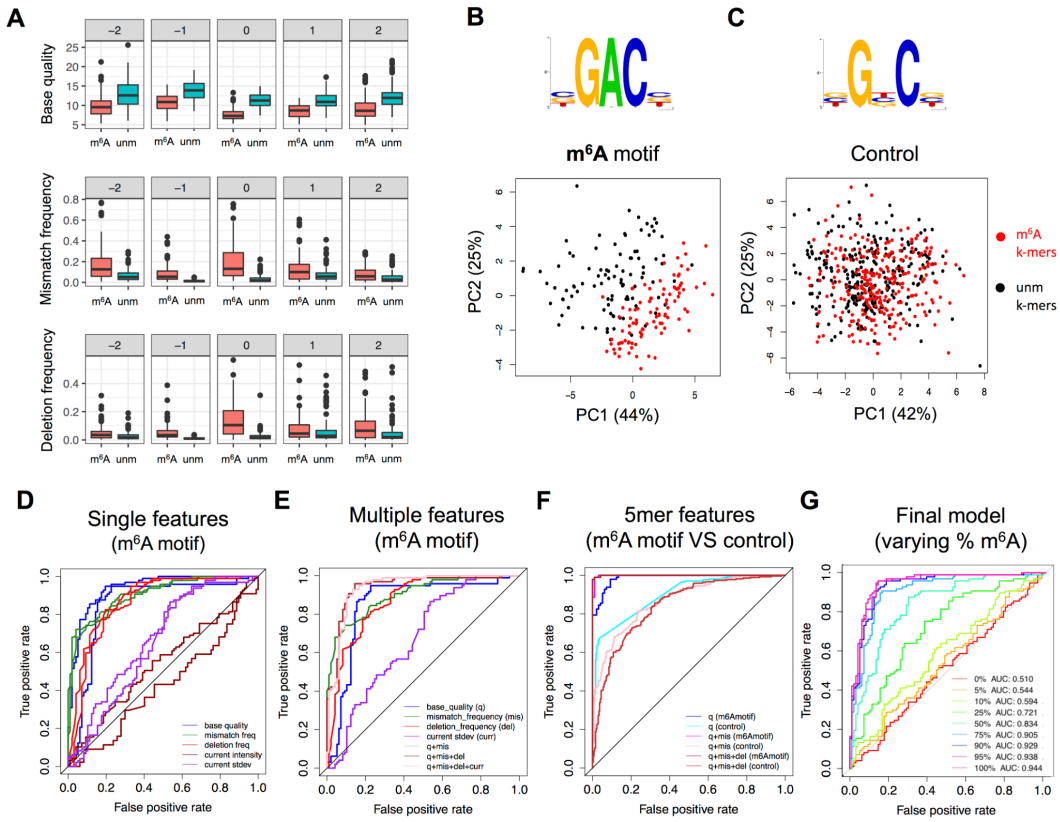


Figure 3.5 - Base-calling 'errors' alone can accurately identify m^{6}A RNA modifications.

(**A**) Base-called features (base quality, insertion frequency and deletion frequency) of m^{6}A motif 5-mers, and for each position of the 5-mer, are shown. The features of the m^{6}A-modified transcripts ('m^{6}A') are shown in red, whereas the features of the unmodified transcripts ('unm') are shown in blue. **(B, C**) Principal component analysis (PCA) scores the plot of the two first principal components, using 15 features (base quality, mismatch frequency, deletion frequency, for each of the 5 positions of the k-

mer) as input. The logos of the k-mers used in the m^{6}A-motif RRACH set (left) and control set (right) are also shown. Each dot represents a specific k-mer in the synthetic sequence, and has been coloured depending on whether the k-mer belongs to the m^{6}A-modified transcripts (red) or the unmodified transcripts (black). The contribution of each principal component is shown in each axis. (**D, E, F, G**) ROC curves of the SVM predictions using: i) each individual feature separately to train and test each model, at m^{6}A sites (D); ii) combined features at m^{6}A sites, relative to the individual features (E); iii) combined features at m^{6}A sites relative to control sites, where the base-called 'errors' information of neighbouring nucleotides has been included in the model (F); and iv) different mixtures of methylated and unmethylated reads, using the combined features model (G). Error bars indicate s.d.

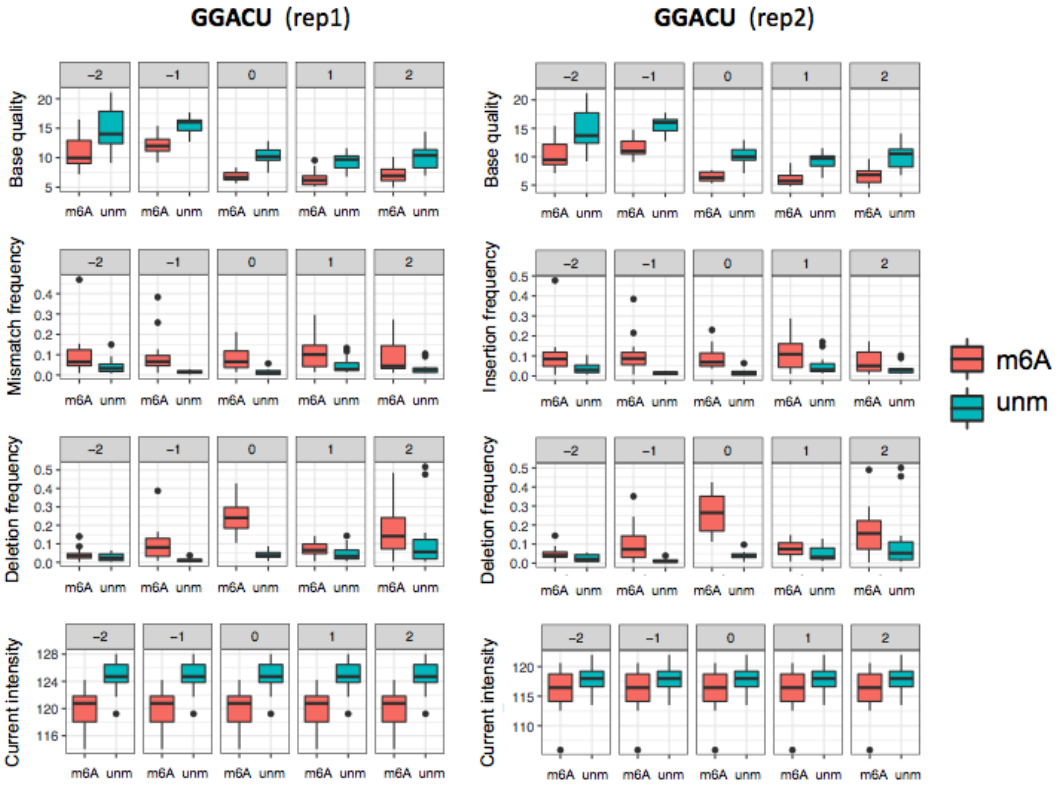


Figure 3.6 - Replicability of the base-called features of GGACU k-mers, for each position of the k-mers.

Base-called features of m^{6}A-modified datasets are depicted in red, whereas those from unmodified datasets are depicted in blue. Error bars indicate s.d.

To statistically determine whether these features could be used to accurately classify a given site into 'm^{6}A-modified' or 'unmodified', multiple Support Vector Machines (SVM) were trained using as input the base-called features from m^{6}A-

containing RRACH k-mers and unmodified RRACH k-mers (see Methods). We first tested whether each individual feature at position 0 (the modified site) was able to classify a given RRACH k-mer into m^{6}A-modified or unmodified. The results show that base quality, deletion frequency and mismatch frequency alone were able to accurately predict the modification status with reasonable accuracy (70-86% accuracy, depending on the feature used) (**Figure 3.5D**, see also Methods). By contrast, the current mean intensity values and current intensity standard deviation were poor predictors of the modification status of the k-mer (43-65% accuracy). As a control, the same set of features in control k-mers (i.e., those with the same sequence context, but centered in C, G or U) were used, finding that the features did not distinguish between m^{6}A-modified and m^{6}A-unmodified datasets (see also **Figure 3.7**).

To improve the performance of the algorithm, we then examined whether a combination of the features might improve the prediction accuracy, finding that the combination of the 3 features (base quality, mismatch and deletion frequency) increased the accuracy of the model (88-91%) (**Figure 3.5E**). We then tested whether the inclusion of features from the neighbouring positions (-2, -1, +1, +2) might further improve the model. Indeed, we find that the inclusion of neighbouring features slightly improves the performance of the algorithm (accuracy = 97-99%), however, this was at the expense of increasing the number of false positives in the control k-mer set -which do not contain the modification- (**Figure 3.5F,** see also **Figure 3.7**), suggesting that features from neighbouring positions should not be employed with this model.

It should be noted that the current algorithm has been trained using either 100% methylated or 100% unmethylated reads; however, in *in vivo* data, this will likely not be the scenario. Previous studies probing the m^{6}A modification status in individual sites have estimated that m^{6}A methylation in mRNAs occurs only partially, with methylation ratios ranging from 6% to 80% [94]. Therefore, we wondered whether the algorithm would be able to detect m^{6}A modifications on mixtures of methylated and unmethylated reads. To test this, reads from both m^{6}A-modified and unmodified datasets were sampled and mixed them in different proportions, to achieve partial methylation ratios of 0% (unmodified), 5%, 10%, 25%, 50%, 75%, 90%, 95% and 100% (fully modified). The algorithm performance is dependent on the proportion of methylated reads (**Figure 3.5G**); however, even at 25% of methylation ratio, it predicts m^{6}A sites with reasonable accuracy, with an area under the curve (AUC) of 0.72 (**Figure 3.5G**).

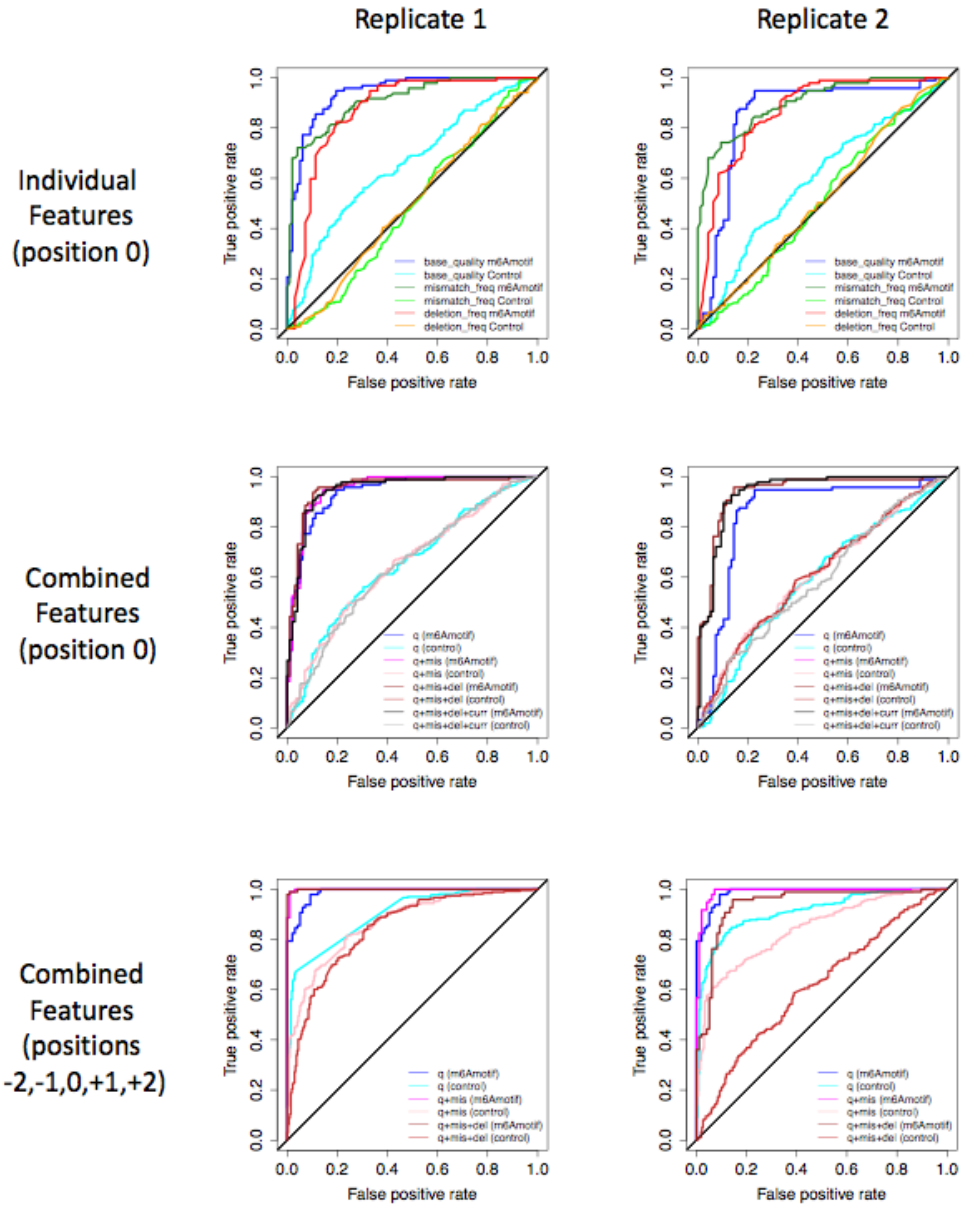


Figure 3.7 - ROC curves of SVM trained with single features compared to combined features.

Performance of each replicate is shown separately in each plot.

3.5. Trained SVM models can predict m^{6}A RNA modifications in *in vivo* datasets

To assess whether these findings could be extended to *in vivo* datasets, I sequenced native polyA(+)-selected RNA from *S. cerevisiae* wild-type (*wt*) and *ime4Δ* knockout strains (**Figure 3.8A** and Methods). *Ime4Δ* yeast strains constitute an excellent background model to identify false positives in m^{6}A analyses [340], as the deletion of *ime4* results in complete elimination of m^{6}A. Biological triplicates of polyA(+)-selected RNA from both *wt* and *ime4Δ* strains were sequenced in independent flow cells (see Methods), producing more than 5 million sequenced reads.

An initial assessment of the quality of the direct RNA sequencing runs showed that these were highly replicable both in terms of per-gene counts (spearman's rho=0.945-0.948) and average per-read quality scores (**Figure 3.8B**, see also **Figure 3.9**). *EpiNano* was then used to extract base-called features for all 6 samples. First, features corresponding to ~1300 known m^{6}A-modified RRACH site, previously identified using antibody immunoprecipitation coupled to next-generation sequencing (m^{6}A-Seq) were analysed [340]. Base-called features at m^{6}A-modified RRACH sites were distinct across yeast strains (*wt* and *ime4Δ*), for all three metrics analysed (base quality, deletion frequency and mismatch frequency) (**Figure 3.8C**). These results were consistent across biological replicates, and are in agreement with our observations using *in vitro* constructs (**Figure 3.3C**). By contrast, this was not observed when comparing unmodified RRACH base-called features across yeast strains (**Figure 3.10**), suggesting that the observed differences were due to the presence of m^{6}A. These results were further confirmed by individual inspection of 'known' m^{6}A-modified sites, where both increased mismatch and deletion frequencies were consistently observed in *wt* m^{6}A-modified positions, but not in their corresponding *ime4Δ* sites (**Figure 3.8D**).

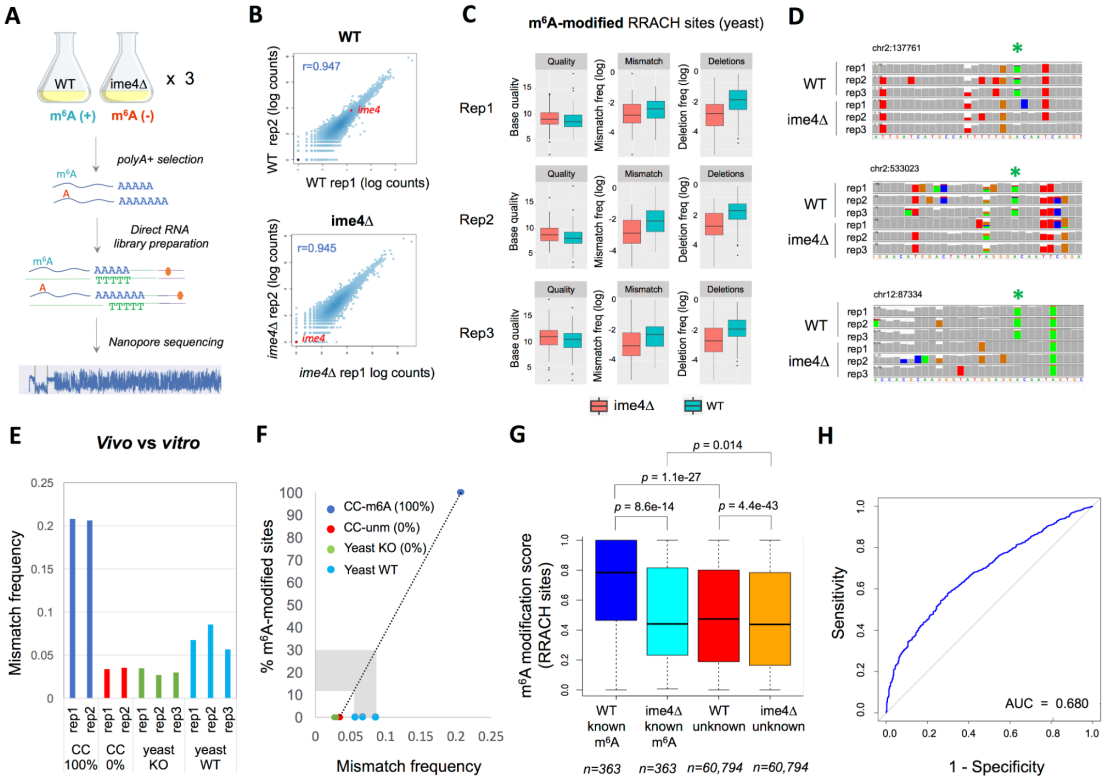


Figure 3.8 - Yeast wild-type and ime4Δ strains show distinct base-called features at known m^{6}A-modified RRACH sites.

(**A**) Overview of the direct RNA sequencing library preparation using *in vivo* polyA(+) RNA from *S. cerevisiae* cultures (**B**) Replicability of per-gene counts using direct RNA sequencing across wild-type yeast strains (top) and ime4Δ strains (middle). The correlation between wild-type and *ime4Δ* strains is also shown (bottom). (**C**) Comparison of the observed mismatch frequencies in the 100% modified *in vitro* transcribed 'sequences (blue), unmodified sequences (red), yeast ime4Δ knockout (green) and yeast wild-type (cyan). Values for each biological replicate are shown. (**D**) Base-called features (base quality, insertion frequency and deletion frequency) of RRACH 5-mers known to contain m^{6}A modifications. Only features corresponding to the modified nucleotide (position 0) are shown. Features extracted from wild-type yeast reads (m^{6}A-modified) are shown in blue, whereas those from ime4Δ (unmodified) for the same set of k-mers are shown in red. (**F**) Genomic tracks of previously reported m^{6}A-modified RRACH sites in yeast, identified using Illumina sequencing. The m^{6}A-modified nucleotide is highlighted with a green asterisk. In these positions, wild-type yeast strains show increased mismatch frequencies as well as decreased coverage -

reflecting increased deletion frequency- in all three biological replicates, whereas these features are not observed in any of the three ime4Δ replicates. (**G**) Predicted m^6A modification scores predicted by the trained SVM at known m^6A-modified (n=363) and unknown (n=60,794) RRACH sites, both for yeast wild-type and ime4Δ datasets. P-values have been computed using Kruskal-Wallis test. A site was included in the analysis if there were mapped reads present in all 6 yeast samples. Sites with more than one "A" in the 5-mer were excluded from the analysis. (**H**) ROC curve depicting the performance of *EpiNano* in yeast datasets (n=61,363 sites). Error bars indicate s.d.

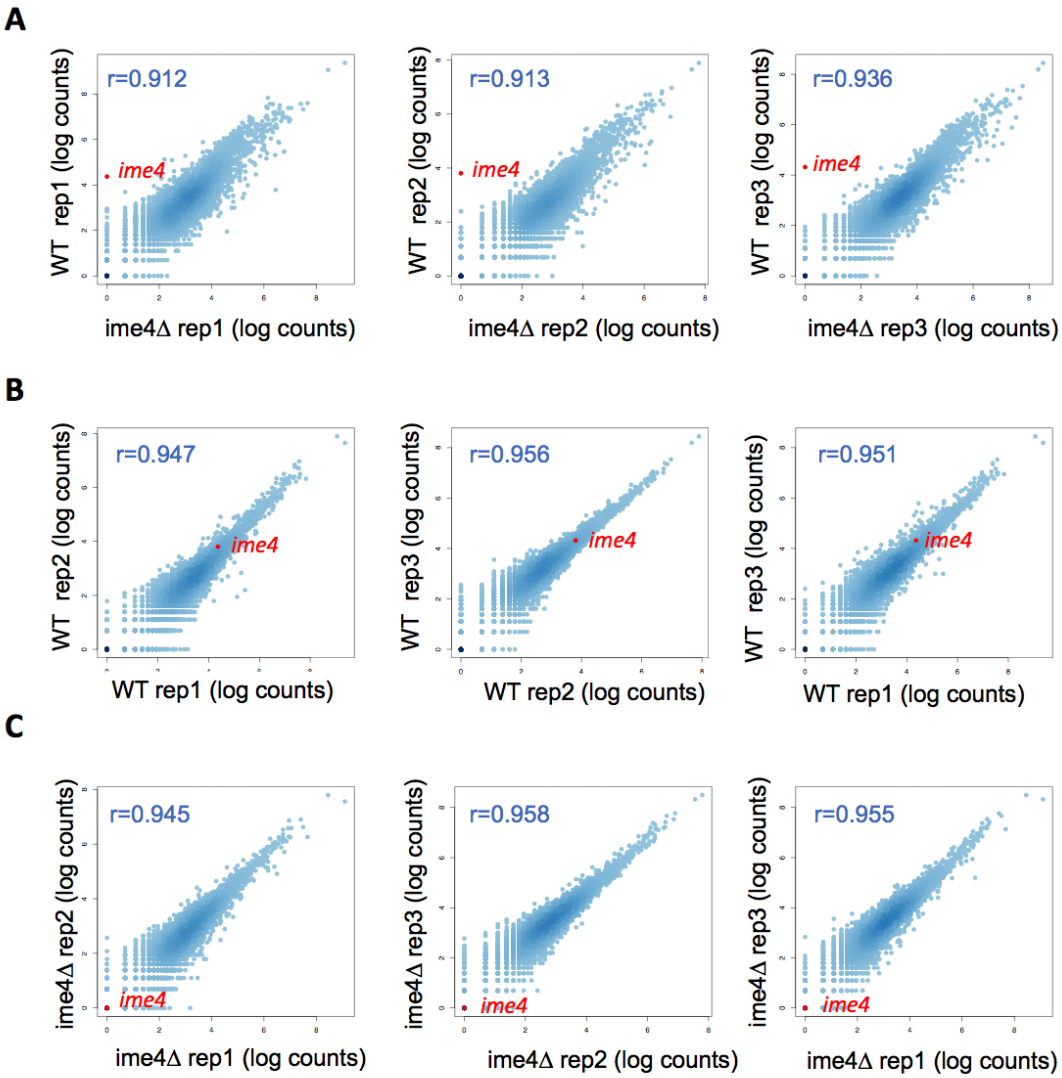


Figure 3.9 - Replicability of the direct RNA sequencing experiments across biological replicates expressed as log counts for each gene.

Each dot represents a gene, and the *ime4*gene has been highlighted in red. Correlation values shown correspond to Spearman's rho.

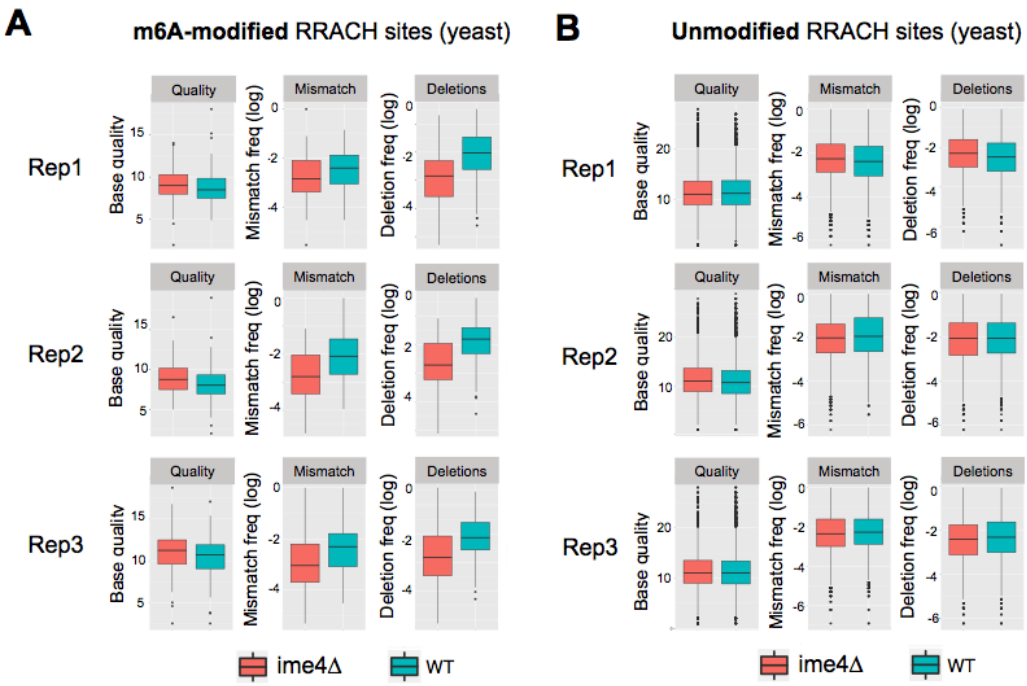


Figure 3.10 - Base-called features (base quality, insertion frequency and deletion frequency) of RRACH 5-mers

Comparison of 5-mers known to contain m^{6}A modifications (left panels) compared to those that are not known to contain m^{6}A modifications (right panels). Only features corresponding to the modified nucleotide (position 0) are shown. Features extracted from *wt* yeast reads (m^{6}A-modified) are shown in blue, whereas those from *ime4*Δ (unmodified) are shown in red. Error bars indicate s.d.

To determine whether the trained SVM could be applied to *in vivo* datasets, we first investigated whether the global *in vivo* base-called features were consistent with those observed *in vitro*. Unmodified *in vitro* sequences (CC 0%) displayed similar mismatch frequencies to those observed in *ime4Δ* strains, which also lack m^{6}A

modifications (**Figure 3.8E**). By contrast, m^{6}A-modified yeast RNAs (*wt*) showed intermediate mismatch frequencies between fully modified (CC 100%) and unmodified (CC 0%, yeast *ime4Δ*) sequences. Using linear regression, the median stoichiometry of m^{6}A modifications in *wt* strains was estimated to be 12-30% (**Figure 3.8F**), which is in agreement with previous works, where m^{6}A was found to be present at levels ranging from 7 to 69% (with a median of 23%) in yeast samples [341]. Altogether, these results reveal that non-random base-called 'errors' present in *in vivo* datasets are replicable, are in agreement with *in vitro* results, and are correlated with the presence of m^{6}A RNA modifications in a given site.

The SVM model that previously trained with m^{6}A-modified and unmodified *in vitro* constructs was used to predict the transcriptome-wide m^{6}A modification status of yeast RRACH sites, both in *wt* and *ime4Δ* datasets. A site was kept for downstream analyses if there was at least 1 read per site in each of the six samples. This criterion was met by 61,163 RRACH sites, from which 363 had been reported as 'm^{6}A-modified', based on Illumina sequencing[340]. Per-site SVM predictions for each biological replicate were then merged into a single 'm^{6}A modification score' (see Methods). It should be noted that low read coverage leads to decreased accuracy (**Figure 3.11**); however, low coverage sites were retained in order to maximise the number of sites included in the analyses. First, the m^{6}A modification scores of known m^{6}A-modified RRACH sites (n=363) in *wt* and *ime4Δ* were compared, finding that modification scores in *wt* were significantly higher than those observed in *ime4Δ* (p = 8e-14), for the same set of sites (**Figure 3.8G**). By contrast, modification scores of *ime4Δ* known m^{6}A-modified RRACH sites (n=363) and unknown sites in the same strain, which do not contain m^{6}A modifications, were relatively similar (p=0.01, Kruskal-Wallis test) (**Figure 3.8G**). Interestingly, the method also identified significant differences in m^{6}A modification scores when comparing *wt* and *ime4Δ* unknown RRACH sites (p = 4e-43; **Figure 3.8G**), suggesting that there might be additional m^{6}A-modified sites present in the transcriptome, apart from those identified using m^{6}A-Seq[340]. Indeed, recent efforts using enzymatic-based m^{6}A detection methods have reported that antibody-based methods severely underestimate the number of m^{6}A sites [341]. Overall, the model identifies m^{6}A modifications in yeast datasets with an overall accuracy of 87.8%,

recovering 32% (117 out of 363) of known m^{6}A-modified sites (**Figure 3.8H**), and with a specificity of 89%.

3.6. EpiNano performance compared to methods relying on direct comparison of raw current intensities

Previous efforts have attempted to identify RNA modifications from direct RNA sequencing samples by performing direct comparison of raw current intensities. This is the case of Tombo [334], a software originally developed for the detection of DNA modifications in nanopore sequencing data, which has recently been extended to detect RNA modifications. Identification of modifications from raw signal typically requires a two-step process: (i) re-squiggling of the raw signal to 'align' all reads mapping to the same genomic location, and (ii) comparison of raw current intensities across reads or samples. First, we found that the re-squiggling step used by Tombo discards ~50% of the reads. From the re-squiggled reads, Tombo is able to identify 220 out of the 363 known m^{6}A-modified sites in yeast *wt*, thus recovering 59.6% of known sites. However, this increased recovery of true positives was at the expense of increased number of false positives (Tombo specificity= 69.8%; EpiNano specificity = 89%). Thus, for the same set of 61,163 sites, Tombo correctly predicts known m^{6}A sites with accuracy of 69% and recovery of 59%, whereas EpiNano predicts them with an accuracy of 87% and recovery of 32%.

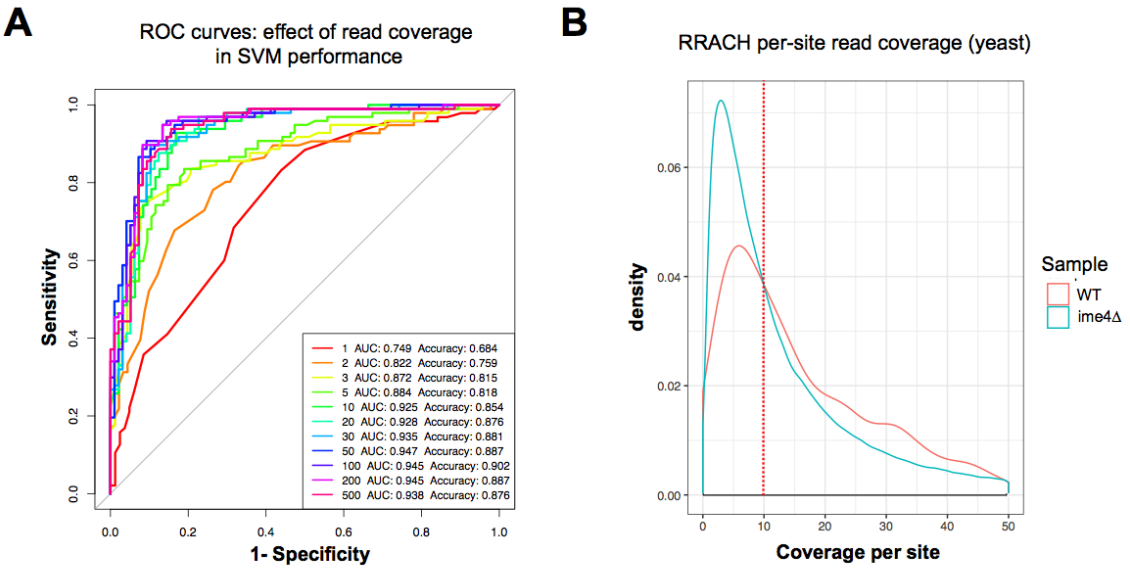


Figure 3.11 - SVM performance is dependent on per-site read coverage.

(**A**) ROC curves depicting the effect of per-site read coverage in SVM performance (number of reads per-site tested: 1, 2, 3, 5, 10, 20, 30, 50, 100, 200 or 500). Different read coverages have been simulated by random subsampling of reads. For each simulated dataset, area under the curve (AUC) and the accuracy are shown. (**B**) Per-site read coverage in yeast wild-type (WT) and ime4Δ knock-out (KO) strains, for each biological replicate. The median coverage for all sites is depicted in red.

Altogether, the *in vivo* analyses validate the findings using *in vitro* m^6A-modified and unmodified sequences, and confirm the use of base-calling 'errors' as a proxy to identify m^6A modifications in direct RNA sequencing datasets. Furthermore, the findings validate the use of *in vitro* constructs, transcribed with and without RNA modifications, as a valid strategy for training direct RNA sequencing base-calling algorithms, suggesting that similar approaches could be envisioned with additional datasets containing distinct RNA modifications in the future.

3.7. Materials and Methods

3.7.1. Synthetic sequence design

Sequences were designed such that they would include all possible 5-mers, while minimizing the secondary RNA structure. For this aim, we employed the software *curlcake* (http://cb.csail.mit.edu/cb/curlcake/), which internally uses RNAshapes version 2.1.6 (http://bibiserv.techfak.uni-bielefeld.de/rnashapes) to predict RNA secondary structure. The final output sequence given by the software was ~10kb long. For synthesis purposes, a total of 4 sequences were designed by splitting the 10kb sequence into smaller sequences of slightly different size (2329bp, 2543bp, 2678bp and 2795bp, which is named 'Curlcake 1', 'Curlcake 2', 'Curlcake 3' and 'Curlcake 4', respectively). Each sequence was designed with an internal strong T7 polymerase promoter, an additional BamHI site at the end of the sequence, and with all EcoRV and BamHI sites removed from the sequence. All 4 sequences were synthesised and cloned in pUC57 vector using blunt EcoRV by General Biosystems. Plasmids were double digested O/N with EcoRV-BamHI-HF, and DNA was extracted with Phenol-Chloroform followed by EtOH precipitation. Plasmid digestion was confirmed by agarose gel. Digestion product quality was assessed with Nanodrop before proceeding to *in vitro* transcription (IVT).

3.7.2. In vitro transcription, capping and polyadenylation

In vitro transcribed (IVT) sequences were produced using the Ampliscribe™ T7-Flash™ Transcription Kit (Lucigen-ASF3507), using 1 ug of purified digestion product as starting material, following manufacturer's recommendations. ATP was replaced by N^6-Methyladenosine-5'-Triphosphate(m^6ATP) (Trilink-N-1013;) for the IVT reaction of m^6A-modified RNA. IVT reaction was incubated for 4 hours at 42°C. *In vitro* transcribed RNA was then incubated with DNAse I (Lucigen), followed by purification using RNeasy Mini Kit (Qiagen-74104). Integrity and quality of the RNA was determined using Agilent 4200 Tapestation, to ensure that a single product band of the correct size had been produced for each IVT product (**Figure 3.2**). Each IVT product was 5' capped using Vaccinia Capping Enzyme (NEB-M2080S) following manufacturer's recommendations. The capping reaction was incubated for 30 minutes at 37 °C. Capped IVT products were purified using RNA Clean XP Beads (Beckman Coulter-A66514). Poly(A)-tailing was performed using *E. coli* Poly(A) Polymerase (NEB-M0276S), following manufacturer's recommendations. Poly(A)-tailed RNAs were purified using RNA Clean XP beads, and the addition of poly(A)-tail was confirmed using Agilent 4200 Tapestation (**Figure 3.2**). Concentration was determined using Qubit Fluorometric Quantitation. Purity of the IVT product was measured with NanoDrop 2000 Spectrophotometer.

3.7.4. Yeast culturing

Sk1 strains used in this study were SAy841 comprising a deletion of NDT80 (hereafter, referred to as 'wild-type'), and SAy966, in which both NDT80 and IME4 were deleted (hereafter, referred to as 'ime4Δ'); These strains are characterised in Agarwala et al [342]. To induce synchronous meiotic entry, cells were grown for 24 hr in 1% yeast extract, 2% peptone, 4% dextrose at 30°C, diluted in BYTA (1% yeast extract, 2% tryptone, 1% potassium acetate, 50 mM potassium phthalate) to OD600 = 0.2 and grown for another 16 h at 30°C, 200 rpm. Cells were then washed twice with water and re-suspended in SPO (0.3% potassium acetate) at OD600 = 2.0 and incubated at 30°C at 190 rpm. Cells were isolated from SPO following 5 hours and collected by 2 min centrifugation at 3000g. Pellets were snap frozen and stored at -80ºC for RNA extraction. Three independent biological replicates for each strain were collected.

3.7.5. Yeast mRNA preparation

Yeast total RNA samples were prepared using a modified protocol of nucleospin® 50 RNA kit (Machery-Nagel, cat 740955.50). Specifically, cells lysis was done in a 1.5ml tube by adding 450µl of lysis buffer containing 1M sorbitol (SIGMA-ALDRICH), 100mM EDTA and 0.45µl lyticase (10U/µl). The sample was incubated in 30°C for 30 minutes to break the cell wall, centrifuged for 10' at 3000 rpm, and the supernatant was removed. From this stage, extraction proceeded as in the protocol of nucleospin® 50 RNA kit, only substituting β-mercaptoethanol with DTT. Enrichment of polyadenylated RNA from total RNA was performed using Oligo(dT) dynabeads mRNA-DIRECT kit (Thermo Scientific, 61012) for small mRNA amounts.

3.7.6. Direct RNA library preparation and sequencing

RNA library for direct RNA Sequencing (SQK-RNA001) was prepared following the ONT Direct RNA Sequencing protocol version DRS_9026_v1_revP_15Dec2016. Briefly, 800 nanograms of Poly(A)-tailed and capped *in vitro* transcribed RNA –in the case of curlcakes– or 500 nanograms of yeast polyA+ RNA were ligated to ONT RT Adaptor (RTA) using concentrated T4 DNA Ligase (NEB-M0202T), and was reverse transcribed using SuperScript III Reverse Transcriptase (Thermo Fisher Scientific-18080044). The products were purified using 1.8X Agencourt RNAClean XP beads (Fisher Scientific-NC0068576), washing with 70% freshly prepared ethanol. RNA Adapter (RMX) was ligated onto the RNA:DNA hybrid, and the mix was purified using 1X Agencourt RNAClean XP beads, washing with Wash buffer (WSB) twice. The sample was then eluted in Elution Buffer (ELB) and mixed with RNA running buffer (RRB) prior to loading onto a primed R9.4.1 flow cell, and ran on a GridION (MinION for the second replicate) sequencer with MinKNOW acquisition software version v1.14.1 (v.1.15.1 for the second replicate in the curlcake experiment). The sequencing was performed in independent days and machines, with two biological replicates for each 'curlcake' experiment condition (non-modified and m^6A-modified RNA, total of 4 flow cells). Each biological replicate and condition was sequenced independently in a different flow cell. For the *in vivo* analysis in *S. cerevisiae*, three biological replicates for each yeast strain (wild-type and ime4Δ) were sequenced, and each biological condition and replicate was sequenced in an independent flow cell (total of 6 flowcells).

3.7.7. Base-calling, filtering and mapping

Reads were locally base-called using Albacore 2.1.7 (Oxford Nanopore Technologies). Base-called reads were filtered using NanoFilt, a component from Nanopack with settings '-q 0 --headcrop 5 --tailcrop 3', and mapped to the 4 synthetic sequences using minimap2 with the settings -ax map-ont. Mapped reads were then converted into mpileup format using Samtools version 1.4. For comparison, reads were also base-called with Albacore 2.3.4 and Guppy 2.3.1, finding that all base-callers showed increased mismatch frequencies in m^6A-modified datasets (with the largest increased in A positions) and decreased qualities (**Figure 3.12**).

3.7.8. Feature extraction

To extract per-site features (mean per-base quality, mismatch frequency, insertion frequency and deletion frequency), BAM alignment files were converted to tab delimited format using sam2tsv from jvarkit. For each individual reference site, the mean quality of the aligned bases, the mismatch, insertion and deletion frequency was computed using in-house scripts (available on github). To extract current intensity information from individual reads, the h5py (version 2.7.0) module in python was used to parse each individual fast5 file. Reference sequences were slided with a window size of 5bp, and mean and standard deviation of current intensities was computed for each sliding window. All in-house python scripts used to extract the features described above are publicly available as part of *EpiNano* (github.com/enovoa/EpiNano).

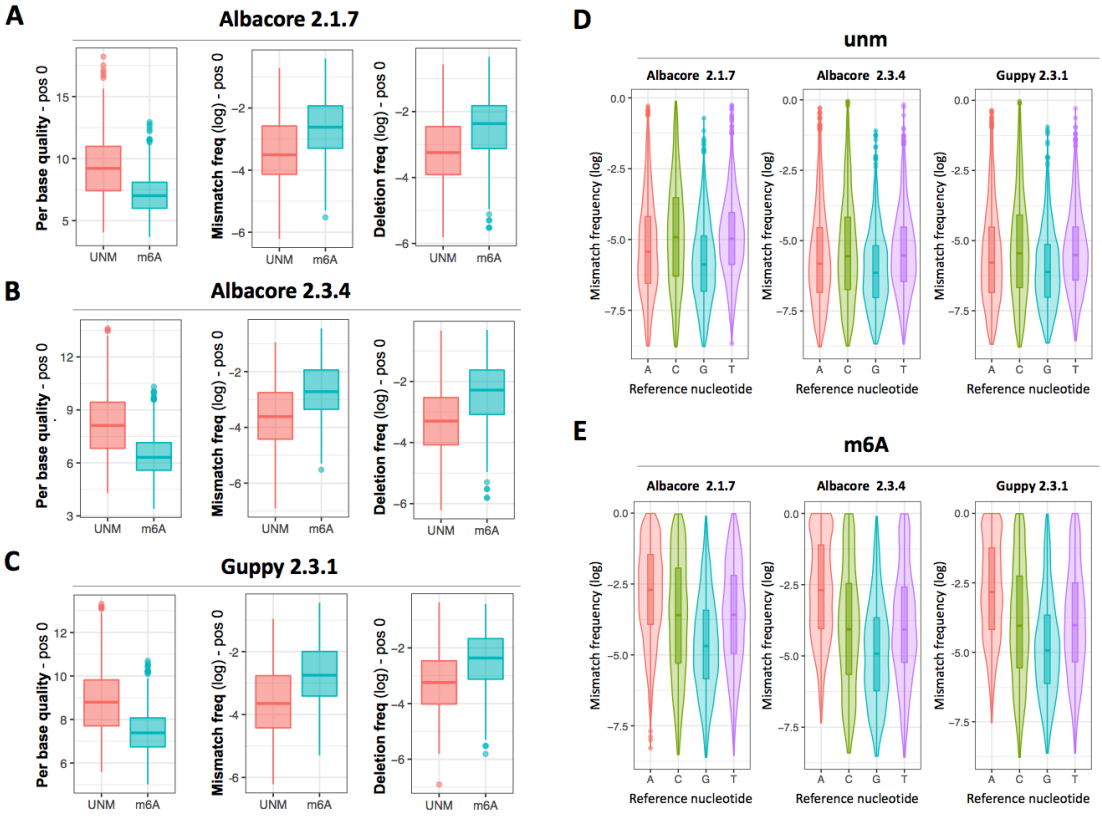


Figure 3.12. Comparison of base-called features using different base-calling algorithms: Albacore 2.1.7, Albacore 2.3.4 and Guppy 2.3.1.

(**A, B, C**) Per-base quality scores, mismatch frequencies and deletion frequencies at position 0, comparing m^{6}A-modified reads (blue) to unmodified reads (red), using either Albacore version 2.1.7 (A), Albacore version 2.3.4 (B) or Guppy 2.3.1 (C). (**D, E**) Comparison of mismatch frequencies at position 0, grouped by the reference nucleotide, both for unmodified datasets (D) and m^{6}A-modified datasets (E). Mismatch frequencies at A positions are consistently increased in m^{6}A-modified datasets, and for all three base-callers tested. Error bars indicate s.d.

3.7.9. Machine learning

The set of extracted features of both m^{6}A-modified and unmodified 'curlcakes' was used as input to train a Support Vector Machine (SVM). Initial training (75% of the sites) and testing (25%) of the SVM was performed with m^{6}A-modified and unmodified curlcake reads from one replicate (rep1). Multiple kernels ('linear', 'poly' and 'rbf') were compared, and the best performing kernel was retained. The model was validated on new sequencing runs of *in vitro* transcribed m^{6}A-modified and unmodified sequences (rep2), which had not been used for initial training or testing of the SVM. The reported accuracy values refer to the predictions on replicate 2. The code to extract the set of features for machine learning from fastq and fast5 reads, the code for building the SVM

models, as well as the trained SVM models, are publicly available in github (github.com/enovoa/EpiNano). We should note that a limitation in utilizing *in vitro* transcription to generate all possible 5-mers is that 5-mers that contain more than one "A" will contain more than one modification in the kmer, e.g. AGACC will in fact be m^6AGm^6ACC, which are unlikely to occur in a biological context. Therefore, 5-mers that contained more than one A have been excluded from the analyses as well as from the training set. Accuracy of the model has been computed as the sum of correct m^6A modification predictions - correctly predicted m^6A-modified k-mers (true positives, TP) and correctly predicted unmodified k-mers (true negatives, TN)- divided by the total number of k-mers tested.

3.7.10. Prediction of m^6A modified sites in yeast using EpiNano

EpiNano was used to extract per-site features (mean per-base quality, mismatch frequency, insertion frequency and deletion frequency) from the mapped BAM files of the six samples sequenced (WTrep1, WTrep2, WTrep3, ime4Δrep1, ime4Δrep2, ime4Δrep3). m^6A-modified RRACH sites with minimum coverage of 5 reads/site were kept and scored using the previously trained SVM model. 5-mers containing more than one "A" in the motif were discarded from downstream analyses, as these k-mers had not been included in the training sets (see above). A total of 61,163 sites were analysed for each sample and replicate, from which 363 corresponded to 'known' m^6A-modified sites, which had been identified using Illumina sequencing [343].

M^6A modification scores for each site were computed by merging the SVM predicted probabilities across replicates. Specifically, if the probability being modified was greater than 0.5 in all three biological replicates (s1, s2, s3), the modification score (M) was set to 1. Otherwise, the modification score was determined by computing the mean of the probabilities (pseudocode 1). Modification scores were obtained for each site, both for wild-type and ime4Δ strains.

Pseudocode 1: *if (s1 ≥ 0.5 & s2 ≥ 0.5 & s3 ≥ 0.5):*

M = 1

else:

M= (s1+s2+s3)/3

To classify a site as "m^{6}A-modified" or "unmodified", the modification scores of each site, obtained for each of the two strains, were compared. Specifically, the modification ratio was calculated by dividing the modification score of the wild-type strains (M_{wt}) and the modification score of the ime4Δ strains (M_{ko}). A site was considered to be modified if the modification ratio was greater than 1.5 and the modification score in wild-type strains (M_{wt}) was greater than 0.5 (pseudocode 2).

Pseudocode 2: *if (M_{wt}/M_{ko})>1.5 & M_{wt}> 0.5:*

status=modified

else:

status=unmodified

Accuracy of the predictions was computed as the sum of correct m^{6}A modification predictions - correctly predicted m^{6}A-modified k-mers (true positives, TP) and correctly predicted unmodified k-mers (true negatives, TN)- divided by the total number of k-mers tested (n=61,163). Positive predictive value (PPV) was computed by dividing the number correctly predicted m^{6}A-modified k-mers (true positives, TP) by the total number of m^{6}A-modified k-mers included in the analysis (n=363).

3.7.11. Prediction of m^{6}A modified sites using Tombo

First, Tombo version 1.5 [334] was run to align the raw signal and the base-called reads sequences (*tombo resquiggle*), both for wild type and ime4Δ samples. Then the Tombo 'canonical sample comparison' method (*tombo model_sample_compare*) was used to identify significant shifts in raw signals in paired datasets (*wt* and *ime4Δ*). To maximise the number of predictions, the parameter --num-most-significant-stored 14000000 was used, which approximates yeast genome size, and --minimum-test-reads 1. M^{6}A modification scores for each site were computed by merging the Tombo predicted probabilities across replicates. Specifically, if the probability being modified was greater than 0.5 in all three biological replicates (s1, s2,

s3), the modification score (M) was set to 1, as previously done for *EpiNano*. Otherwise, the modification score was determined by computing the mean of the probabilities (pseudocode 1). A site was considered as modified if the modification score was greater than 0.5.

3.8. Discussion

The human epitranscriptome is still largely uncharted. Only a handful of the 170 different RNA modifications that are known to exist have been mapped. Importantly, several of these modifications are involved in central biological processes, such as sex determination [49,201,344] or cell fate transition [51], and their dysregulation has been linked to multiple human diseases [52,53,55], including neurological disorders [345–347] and cancers [348–350]. Yet, our understanding of this regulatory layer is restricted to a few RNA modifications, largely due to the lack of a generic methodology to map them in a transcriptome-wide fashion.

The establishment of the ONT platform as a tool to map RNA modifications has great potential to revolutionise our understanding of the epitranscriptome, as in principle, it should be capable of identifying RNA modifications in individual RNA sequences, and with single nucleotide resolution. Such ability would allow us to study the functions of the epitranscriptome in ways that, until now, have not been possible. Unfortunately, currently there is no software available that can predict RNA modifications from direct RNA sequencing reads with sufficient accuracy, limiting the applicability of direct RNA sequencing as a tool to identify RNA modifications. To tackle this limitation, a novel strategy was provided to identify RNA modifications from base-called features, without the need of squiggling realignments or manipulation of raw current intensity datasets.

Here it is shown that RNA modifications can be identified in the form of systematic and reproducible base-calling ‘errors’ in direct RNA sequencing datasets. These ‘errors’ can be detected in the form of altered per-base qualities, mismatch frequencies and deletion frequencies at the modified site. The method accurately detects modifications both *in vitro* (90% accuracy) and *in vivo* (87% accuracy), with an overall recovery of 32% of known sites. Despite the promising results, it is important to note, however, that the current method have several limitations as well as ample room for improvement.

Firstly, the current algorithm does not predict RNA modifications in individual RNA molecules, but rather employs information from all the reads mapping to a specific site to determine whether a given position is modified or unmodified. Secondly, the

algorithm does not distinguish between different types of RNA modifications (e.g. m^1A versus m^6A). Future work will be needed to decipher whether different types of RNA modifications can be associated to distinct 'error signatures', which could be potentially used to identify the underlying RNA modification type. Thirdly, although m^6A-modified RRACH k-mers globally display altered base qualities, mismatch frequencies and deletion frequencies, it should be noted that the contribution of each feature varies across different k-mers. For example, the presence of m^6A in GGACA and GGACT k-mers mainly affects the mismatch and deletion frequency, whereas in the case of GGACC, base quality and deletion frequency are the most altered features by the presence of m^6A modifications (**Figure 3.13)**. Future models that include k-mer specific training and testing could potentially improve the accuracy of prediction of modified sites, as well as reduce the number of false positives. In this regard, it is expected that by making the m^6A-modified and unmodified datasets publicly available -both base-called fastq and raw fast5-, these can be employed by the community to train different machine learning algorithms (e.g., signal-based machine learning, base-caller training, etc.), and thus lead to improved m^6A RNA modification base-callers for the whole community.

Overall, these results show that base-calling 'errors' can be used as an accurate and computationally simple solution to identify m^6A modifications, which does not require the manipulation of raw current intensities or squiggle alignments. Moreover, the findings were extended to an *in vivo* system, showing that our algorithm can capture m^6A-dependent changes that are present in wild-type SK1 yeast strains, while these are not observed in their *ime4*Δ counterparts. Future work will be needed to achieve single read RNA modification detection, as well as to expand our findings to other RNA modifications.

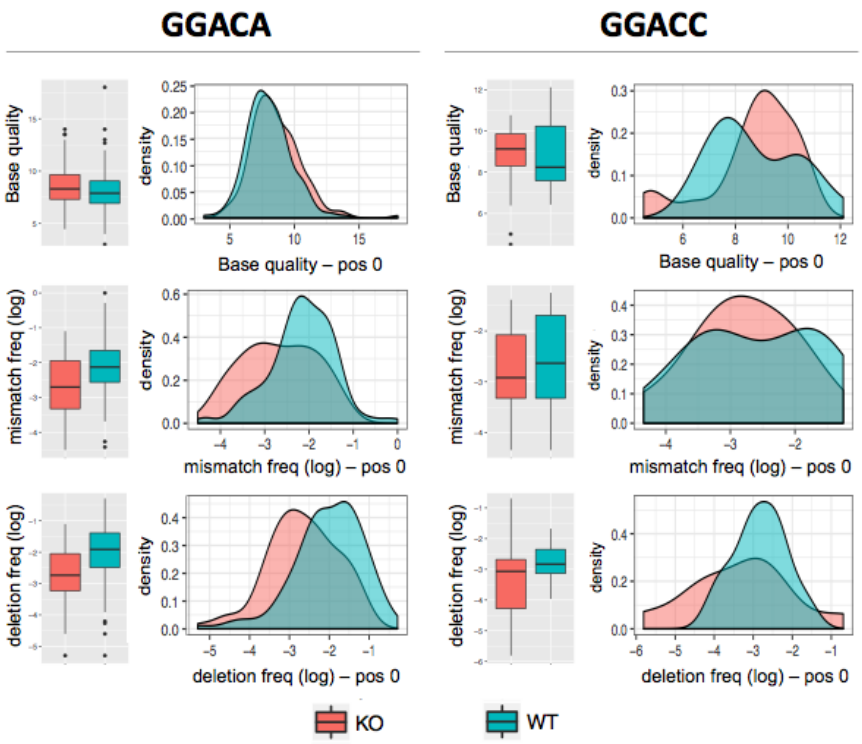


Figure 3.13 - Comparison of base-called features at position 0 in two different RRACH k-mers (GGACA and GGACC).

Base-called features from ime4Δ (red) and wild-type (blue) strains are shown, both in the form of boxplots and density histograms. Only known m^6A-modified RRACH sites in *S. cerevisiae* have been included in these plots. Error bars indicate s.d.

4. Quantitative profiling of pseudouridylation dynamics in native RNAs with nanopore sequencing

This chapter contains material described in the publication published in Nature Biotechnology (Begik and Lucas et al 2021) [351].

I performed most of the wet-lab experiments and data analyses in this study.

The co-first author of the paper Morghan C. Lucas performed the synthesis and sequencing of *in-vitro* constructs for Figure 4.1 and 4.2. She also contributed to making the figures using Adobe Illustrator.

Leszek P. Pryszcz performed the analyses for Figure 4.8D, 4.9, 4.10, 4.15B-H-I-J-K. He also established an RNA modification stoichiometry prediction algorithm.

Jose Miguel Ramirez analysed the sequencing data of *in-vitro* constructs for Figure 4.1 and Figure 4.2 with the help of Morghan C. Lucas.

Rebeca Medina performed the yeast culturing experiments in the presence of stress conditions.

Ivan Milenkovic and Helaine Graziele Santos Vieira performed the yeast culturing in the presence of oxidative stress and sucrose gradient experiments.

Sonia Cruciani and Ivan Milenkovic contributed to the sequencing of snR-60,61 and 62 KO yeast strains.

Eva Maria Novoa and John S. Mattick supervised this study.

4.1. Introduction

As mentioned earlier in the general introduction, algorithms to detect RNA modifications have been made available in the last few months [255,257,352], many of which rely on the use of systematic base-calling 'errors' caused by the presence of RNA modifications. However, to date the vast majority of efforts have been devoted to the detection of m^6A modifications [255,257,258,352–355], and it is largely unknown whether other modifications of RNA bases may be distinguishable from their unmodified counterparts using this technology. Thus, a systematic, multiplexed and unbiased approach that can map and quantify diverse RNA modifications simultaneously in full-length molecules is currently missing.

Here, the *S. cerevisiae* coding and non-coding transcriptome was examined at single molecule resolution using native RNA nanopore sequencing. Most RNA modifications cause systematic base-calling errors, and that the signature of these base-calling 'errors' can be used to identify the underlying RNA modification type. For example, pseudouridine typically appears in the form of U-to-C mismatches, whereas m^5C modifications appear in the form of insertions. I then exploit the identified signatures to *de novo* predict RNA modifications in rRNAs, identifying two previously unreported Ψ modifications in mitochondrial rRNA, which I confirm using CMC-probing coupled to nanopore sequencing (nanoCMC-seq). I demonstrate that one of these Ψ modifications (15S:Ψ854) is placed by the enzyme Pus4, which was previously thought to pseudouridylate only mRNAs and tRNAs[121]. Moreover, once RNA modifications have been accurately predicted using base-calling 'errors', the stoichiometry of a given Ψ- or Nm-modified site can be estimated by clustering per-read features (current intensities and trace) of the modified regions.

Then the dynamics of RNA modifications present in non-coding RNAs were explored. It has been proposed that differential rRNA modifications may constitute a source of ribosomal heterogeneity [356–358]. Indeed, previous studies have shown that temperature changes affect rRNA pseudouridylation levels, suggesting that cells may be able to generate compositionally distinct ribosomes in response to environmental cues [121,359,360]. Similarly, alterations in the stoichiometry of 2'-O-methylation (Am, Cm, Gm, Um) [233,361,362] and pseudouridylation (Ψ) [356–358] can affect translation initiation of mRNAs containing internal ribosome entry sites (IRES) [363,364]. Here we re-examine this question using direct RNA sequencing, and characterise the RNA modification dynamics in rRNAs, snRNAs and snoRNAs upon a battery of environmental cues, translational repertoires and genetic strains. Contrary to expectations, none of the environmental stresses tested lead to significant changes in

the ribosomal epitranscriptome. By contrast, this method does recapitulate previously reported heat-dependent Ψ snRNA modifications, as well as identifies previously unreported heat-sensitive sites in snRNAs and snoRNAs.

Finally, we developed an algorithm named nanoRMS, which can predict Ψ RNA modifications *de novo*, and estimate the stoichiometry of modification both in highly- and lowly-modified Ψ and Nm sites across diverse types of RNA molecules, including rRNAs, sn/snoRNAs and mRNAs. This approach recapitulates known Pus1-dependent, Pus4-dependent and heat stress-dependent mRNA sites, as well as reveals Ψ mRNA sites that had not been previously reported. Altogether, this work establishes a framework for the study of RNA modification dynamics using direct RNA nanopore sequencing, opening avenues to study the plasticity of the epitranscriptome at single molecule resolution.

4.2. RNA modification detection depends on base-calling and mapping algorithms

Previous studies have shown that m^6A RNA modifications can be detected in the form of non-random base-calling 'errors' in direct RNA sequencing datasets [255,257,258,353,354]. However, it is unclear how these 'errors' may vary with the choice of base-calling and mapping algorithms, and consequently, affect the ability to detect RNA modifications. Here, the performance of commonly used base-calling and mapping algorithms were compared on *in vitro* transcribed RNA sequences that contained all possible combinations of 5-mers, referred to as 'curlcakes' (CCs) [255], that included: (i) unmodified nucleosides (UNM), (ii) m^6A, (iii) pseudouridine (Ψ), (iv) m^5C, and (v) 5-hydroxymethylcytosine (hm^5C) (**Figure 4.1A**). In addition, a sixth dataset containing unmodified short RNAs (UNM-S), with median length of 200 nucleotides, was included in the analysis to assess the effect of input sequence length in base-calling (see *Methods*). Each dataset was base-called with two distinct algorithms (*Albacore* and *Guppy*), and using two different versions for each of them, namely: (i) *Albacore* version 2.1.7 (AL 2.1.7); (ii) its latest version, *Albacore* 2.3.4 (AL 2.3.4); (iii) *Guppy* 2.3.1 (GU 2.3.1); and (iv) a more recent version of the latter base-caller, *Guppy* 3.0.3 (GU 3.0.3), which employs a flip-flop algorithm. The latest version of *Albacore* (2.3.4) base-called 100% of sequenced reads in all 6 datasets, whereas its previous version did not (average of 90.8%) (**Figure 4.1B**). By contrast, both versions of *Guppy* (2.3.1 and 3.0.3) produced similar results in terms of percentage of base-called reads (99.96% and 100%, respectively).

In order to assess whether the choice of mapper might affect the ability to detect RNA modifications, two commonly used long-read mappers, *minimap2 [365]* and *GraphMap [366]*, were employed using either 'default' or 'sensitive' parameter settings (see *Methods*). Notably, both the choice of mapper and parameters used severely affected the number of mapped reads (**Figure 4.1C**). The most extreme case was observed with the Ψ-modified dataset, where *minimap2* was unable to map the majority of the reads (0-0.3% mapped reads) (**Figure 4.1C,D,** see also **Figure 4.2A**). By contrast, *GraphMap* 'sensitive' was able to map 35.5% of Ψ-modified base-called reads, with only a minor loss in accuracy (3%) (**Figure 4.2B)**, proving to be a more appropriate choice for highly modified datasets.

4.3. Base-calling 'error' signatures can be used to predict RNA modification type

While base-calling 'errors' can be used to identify m^6A RNA modified sites [255,257,258], whether this approach is applicable for the detection of other RNA modifications, and whether these signatures could be employed to distinguish among distinct RNA modification types, is largely unknown. To this end, the base-calling errors caused by the presence of m^6A, Ψ, m^5C and hm^5C were systematically characterised. Regardless of the base-caller and mapper settings used, modified RNA sequences presented decreased quality scores (**Figure 4.2C-E**) and higher mismatch frequencies (**Figure 4.1E**), being these differences more prominent in Ψ-modified datasets. Principal component analysis of base-calling 'errors' of each modified dataset (m^6A, Ψ, m^5C and hm^5C) -relative to unmodified- showed that this difference was greatest in Ψ-modified datasets (**Figure 4.1F**), and maximised in datasets that were base-called with GU 3.0.3. Thus, all four RNA modifications can be detected in direct RNA sequencing data; however, their detection is severely affected by the choice of both base-calling and mapping algorithms, and varies depending on the RNA modification type.

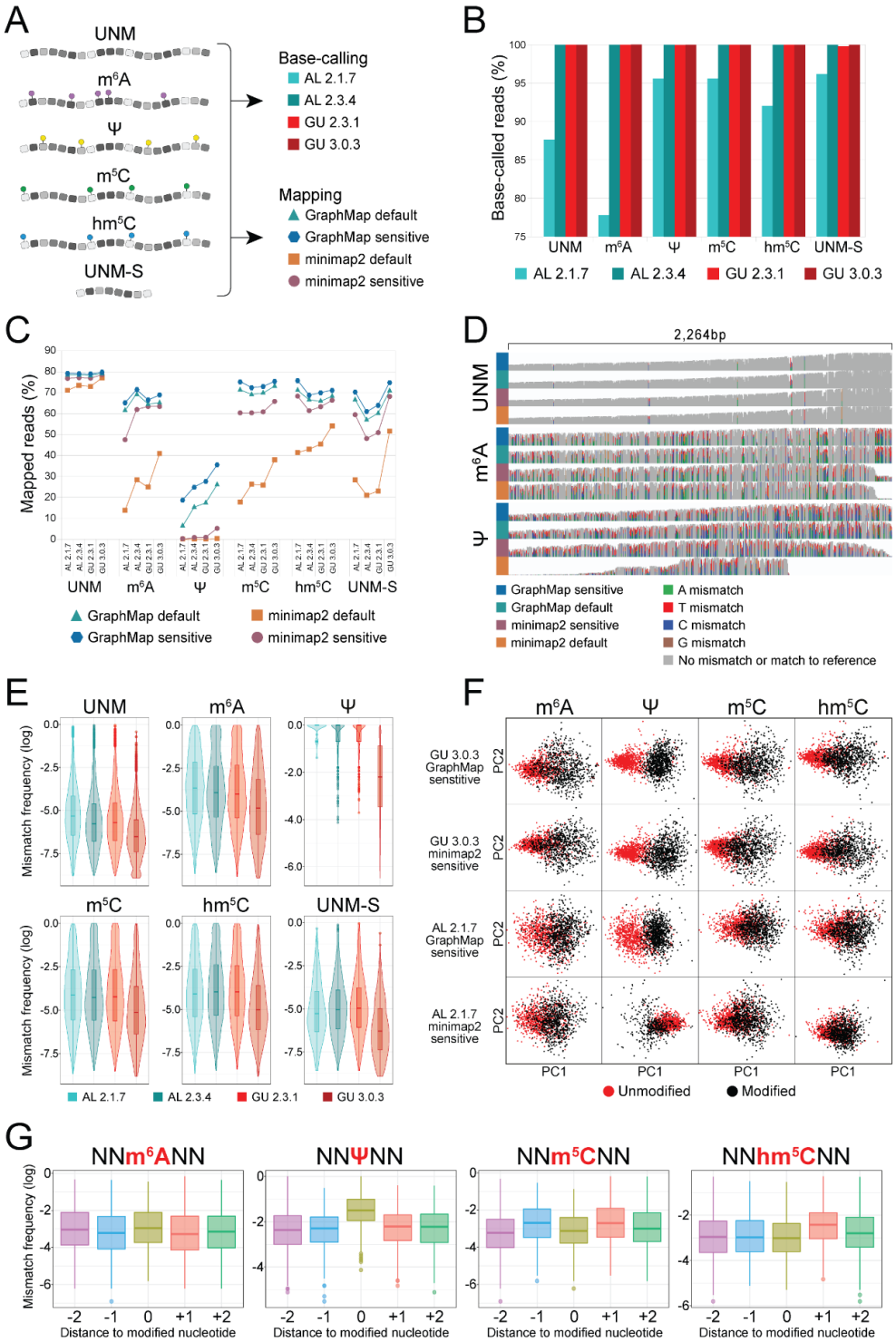


Figure 4.1 - Systematic analysis of base-calling and mapping algorithms for the detection of RNA modifications in direct RNA sequencing datasets.

(A) Overview of the synthetic constructs used to benchmark the algorithms, which included both unmodified (UNM and UNM-S) and modified (m^6A, m^5C, hm^5C and Ψ)

sequences. For each dataset, we performed: i) comparison of base-calling algorithms, ii) comparison of mapping algorithms, iii) detection of RNA modifications using base-called features and iv) comparative analysis of features to distinguish similar RNA modifications. **(B)** Barplots comparing the percentage of base-called reads using 4 different base-calling algorithms in 6 different unmodified and modified datasets. **(C)** Relative proportion of base-called and mapped reads using all possible combinations (16) of base-callers and mappers included in this study, for each of the 6 datasets analysed. **(D)** IGV snapshots illustrating the differences in mapping for 3 distinct datasets: UNM, m^6A-modified and Ψ-modified when base-called with GU 3.0.3. Positions with mismatch frequencies greater than 0.1 have been colored, gray represents match to reference. **(E)** Comparison of global mismatch frequencies using different base-calling algorithms, for the 6 datasets analysed. Box, first to last quartiles; whiskers, 1.5x interquartile range; center line, median; points, outliers; violin, distribution of density. **(F)** Principal Component Analysis (PCA) using as input the base-calling error features of quality, mismatch frequency and deletion frequency in positions -2, -1, 0, 1 and 2, for all datasets base-called with GU 3.0.3 and AL 2.1.7 and mapped with GraphMap and minimap2 on sensitive settings. Only k-mers that contained a modification at position 0, and no other modifications in the 5-mer, were included in the analysis, and the equivalent set of unmodified k-mers was used as a control. **(G)** Mismatch frequency of each position of the 5-mers centered in the modified position (position 0). Box, first to last quartiles; whiskers, 1.5x interquartile range; center line, median; points, outliers.

We then examined whether the base-called 'errors' observed in modified and unmodified datasets occurred in the modified position. We found that both m^6A and Ψ modifications led to increased mismatch frequencies at the modified site (**Figure 4.1G**), mainly in the form of U-to-C mismatches in the case of Ψ modifications (**Figure 4.2F**). By contrast, m^5C and hm^5C modifications did not appear in the form of increased mismatch frequencies at the modified site; rather, these modifications appeared in the form of increased mismatch frequencies in the neighbouring residues (position -1 and +1 in the case of m^5C modifications; position +1 in hm^5C) (**Figure 4.1G**). Moreover, the base-called 'error' signatures of m^5C and hm^5C were also dependent on the sequence context (**Figure 4.2G**). Altogether, all four RNA modifications studied (m^6A, m^5C, hm^5C and Ψ) cause base-calling 'errors', and that these 'errors' follow specific patterns that depend on the RNA modification type.

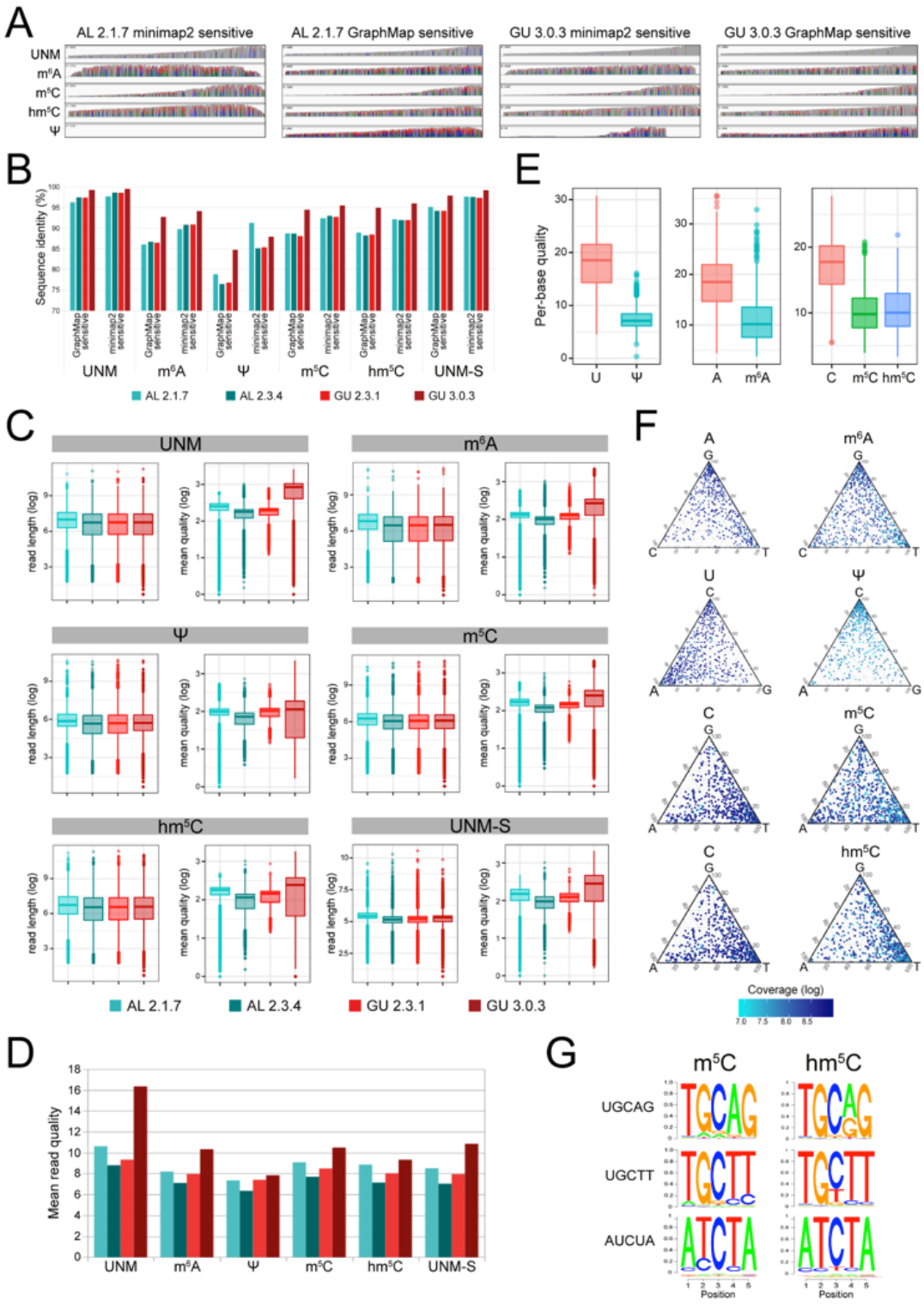


Figure 4.2 - Bench-marking of base-calling and mapping algorithms enables dissection of RNA modification base-calling 'error' signatures and reveals their sequence context-dependence.

(A) IGV snapshots of unmodified (UNM), m^6A-modified (m^6A), m^5C-modified (m^5C), hm^5C-modified (hm^5C) or Ψ-modified (Ψ) *in vitro* transcribed sequence Curlcake 1, base-called using either Albacore 2.1.7 (AL 2.1.7) or Guppy 3.0.3 (GU 3.0.3), and then mapped using minimap2 or GraphMap in 'sensitive' mode. Nucleotides with mismatch frequencies greater than 0.1 have been colored. **(B)** Mean sequence identity of different combinations of base-calling and mapping algorithms, for each of the 6 *in vitro* transcribed datasets analysed. **(C)** Comparison of read lengths and per-read mean quality scores in different *in vitro* transcribed datasets (UNM, m^6A, Ψ, m^5C, hm^5C and UNM-S) when base-called using different algorithms (AL 2.1.7, AL 2.3.4, GU 2.3.1 or GU 3.0.3). Results show that read lengths do not largely vary across base-callers. By contrast, per-read quality strongly varies depending on the choice of base-calling algorithm. Box, first to last quartiles; whiskers, 1.5x interquartile range; center line, median; points, outliers. **(D)** Barplots of mean per-read quality show that per-read qualities are slightly decreased in all modified datasets, relative to unmodified ones, with this difference being most evident in GU 3.0.3 base-called data. **(E)** Boxplots of mean per-base quality of reads base-called with GU 3.0.3 show that per-base qualities are decreased in all modified datasets, relative to unmodified ones. Box, first to last quartiles; whiskers, 1.5x interquartile range; center line, median; points, outliers. **(F)** Ternary plots depicting the mismatch distribution of the unmodified (left) and modified (right) positions colored by log coverage, in 5 different datasets: unmodified (all left panels), m^6A-modified (m^6A), Ψ-modified (Ψ), m^5C-modified (m^5C), hm^5C-modified (hm^5C). Only modified nucleotides, and their relative unmodified counterparts in the UNM dataset, are shown. Each dot represents a different nucleotide in the reference. **(G)** Logo representations of the mismatch signatures generated by m^5C and hm^5C. Results show that the signatures are different depending on the modification, however, these also vary depending on the 5-mer sequence (reported on the left).

4.4. Ψ modifications can be detected as U-to-C mismatches

We then examined whether the results obtained using *in vitro* transcribed constructs would be applicable to *in vivo* RNA sequences. To this end, total RNA from *S. cerevisiae* was prepared for direct RNA sequencing (see *Methods*). Visual inspection of the mapped reads revealed a high proportion of base-calling errors present in 25S and 18S rRNAs, as could be expected from sequences that are highly enriched in RNA modifications (**Figure 4.3A**). By contrast, 5S and 5.8S rRNAs did not show such base-calling errors, in agreement with their low level of modification.

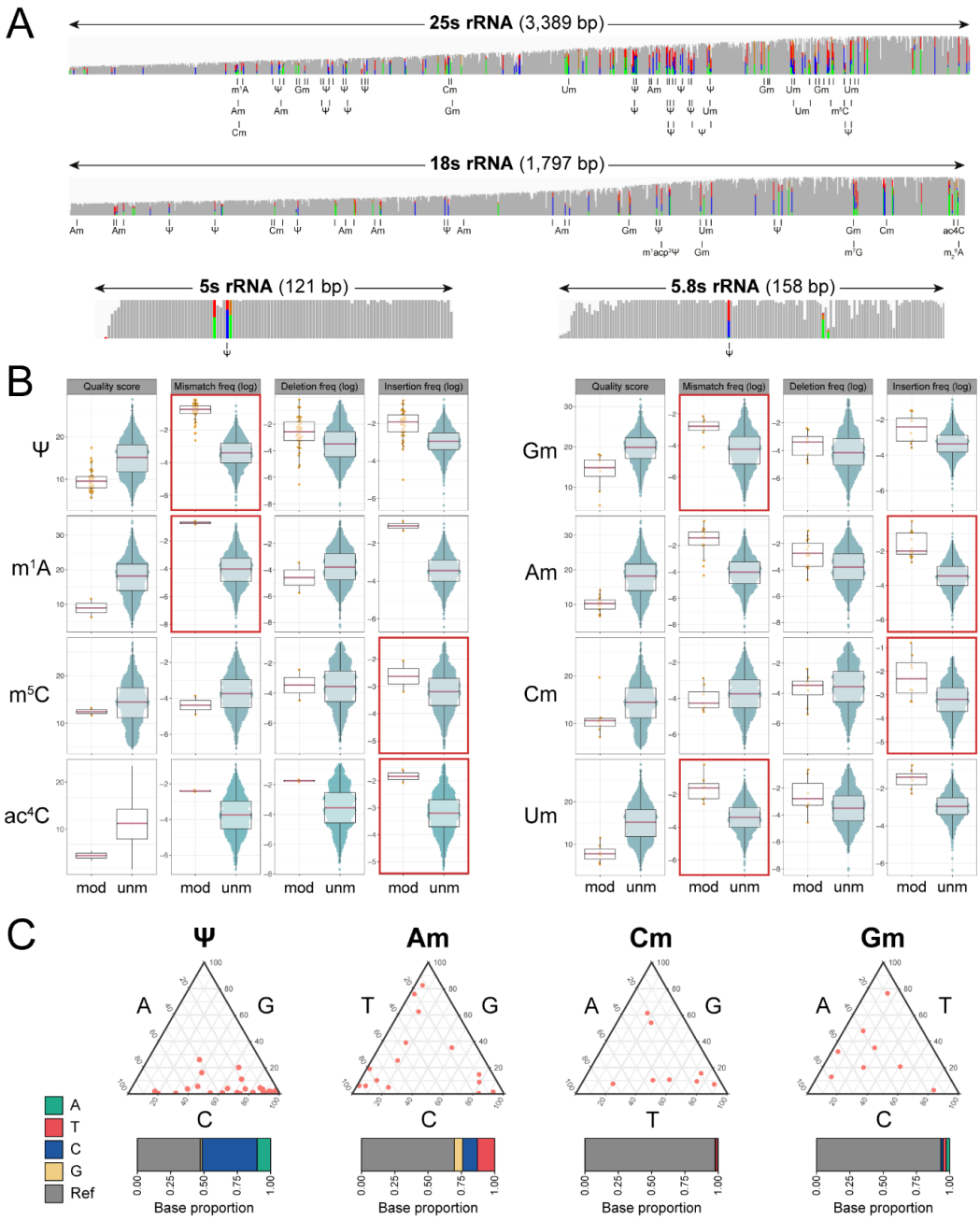


Figure 4.3 - RNA modifications can be detected in yeast ribosomal RNA in the form of base-calling errors, and each RNA modification type shows a distinct 'error' signature.

(A) IGV snapshots of yeast ribosomal subunits 5S, 5.8S, 18S and 25S. Known modification sites are indicated below each snapshot and nucleotides with mismatch frequencies greater than >0.1 have been colored and gray represents match to reference or no mismatch **(B)** Comparison of base-calling features (base quality, mismatch, deletion and insertion frequency) from distinct RNA modification types present in yeast ribosomal RNA. The most descriptive base-calling error per modification is outlined in red. Only RNA modification sites without additional

neighboring RNA modifications in the 5-mer were included in the analysis: Ψ (n=37), Gm (n=8), m^1A (n=2), Am (n=14), m^5C (n=2), Cm (n=8), ac^4C (n=2), Um (n=7). Box, first to last quartiles; whiskers, 1.5x interquartile range; center line, median; dots: individual data points. (**C**) Ternary plots and barplots depicting the mismatch directionality for selected rRNA modifications (Ψ, Am, Cm, Gm). Ψ rRNA modifications tend towards U-to-C mismatches while Am, Cm and Gm modifications did not show specific mismatch directionality patterns.

Then, I systematically analysed base-called features (mismatch, deletion, insertion and per-base qualities) of rRNA modified sites relative to unmodified ones (**Figure 4.3B**), and found that all rRNA modification types consistently led to decreased per-base qualities at modified sites, suggesting that per-base qualities can be employed to identify RNA modifications, but not the underlying RNA modification type. I found that Ψ modifications caused significant variations in mismatch frequencies, in agreement with the observations using *in vitro* constructs. By contrast, other RNA modifications, such as 2'-O-methylcytidine (Cm) or m^5C did not appear in the form of increased mismatch frequencies at modified sites, but rather, in the form of increased insertions. In addition, Ψ modifications typically appeared in the form of U-to-C mismatches (**Figure 4.3C,** see also **Figure 4.4**), in agreement with the *in vitro* observations, whereas other RNA modifications such as 2'-O-methyladenosine (Am) did not cause mismatches with unique directionality. In conclusion, rRNA modification types can be detected in the form of altered base-called features *in vivo*, and that their base-calling 'error' signature is dependent on the RNA modification type.

To confirm that the detected signal (U-to-C mismatches) in Ψ positions was caused by the presence of the Ψ modification, I compared rRNA modification profiles from wild type *S. cerevisiae* to those from snoRNA-knockout strains (snR3, snR34 and snR36) (**Figure 4.5A,** see also **Table 5.1**). These results show that changes in rRNA modification profiles were consistently and exclusively observed in those Ψ positions reported as targets of each snoRNA. Moreover, the remaining Ψ-modified positions were not significantly altered by the lack of Ψ modifications guided by snR3, snR34 or snR36 (**Figure 4.5B**).

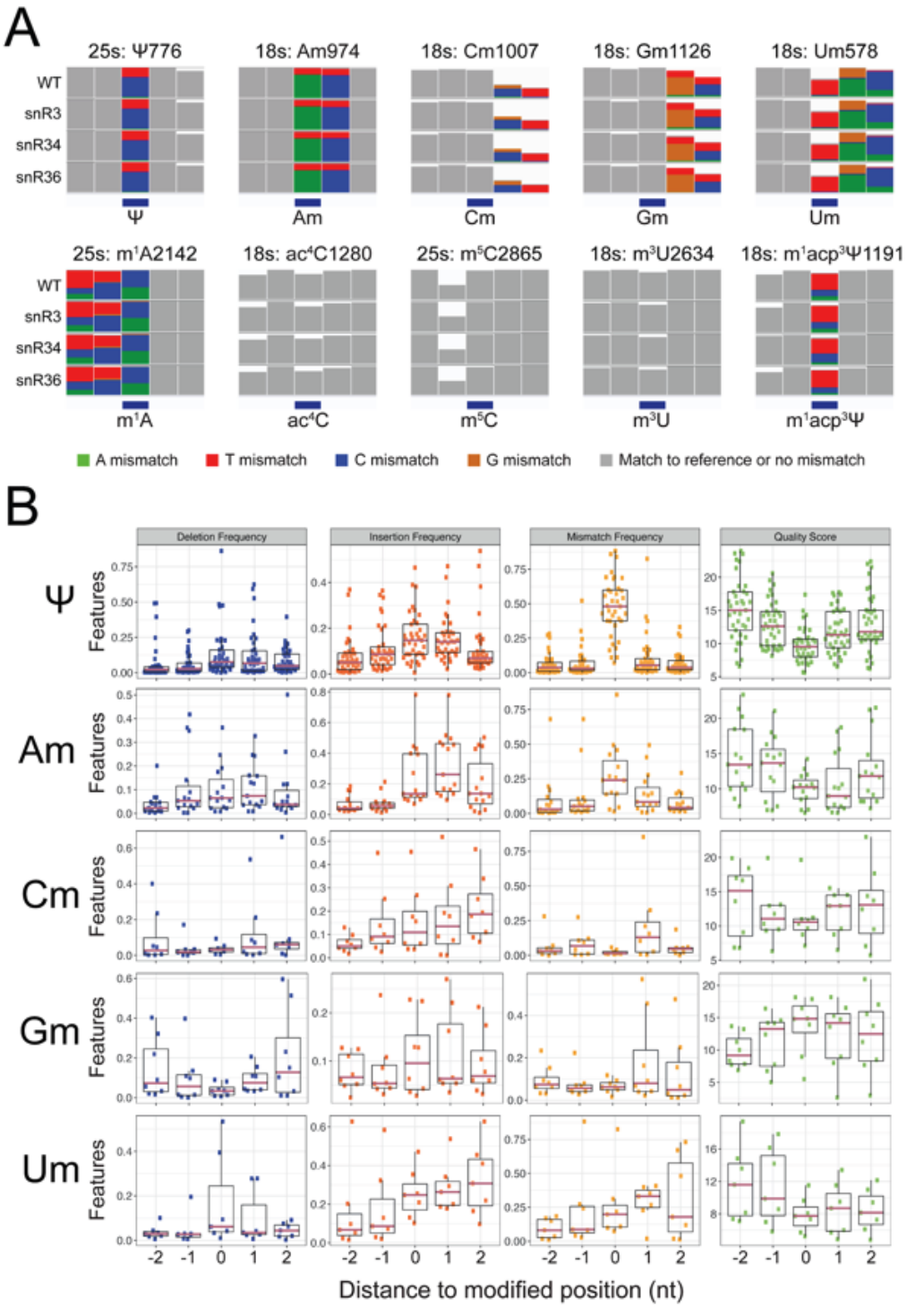


Figure 4.4 - Known yeast ribosomal RNA modifications show distinct base-calling 'error' signatures.

(A) IGV snapshots centered in distinct yeast ribosomal RNA modifications in 4 different yeast strains (wild type, snR3-KO, snR34-KO, snR36-KO, in descending order). Known rRNA modification sites are indicated below each snapshot. Nucleotides with mismatch frequencies greater than 0.15 have been colored. **(B)** Dotplots of base-calling errors

(deletion frequency, insertion frequency, mismatch frequency, and per-base quality) observed in modified 5-mers, centered in the modified position. Each dot corresponds to a different 5-mer. The total number of 5-mers included in the analysis varies depending on the abundance of each rRNA modification type in yeast rRNAs: Ψ (n=46), Am (n=14), Cm (n=10), Gm (n=15) and Um (n=9). 5-mers that contain more than one modification in the 5-mer region were excluded from the analysis. Box, first to last quartiles; whiskers, 1.5x interquartile range; center line, median; points, individual data points.

Strain	snoRNA	Modification	Sites	Predicted stoichiometry based on LC-MS/MS (%)*
snR3-KO	snR3	Ψ	25S:Ψ2129	25S:2129 - 91%
			25S:Ψ2133	25S:2133 - 100%
			25S:Ψ2264	25S:2264 - 95%
snR34-KO	snR34	Ψ	25S:Ψ2826	25S:2826 - 87%
			25S:Ψ2880	25S:2880 - 100%
snR36-KO	snR36	Ψ	18S:Ψ1187	18S:1187 - 95%
snR60-KO	snR60	Am,Gm	25S:Am817	25S:817 - 89%
			25S:Gm908	25S:908 - 100%
snR61-KO	snR61	Am	25S:Am1133	25S:1133 - 88%
snR62-KO	snR62	Um	25S:Um1888	25S:1888 - 95%

Table 4.1 - List of snoRNA mutant yeast strains used in this work, including their described rRNA targets. Data taken from Taoka et al.,2016 [367]

3 additional *S. cerevisiae* strains depleted of snoRNAs (snR60, snR61 and snR62, respectively) guiding 2'-O-methylation (Nm) at specific positions were then sequenced (**Table 4.1**). In contrast to Ψ modifications, I found that 2'-O-methylations often caused increased errors not only at the modified position, but also at neighbouring positions (**Figure 4.5C,** see also **Figure 4.6A**). These errors disappeared in the knockout strain, confirming that neighbouring base-calling errors were indeed caused by the 2'-O-methylation (**Figure 4.5C**). On the other hand, while Ψ

modifications mainly affected mismatch frequency, I observed that Nm modifications often affected several base-called 'error' features (mismatch, insertion and deletion frequency) (**Figure 4.6B**). Thus, we reasoned that combining all three features might improve the signal-to-noise ratio for the detection of 2'-O-methylated sites, and found that the combination of features led to improved detection of Nm-modified sites (**Figure 4.5D**).

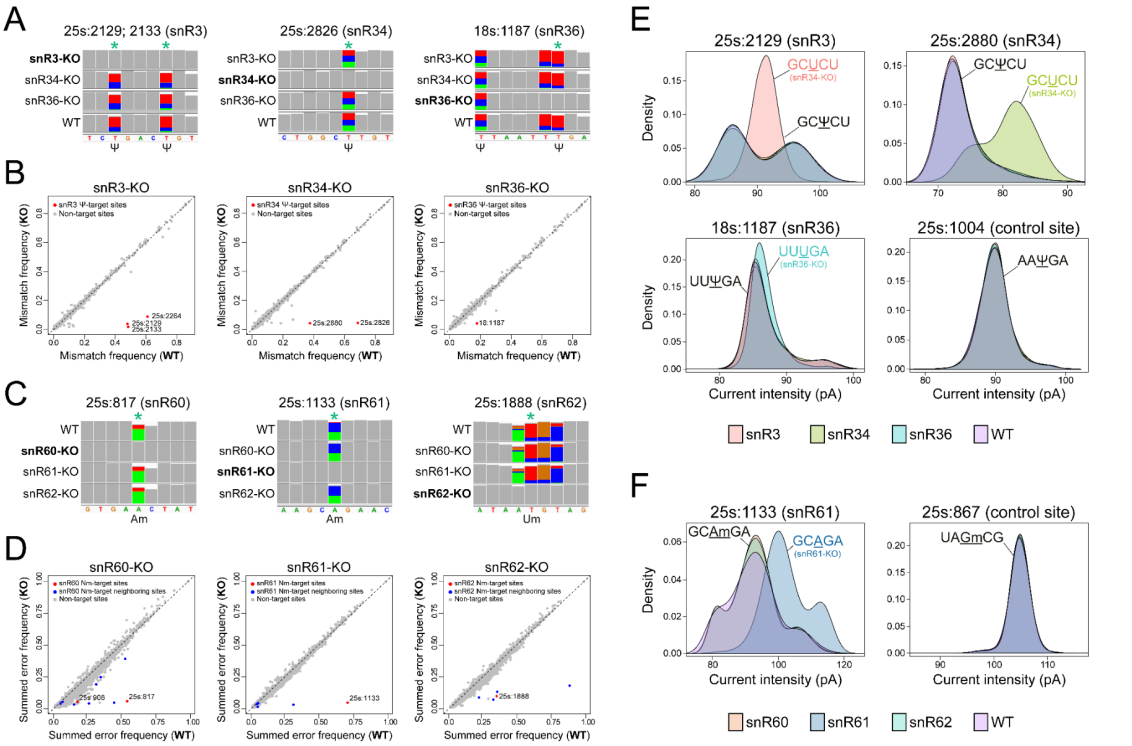


Figure 4.5 - Pseudouridylation and 2'-O-methylations cause systematic base-calling 'errors' as well as altered current intensities, and their signature disappears upon depletion of snoRNAs guiding the modification.

(A) IGV snapshots of wild type and three snoRNA-depleted strains depicting the site-specific loss of base-called errors at known Ψ target positions (indicated by asterisks). Nucleotides with mismatch frequencies greater than 0.1 have been colored. **(B)** Comparison of snoRNA knockout mismatch frequencies for each base, relative to wild type, with snoRNA targets sites indicated in red, and non-target sites in gray. **(C)** IGV snapshots of wild type and three snoRNA knockout yeast strains depicting the site-specific loss of base-calling errors at known Nm target positions. Nucleotides with mismatch frequencies greater than 0.1 have been colored. **(D)** Comparison of snoRNA knockout summed error frequencies for each base, relative to wild type, with snoRNA targets sites indicated in red, neighboring sites in blue and non-target sites in gray. **(E,F)** Distributions of per-read current intensity at known Ψ-modified (E), 2'-O-methylated (F) and negative control sites. Current intensities at Ψ and 2'-O-methylated positions were altered upon deletion of specific snoRNAs relative to wild type, whereas no shift was observed in control sites.

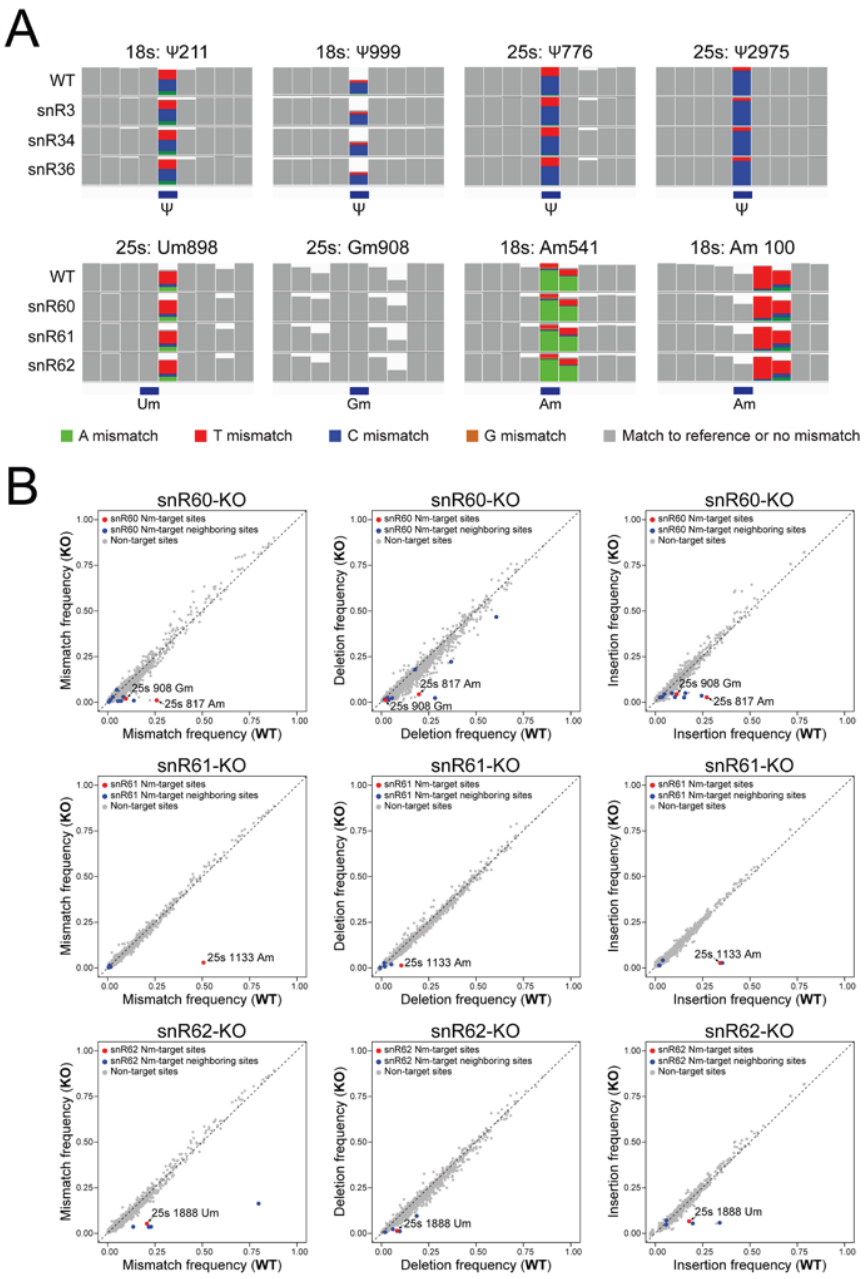


Figure 4.6 - Base-calling signature of 2'-O-methylations often alter the neighboring positions, whereas Ψ modifications mainly affect the modified site.

(A) IGV snapshots centered on known yeast rRNA modified sites: Ψ-modified sites are shown in the upper panels, whereas 2'-O-methylated sites are shown in the bottom panels. Nucleotides with mismatch frequencies greater than 0.15 have been colored. **(B)** Comparison of base-calling 'errors' (mismatch, deletion and insertion frequency) observed in snoRNA-depleted strains (snR60, top panels; snR61, middle panels, snR62, bottom panels) relative to wild type, with snoRNA target sites indicated in red, neighboring sites indicated in blue and non-target sites in gray.

4.5. Current intensity variations cannot accurately predict the modified site

We then wondered whether Ψ and Nm sites would also be detected at the level of current intensity changes. Certain Ψ and Nm-modified sites, such as 25S:Ψ2129 or 25S:Am1133, showed drastic alterations in their current intensity values in the snoRNA-depleted strain, while no significant alteration was observed in control sites (**Figure 4.5E,F**). However, in other sites the distribution of current intensities did not significantly change in the knockout strain (18S:Ψ1187, **Figure 4.5E** lower panel) or did not differ in their mean (25S:Ψ2133, **Figure 4.7A**).

We hypothesised that deviations in current intensity alterations might not always be maximal in the modified site, but might sometimes appear in neighboring sites. To test this, I examined the difference in current intensity values along the rRNA molecules for each wild type-knockout pair (**Figure 4.8A**, see also **Figure 4.7B**). However, the highest deviations in current intensity were often not observed at the modified sites (**Figure 4A** lower panel)**.** From all 6 Ψ sites that were depleted in the 3 knockout strains studied, only 2 of them (25S:Ψ2826 and 25S:Ψ2880) showed a maximal deviation in current intensity in the modified site (**Figure 4.8B,** see also **Figure 4.7C**). Similarly, depletion of Nm sites led to changes in current intensity values, but the largest deviations were not observed at the modified site (**Figure 4.7C**). In conclusion, current intensity-based methods can detect both Ψ and Nm RNA modifications; however, base-calling errors are a better choice to achieve single nucleotide resolution, at least in the case of Ψ RNA modifications.

4.6. Detection of Ψ and Nm modifications in individual reads

Direct RNA sequencing produces current intensity measurements for each individual native RNA molecule. Thus, modification stoichiometries can be, in principle, estimated by identifying the proportion of reads with altered current intensity at a given site. To this end, I first examined the per-read current intensity values of wild type and knockout strains at the Ψ- and Nm-depleted sites. Despite the significant variability of current intensities across reads, robust differences in current intensities across strains at the depleted RNA modified sites at the per-read level were observed (**Figure 4.8C,** upper panel). As a control, I performed the same analysis in Ψ sites unaffected by snoRNA depletion, finding no significant differences between wild type and knockout

strains (**Figure 4.8C,** lower panel). However, in some sites such as 18S:1187, the per-read shifts in current intensity between the wild type and knockout strain were far more modest (**Figure 4.7D**). Principal Component Analysis (PCA) of the current intensity values of 15-mer regions that contained the modified site showed that the reads clustered into two distinct populations: the first population mainly comprised unmodified reads from the snoRNA-depleted strain, whereas the second comprised reads from the 3 other strains, which are mostly modified (**Figure 4.8C** right panels, see also **Figure 4.7E**).

Suprisingly, *Nanopolish* software did not resquiggle the reads evenly across sites. For example, it failed at resquiggling the majority of reads in the region surrounding 25S:Ψ2264 **(Figure 4.7D**). Thus, we wondered whether *Tombo*, which uses global resquiggling instead of local resquiggling, might overcome this limitation. *Tombo* resquiggling led to a global increase in the proportion of resquiggled reads (**Figure 4.9A**). Moreover, *Tombo* showed a uniform proportion of resquiggled reads along the same transcript, whereas *Nanopolish* showed a variable proportion of resquiggled reads depending on the site. Notably, *Tombo* was equally effective at resquiggling both modified and unmodified reads, whereas *Nanopolish* preferentially resquiggled unmodified reads relative to modified ones, biasing the unmodified:modified proportion up to 7:1 (**Figure 4.9B**). This uneven resquiggling from *Nanopolish* implies that using *Nanopolish* for predicting RNA modification levels at individual sites may cause a dramatic bias in the predicted stoichiometry of individual sites. Thus, based on these results, *Tombo* resquiggling was adopted instead of *Nanopolish* resquiggling for the prediction of RNA modification stoichiometries from individual RNA reads in all our downstream analyses.

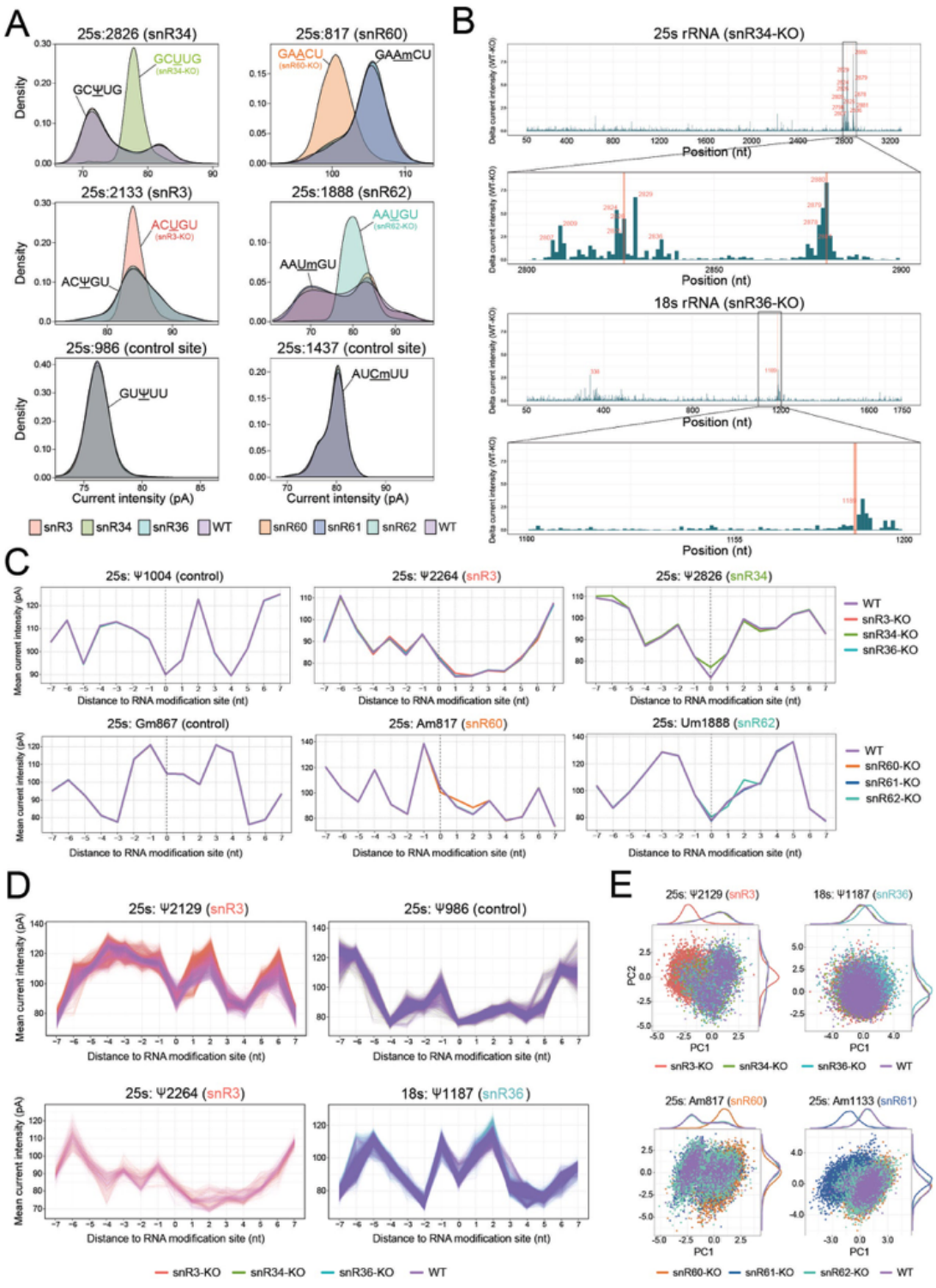


Figure 4.7 - Pseudouridylations and 2'-O-methylations can be detected in the form of altered current intensities.

(A) Distributions of per-read current intensity at known Ψ-modified, 2'-O-methylated and negative control sites. Ψ and 2'-O-methylated positions were altered upon deletion of specific snoRNAs relative to wild type, whereas no shift was observed in control sites. **(B)** Absolute differences in current intensity along the 25S rRNA and 18S rRNAs

upon depletion of snR34 and snR36, respectively, relative to the wild type strain. Red vertical lines indicate the KO pseudouridylation positions. **(C)** Comparison of mean current intensity changes for Ψ and 2'-O-methyl knockout sites across each of the snoRNA knockout strains. The dotted vertical line indicates the modified position. **(D)** Per-read analysis of current intensities centered at 3 different Ψ modified sites targeted by the snoRNAs depleted in each knockout strain (25S:Ψ2129, 25S:Ψ2264 and 18S:Ψ1187). In each panel, the per-read current intensities centered in the modified site are shown, both for the wild type (purple) and knockout strain (red: snR3; green: snR34; cyan: snR36). As a control, the same analysis was performed at a control site (25S:Ψ986), using reads from wild type (purple) and snR34 knockout strain (green) showing no differences between the read populations. Each line indicates a single read. **(E)** Principal Component Analysis of the current intensity values of the 15-mer regions was performed, and the corresponding scatterplots of the two first principal components (PC1 and PC2) are depicted for 4 different Ψ and Nm sites. Each dot corresponds to a different read, and is colored according to the strain.

4.7. Stoichiometry prediction using signal intensity, dwell time and trace

Ψ and Nm modifications can lead to significant alterations in the current intensity profiles at the modified region (e.g. 25S:Ψ2880, **Figure 4.8B-C**). However, in other sites such as 18S:Ψ1187, current intensity alone was insufficient to bin the reads into two separate clusters (**Figure 4.7D,E**), suggesting that, in addition to current intensity, other features might be needed to distinguish modified from unmodified reads.

Previous works predicting DNA modifications from individual nanopore reads typically relied on features such as signal intensity or dwell time to distinguish modified and unmodified read populations [368–371]. Here, in addition to these two features, we explored whether the use of 'trace' would improve our ability to predict RNA modification stoichiometry. Trace (also termed 'base probability') represents the probability that a given signal intensity chunk may be originating from each of the 4 canonical bases (A, C, G and T/U). To this end, we first examined how the presence of Ψ and Nm modifications altered each of the features (signal intensity, dwell time and trace) in Ψ and Nm modified sites, both at snoRNA-targeted positions and control sites (**Figure 4.10**). In addition to signal intensity, base probability (trace) was significantly different in all examined sites. Moreover, in some sites such as 25S:Ψ2264, trace was the most altered feature from those examined. By contrast, dwell time was not consistently different in snoRNA-targeted sites relative to wild type (e.g. 25S:Ψ2264, 25S:Ψ2826, 18S:Ψ1187).

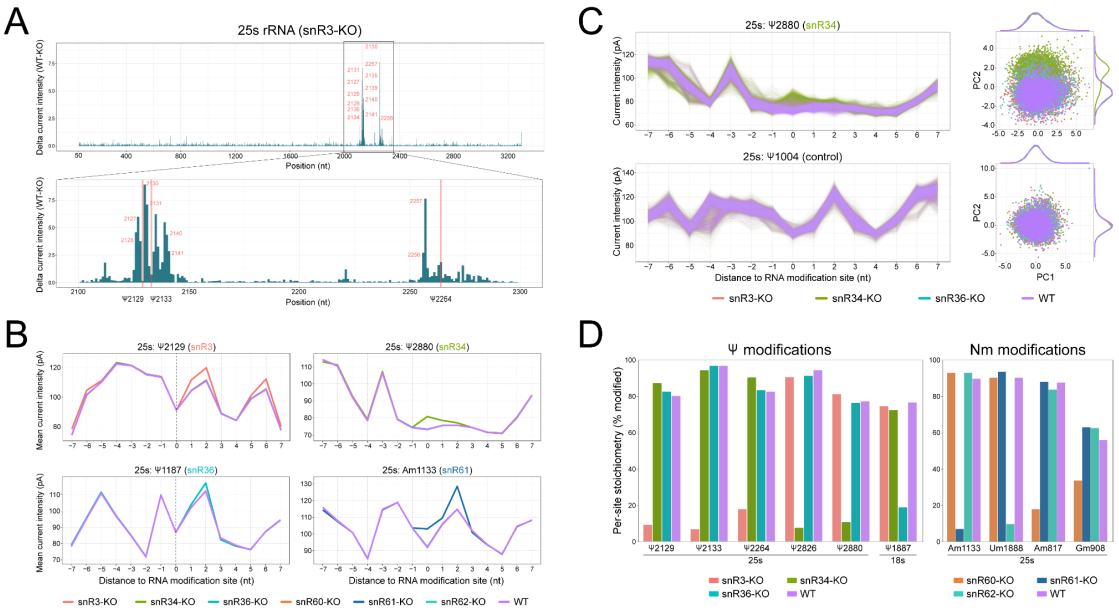


Figure 4.8 - Loss of specific Ψ rRNA modifications causes deviations in current intensity in regions surrounding the Ψ sites.

(A) Current intensity changes along the 25S rRNA molecule upon snR3 depletion, relative to the wild type strain. In the lower panel, a zoomed subset focusing on the two regions with the most significant current intensity deviations is shown; the first one comprising the 25S:Ψ2129 and 25S:Ψ2133 sites, and the second one comprising the 25S:Ψ2264 site. **(B)** Comparison of current intensities in the 15-mer regions surrounding Ψ and 2'-O-methyl knockout sites, for each of the 4 strains. The dotted vertical line indicates the modified position. See also Figure S4 for current intensity changes in other knockout strains and sites. **(C)** Per-read current intensity analysis centered at the 25S:Ψ2880 site targeted by snR34 (upper panel) and a control site, 25S:Ψ2880, which is not targeted by any of the knockouts (lower panel). For each site, Principal Component Analysis was performed using 15-mer current intensity values, and the corresponding scatterplot of the two first principal components (PC1 and PC2) is shown on the right, using as input the same read populations as in the left panels. Each dot corresponds to a different read, and is colored according to the strain. **(D)** Predicted stoichiometry of Ψ- and Nm-modified sites using a k-nearest neighbors (KNN) algorithm trained to classify the reads into 2 classes: modified or unmodified. The features used to predict modifications status of every read from which stoichiometry was calculated were signal intensity (positions -1,0,+1) and trace (positions -1,0,+1).

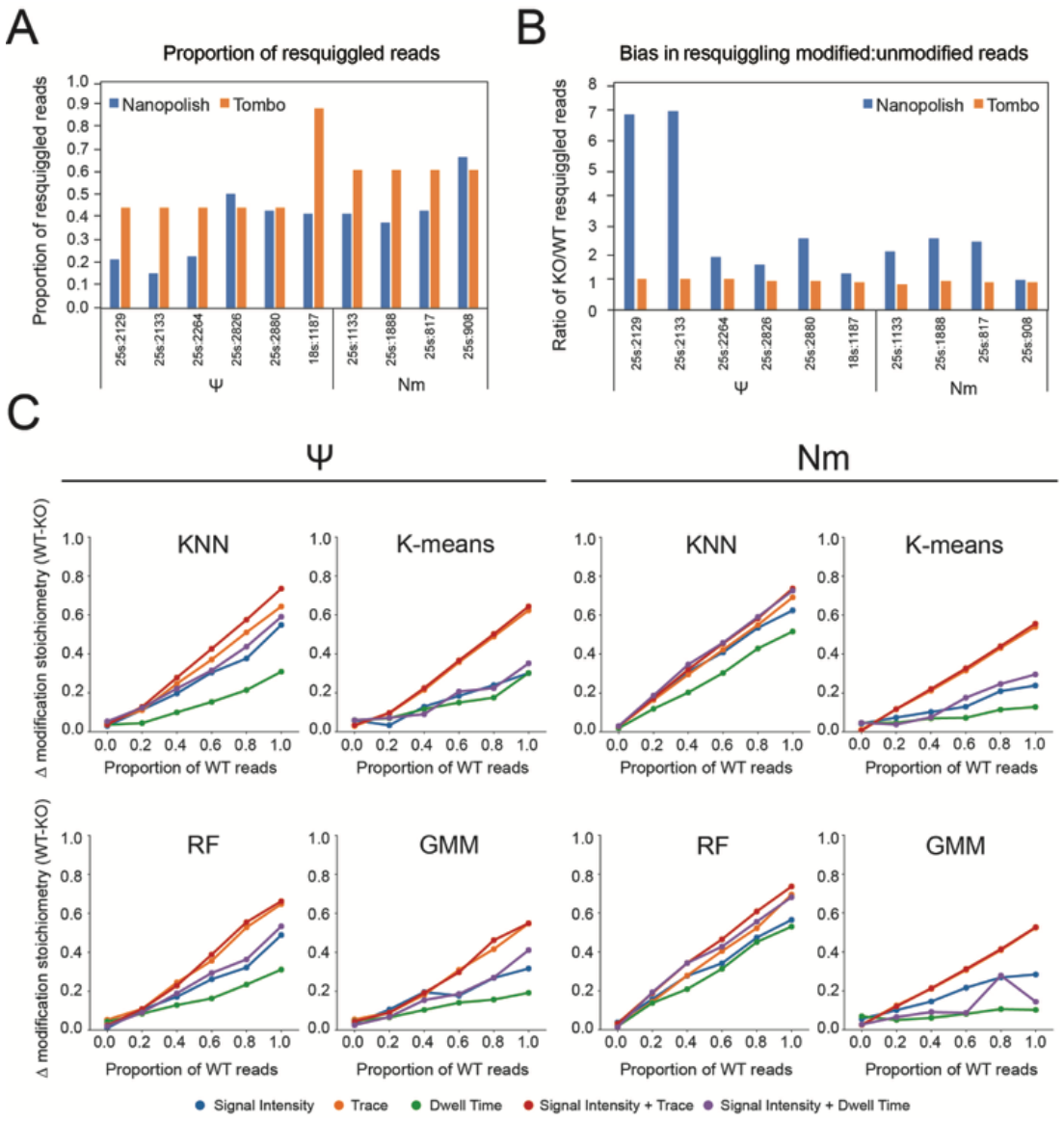


Figure 4.9 - Systematic benchmarking of resquiggling softwares, machine learning algorithms and distinct feature sets for the prediction of RNA modification stoichiometry from individual RNA reads.

(A) Comparative analysis of read resquiggling using *Nanopolish* and *Tombo,* depicting the relative proportion of resquiggled reads for each algorithm at each individual site. The Ψ-modified sites and 2'-O-methylated sites were analysed independently, as they come from independent flowcells. *Tombo* shows uniform proportion of resquiggled reads along the same transcript, whereas *Nanopolish* shows variable proportion of resquiggled reads depending on the site. **(B)** Comparative analysis of read resquiggling using *Nanopolish* and *Tombo*, depicting the relative proportion of resquiggled reads from KO strains for a given position (relative to WT), using as input 1000 reads for each strain, and for each algorithm. If there is no difference in resquiggling depending on the presence or absence of modification, the expected proportion of KO:WT reads is 1. **(C)** Line chart of expected (X-axis) and observed (Y-axis) modification frequency for Ψ-modified and 2'-O-methylated positions. The absolute modification frequency difference

was estimated between two samples: KO (no modified reads) and WT (simulating varying levels of modification frequency: 0.0, 0.2, 0.4, 0.6, 0.8 and 1.0). Modification stoichiometry was calculated using four different machine learning methods: two supervised (k-nearest neighbour (KNN) and random forest (RF)) and two unsupervised (K-means and gaussian mixture model (GMM). Five sets of distinct feature combinations have been tested for each algorithm and RNA modification type: current intensity (blue), trace (orange), dwell time (green), the combination of current intensity and trace (red) and the combination of current intensity and dwell time (purple).

The use of distinct features for RNA modification stoichiometry was then systematically benchmarked. To this end, *nanoRMS*, a software that extracts signal intensity, trace and dwell time from individual reads, and then predicts RNA modification stoichiometry by using distinct feature combinations as well as various machine learning algorithms was built. Firstly, different mixes of modified (wild type) and unmodified (knockout) reads were generated to simulate varying read stoichiometry (0, 20, 40, 60, 80 and 100%), for each of the Ψ and Nm positions for which knockouts were available (**Table 4.1**). Then, we examined how different supervised and unsupervised algorithms would predict the stoichiometry of each of the sites, and using distinct combinations of the 3 features (signal intensity, trace and dwell time) for each individual site (**Figure 4.9C**). These results show that the combination of signal intensity and trace outperformed all the other feature combinations for predicting both Ψ and Nm modification stoichiometry, and that the supervised k-nearest neighbor (KNN) was the best performing algorithm. The k-means clustering algorithm (KMEANS) was the best-performing algorithm among the unsupervised clustering methods tested, although its performance in predicting Ψ modification stoichiometry was slightly better than in the case of Nm modification stoichiometry predictions. Overall, *nanoRMS* can accurately predict Ψ and Nm RNA modification stoichiometry from individual RNA reads (**Figure 4.8D**), with predicted stoichiometry values that are similar to those that have been previously reported by Mass Spectrometry [367].

4.8. *De novo* prediction reveals a Pus4-dependent mitochondrial Ψ rRNA modification

The identification of RNA modification-specific signatures allows us to perform *de novo* prediction of Ψ RNA modifications transcriptome-wide using direct RNA sequencing. In this regard, *S. cerevisiae* mitochondrial rRNAs remains much less characterised than cytosolic rRNAs, with only 3 modified sites identified so far in *S. cerevisiae* LSU (21s) [372], and none in SSU (15S) rRNAs. Thus, we hypothesised that

direct RNA might reveal previously uncharacterised Ψ-modified sites in mitochondrial rRNAs. To this end, I first determined the 'error'-based thresholds (mismatch frequency and C mismatch frequency) that would distinguish unmodified uridines from pseudouridines in cytosolic rRNAs (**Figure 4.11A**). I then applied this filter to predict Ψ modifications on 15S rRNA and 21s rRNA, identifying two novel candidate Ψ sites (15S:854 and 15S:579) that displayed high modification frequency as well as U-to-C mismatch signature (**Figure 4.11B,C**).

To further confirm that the two predicted 15S rRNA sites are pseudouridylated, I developed nanoCMC-seq, a protocol that identifies Ψ modifications by coupling CMC probing with nanopore cDNA sequencing. This method allows capturing reverse-transcription drop-off information by sequencing only the first-strand cDNA molecules of CMC-probed RNAs using a customised direct cDNA sequencing protocol (**Figure 4.11D,** see also *Methods*). NanoCMC-seq captured known sites in cytoplasmic rRNA with a very high signal-to-noise ratio, as well as confirmed the existence of Ψ in position 854 and 579 of 15S rRNA, validating the *de novo* predictions using direct RNA sequencing (**Figure 4.11E,** see also **Figure 4.12A**).

I observed that 15S:Ψ854 was embedded in a similar sequence context and structure as the t-arm of tRNAs, which contains a pseudouridylated (Ψ55) position placed by Pus4 (**Figure 4.11F**). Given the resemblance between these two sequences and structures, I hypothesised that Pus4 might be responsible for this modification. To validate this hypothesis, I sequenced total RNA from a Pus4-deficient *S. cerevisiae* strain, finding that the 15S:854 position loses its mismatch signature upon deleting Pus4 gene, confirming that this site is not only pseudouridylated, but also that it is Pus4-dependent (**Figure 4.11G,** see also **Figure 4.12B)**. Additionally, I observed that previously reported Pus4 target sites (TEF1:239,TEF2:239) [121,122,126] completely lost their mismatch signature in Pus4 knockout cells **(Figure 4.12B,C)**, confirming that this method is able to capture previously reported Pus4-dependent Ψ sites, in addition to previously unknown ones.

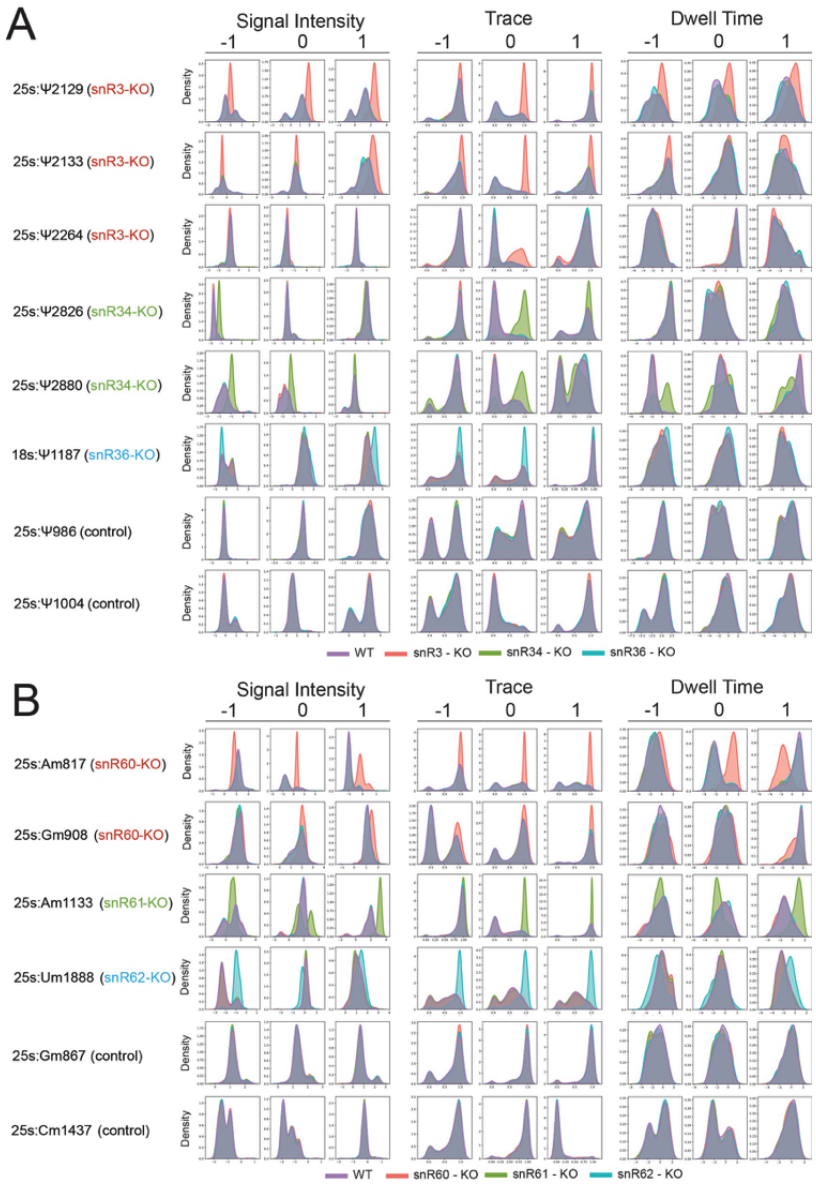


Figure 4.10 - Density plots of the per-read current intensity, trace and dwell time features in selected Ψ and 2'-O-methylated rRNA sites.

(A,B) Per-read distributions of current intensity (SI), trace (TR) and dwell time (DT) between respective mutants and wild type at Ψ-modified **(A)**, and 2'-O-methylated sites **(B)**. Control sites, which are not affected by any of the knockouts, have been included in the analysis as negative controls. The distributions are plotted for the positions of interest (0) and two neighbouring positions: downstream (-1) and upstream (+1). The density distribution of each feature has been colored depending on the strain. X-axis scale depends on the feature type: SI is reported as median absolute deviation normalised signal intensity as reported by *Tombo*; TR is reported as reference-base probability (0-1 scaled), and DT is reported as log2 (observed/expected), where expected is dwell time mean value per base calculated for every read.

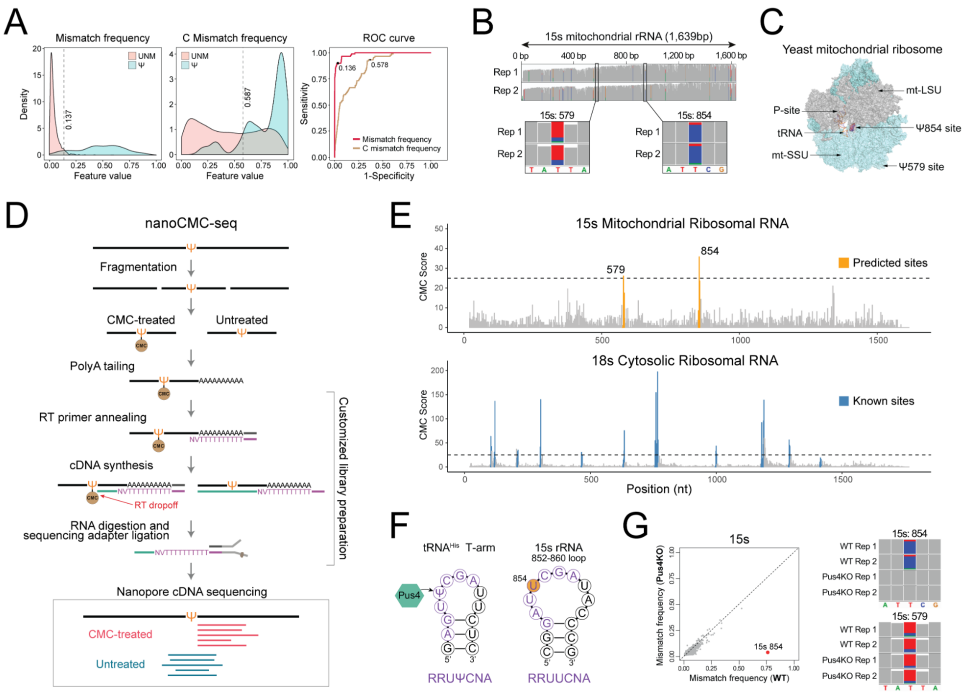


Figure 4.11 - ***De novo* prediction of Ψ modifications reveals a novel Pus4-dependent mitochondrial rRNA modification**. **(A)** Density distributions of mismatch and C mismatch frequency in unmodified uridine (red) and pseudouridine (cyan) positions. The dashed lines represent the optimal cutpoints between two groups determined by maximizing the Youden-Index. In the right panel, the ROC curve illustrates the sensitivity and specificity at these two cutpoints. **(B)** IGV coverage tracks of the 15S mitochondrial rRNA, including a zoomed version showing the tracks centered at the 15S:854 and 15S:579 sites, in two biological replicates. Nucleotides with mismatch frequencies greater than 0.15 have been colored. **(C)** Location of the putative Ψ854 modified site in the yeast mitochondrial ribosome. The PDB structure shown corresponds to 5MRC. **(D)** Validation of putative Ψ sites with nanoCMC-Seq, which combines CMC treatment with Nanopore cDNA sequencing in order to capture RT-drops that occur at Ψ-modified sites upon CMC probing. RT-drops are defined by counting the number of reads ending (3') at a given position. **(E)** Predicted Ψ sites U854 and U579 (orange) in the 15S rRNA are validated using nanoCMC-seq (upper panel). Dashed lines indicate the CMC-score threshold used for determining the positive sites (upper panel). As a control, we analysed the nanoCMC-seq results in other rRNAs (lower panel), finding that all positions with a significant CMC Score (>25) correspond to known Ψ rRNA modification sites (blue). See also Figure S7A for CMC scores in additional rRNA transcripts. **(F)** The candidate Ψ854 site is located at the 852-860 loop of the 15S rRNA, which resembles the t-arm of the tRNAs that is modified by Pus4. The binding motif of Pus4 (RRUUCNA) matches the motif surrounding the 854U site[121]. **(G)** Scatterplot of mismatch frequencies in WT and Pus4KO cells, showing that the only significant position affected by the knockout of Pus4 is 15S:U854 (left panel). IGV coverage tracks showing that Pus4 knockout leads to depletion of the mismatch signature in the 15S:854 position (right panel), but not at the 15S:579 position.

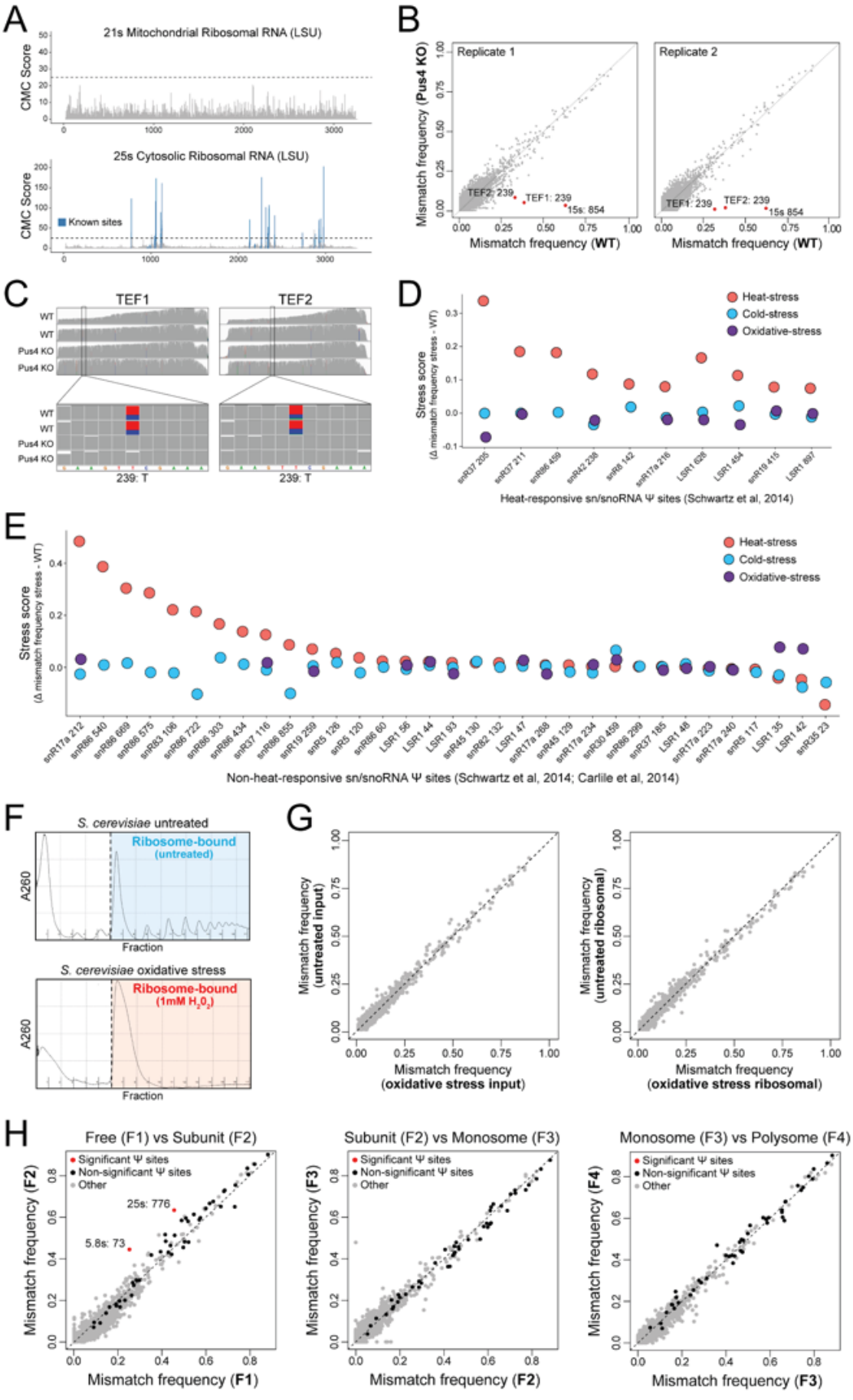

A
21s Mitochondrial Ribosomal RNA (LSU)
CMC Score
25s Cytosolic Ribosomal RNA (LSU)
Known sites
B
Replicate 1
Replicate 2
Mismatch frequency (Pus4 KO)
Mismatch frequency (WT)
TEF2: 239
TEF1: 239
15s: 854
C
TEF1
TEF2
WT
Pus4 KO
239: T
D
Heat-stress
Cold-stress
Oxidative-stress
Stress score
(Δ mismatch frequency stress - WT)
Heat-responsive sn/snoRNA Ψ sites (Schwartz et al, 2014)
E
Non-heat-responsive sn/snoRNA Ψ sites (Schwartz et al, 2014; Carlile et al, 2014)
F
S. cerevisiae untreated
Ribosome-bound
(untreated)
A260
Fraction
S. cerevisiae oxidative stress
Ribosome-bound
(1mM H2O2)
G
Mismatch frequency
(untreated input)
Mismatch frequency
(oxidative stress input)
Mismatch frequency
(untreated ribosomal)
Mismatch frequency
(oxidative stress ribosomal)
H
Free (F1) vs Subunit (F2)
Subunit (F2) vs Monosome (F3)
Monosome (F3) vs Polysome (F4)
Significant Ψ sites
Non-significant Ψ sites
Other
25s: 776
5.8s: 73
Mismatch frequency (F1)
Mismatch frequency (F2)
Mismatch frequency (F3)

Figure 4.12 - *De novo* prediction of Ψ modifications reveals a novel Pus4-dependent modification (15S:Ψ854) in yeast mitochondrial rRNAs, and captures previously reported Pus4-dependent mRNA modifications.

(A) NanoCMC-seq scores along the 21s mitochondrial LSU rRNA (upper panel) and the 25S cytosolic LSU rRNA (bottom panel). Dashed lines indicate the CMC-score threshold used for determining the positive sites. All the positions with a significant CMC Score (>25) correspond to known Ψ rRNA modification sites (blue). **(B)** Comparison of mismatch frequency for each base in Pus4 knockout strains, relative to wild type, in positions mapped to yeast genome and rRNA, in two independent biological replicates. **(C)** IGV snapshots of wild type (rep1 and rep2) and Pus4 knockout (rep1 and rep2) yeast strains with zoomed subsets depicting the site-specific loss of mismatch at known target positions. Nucleotides with mismatch frequencies greater than 0.15 have been colored. **(D)** Stress scores in previously reported heat-responsive sn/snoRNA pseudouridylated sites (as defined by Schwartz et al 2014). Stress scores are calculated by taking the difference between mismatch frequency in stress (heat-shock, cold-shock, and oxidative) and normal conditions. (**E)** Stress scores in sn/snoRNA Ψ sites that were not previously reported as heat-responsive. Our analysis identifies some of these sites as responsive to heat-stress. **(F)** Polysome profiles of ribosomal-bound RNA fractions isolated from untreated and stressed H_2O_2-treated yeast cells. **(G)** Comparison of mismatch frequency for untreated vs H_2O_2-treated input RNA (upper panel) and untreated vs H_2O_2-treated ribosome-bound RNA (lower panel). **(H)** Comparison of mismatch frequencies of ribosomal RNAs for different fractions (F1: Free, F2: Subunit, F3: Monosome, F4: Polysome). Each dot represents a base in the rRNA, with significantly altered Ψ sites reproducible across biological replicates highlighted in red. The remaining Ψ sites are shown in black and the rest of the sites in gray. All rRNA bases from cytosolic rRNAs were included in the analysis and plots.

4.9. rRNA modification profiles do not vary upon oxidative or thermal stress

Ribosomal RNAs are extensively modified as part of their normal maturation, and their modification landscape is relatively well-defined for a series of organisms [39,53–57]. Despite the central role that rRNA molecules play in protein translation, recent evidence has shown that rRNA modifications are in fact dynamically regulated [373,374], and that their alterations can lead to disease states [232,233,362,375–380]. Moreover, the stoichiometry of some pseudouridylated and 2′-O-methylated rRNA sites is cell-type dependent, suggesting that rRNA modifications may be an important source of ribosomal heterogeneity [363,367,381–384]. However, a systematic and comprehensive analysis of which environmental cues may lead to changes in rRNA

modification stoichiometries, which RNA modifications may be subject to this tuning, and to which extent, is largely missing.

To assess whether rRNA modification profiles change in response to environmental stimuli, *S. cerevisiae* cells were treated with diverse environmental cues (oxidative, cold and heat stress) and their RNA was sequenced using direct RNA sequencing. Firstly, I confirmed that the rRNA modification profiles from independent biological replicates were highly reproducible (Pearson r2=0.976-0.996, see also **Figure 4.13**). Then, I examined whether exposure to stress would lead to significant changes in base-calling 'errors' in rRNA molecules, finding no significant differences in rRNA modification profiles between normal and stress conditions (**Figure 4.14A**). In contrast, I recapitulated previously reported changes in snRNA Ψ modifications upon exposure to environmental cues [121] (**Figure 4.14B,** see also **Figure 4.12D**), as well as identified 8 additional Ψ modification sites in snRNAs and snoRNAs whose stoichiometry varies upon heat exposure, which had not been previously described (**Figure 4.14,** see also **Figure 4.12E)** [121,122,359,385]. Overall, this approach confirmed previous reports and predicted novel Ψ sites in ncRNAs whose modification levels vary upon heat shock exposure (**Figure 4.14B-D,** see also **4.12D-E**), but did not identify any rRNA modified site to be varying in its stoichiometry upon any of the tested stress conditions.

4.10. rRNA modification profiles do not vary across translational repertoires

Next, we questioned whether pseudouridylation changes in distinct translational repertoires may be more nuanced, in that Ψ levels may differ between rRNAs present in different translational fractions along a polysome gradient, which would not be detected when examining rRNAs as a whole. To test this, both total (input) and polysomal rRNAs from untreated and H_2O_2-treated yeast cells were sequenced (**Figure 4.12F**). However, I observed no significant changes in Ψ rRNA modification profiles when comparing rRNAs from actively translating ribosomes in untreated versus H_2O_2-treated cells (**Figure 4.12G**).

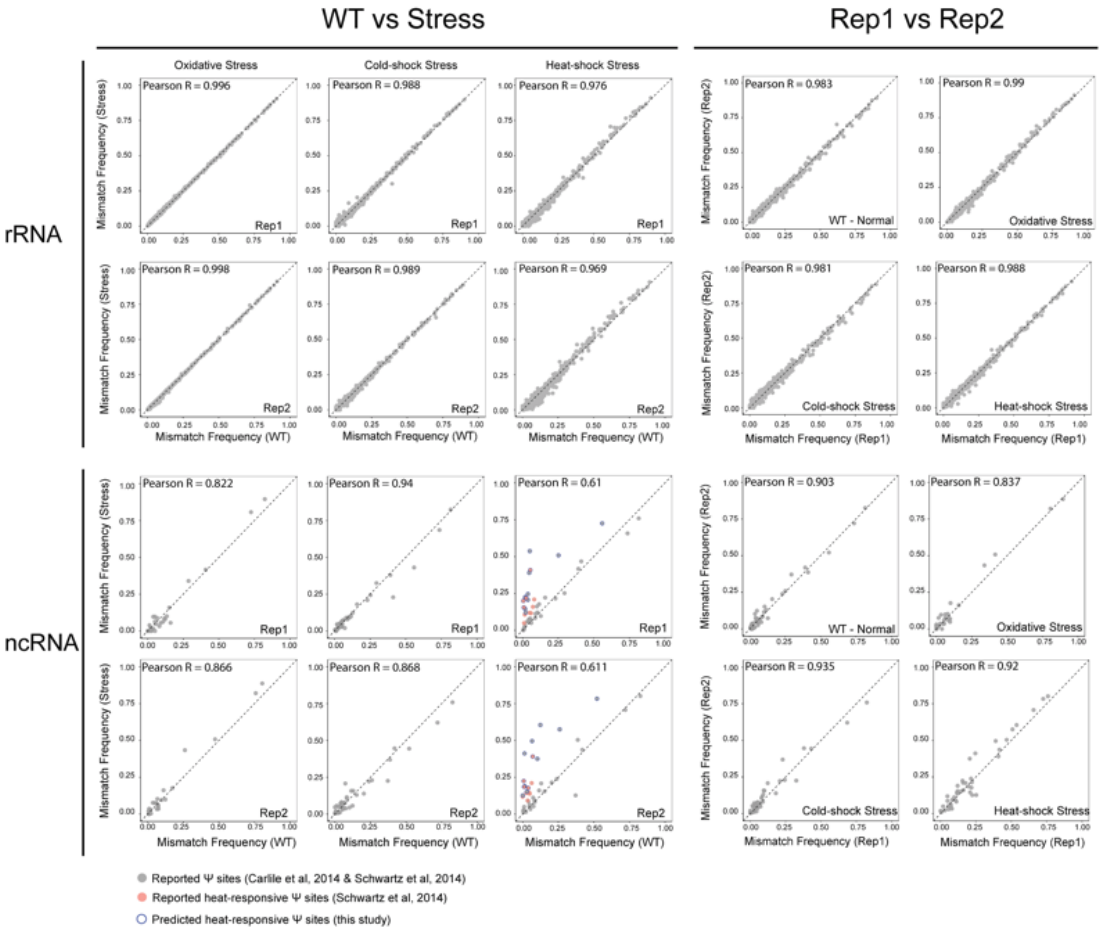


Figure 4.13 - rRNA and ncRNA modification profile reproducibility across biological replicates of *S. cerevisiae* cells under diverse environmental cues.

Pearson correlation coefficients are included for each pairwise comparison. Median Pearson across replicates was 0.985 for rRNA and 0.915 for ncRNAs. By contrast, the median pearson correlation between WT-heat in ncRNAs was 0.611 (replicate 1) and 0.611 (replicate 2), illustrating that there are changes in the ncRNA mismatch scores that are not due to low replicability of the biological replicates. Only sites with coverage >30 reads were included in the analysis. A site was considered to be stress-responsive if the mismatch score is significantly changing in both replicates, relative to the control condition. Each dot represents a uridine base. For rRNAs, all uridine bases were included in the plot, whereas for ncRNAs, only the reported ncRNA Ψ sites were included.

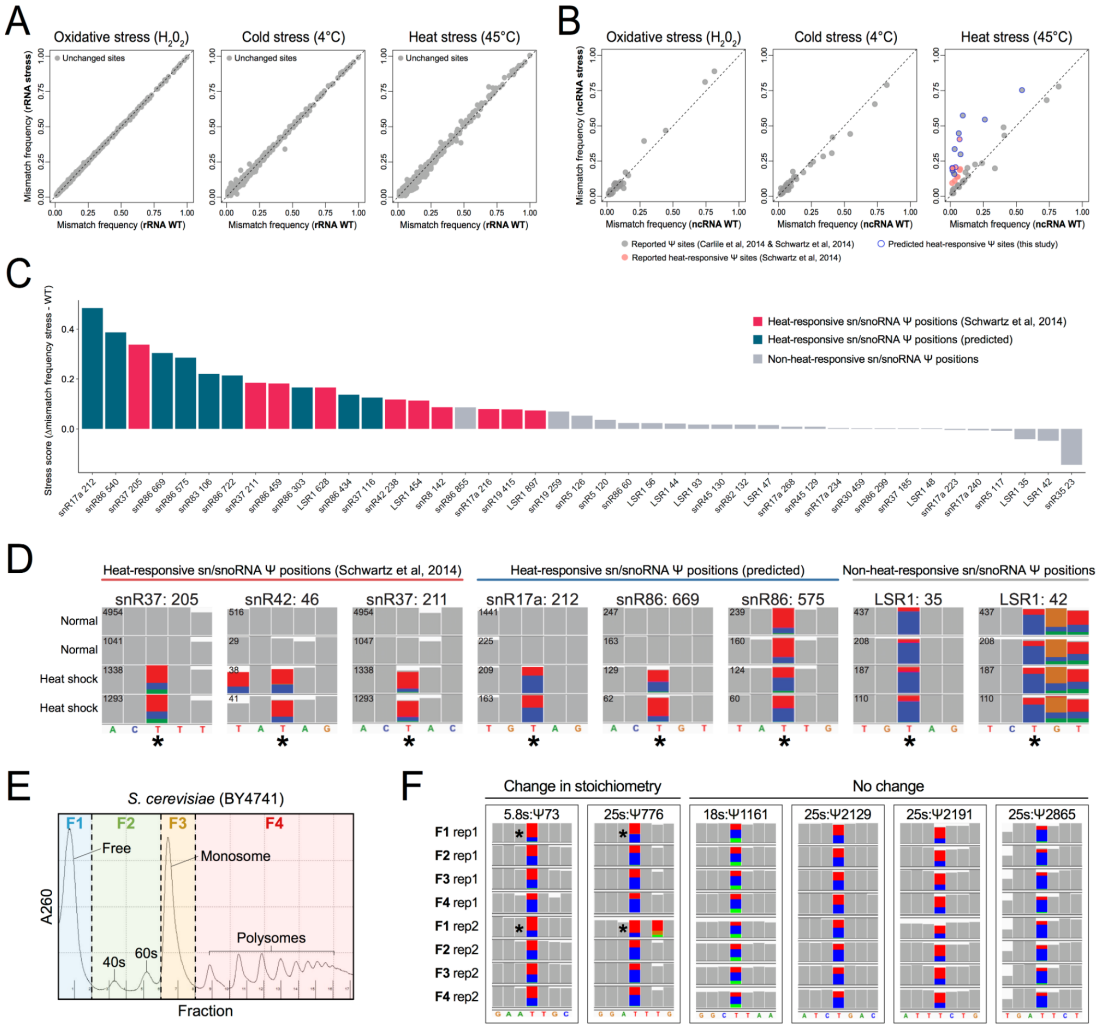


Figure 4.14 - Comparative analysis of yeast rRNA and snRNA Ψ modifications upon distinct environmental stresses identifies known and previously unknown heat-sensitive snRNA and snoRNA Ψ modifications.

(A) Comparison of mismatch frequencies for all rRNA bases from untreated or yeast exposed to oxidative stress (H_2O_2, left panel), cold stress (4ºC, middle panel) or heat stress (45ºC, right panel). Each dot represents a uridine base. All rRNA bases from cytosolic rRNAs were included in the analyses. **(B)** Comparison of mismatch frequencies in untreated versus stressed-exposed yeast cells (oxidative, cold or heat), in previously reported ncRNA Ψ sites[121,122]. **(C)** Stress scores in sn/snoRNA Ψ sites calculated by Δ mismatch frequency between heat shock and WT. **(D)** IGV snapshots of normal condition (rep1 and rep2) and heat shock condition (rep1 and rep2) yeast cells zoomed into the known sn/snoRNA Ψ positions (indicated by an asterisk). Nucleotides with mismatch frequencies greater than 0.1 have been colored. Coverage for each position/condition is given on the top left of each row. **(E)** Profiles of ribosomal fractions isolated from yeast grown under normal conditions, using sucrose gradient fractionation, including free rRNAs which are not assembled into ribosomal subunits (F1), rRNAs from 40s and 60s subunits (F2), rRNAs extracted from monosomal

fractions (F3) and polysome fractions (F4). **(F)** IGV snapshots of the two Ψ sites that change stoichiometry between translational fractions and four representative Ψ sites that show no significant change. Nucleotides with mismatch frequencies greater than 0.1 have been colored.

In an attempt to further dissect the different translational repertoires into a higher number of rRNA pools, we sequenced: i) rRNAs from unassembled free rRNA fractions (F1), ii) rRNAs from 40s and 60s subunits (F2), iii) rRNAs from monosomal fractions (F3) and iv) rRNAs from polysomal fractions (F4) (**Figure .14E**). While two positions showed slightly decreased levels of Ψ (5.8S:Ψ73 and 25S:Ψ776) in the free rRNA fraction (F1) compared to assembled ribosomes and/or subunits, no significant changes were observed across the other translational fractions (**Figure 4.14F,** see also **Figure 4.12H**). Globally, these results indicate that differential rRNA modifications are likely not the mechanism employed by yeast cells to adapt to environmental stress conditions, in agreement with previous observations [122].

4.11. *De novo* prediction of Ψ modifications in mRNAs

Ribosomal RNAs are modified at very high stoichiometries[367,384]. By contrast, other molecules such as mRNAs are modified at lower stoichiometries, making the detection of their RNA modifications a much more challenging task[248]. To ascertain whether this methodology would be applicable to lowly modified RNA sites, such as those present in mRNAs, the performance of *nanoRMS* was first assessed in RNA molecules that contained Ψ RNA modifications at low RNA modification stoichiometries (0, 3, 7 and 20%) (**Figure 4.15A**, see also *Methods*). The relative incorporation of Ψ RNA modifications was validated using Mass Spectrometry. Then the quantitative performance of *nanoRMS* under low stoichiometry conditions using both KNN and k-means were assessed, finding that the combination of signal intensity and trace features yielded the most accurate results in terms of stoichiometry prediction (**Figure 4.15B**), in agreement with the previous results (**Figure 4.9C**).

Next, I sequenced polyA(+)-selected RNA from *S. cerevisiae* wild type, Pus1 knockout, Pus4 knockout and heat stress-exposed strains using direct RNA sequencing, in biological duplicates. Considering that mRNA sites are lowly modified, I restricted the *de novo* identification of mRNA Ψ sites to those whose base-calling 'error' features significantly changed between pairwise conditions (**Figure 4.15C**, see also *Methods*), met the pseudouridine 'error' signature, and had a minimum coverage of 30 reads in both conditions and biological replicates (see *Methods*). Through this approach, I predicted 13 Pus1-dependent Ψ mRNA modifications, 14 Pus4-dependent

Ψ mRNA modifications, 17 heat stress-dependent Ψ mRNA modifications and 16 heat stress-dependent Ψ ncRNA modifications, respectively (**Figure 4.15D-G** left panels), some of which were not previously reported to be Ψ-modified.

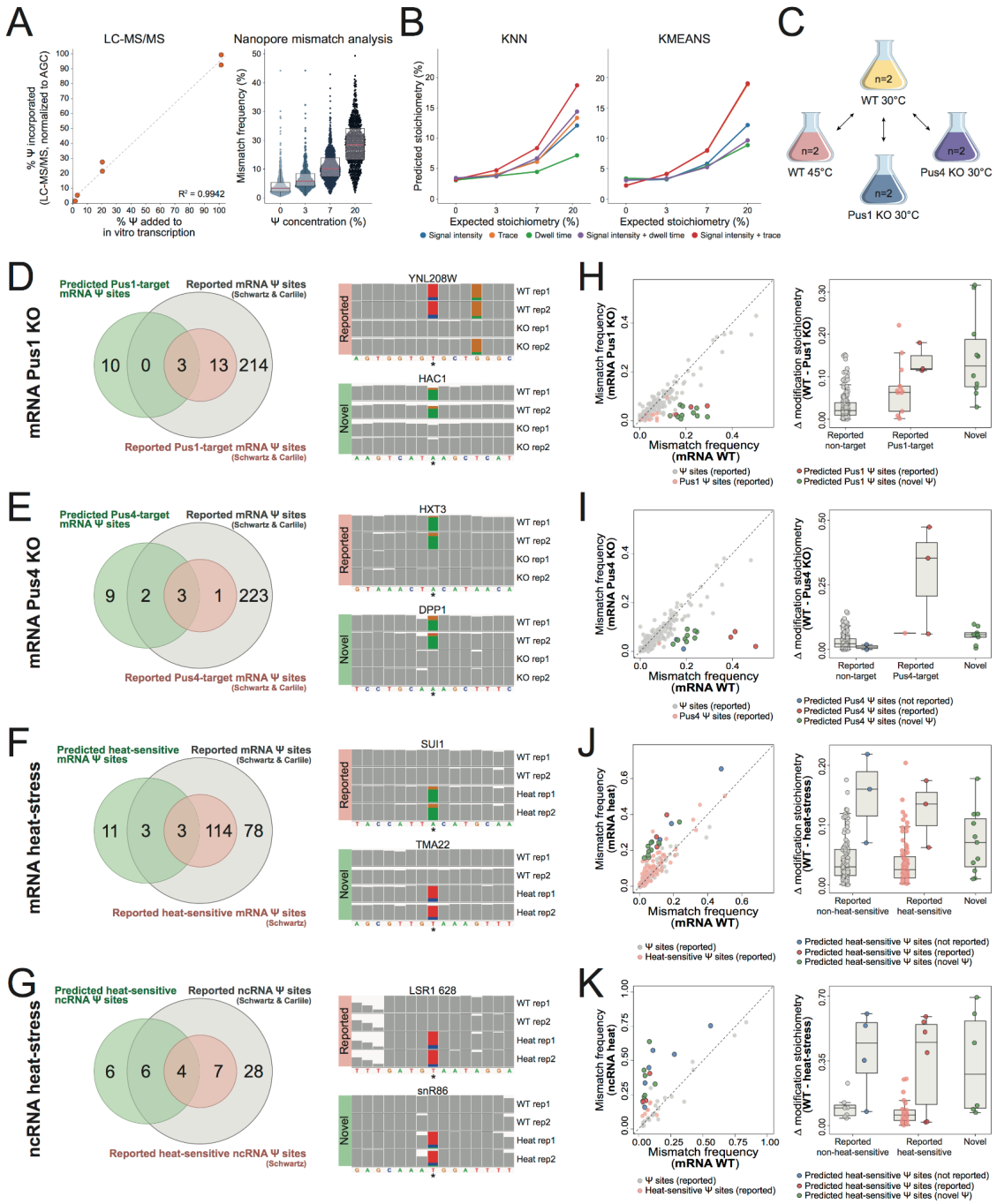


Figure 4.15 - Quantitative prediction of pseudouridine stoichiometry transcriptome-wide and systematic benchmarking of *nanoRMS* using RNA molecules with diverse modification stoichiometries.

(A) LC-MS/MS validation of pseudouridine incorporation at different proportions (0%, 3%, 20%, 100%) in the *in vitro* transcribed products, relative to the expected incorporation (% ΨTP relative to UTP) (left panel). Dotplot illustrates the mismatch frequency distribution of the uridine positions in the *in vitro* transcribed products incorporated with different concentrations of Ψ (right panel). Each dot represents one uridine position. **(B)** Stoichiometry predictions of the Ψ incorporated *in vitro* transcription products using two different algorithms (KNN and k-means) with different current information (middle right and right panels). (**C**) Conditions and strains used to predict Ψ mRNA modifications transcriptome-wide. (**D-K)** Transcriptome-wide Ψ RNA modification predictions and predicted stoichiometries in mRNAs and ncRNAs, for Pus1-dependent mRNA Ψ sites (D,H), Pus4-dependent mRNA Ψ sites (E,I), heat stress-dependent mRNA Ψ sites (F,J) and heat stress-dependent ncRNA Ψ sites (G,K). (**D-G**) Venn diagrams depict the overlap between Ψ sites predicted by our analysis and the previously reported pseudouridine sites. IGV snapshots of reported and not previously reported predicted sites illustrate the absence of the mismatch signature in the Pus1 (D) or Pus4 (E) knockout samples as well as under normal conditions, relative to heat stress conditions in mRNA (F) and ncRNA (G). The reported or predicted Ψ site is indicated by an asterisk. Nucleotides with mismatch frequencies greater than 0.15 have been colored. It should be noted that IGV snapshots that show a reference “A” with mismatch signature to G are genes that are in the minus strand (and thus are in reality positions showing U-to-C mismatch signatures). (**H-K**) Quantitative analysis of previously reported and *de novo* predicted Ψ sites in mRNAs and ncRNAs. In the left panels, comparative scatterplots of mismatch frequency illustrate differentially modified sites of reported and *de novo* predicted Ψ sites. In the right panels, stoichiometry prediction differences between WT and knockout strains (H-I) or between normal and heat stress conditions (J-K) are shown as boxplots. Box, first to last quartiles; whiskers, 1.5x interquartile range; center line, median; points, individual Ψ sites.

NanoRMS recovered 11% of previously reported Pus1-dependent Ψ sites as well as 75% Pus4-dependent Ψ sites, in addition to predicting 10 not previously reported Pus1 and 11 not previously reported Pus4-dependent mRNA Ψ-modified sites . These novel predicted Ψ mRNA sites displayed similar mismatch signatures to those observed in previously reported Ψ sites (**Figure 4.15D-E**, right panels), were highly replicable across biological replicates, and their signature disappeared in Pus1 or Pus4 knockout strains. Similarly, *nanoRMS* was able to capture previously reported heat-responsive Ψ sites present in mRNAs and ncRNAs, which resulted in predicting 17 heat-responsive Ψ mRNAs sites, among which 6 of them were previously reported Ψ sites (**Figure 4.15F**), as well as 16 heat-responsive Ψ ncRNAs sites, from which 10 were previously reported Ψ sites (**Figure 4.15G**).

Surprised by the relatively poor overlap between our predictions and previously reported Pus1 mRNA Ψ-modified sites (3 out of 16 sites), as well as between predicted and previously reported heat stress-dependent sites (7 out of 128 sites), the individual per-read features at previously reported Pus1- and heat stress-dependent sites were inspected (**Figure 4.16A,B)**. Indeed, the Ψ sites that *nanoRMS* did not report as Pus1 or heat stress-dependent were not significantly different for any of the features examined (current intensity, dwell time or trace). Thus, we wondered whether some of these sites might have been misassigned as Pus1 or heat stress-dependent by previous works. A closer examination of the overlap between Ψ sites predicted by the two previously published studies using CMC probing coupled to Illumina sequencing [121,122], which is used to define the set of 'previously reported Pus1-, Pus4- and heat stress-dependent Ψ sites', showed that the overlap was in fact very poor (**Figure 4.16C**), both when examining the set of predicted mRNA and ncRNA Ψ sites (7% and 17%, respectively), as well as when examining the sets of predicted Pus1- and Pus4-dependent mRNA and ncRNA Ψ sites (6% and 50%, respectively). Altogether, our approach detected 100% of Pus1- and Pus4-dependent sites that were identified by both studies, but very few of those that were identified by only one of the studies. In conclusion, the poor overlap between our results and previously reported Ψ sites is in fact a direct consequence of the poor overlap between the set of predicted Pus1-, Pus4- and heat stress-dependent mRNA and ncRNA Ψ sites by the two previous studies (**Figure 4.16C**).

Finally, *nanoRMS* was applied to predict the modification stoichiometry of all Ψ sites predicted in mRNAs and ncRNAs. Reads were classified based on the per-read signal intensity and trace features from positions -1, 0, and +1 using the k-means unsupervised clustering algorithm (**Figure 4.15H-K**). As expected, per-read stoichiometry predictions were low in non-targeted Ψ sites. By contrast, predicted Ψ Pus1/Pus4/heat stress-dependent sites (which included all Ψ sites) typically showed significant RNA modification stoichiometry changes, ranging from 5 to 50% change in their Ψ modification stoichiometries between the two conditions.

Altogether, differential 'error' Ψ signatures are a useful approach to identify dynamic Ψ RNA modifications across two conditions even at low stoichiometry sites, and that *nanoRMS* can be used to *de novo* predict and quantify the RNA modification stoichiometry dynamics, both in previously reported Ψ sites as well as in *de novo* predicted Ψ sites.

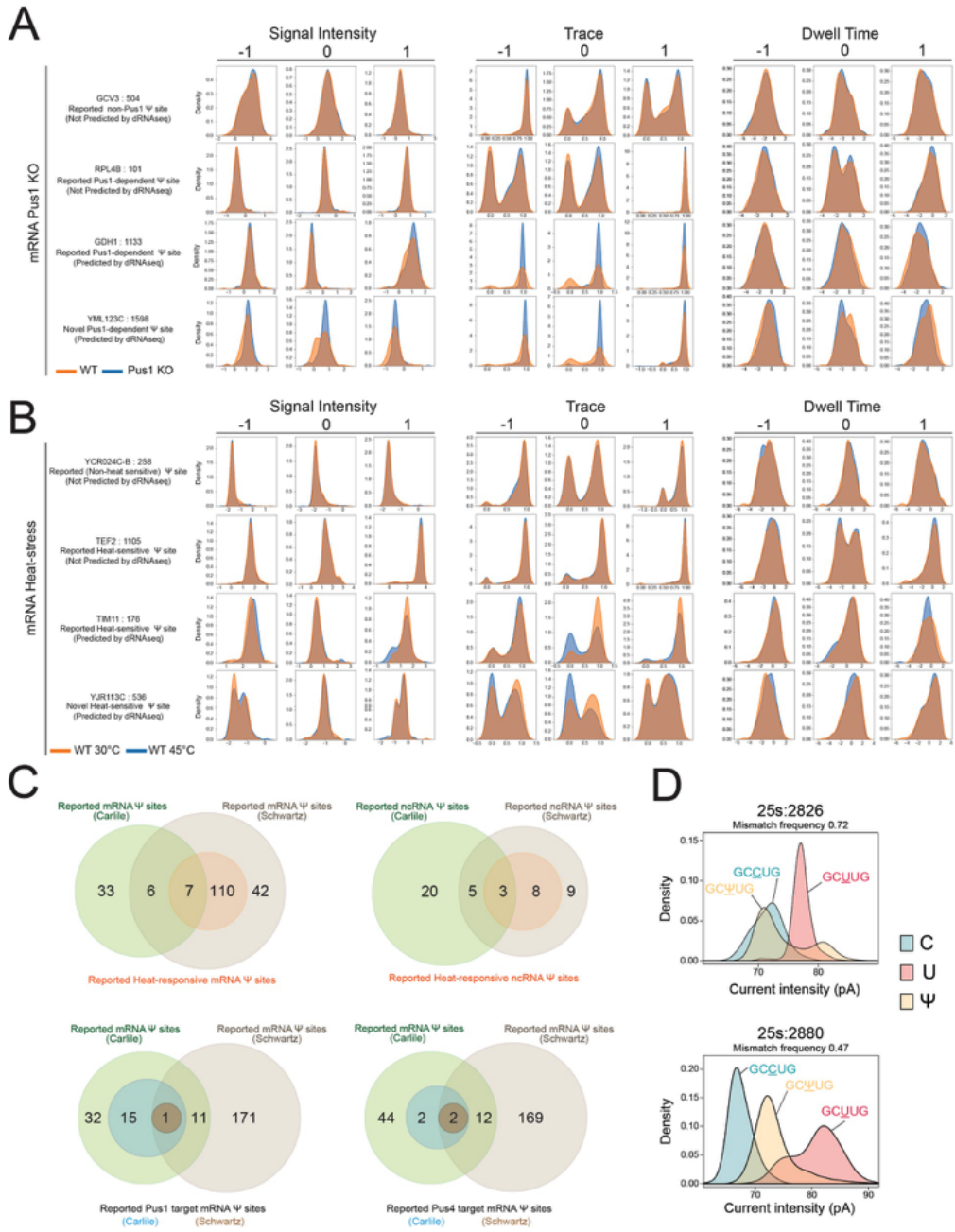


Figure 4.16 - Analysis of features in previously reported and novel mRNA Ψ sites.

(A,B) Per-read distributions of current intensity (SI), trace (TR) and dwell time (DT) in Pus1 KO and wild type *S. cerevisiae* strains (A) and Pus4 KO and wild type strains (B), for 4 different Ψ sites (reported and novel). The distributions are shown for the features observed at the modified site (0) as well as at the two neighbouring positions: downstream (-1) and upstream (+1), both in predicted and not predicted Ψ sites. The density of each feature has been colored depending on the sample. X-axis scale depends on the feature type: SI is reported as median absolute deviation normalised signal intensity as reported by *Tombo*; TR is reported as reference-base probability (0-

1 scaled), and DT is reported as log2 (observed/expected), where expected is dwell time mean value per base. **(B)** Comparison of per-read distributions of current intensity (SI), trace (TR) and dwell time (DT) in normal (30°C) and heat (45°C) conditions, at 4 different Ψ sites (reported and novel). **(C)** Venn diagrams illustrate the overlap of predicted Ψ sites in mRNAs and ncRNAs between two studies (Schwartz et al, 2014 and Carlile et al, 2014). **(D)** Current intensity density plots showing altered current intensity distribution in positions 25S:Ψ2826 (left panel) and 25S:Ψ2880 (right panel), in the wild type strain (which corresponds to Ψ-centered k-mers) and snR34 knockout strain (which corresponds to U-centered k-mers). Current intensity distribution of the equivalent 5-mers with C in the middle position are shown in blue.

4.12. Materials and Methods

4.12.1. Yeast culturing

Saccharomyces cerevisiae (strain BY4741) was grown at 30ºC in standard YPD medium (1% yeast extract, 2% Bacto Peptone and 2% dextrose). The deletion strains snR3Δ, snR34Δ and snR36Δ were generated on the background of the BY4741 strain by replacing the genomic snoRNA sequence with a *kanMX4* cassette as detailed in Parker et al. [386]. 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. For thermal stress, *Saccharomyces cerevisiae* BY4741 cultures were grown in 4 mL of YPD overnight at 30ºC. The next day, cultures were diluted to 0.0001 OD600 in 200 mL of YPD and grown overnight at 30ºC shaking (250 rpm). When the cultures reached an OD600 of 0.4-0.5, the cultures were divided into 3 x 50 mL subcultures, which were then incubated at 30ºC (control), 45ºC (heat shock) or 4ºC (cold shock) for 1 hour. Cells were collected by pelletting and snap freezing. For the analysis of rRNAs modifications across polysomal fractions, yeast BY4741 starter cultures were grown in 6 mL YPD medium at 30ºC with shaking (250 rpm) overnight. 100 mL of fresh YPD medium was inoculated with 10 µL of the stationary culture in a 250 mL erlenmeyer flask, in biological duplicates. Cells were incubated at 30ºC with shaking (250 rpm) until the cultures reached mid-exponential growth phase ($O.D_{660}$.~ 0.4-0.6). Yeast cells were then treated with 1 mM H_20_2 or left without treatment (control) for 30 minutes. 1 mL of cycloheximide stock solution (10 mg/mL) was added to each culture. Pus4 knockout strains (BY4741 MATa pus4::KAN) and its parental strain were obtained from the Yeast Knockout Collection (Dharmacon) and grown under standard conditions in YPD (1%

[w/v] yeast extract, 2% [w/v] peptone supplemented with 2% glucose) at 30°C unless stated otherwise.

4.12.2. Total RNA extraction from yeast cultures

Saccharomyces cerevisiae BY4741 cells (strains: snR3Δ, snR34Δ snR36Δ, snR60Δ, snR61Δ, snR62Δ and WT) were harvested via centrifugation at 3000 rpm for 1 minute, followed by two washes with water. RNA was purified from pelleted cells using a MasterPure Yeast RNA extraction kit (Lucigen, MPY03100), according to manufacturer's instructions. Total RNA was then treated with Turbo DNase (Thermo, #AM2238) with a subsequent RNAClean XP bead cleanup prior to starting the library preparation. For stress conditions and the Pus4KO strain, 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.

4.12.3. mRNA extraction from yeast cultures

Saccharomyces cerevisiae BY4741 (strains: BY4741 MATa pus4::KAN, BY4741 MATa pus1::KAN and BY4741 MATa) were cultured up to log phase at 30ºC. The cultures were then divided into two flasks and cultivated at 30ºC or 45ºC for 1 hour. The cells were harvested via centrifugation at 3,000 rpm for 5 minutes and snap frozen. Total RNA was purified from pelleted cells using a MasterPure Yeast RNA extraction kit (Lucigen, MPY03100), according to manufacturer's instructions. Total RNA was then DNAse-treated (Ambion, AM2239) at 37ºC for 20 minutes with a

subsequent clean up using RNeasy MinElute Cleanup Kit (Qiagen, 74204). 70-100 ug of total RNA was subjected to double polyA-selection using Dynabeads Oligo(dT)25 (Invitrogen, 61002) and finally eluted in ice-cold 10 mM Tris pH 7.5.

4.12.4. Polysome gradient fractionation and rRNA extraction

Yeast pellets from 100 mL cultures were washed with 6 mL of ice-cold Polysome Extraction Buffer (PEB), which contained 20 mM Tris-HCl pH 7.4, 100 mM KCl, 10 mM $MgCl_2$, 0.5 mM DTT, 0.1 mg/mL cycloheximide and 100 U/mL RNAse inhibitors (RNaseOUT, Invitrogen, #18080051). Cells were centrifuged for 5 minutes at 3,000 g at 4ºC. Washing was repeated by adding 6 mL of ice-cold PEB, followed by centrifugation. Cells were then resuspended in 700 µL of ice-cold PEB, and transferred into pre-chilled 2 mL Eppendorf tubes containing 450 µL of pre-chilled RNAse-free 425-600 µm diameter glass beads (Sigma G8772). Cells were lysed by vortexing at maximum speed for 5 minutes at 4ºC, followed by centrifugation also at maximum speed at bench centrifuge for 5 minutes at 4ºC. 10% of the supernatant was aliquoted into Trizol for total RNA isolation, and kept at -80ºC, which was later used as input. The remaining volume, corresponding approximately to 8 x 10^8 cells, was subsequently loaded onto the sucrose gradient. Linear sucrose gradients of 10-50% were prepared using the Gradient Station (BioComp). Briefly, SW41 centrifugation tubes (Beckman, Ultra-ClearTM 344059) were filled with Gradient Solution 1 (GS1), which consisted of 20 mM Tris-HCl pH 7.4, 100 mM KCl, 10 mM $MgCl_2$, 0.5 mM DTT, 0.1 mg/mL cycloheximide and 10% w/v RNAse-free sucrose. Solutions GS1 and GS2 were prepared with RNase-DNase free UltraPure water and filtered with a 0.22 µM filter. The tube was then filled with 6.3 mL of Gradient Solution 2 (GS2) layered at the bottom of the tube, which consisted of 20 mM Tris-HCl pH 7.4, 100 mM KCl, 10 mM $MgCl_2$, 0.5 mM DTT, 0.1 mg/mL cycloheximide and 50% w/v RNAse-free sucrose. The linear gradient was formed using the tilted methodology, with the Gradient Station Maker (Biocomp). Once the gradients were formed, 350 µL of each lysate was carefully loaded on top of the gradients, and tubes were balanced in pairs, placed into pre-chilled SW41Ti buckets and centrifuged at 4ºC for 150 minutes at 35,000 rpm. Gradients were then immediately fractionated using the Gradient Station, and 20 x 500 µL fractions were collected in 1.5 mL Eppendorf tubes, while absorbance was monitored at 260 nm continuously. Fractions were combined in the following way: the free rRNA (F1, fractions 1 and 2), the unassembled subunits (F2, fractions 3-6), the lowly-translating monosomes (F3, fractions 7-10) and the highly-translating polysomes

(F4, fractions 12-17). The pooled fractions were then concentrated using Amicon-Ultra 100K columns (Millipore), and washed two times with cold PEB. The final volume was brought down to 200 µL, and RNA was extracted using TRIzol reagent. Purity of the RNA was measured with a NanoDrop 2000 Spectrophotometer.

4.12.5. In vitro transcription of modified and unmodified RNAs

The synthetic 'curlcake' sequences [255] used in this study are designed to include all possible 5-mers while minimizing the secondary RNA structure, and consist in 4 *in vitro* transcribed constructs: (i) Curlcake 1, 2244 bp; (ii) Curlcake 2, 2459 bp; (iii) Curlcake 3, 2595 bp, and (iv) Curlcake 4, 2709. The curlcake constructs were *in vitro* transcribed using Ampliscribe™ T7-Flash™ Transcription Kit (Lucigen-ASF3507) with either unmodified rNTPs (UNM), N6-methyladenosine triphosphate (m^6ATP), 5-methylcytosine triphosphate (m^5CTP), 5-hydroxymethylcytosine triphosphate (hm^5CTP) or pseudouridine triphosphate (ΨTP). All modified NTPs were purchased from TriLink. The sequences included in the short unmodified dataset (UNM-S), which included *B. subtilis* guanine riboswitch, *B. subtilis* lysine riboswitch and *Tetrahymena* ribozyme, were also produced by *in vitro* transcription using Ampliscribe™ T7-Flash™ Transcription Kit (Lucigen-ASF3507). All constructs were 5' capped using vaccinia capping enzyme (NEB-M2080S) and polyadenylated using E. coli Poly(A) Polymerase (NEB-M0276S). Poly(A)-tailed RNAs were purified using RNAClean XP beads, and the addition of poly(A)-tail was confirmed using Agilent 4200 Tapestation. Concentration was determined using Qubit Fluorometric Quantitation. Purity of the IVT product was measured with NanoDrop 2000 Spectrophotometer.

4.12.6. Direct RNA library preparation and sequencing of *in vitro* transcribed constructs

The RNA libraries for direct RNA Sequencing (SQK-RNA001) were prepared following the ONT Direct RNA Sequencing protocol version DRS_9026_v1_revP_15Dec2016, which corresponds to the flowcell FLO-MIN106. Briefly, 800 ng of Poly(A)-tailed and capped RNA (200 ng per construct) was ligated to ONT RT Adaptor (RTA) using concentrated T4 DNA Ligase (NEB-M0202T), and was reverse transcribed using SuperScript III RT (Thermo Fisher Scientific-18080044). The products were purified using 1.8X Agencourt RNAClean XP beads (Fisher Scientific-NC0068576), washing with 70% freshly prepared ethanol. RNA Adapter (RMX) was ligated onto the RNA:DNA hybrid, and the mix was purified using 1X Agencourt RNAClean XP beads, washing with Wash buffer (WSB) twice. The sample was then eluted in Elution Buffer (ELB) and mixed with RNA running buffer (RRB) prior to loading

onto a primed R9.4.1 flowcell, and ran on a MinION sequencer with MinKNOW acquisition software version 1.15.1. The sequencing was performed in independent days and using a different flowcell for each sample (UNM, m^6A, m^5C, hm^5C, Ψ, UNM-S).

4.12.7. Direct RNA library preparation and sequencing of yeast total RNAs and mRNAs

Here we performed direct RNA sequencing of two types of *S. cerevisiae* RNA inputs: i) total RNA from *S. cerevisiae*, and ii) polyA-selected RNA from *S. cerevisiae*. Yeast total RNAs were polyadenylated using *E. coli* Poly(A) Polymerase (NEB, M0276S), following the commercial protocol, prior to starting the library prep. Yeast polyA-selected RNA was directly used as input to start the libraries since they already contain poly(A) tail. Four different direct RNA libraries were barcoded according to the recent protocol that our group recently published [387]. Custom RT adaptors (IDT) were annealed using following conditions: custom Oligo A and B were mixed in annealing buffer (0.01 M Tris-Cl pH 7.5, 0.05M NaCl) to the final concentration of 1.4 µM each in a total volume of 75 µL. The mixture was incubated at 94°C for 5 minutes and slowly cooled down (-0.1°C/s) to room temperature. RNA library for direct RNA Sequencing (SQK-RNA002) was prepared following the ONT Direct RNA Sequencing protocol version DRS_9080_v2_revI_14Aug2019 with half reaction for each library until the RNA Adapter (RMX) ligation step. Per reaction (half), 250 ng total of yeast RNAs were ligated to pre-annealed custom RT adaptors (IDT) [387] using concentrated T4 DNA Ligase (NEB-M0202T), and was reverse transcribed using Maxima H Minus RT (Thermo Scientific, EP0752), without the heat inactivation step. The products were purified using 1.8X Agencourt RNAClean XP beads (Fisher Scientific-NC0068576) and washed with 70% freshly prepared ethanol. 50 ng of reverse transcribed RNA from each reaction was pooled and RMX adapter, composed of sequencing adapters with motor protein, was ligated onto the RNA:DNA hybrid and the mix was purified using 1X Agencourt RNAClean XP beads, washing with Wash Buffer (WSB) twice. The sample was then eluted in Elution Buffer (EB) and mixed with RNA Running Buffer (RRB) 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.

4.12.8. NanoCMC-seq

CMC treatment was adapted from Schwartz et al [121] with minor changes. Briefly, 20 ug total RNA was incubated in NEBNext® Magnesium RNA Fragmentation

Module at 94°C for 1.5 minutes. The fragmented RNA was then incubated with either 0.3 M CMC dissolved in 100 µL TEU buffer (50 mM Tris pH 8.5, 4 mM EDTA, 7 M Urea) or 100 µL TEU buffer (no CMC) for 20 minutes at 37°C. Reaction was stopped with 100 µL of Buffer A (0.3 M NaOAc and 0.1 mM EDTA, pH 5.6), 700 µL absolute ethanol, and 1 µL GlycoBlue (Thermo Scientific, AM9515). RNA in the stop solution was chilled on dry ice for 5 minutes, and then centrifuged at maximum speed for 15 minutes at 4°C. Supernatant was removed and the pellet was washed with 70% ethanol. After air drying for a few minutes, the pellet was dissolved in 100 µL Buffer A and mixed with 300 µL absolute ethanol and 1 µL GlycoBlue. After chilling on dry ice for 5 minutes, the solution was then centrifuged at maximum speed for 15 minutes at 4°C. Supernatant was removed, and the pellet was washed with 70% ethanol. After washing, the pellet was air dried, and resuspended in 40 µL of 50 mM sodium bicarbonate, pH 10.4, and incubated at 37°C for 3 hours. Furthermore, RNA was mixed with 100 µL Buffer A, 700 µl ethanol, and 1 µL Glycoblue overnight at -20°C. The next day, the solution was centrifuged at maximum speed for 15 minutes at 4°C and the pellet was washed with 70% ethanol and dissolved in the appropriate amount of water after air drying. Unprobed and probed RNAs were treated with T4 Polynucleotide Kinase (PNK) (NEB, M0201S) as described above before proceeding with ONT Direct cDNA sequencing.

Before starting the library preparation, 9 µL of 100 µM Reverse-transcription primer (Original ONT VNP: 5’ /5Phos/ACTTGCCTGTCGCTCTATCTTCTTTTTTTTTTTTTTTTTTTVN 3’) and 9 µL of 100 µM complementary oligo (CompA: 5’ GAAGATAGAGCGACAGGCAAGTA 3’) were mixed with 1 µL 0.2 M Tris pH 7.5 and 1 µL 1 M NaCl. 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, 1 µL 10 mM dNTP mix, 4 µL 5X RT Buffer, 1 µL RNasin® Ribonuclease Inhibitor (Promega, N2511), 1 µL Maxima H Minus RT (Thermo Scientific. EP0742) and nuclease-free water up to 20 µL. The reverse-transcription mix was incubated at 60°C for 60 minutes and inactivated by heating at 85°C for 5 minutes before moving ontoice. 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 1.2X 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 2.5 µL Native Barcode (EXP-NBD104) and 25 µL Blunt/TA Ligase Mix (NEB, M0367S) and incubated in room temperature for 10 minutes. The reaction was cleaned up using 1X AMPure XP beads and the libraries were pooled into one tube that finally contains 200 fmol library. The pooled library was then ligated to the sequencing adapter (AMII) using Quick T4 DNA Ligase (NEB, M2200S) in room temperature for 10 minutes, followed with 0.65X AMPure XP Bead cleanup 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.

4.12.9. Analysis of nanoCMC-seq

Reads were base-called with stand-alone Guppy version 3.6.1 with default parameters running in GPU, with built-in demultiplexing tool of Guppy. Unclassified reads were then demultiplexed further using Porechop with --barcode_threshold 50 option (https://github.com/rrwick/Porechop). Then all the merged classified reads were mapped to cytosolic and mitochondrial ribosomal RNA sequences in S. cerevisiae using minimap2 default. Furthermore, a custom script was used to extract RT-drop signatures and the RT-drop scores were plotted using ggplot2. All scripts used to process nanoCMC-seq data with RT-Drop information have been made available in GitHub (https://github.com/novoalab/yeast_RNA_Mod). Notably, due to the 5' end truncation of the nanopore sequencing reads by ~13 nt, RT-drop positions were shifted by 13 nt to accurately determine the exact RT-drop positions. To identify significant RT drops in a given transcript, I first computed RT-drop scores at each site, which took the difference in the coverage at a given position (0) relative to the previous position (-1). I then computed the difference (delta RT drop-off score) in RT-drop scores between CMC-probed and unprobed conditions. Lastly, I normalised the delta RT drop-off score at each position by the median RT drop-off per transcript, leading to final CMC-Scores, which can be compared across transcripts. Positions with CMC-Score greater than 25 were considered significant, i.e. to contain a pseudouridine. It should noted that the nanoCMC-seq signal-to-noise ratio is dependent on the coverage of the individual transcript.

4.12.10. Demultiplexing direct RNA sequencing

Demultiplexing of the barcoded direct RNA sequencing libraries was performed using DeePlexiCon with default parameters [387]. Reads with demultiplexing confidence scores greater than 0.95 were kept for downstream analyses. I used a lower score in the case of polysomal fractions and mRNA runs (0.8), due to the low read coverage of some fractions and/or genes. It should be noted that the dataset was also analysed using 0.95 threshold, and results and conclusions of the analysis did not change, compared to those obtained using 0.80 threshold.

4.12.11. Base-calling direct RNA sequencing

Reads were base-called with stand-alone Albacore versions 2.1.7 and 2.3.4 with the --disable_filtering parameter, and stand-alone Guppy versions 2.3.1 and 3.0.3 with default parameters running in CPU. In-house scripts were used for computing the number of unique and common base-called reads between the different approaches, as well as to compare the tendency of each base-caller regarding read lengths and qualities. Both Albacore and Guppy are available to ONT customers via their community site (https://community.nanoporetech.com/). Differences between the base-called features using distinct base-callers were determined using Kruskal-Wallis test with Bonferroni correction for pairwise comparisons, whereas differences between unmodified and modified sites were assessed using Mann-Whitney-Wilcoxon test.

4.12.12. Mapping algorithms and parameters

Reads were mapped using either *Minimap2* [365] or *GraphMap* [366]. *Minimap2* version 2.14 was run with two different parameter settings: (i) minimap2 -ax map-ont, which is the recommended setting for direct RNA sequencing mapping, and thus is referred to as ‘default’, and (ii) minimap2 -ax map-ont -k 5, which is referred to as ‘sensitive’. *GraphMap* version 0.5.2 was also run with two different parameter settings, for comparison, (i) graphmap align, using ‘default’ parameters, and (ii) graphmap align --rebuild-index -v 1 --double-index --mapq -1 -x sensitive -z 1 -K fastq --min-read-len 0 -A 7 -k 5, which is expected to increase the tolerance to errors that may occur under the presence of RNA modifications, and thus is referred to as ‘sensitive’. Yeast total RNA runs were mapped to ribosomal RNAs and non-coding RNA transcripts using graphmap with default settings. Yeast poly(A)-selected runs were mapped to the yeast genome (SacCer3) using minimap2 with -ax splice -k14 -uf parameters. The scripts can be found in the GitHub repository https://github.com/novoalab/yeast_RNA_Mod.

4.12.13. Analysis of base-called features in curlcakes

Sam files were transformed into bam files using Samtools version 1.9 [388], and were then sorted and indexed in order to visualise the data using the Integrative Genomics Viewer (IGV) version 2.4.16 [389]. Base-called features were extracted with *EpiNano* version 1.1 (https://github.com/enovoa/EpiNano). Principal Component Analysis (PCA) was used to reduce the dimensionality of the base-calling error data to visually inspect for base-calling differences, using as input the base-called features (mismatch frequency, deletion frequency and per-base quality) from all 5 positions of each k-mer. Only k-mers that contained a given modification once in the 5-mer were included in the analysis. All scripts used to analyse *in vitro* transcribed sequences using different base-calling algorithms and mappers, as well as to generate the Figures related to their analysis are available in https://github.com/novoalab/Best_Practices_dRNAseq_analysis.

4.12.14. Analysis of base-called features in yeast RNAs

Sam files were transformed into bam files using Samtools version 1.9 [388], then sorted and indexed in order to visualise the data using the Integrative Genomics Viewer (IGV) version 2.4.16 [389]. Base-called features were extracted using *EpiNano* version 1.1 with minor modifications, which consisted in including in the output csv file the directionality of mismatched bases (C_frequency, G_frequency, A_frequency, U_frequency). The modified *EpiNano* script can be found at https://github.com/novoalab/yeast_RNA_Mod. Scripts for the analysis and visualization of base-called features are also included in the same GitHub repository.

4.12.15. Visualization per-read current intensities using *Nanopolish*

Nanopolish eventalign output was processed to extract the current intensity values corresponding to the 15-mer regions centered in the modified sites, for the following sites: (i) 6 Ψ rRNA sites for which knockout data was available (25S:2133, 25S:2129, 25S:2826, 25S:2880, 25S:2264, 18S:1187), for all 4 sequencing datasets (wild type, snR3-KO, snR34-KO, snR36-KO); (ii) 4 Nm sites for which knockout data was available (25S:817, 25S:908, 25S:1133, 25S:1888), for all 4 sequencing datasets (wild type, snR60-KO, snR61-KO, snR62-KO); (iii) 7 Ψ snRNA/snoRNA sites which were identified as heat-sensitive, for which there was a minimum of 100 reads of coverage. Reads with empty values in the 15-mer region in the *Nanopolish* eventalign output were omitted from the analysis.

4.12.16. Analysis of current intensity, dwell time and trace

In this work, two different softwares to extract current intensity were used, namely Nanopolish [390] and Tombo [371]. Nanopolish was used to extract the aligned current intensity values per read and position, using the option *--scale-events*. Mean current intensity per-position was computed by summing the current intensities of all reads aligned to the same position, divided by the total number of reads mapping at a given position. All scripts used to process *Nanopolish* event align output, including scripts to display mean current intensity values along transcripts have been made available in GitHub (https://github.com/novoalab/nanoRMS).

Signal intensity, dwell time and trace were retrieved using get_features.py script, which is available as part of *nanoRMS*. This program internally uses: minimap2 (read alignment), Tombo (calculation of signal intensity and dwell time) and ont-fast5-api (retrieval of trace). Trace represents the probability that a given signal intensity chunk may be originating from each of the 4 canonical bases (A, C, G and T/U), and it is reported relative to the reference base. For example, in a T reference position that is incorrectly reported as C (common base-calling error observed for Ψ sites), the trace value will be reported for the reference base (T in this case). Then, the final read alignment and all the features are stored into sorted BAM files. All scripts necessary to retrieve and store per-read, per-position features and plot/calculate results are available within the *nanoRMS* GitHub repository (https://github.com/novoalab/nanoRMS).

4.12.17. *De novo* prediction of pseudouridine modifications on yeast mitochondrial rRNAs

To systematically identify Ψ sites *de novo* based on the Ψ base-calling signatures, I first extracted the mismatch frequency and per-base mismatch frequency (C_freq, A_freq, U_freq, G_freq) from both unmodified (U) and modified (Ψ) sites from cytosolic ribosomal RNAs, from three biological replicates. As expected, C mismatch frequency (C_freq) and global mismatch frequency (mis_freq) showed clearly distinct distributions when comparing unmodified and Ψ-modified sites (**Figure 4.11A**). I then determined the optimal cut-points for these two features using the *cutpointr* package in R with oc_youden_kernel method, which applies Kernel smoothing and maximises the Youden-Indexing. This approach predicted C_freq=0.137 and mis_freq= 0.587 as optimal cut-offs. For the mitochondrial ribosomal RNA, I filtered the uridine sites based on the selected features and assigned those that are replicable in three biological replicates as “candidate” pseudouridine sites.

4.12.18. *De novo* prediction of pseudouridine modifications in yeast mRNAs and non-coding RNAs

Due to the lower stoichiometry of modification of noncoding RNAs (snRNA and snoRNAs) and mRNAs, I focused on analysis of the *de novo* detection of Ψ sites whose pseudouridylation levels would be changing between two conditions, either by comparing normal and stress (heat-shock) conditions, or by comparing the base-calling 'error' patterns of wild type strains and Pus1 or Pus4-deficient strains. Only sites which passed the coverage filter (n>30 reads) in both biological replicates from both conditions were considered in the analysis. Sites with minimal mismatch frequency difference of 0.1 between the two conditions in both replicates that met the identified Ψ signature (C_freq=0.137 and mis_freq= 0.587) were considered as true Ψ sites that were either heat-sensitive, Pus1-dependent, or Pus4-dependent, respectively.

4.12.19. Prediction of RNA modification stoichiometry using *nanoRMS*

Per-position features from individual reads were stored in BAM files using pysam (https://github.com/pysam-developers/pysam) and stored them either in Numpy arrays (https://numpy.org/) or Pandas DataFrames (https://pandas.pydata.org/) using the script get_features.py, which is available as part of *nanoRMS*. Models were trained with combinations of features with diverse ranges of sequence contexts surrounding the modified sites (k=1-15). Features used to predict stoichiometry included: (i) current intensity (SI), (ii) dwell time in the centre of the pore (at position 0, DT/DT0), (iii) dwell time at helicase centre (shifted by 10 positions, DT10) and (iv) base probability (trace, TR). Estimation of modification frequency was performed using unsupervised (GMM, KMEANS, IsolationForest, OneClassSVM) and supervised (KNN, RandomForest) machine learning methods implemented in sklearn (https://sklearn.org/). Plots were built using matplotlib and seaborn (https://seaborn.pydata.org/).

Trained models were first benchmarked with unmodified (KO) and modified (WT) reads from rRNA mutants dataset, to identify which machine learning methods and which combination of features discriminated between modified and unmodified reads. Then, we tested how the diverse models would perform at diverse stoichiometries of modification. To this end, we simulated samples with varying levels of modification: 0%, 20%, 40%, 60%, 80% and 100% (using mixes of KO and WT reads) and estimated the modification level in those simulated samples by comparing them to KO (**Figure 4.9C**).

NanoRMS performed best when trained with signal intensity (SI) + trace (TR) as features, and when using KNN supervised models or KMEANS unsupervised models, both for Ψ and Nm-modified sites. For mRNA and ncRNA analysis, only sites with more than 30 reads of coverage in all conditions and replicates were included for predicting RNA modification stoichiometry. Prediction of RNA modification stoichiometry in mRNAs and non-coding RNAs was performed using signal intensity + trace as features, and k-means as classification algorithm. Stoichiometry changes were reported as the difference in predicted stoichiometry between the two conditions. All code and examples to predict RNA modification stoichiometry are available as part of the *nanoRMS* GitHub repository (https://github.com/novoalab/nanoRMS).

4.13. Discussion

RNA modifications regulate a wide range of biological processes [13,391,392]. They can modulate the fate of RNA molecules by altering mRNA splicing [234,393,394] or mRNA decay [395,396], as well as affect major cell and organism-level decisions, such as cellular differentiation [199,397] and sex determination[49,201,344]. While the biological relevance of RNA modifications is out of question, a major difficulty in studying them has been the need for tailored protocols to map each modification individually [48,249]. In this context, direct RNA nanopore sequencing can overcome many of the limitations that NGS-based methods suffer from, as it can sequence full-length native RNA molecules, including their RNA modifications.

Direct RNA nanopore sequencing has been successfully applied in a wide variety of organisms [255,257,353,398–401]. However, the detection of distinct RNA modification types in individual native RNA molecules is still an unsolved challenge. While both current intensity-based and 'error'-based methods have proven useful strategies to detect RNA modifications, these have been mainly focused on the detection of m^6A [255,257,353,354]-, and are typically unable to predict which RNA modification type they are in fact detecting (e.g. m^6A, Ψ, Am or m^5C) [352,371]. Moreover, current algorithms to study RNA modifications using direct RNA sequencing are not quantitative.

To overcome these limitations, here we first explored how distinct RNA modifications may affect direct RNA nanopore signals and base-calling 'errors'. Different RNA modification types (e.g. Ψ versus m^5C) produce distinct yet characteristic base-calling 'error' signatures, both *in vitro* (**Figure 4.1, 4.2F**) and *in vivo* (**Figure 4.3**). Consequently, base-calling errors can be used not only to predict whether a given site is modified or not, but also to identify the underlying RNA modification type. While

base-calling signatures depend to some extent on the surrounding sequence context, I find that Ψ modifications lead to robust U-to-C mismatch signatures, which can be exploited for *de novo* prediction of Ψ modifications (**Figure 4.11**). Through this approach, I identified two previously unreported Ψ modifications in yeast 15S mitochondrial rRNA (15S:579 and 15S:Ψ854), as well as confirmed reported Ψ-modified sites in rRNAs, snRNAs and mRNAs (**Figures 4.5-4.15**). Moreover, I revealed that Pus4, which was previously thought to modify only tRNAs and mRNAs, is the enzyme responsible for placing Ψ854 in mitochondrial rRNA. These findings were further validated using nanoCMC-seq, a novel orthogonal method that can detect Ψ modifications with single nucleotide resolution by coupling CMC probing to nanopore cDNA sequencing (**Figure 4.11**).

While Ψ modifications can be detected both in the form of base-calling 'errors' and altered current intensities (**Figures 4.5-4.8**), the latter does not provide single nucleotide resolution, with maximal current intensity shifts often seen a few nucleotides away from the real modified site. Thus, current intensity-based methods alone may suffer from imprecisions in the assignment of the RNA-modified site. Here the combination of both approaches were proposed to be the optimal design to obtain stoichiometric information of Ψ-modified sites with single nucleotide resolution. Specifically, once the site has been located using base-calling error features, per-read features (current intensity and trace) from the regions surrounding Ψ or Nm-modified site are sufficient to robustly bin the reads into two separate clusters (modified and unmodified), and provide good estimates of Ψ and Nm modification stoichiometries (**Figure 4.8** and **4.15**).

One surprising feature of base-calling 'errors' is that fully modified sites do not always lead to same mismatch frequencies, suggesting that mismatch frequencies alone cannot be used per se as an estimation of the stoichiometry of the site (**Figure 4.3B**). While within the same sequence context, higher mismatch frequencies correspond to higher modification levels, this same rule cannot be used to compare across distinct RNA-modified sites. It is speculated that the differences observed in mismatch frequency across different sites might be in fact a consequence of the distinct deviations in current intensity of the modified k-mer relative to unmodified counterparts (**Figure 4.16D**).

Finally, it should be noted that while *nanoRMS* allows predicting and studying the dynamics of diverse RNA modifications in a quantitative manner, there are caveats and limitations, leaving ample room for future improvements. First, not all RNA modifications lead to strong alterations in the base-calling features and/or current

intensity patterns, such as 2'-O-methylcytosine (Cm), which is poorly detected in direct RNA sequencing datasets, compared to other RNA modifications (**Figure 4.3C**). Second, the detection of RNA modifications is partly dependent on the sequence context; for example, detect 25S:Gm908 could not be detected (**Figure 4.5**). Similarly, some Ψ-modified sites, such as 18S:Ψ1187, cause weaker alterations in base-calling features and current intensity shifts than other Ψ-modified positions (**Figures 4.8-4.8**), although this limitation can be alleviated by the incorporation of additional features into the model (**Figure 4.9C**). Third, not all RNA modifications lead to base-calling errors with single nucleotide resolution, as with pseudouridine. For example, 2'-O-methylations often affect neighboring bases (**Figure 4.5C** and **4.7A**), making it challenging to *de novo* predict modified sites without any prior information. Fourth, stoichiometry prediction is heavily affected by the choice of resquiggling algorithms (**Figure 4.9** and **4.17**). For example, stoichiometry in 25S:Ψ2264 could not be predicted when using resquiggling due to the low number of reads that the *Nanopolish* algorithm was able to resquiggle (**Figure 4.7E**); however, this limitation could be overcome when using *Tombo* resquiggling, leading to stoichiometry predictions similar to those observed using Mass Spectrometry (**Figure 4.8D**). Finally, it should be noted that while *nanoRMS* was successful at detecting RNA modification stoichiometry changes as low as 5-10% (**Figure 4.15**), the detection of RNA modification changes in low modification stoichiometry sites was only possible when using pairwise comparisons.

Despite these challenges and limitations, this work provides a framework for the systematic and comprehensive analysis of the epitranscriptome with single molecule resolution, showing that direct RNA sequencing can be employed to estimate Ψ and Nm modification stoichiometry as well as to *de novo* predict Ψ RNA modifications transcriptome-wide, in rRNAs, ncRNAs and mRNAs. Future work will be needed to functionally dissect the biological roles and dynamics of RNA modifications, to better comprehend how and when the epitranscriptome is tuned to regulate diverse cellular functions.

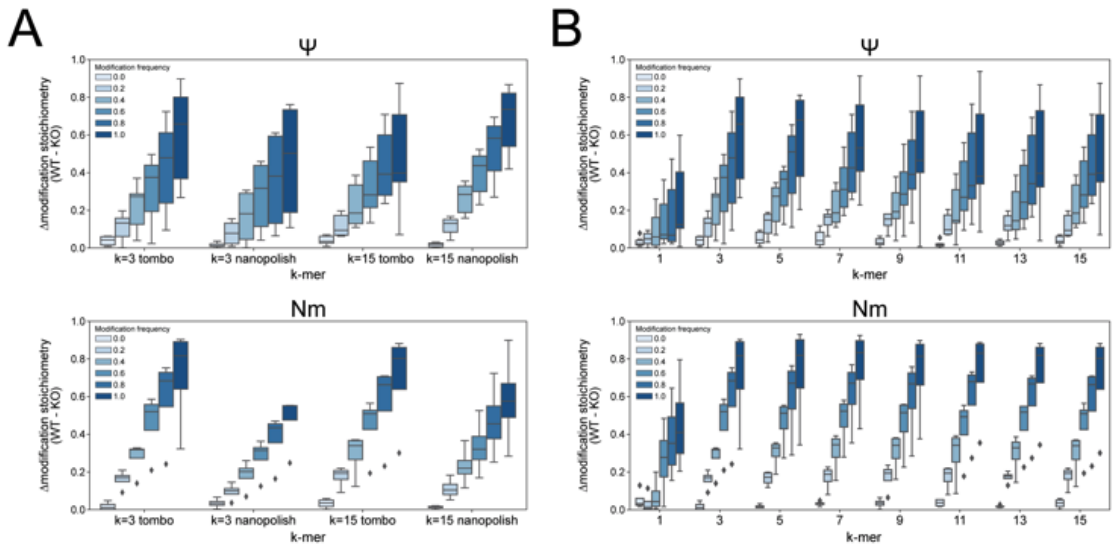


Figure 4.17 - Benchmarking Ψ and Nm stoichiometry predictions using signal intensity features from varying k-mer sizes and resquiggling softwares.

(A) Comparison of stoichiometry prediction using signal intensity features of 3-mers (k=3) or 15-mers (k=15) centered in the modified site, using either *Tombo* or *Nanopolish* resquiggling. **(B)** Comparison of stoichiometry prediction using signal intensity features of 1, 3, 5, 7, 9, 11, 13 and 15-mers centered in the modified site from Tombo resquiggled reads. For each k-mer and resquiggling algorithm, we computed the stoichiometry using mixtures of reads from wild type and knockout strains for the 6 pseudouridine (upper panels) and 4 2'-O-methylated sites (lower panels) used in this work. The colors of the boxplots correspond to the proportion of reads from the wild type strain, ranging from 0% (light blue, all reads come from knockout strain) to 100% (darkest blue, all reads come from wild type strain). Box, first to last quartiles; whiskers, 1.5x interquartile range; center line, median; points, outliers.

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

This chapter contains material described in the manuscript that is submitted for publication (Begik et al., 2021) [402].

I have performed most of the wet-lab experiments and data analyses with the help of other authors in the publication. I also drafted the manuscript with Eva Maria Novoa.

Huanle Liu contributed with custom scripts for the analysis of tails.

Anna Delgado-Tejedor analysed the direct RNA sequencing of the zebrafish, which contributed to making the Figure 5.3 E, F.

Cassandra Kontur, Antonio J. Giraldez and Jean-Denis Beaudoin provided the zebrafish embryo RNA samples used in the study.

Eva Maria Novoa and John S. Mattick supervised this study.

5.1. Introduction

One context where polyadenylation has been shown to play a major role in determining RNA fate and decay is vertebrate embryogenesis [403]. Indeed, vertebrate embryos undergo a major cellular reprogramming in the first hours post-fertilization, known as the Maternal-to-Zygotic Transition (MZT) [404]. 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 [405,406]. Because the MZT begins in a transcriptionally silent embryo, it relies heavily on post-transcriptional regulatory mechanisms [404], including the modulation of the polyadenylation status of the RNA molecules [42,403]. 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 [42,265]. While these methods have been successfully employed to characterise 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 [272]. To sequence native RNAs using dRNA-seq, polyA-tailed RNA molecules are ligated to a 3' adapter that contains an oligo(dT) overhang (**Figure 5.1A**). 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 5.1A**). 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).

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 is termed **Nano**pore **3 P**rime end-capture **seq**uencing (Nano3P-Seq) (**Figure 5.1A**). 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. Nano3P-Seq can capture any type of RNA molecule regardless of its 3' sequence, including polyA-tailed and non-tailed RNAs. Moreover, 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.

5.2. 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 5.1A**), previous efforts have opted to perform an *in vitro* polyadenylation reaction of the total RNA [351] 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 5.1A**).

In order to assess the ability of Nano3P-seq to sequence both polyA(+) and polyA(-) RNAs, I first sequenced two synthetic RNAs, one lacking a polyA tail, and a second that had been *in vitro* polyadenylated (see *Methods*) (**Figure 5.2A-C**). The 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 5.2C**). I 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 [407] (see *Methods*). I 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 5.1B,** see also **Figure 5.2D**). In addition, the results confirmed that polyA tail length information was retained in

individual reads. Specifically, the majority of reads corresponding to mRNAs had polyA tails (**Figure 5.1C,** see also **Figure 5.2E,F**), whereas non-coding RNAs such as snoRNAs (**Figure 5.2F)** or snRNAs did not have polyA tails (**Figure 5.2G**), as could be expected.

To assess the accuracy and reproducibility of Nano3P-seq to quantify RNA abundances, I then examined the performance of Nano3P-seq in synthetic RNA mixes (sequins) [408] that had been spiked into the samples in independent flow cells. The results showed that Nano3P-seq provided accurate estimates of RNA abundances (Pearson's r^2: 0.976) (**Figure 5.1D**), and that these quantifications were highly reproducible across replicables (Pearson's r^2: 0.995) (**Figure 5.2H**).

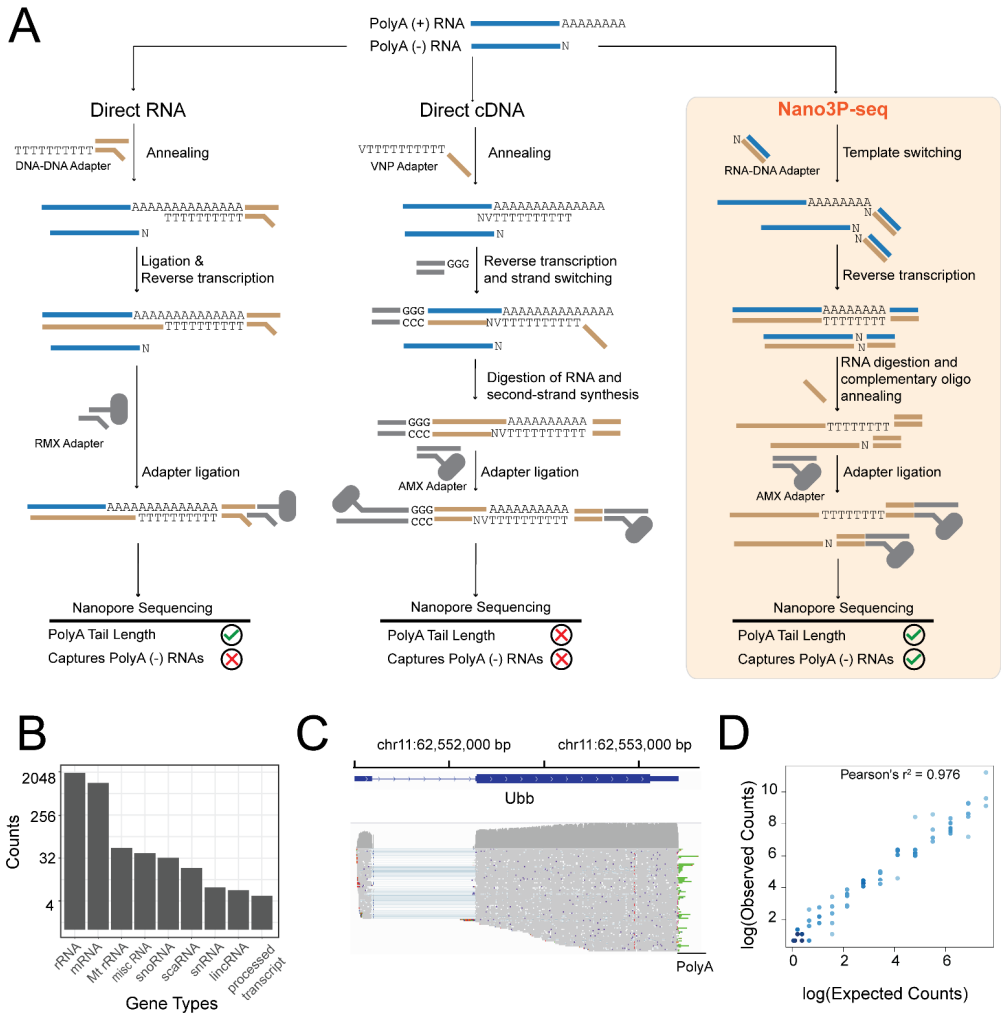


Figure 5.1 - Nano3P-seq captures polyadenylated and non-polyadenylated RNAs, while retaining polyA tail length information.

(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 5.2. 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).

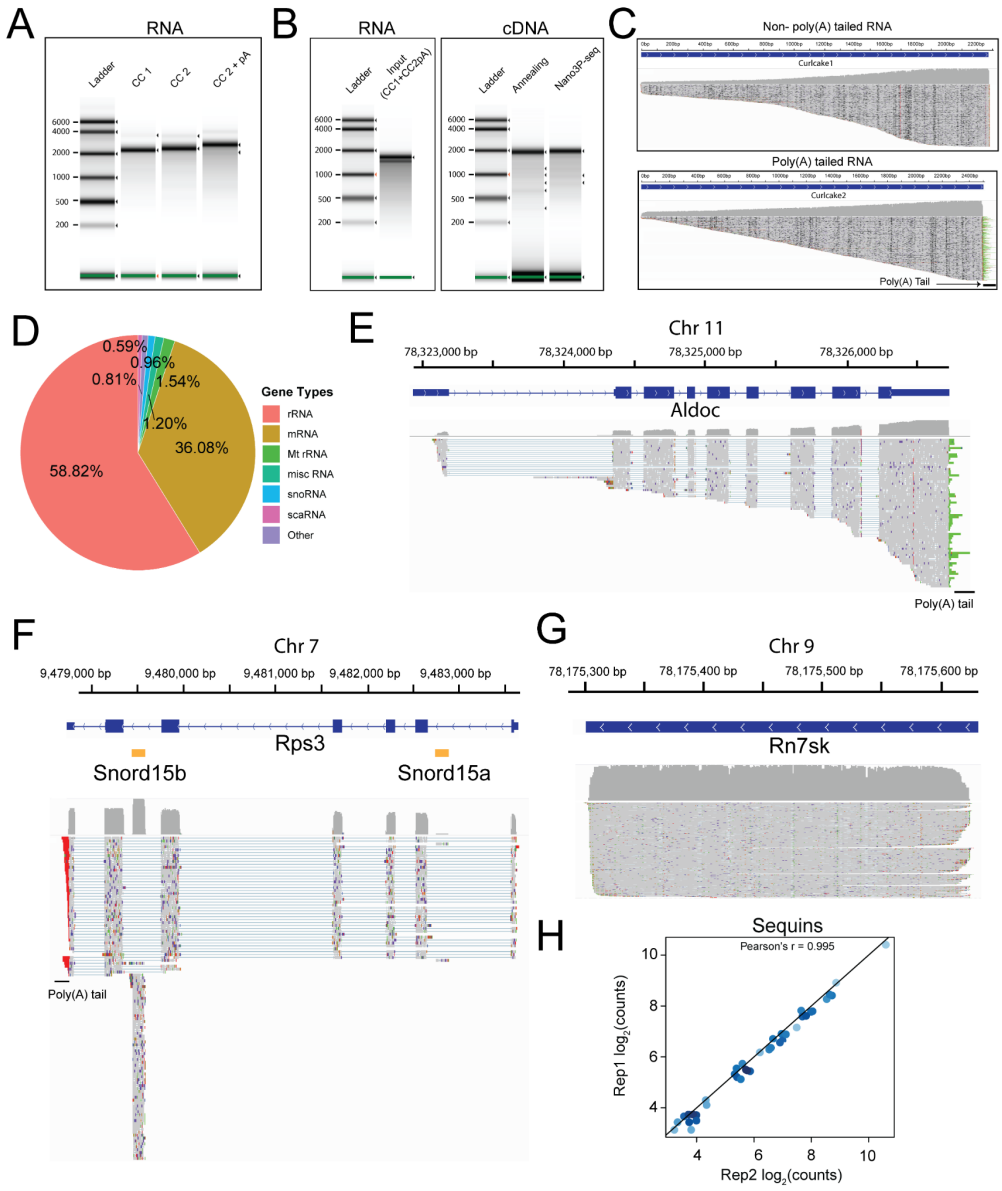


Figure 5.2 - Nano3P-seq captures non-poly(A)-tailed and poly(A)-tailed RNAs

(A) Tapestation profiles of synthetic RNAs ('curlcakes') after being *in-vitro* transcribed and poly(A) tailed (pA). **(B)** Tapestation profiles of the input RNA (curlcake mix) for

reverse-transcription and cDNA produced after annealing based or template-switching based (Nano3P-seq) reverse-transcription **(C)** IGV snapshots of synthetic RNAs illustrating that Nano3P-seq captures both non-polyadenylated (left) and polyadenylated (right) RNAs. In addition, a diversity of poly(A) tail lengths are also captured by Nano3P-seq, which are shown in green (right panel).**(D)** Pie chart showing the abundance of different RNA types in Nano3P-seq of nuclear/mitochondria enriched RNA. **(E)** IGV snapshot of reads mapping to Aldoc gene with poly(A) tail shown in green. **(F)** IGV snapshot of reads mapping to Rps3 and Snord15b genes. Poly(A) tail can be seen in green on the reads mapping to Rps3 mRNA, while it can't be seen in Snord15b snoRNA. **(G)** IGV snapshot of reads mapping to Rn7sk miscRNA, which are not expected to contain poly(A) tails. **(H)** Correlation of per-gene counts observed in synthetic sequins that were used as spike-ins in the sequencing runs.

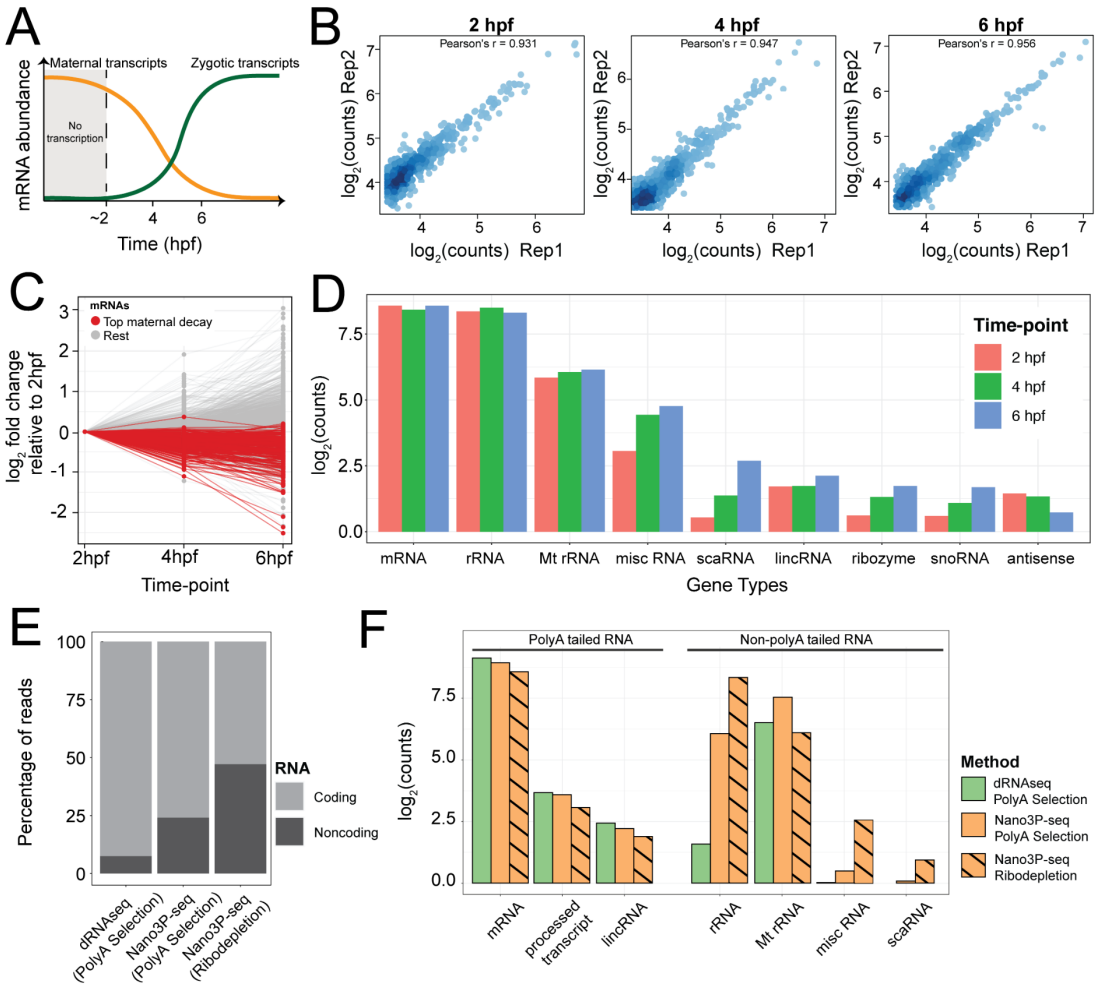
5.3. Nano3P-seq recapitulates the dynamics of coding and non-coding RNAs during vertebrate embryogenesis

Next, I examined the RNA dynamics that occurs during the MZT (**Figure 5.3A**) 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 5.4A-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 5.3B)**.

A comparative analysis of RNA population dynamics across time points showed that Nano3P-seq recapitulated the transcriptomic switch that occurs during the MZT [404,405], with a drastic decay of maternal mRNAs (**Figure 5.3C**), in agreement with previous studies [43,409]. 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 5.3D**). By contrast, much fewer non-coding RNA populations were globally captured when using direct RNA nanopore sequencing on the same samples (**Figure 5.3E**).

It's noted, however, that mitochondrial rRNAs were not enriched in Nano3P-seq datasets relative to dRNAseq datasets (**Figure 5.3F**). 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, I 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 5.4C-E**), in agreement with previous reports



[410].

Figure 5.3 - 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 5.4.

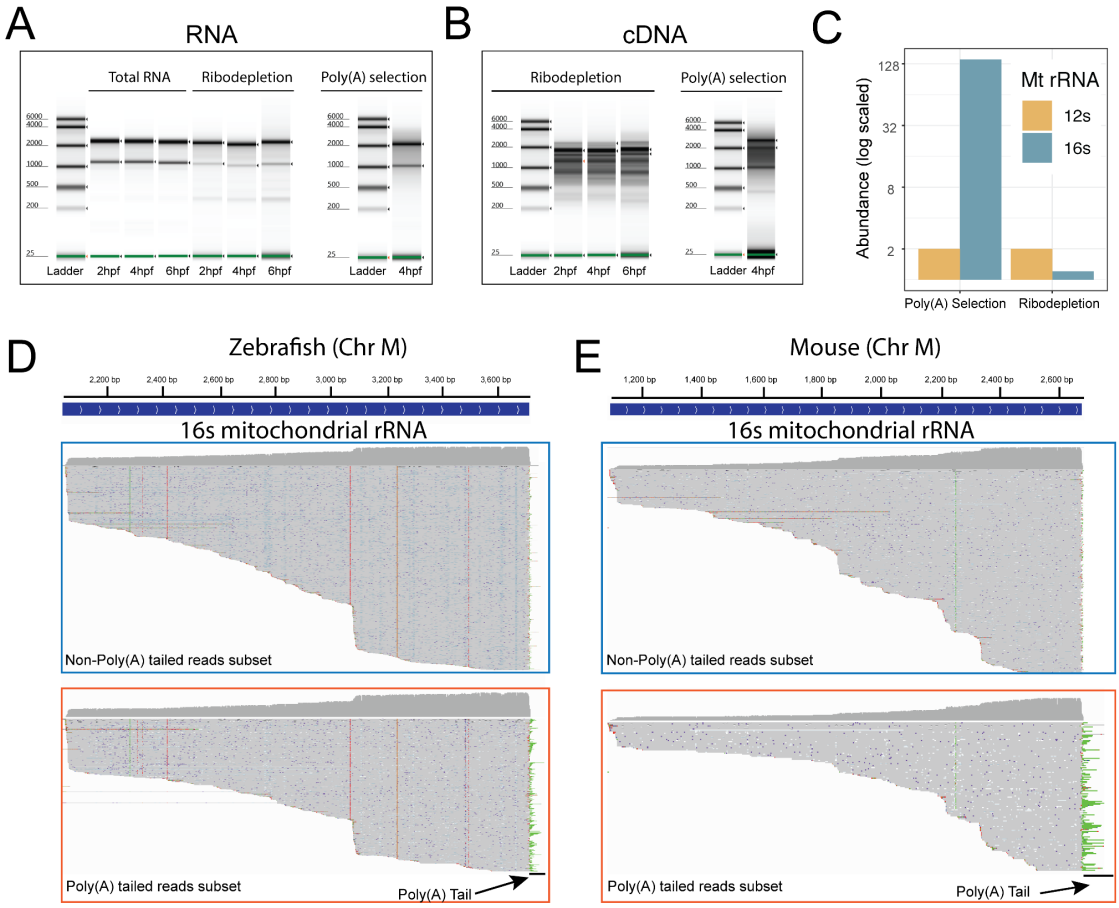


Figure 5.4 - Analysis of abundances and poly(A) tails in mitochondrial rRNAs.

(A) Tapestation profiles of RNAs from zebrafish embryos at different developmental time points (2,4,6 hours post-fertilization). Profiles include total RNA, ribodepleted RNA and poly(A)+ selected RNA. **(B)** Tapestation profiles of the reverse-transcription products of ribodepleted (left) and poly(A)+ selected (right) samples from zebrafish embryos collected at different developmental time points (2,4,6 hours post-fertilization). **(C)** mRNA abundances (log scaled) of 12s and 16s mitochondrial rRNAs in poly(A)+ selected (left) and ribodepleted (right) samples. **(D)** IGV snapshot of reads mapping to zebrafish 16s mitochondrial rRNA, where reads have been grouped as non-poly(A) tailed and poly(A) tailed based on their predicted poly(A) tail length. Poly(A) tail region is shown with an arrow. **(E)** IGV snapshot of reads mapping to mouse 16s mitochondrial rRNA, where reads have been grouped as non-poly(A) tailed and poly(A)

tailed based on their predicted poly(A) tail length. Poly(A) tail region is shown with an arrow.

5.4. 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. It should noted that algorithms to detect polyA tails in native RNA nanopore sequencing reads are well established and benchmarked [272,390,401,411], but their applicability to cDNA reads, such as those from Nano3P-seq, remains unclear. To this end, I first examined whether the *tailfindR* polyA tail prediction software [411] 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 5.5A)**. Then, I assessed the accuracy of the polyA tail length predictions of *tailfindR* in Nano3P-Seq datasets that included a battery of synthetic RNAs (sequins) [408] with known polyA tail lengths. The results showed that polyA tail length estimations in Nano3P-seq data were highly reproducible across replicates (r^2= 0.993, **Figure 5.6A,B**), and with an accuracy similar to that observed when performing polyA tail length estimations in sequins that had been sequenced using dRNAseq (**Figure 5.5B**). 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 5.6C,D**). Finally, I performed a comparative analysis of mouse mRNA polyA tail lengths with those from yeast and zebrafish. I 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 5.5C**), in agreement with previous studies [42].

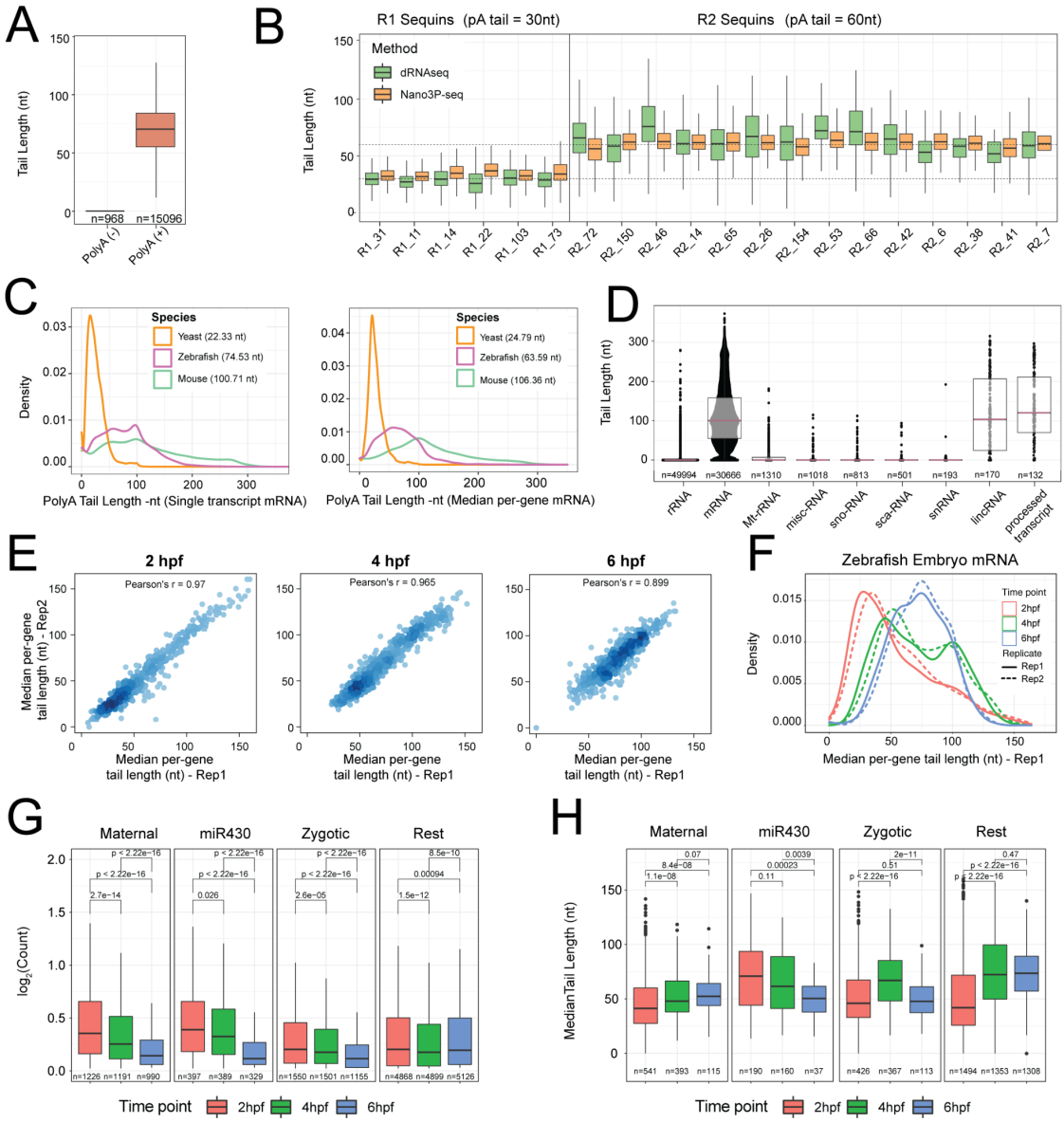


Figure 5.5 - Nano3P-seq can be used to accurately estimate polyA tail lengths in individual molecules.

(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 analysed (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 5.6 and 5.7.

5.5. 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*. I 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. PolyA tails were mainly predicted on mRNAs, but also in lincRNAs and processed transcripts, which are also known to be polyadenylated [412,413] (**Figure 5.5D**).

I next analysed 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 5.5E**). I observed an overall increase in mean mRNA polyA tail lengths during the MZT (**Figure 5.5F**, see also **Figure 5.6E,F**), in agreement with previous reports (**Figure 5.7A-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 5.6G,** see also **Table S2**), in agreement with previous studies reporting their non-polyadenylated status [414]. Moreover, these findings show that Nano3P-seq is able to capture RNA molecules with structured 3'ends, such as those found in histones [415].

Finally, I 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 [416]. I 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 5.5G**), as could be expected. However, the patterns of polyA tail length dynamics heavily varied depending on the decay mode of the transcript (**Figure 5.5H**). Specifically, I 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 [43,417]. 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 5.7C-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.

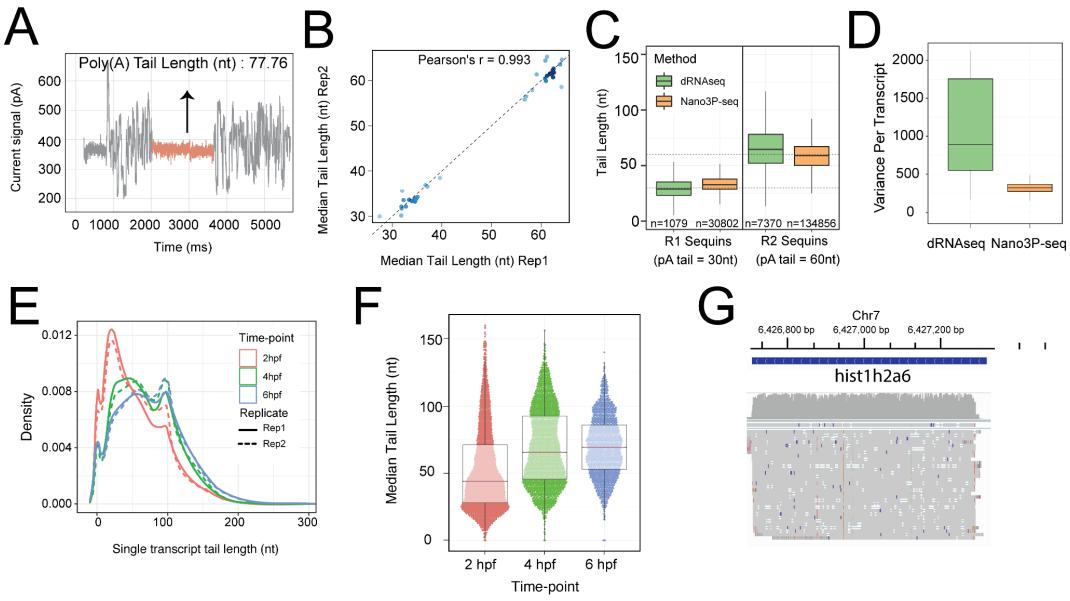


Figure 5.6 - Analysis of poly(A) tail lengths using Nano3P-seq.

(A) Current intensity plot of a synthetic poly(A)+ read, obtained using Nano3P-seq. The homopolymeric poly(T) region is highlighted in orange. **(B)** Replicability of median per-gene poly(A) tail length estimation in sequins captured with Nano3P-seq. The poly(A) tail length of synthetic sequins is 30nt (R1_sequins) or 60nt (R2_sequins). **(C)** Overall comparison of poly(A) tail length estimation of R1 and R2 sequins which contain 30 nt and 60 nt poly(A) tail lengths, respectively, obtained using dRNAseq (green) and Nano3P-seq (orange). **(D)** Per-transcript variance of poly(A) tail length estimations of

sequins obtained using dRNAseq (green) and Nano3P-seq (orange). **(E)** Distribution of poly(A) tail lengths in mRNAs across zebrafish developmental stages (2, 4 and 6hpf, shown in red, green and blue respectively) in two biological replicates (shown as full lines and dashed lines, respectively). **(F)** Median per-gene poly(A) tail length distribution of mRNAs during the zebrafish MZT (t=2, 4 and 6hpf). **(G)** IGV snapshot of reads mapping to hist1h2a6 mRNA, which do not contain poly(A) tails.

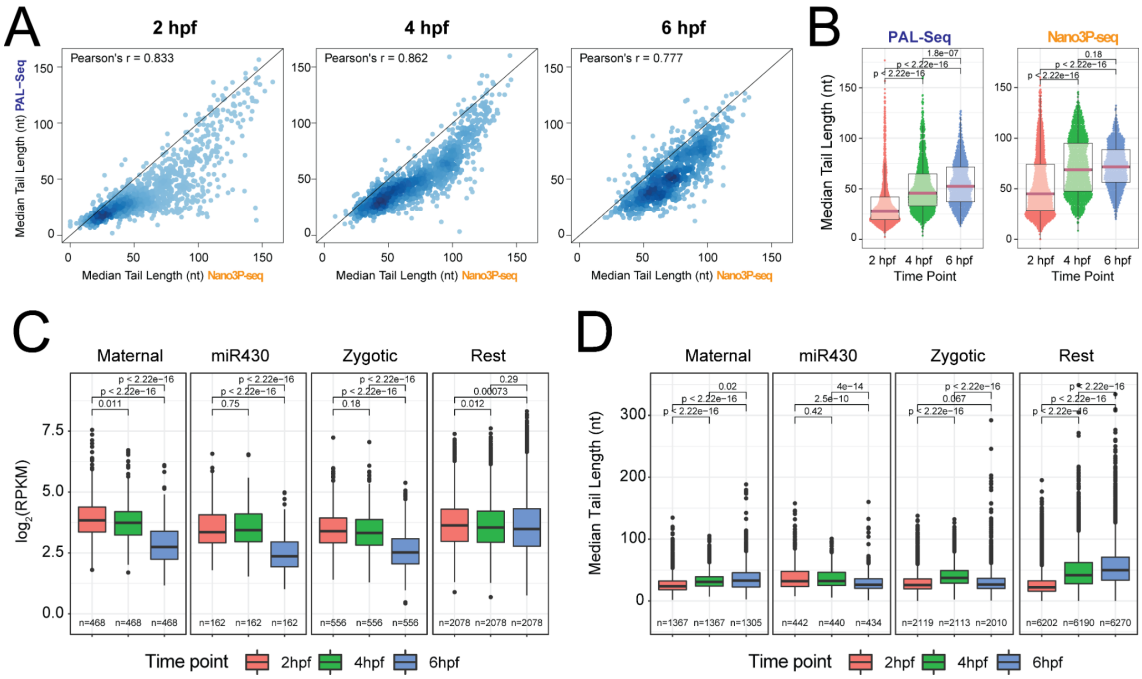


Figure 5.7 - Comparison of poly(A) tail length estimations using PAL-Seq and Nano3P-seq.

(A) Scatterplots of per-gene poly(A) tail length estimations using Nano3P-seq and PAL-seq from zebrafish mRNAs at 2 hpf (left), 4 hpf (middle) and 6 hpf (right). Each dot represents the median poly(A) tail length of a given gene. **(B)** Boxplots depicting the distribution of poly(A) tail lengths during the zebrafish MZT, estimated using PAL-Seq (left) or Nano3P-seq (right). Statistical comparison of means was made using Kruskal Wallis test. **(C)** Comparative analysis of the mRNA abundance (shown as $\log_2$ RPKM) for the 4 groups of zebrafish mRNAs (maternal, miR-430, zygotic and rest) during embryogenesis (t= 2, 4 and 6 hpf) using PAL-seq data. Statistical comparison of means was performed using Kruskal-Wallis test. **(D)** Median tail length estimations of the 4 groups of zebrafish mRNAs (maternal, miR-430, zygotic and rest) at 2, 4 and 6 hpf using PAL-seq data. Statistical analyses were performed using Kruskal-Wallis test.

5.6. 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. I first compared isoform-specific polyA tail lengths across isoforms encoded by the same gene, finding that 5.2% of analysed genes presented significant differences in their polyA tail lengths across isoforms , and that these differences were often conserved across the different time points analysed (**Figure 5.8A**). However, in other cases, the behaviour of polyA tails across isoforms was markedly distinct as the MZT progressed (**Figure 5.8B,C,** see also **Figure 5.9**), 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.

I then examined whether polyA tail lengths significantly diverged across time points at per-isoform level, finding that 11.7% of analysed transcripts significantly varied their polyA tail lengths during the MZT (**Figure 5.8D,E**). These results show that polyA tail length dynamics is not only 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.

5.7. 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 [13,200,256,418]. 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 [115,127,244,419].

I 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 5.10A,B**). 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 5.10C,D**). Therefore, the 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 [420]. Altogether, the 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.

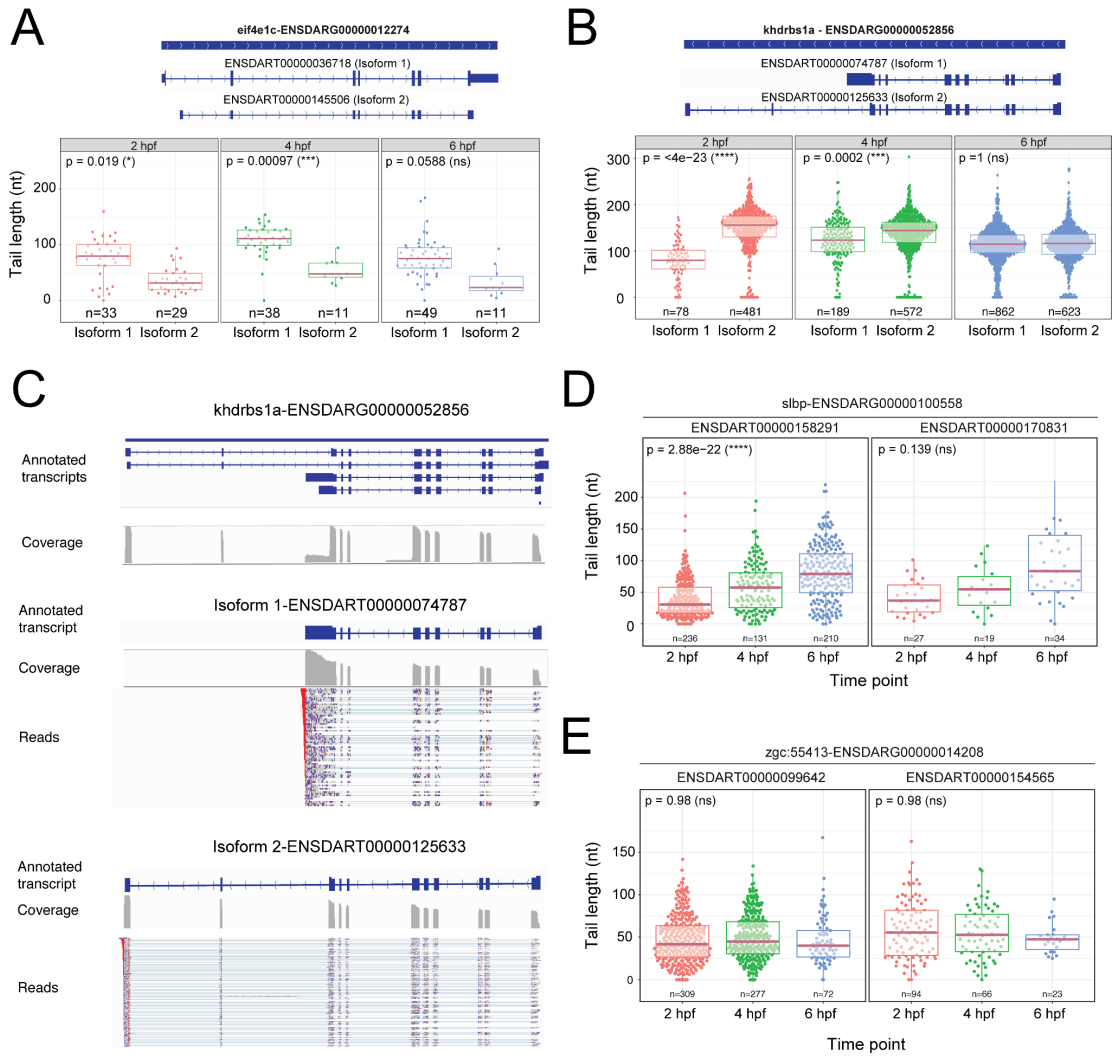


Figure 5.8 - Isoform-specific polyA tail dynamics can be captured using Nano3P-seq.

(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 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 5.9

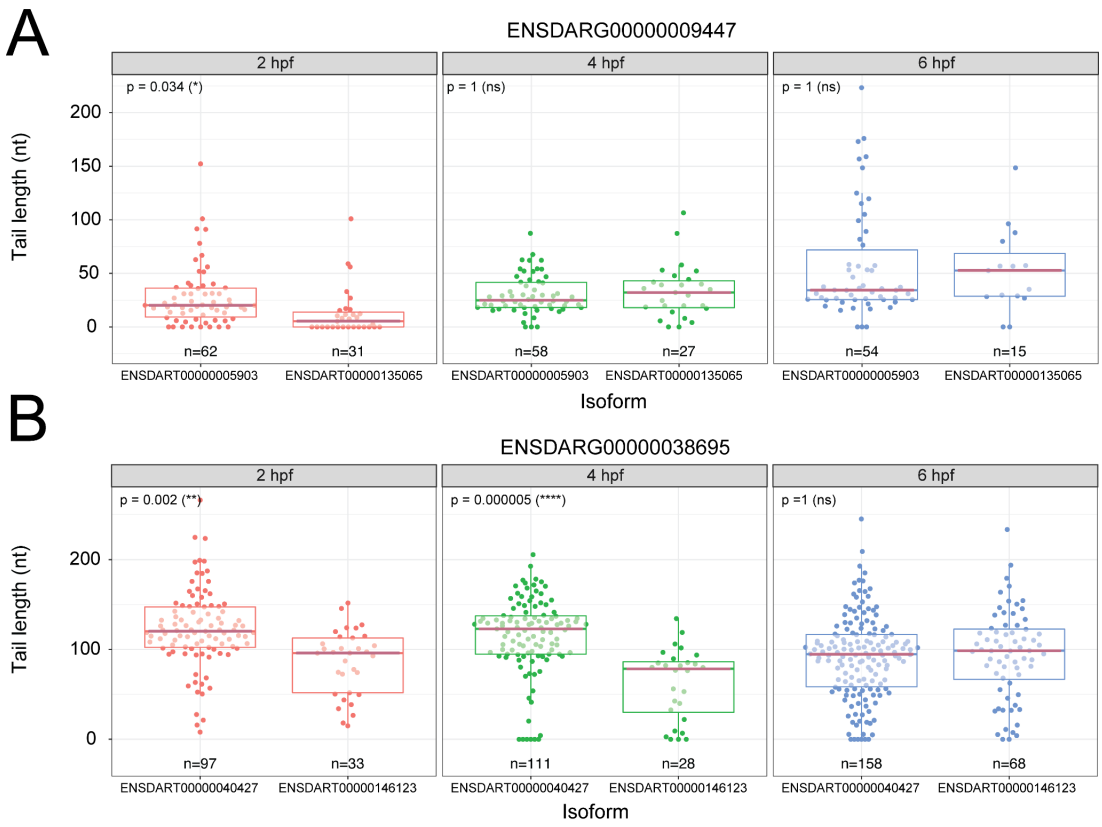


Figure 5.9 - Analysis of Isoform-specific poly(A) tail dynamics using Nano3P-seq.

Examples of genes with differentially polyadenylated isoforms between 2 and 6 hpf.

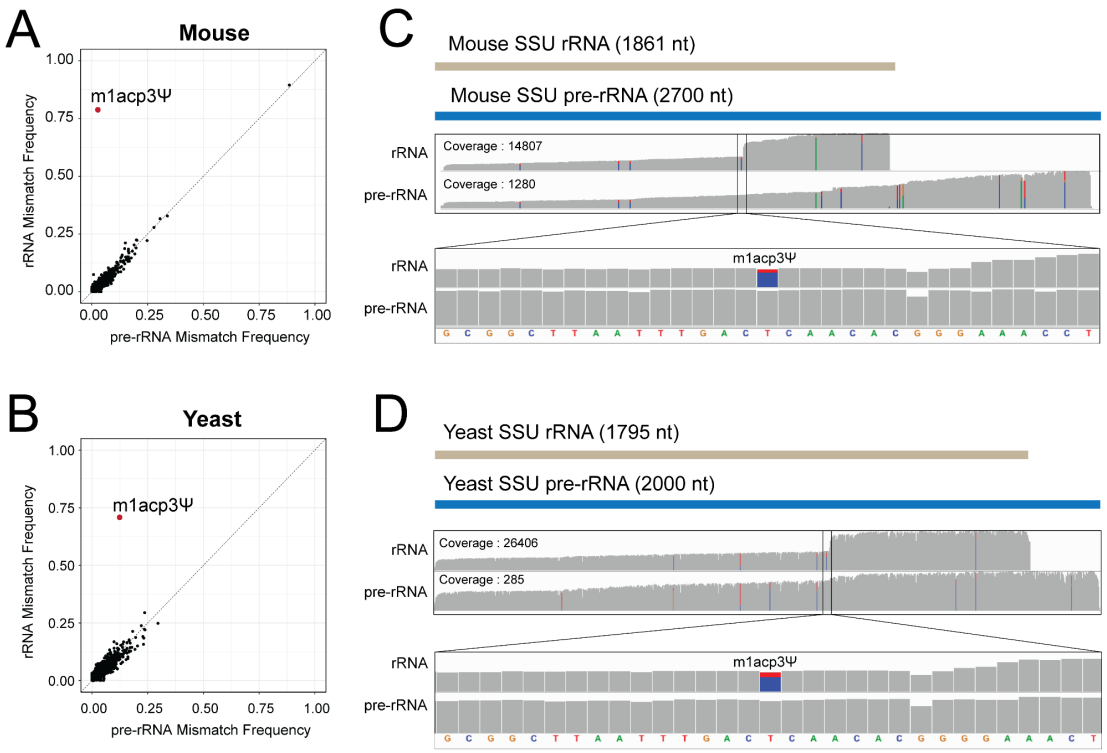


Figure 5.10 - Nano3P-seq identifies differential RNA modified sites in pre-rRNAs and mature rRNAs.

(A) Comparison of the per-site mismatch frequencies observed in reads mapping to mouse (A) and yeast (B) 18S pre-rRNA, relative to 18S rRNA, showing that the unique outlier identified is m^1acp^3Y. **(C,D)** IGV coverage tracks of reads mapping to mouse (C) and yeast (D) 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. Positions with mismatch frequency lower than 0.1 are shown in gray.

5.8. Materials and Methods

5.8.1. *In vitro* transcription of RNAs

The synthetic 'curlcake' sequences [255] (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 5.1A**).

Concentration was determined using Qubit Fluorometric Quantitation. Purity of the IVT products was measured with NanoDrop 2000 Spectrophotometer.

5.8.2. 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.

5.8.3. RNA isolation from mouse brain

In order to isolate nuclear/mitochondrial-enriched RNA from the mouse (*Mus musculus*) brain, I followed previously published protocols [421] 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 homogenised using a dounce, and the homogenate was 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.

5.8.4. 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.

5.8.5. 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.

5.8.6. 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 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 hybridised 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).

5.8.7. 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 software version v.3.5.5.

5.8.8. 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 5.1A**). 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.

5.8.9. Analysis of dRNA datasets

Barcoded direct RNA sequencing (dRNA-seq) run was demultiplexed using *DeePlexiCon* [387]. 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 [408] with minimap2 with -ax splice -k14 -uf --MD parameters [365]. 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 [422] and then to the genome (GRCz11) with minimap2 [365] with -ax splice -k14 -uf --MD parameters. Mapped reads were intersected with ENSEMBL version 103 annotation (Danio_rerio.GRCz11.103.2.gtf) using bedtools intersect option [423].

5.8.10. 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 [365] 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 [255], 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, I 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 [408], I 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, I 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.

5.8.11. Estimation of polyA tail lengths

For direct RNA sequencing reads, polyA tail length estimation was performed using *NanoTail*, a module from *Master of Pores [424]*, a nextflow workflow for the analysis of direct RNA datasets, which uses internally Nanopolish v0.11.1 [390]. 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* [411] with default parameters. I 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.

5.8.12. 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).

5.9. Discussion

In the last few years, a variety of NGS-based high-throughput methods have been developed to characterise the 3' ends of RNA molecules at a transcriptome-wide scale, including methods to reveal polyA tail sites (e.g., 3P-seq [403], PAS-seq [425], PAT-seq [426]) and to estimate polyA tail lengths (e.g., PAL-seq [42], TAIL-seq [265], mTAIL-seq [427]). 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 [266] and PAIso-seq [267]. 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 [268–270]. Moreover, PacBio typically requires expensive sequencing instruments that are not widely available. On the other hand, direct RNA nanopore sequencing [271] has also been proposed as an alternative long-read sequencing technology to study polyA tail lengths [272]; 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 5.1A**), 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 maximises 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 [428,429].

Nano3P-seq provides quantitative measurements of RNA abundances (**Figure 5.1D**) as well as captures diverse RNA biotypes regardless of their tail end composition (**Figure 5.3D**). It can be applied to diverse species with a distinct range of polyA tail lengths (**Figure 5.5B,C**), and can be used to study the dynamics of polyadenylation (Figure **5.5F-H** and **5.8**). Specifically, I demonstrate how Nano3P-seq provides per-read resolution transcriptome-wide maps of RNA abundance and polyadenylation dynamics during the zebrafish MZT. The 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 5.5G,H**). Moreover, I 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 5.8A,B**, see also **Figure 5.9**). Overall, 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, I 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, Ifind that ribo-depletion captures a larger variety of RNA biotypes compared to polyA+ selection, including several non-polyA tailed RNA biotypes, as expected (**Figure 5.3F**). However, I did not observe a significant difference in the distribution of mRNA polyA tail lengths between the two methods (**Figure 5.11**), suggesting that oligo(dT) enrichments do not significantly bias the mRNA populations by preferentially enriching for those having longer polyA tails.

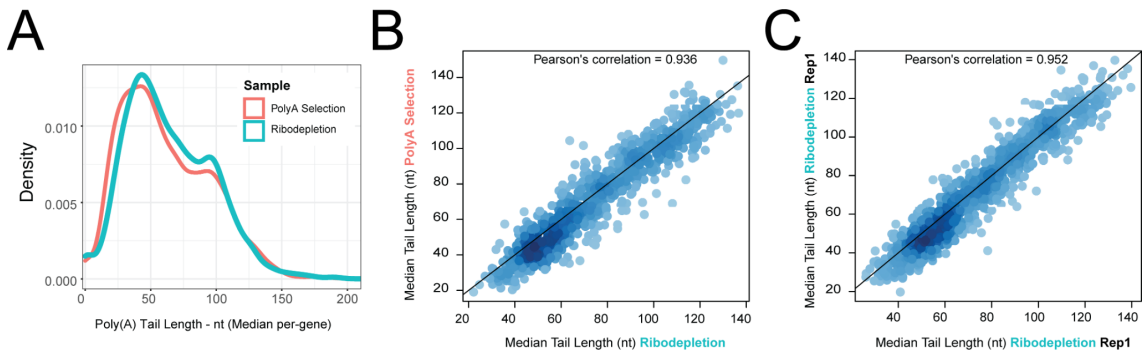


Figure 5.11 - Comparison of poly(A) tail length estimations using poly(A)-selected and ribodepleted samples.

(A) Distribution of per-gene mRNA poly(A) tail lengths from 4 hpf zebrafish embryos, isolated using either poly(A) selection (red) or ribo-depletion (cyan). **(B)** Comparison of median per-gene poly(A) tail length estimation between poly(A) selected and ribo-depleted zebrafish mRNAs isolated at 4 hpf. Each dot represents a gene. **(C)** Comparison of median per-gene poly(A) tail length estimation of mRNAs in zebrafish ribo-depleted samples (replicate 1 and 2) isolated at 4 hpf. Each dot represents a gene.

In addition, I performed a comparative analysis of zebrafish polyA tail lengths in libraries sequenced using either Nano3P-seq or direct RNA sequencing (dRNA-seq). I show that Nano3P-seq captures RNA molecules regardless of their tail ends, resulting in the capture of diverse RNA biotypes (**Figure 5.1B** and **5.2C-G**) including deadenylated mRNA molecules (**Figure 5.6E-F** and **5.12A**). 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, I observed that Nano3P-seq captured mRNAs with predicted zero tail lengths, whereas dRNA-seq only captured reads with longer tails (**Figure 5.12**).

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

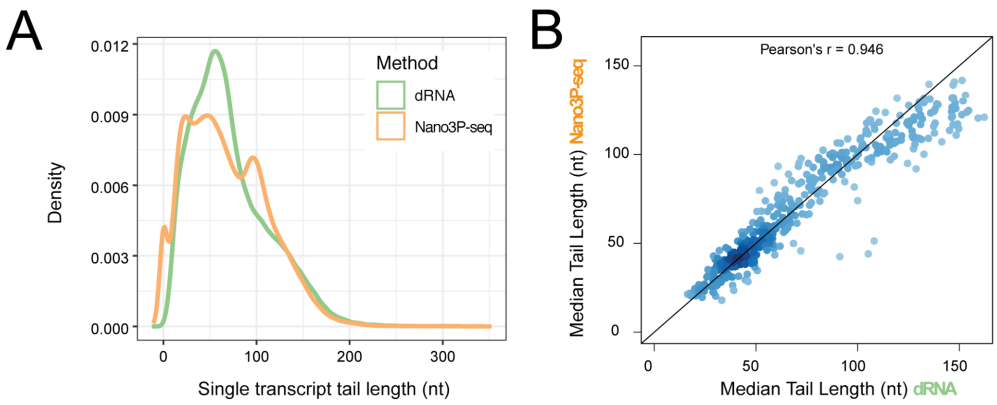


Figure 5.12 - Comparison of poly(A) tail length estimations between dRNAseq and Nano3P-seq.

(A) Distribution of per-gene mRNA poly(A) tail lengths from 4 hpf zebrafish embryos isolated using poly(A) selection and sequenced with dRNAseq (green) or Nano3P-seq (orange). **(B)** Comparison of median per-gene poly(A) tail length estimation of poly(A)-selected mRNAs isolated at 4 hpf with dRNAseq (green) or Nano3P-seq (orange). Each dot represents an mRNA.

6. Concluding Remarks

Post-transcriptional regulation of RNA is essential for the functionality of these molecules. In order to be able to comprehend the range of post-transcriptional regulations, it is important to develop new approaches that provide more information. This thesis describes the adaptation of third-generation sequencing technologies to characterise the RNA modification and polyadenylation landscape at single molecule resolution.

RNA modifications play central roles in cellular fate and differentiation. However, the machinery responsible for placing, removing, and recognizing more than 170 RNA modifications remains largely uncharacterised and poorly annotated, and we currently lack integrative studies that identify which RNA modification-related proteins (RMPs) may be dysregulated in each cancer type. The work in this thesis first aimed to characterise the RNA modification-related proteins. In order to do so, I performed a comprehensive annotation and evolutionary analysis of human RMPs, as well as an integrative analysis of their expression patterns across 32 tissues, 10 species, and 13,358 paired tumor-normal human samples. This analysis reveals an unanticipated heterogeneity of RMP expression patterns across mammalian tissues, with a vast proportion of duplicated enzymes displaying testis-specific expression, suggesting a key role for RNA modifications in sperm formation and possibly intergenerational inheritance.

I also uncovered many RMPs that are dysregulated in various types of cancer, and whose expression levels are predictive of cancer progression. Surprisingly, I found that several commonly studied RNA modification enzymes such as METTL3 or FTO are not significantly upregulated in most cancer types, whereas several less-characterised RMPs, such as LAGE3 and HENMT1, are dysregulated in many cancers. Overall, this part of the thesis provides novel targets for future cancer research studies targeting the human epitranscriptome, as well as foundations to understand cell type-specific behaviors that are orchestrated by RNA modifications.

The epitranscriptomics field has undergone an enormous expansion in the last few years; however, a major limitation is the lack of generic methods to map RNA modifications transcriptome-wide. This thesis aimed to provide a proof of concept of using ONT direct RNA sequencing in order to detect m^6A modification. This work shows that using direct RNA sequencing, *N*6-methyladenosine (m6A) RNA modifications can be detected with high accuracy, in the form of systematic errors and decreased base-calling qualities. Specifically, our algorithm, trained with m6A-modified

and unmodified synthetic sequences, can predict m6A RNA modifications with ~90% accuracy. This work then extends our findings to yeast data sets, finding that the method can identify m6A RNA modifications *in-vivo* with an accuracy of 87%. Moreover, this work further validates the method by showing that these 'errors' are typically not observed in yeast *ime4*-knockout strains, which lack m6A modifications. These results open avenues to investigate the biological roles of RNA modifications in their native RNA context.

Expanding on the ability of nanopore sequencing to detect *N*6-methyladenosine, this thesis also shows that other modifications, in particular pseudouridine (Ψ) and 2′-*O*-methylation (Nm), also result in characteristic base-calling 'error' signatures in the nanopore data. Focusing on Ψ modification sites, I detected known and uncovered previously unreported Ψ sites in mRNAs, non-coding RNAs and rRNAs, including a Pus4-dependent Ψ modification in yeast mitochondrial rRNA. To explore the dynamics of pseudouridylation, yeast cells were treated with oxidative, cold and heat stresses and detected heat-sensitive Ψ-modified sites in small nuclear RNAs, small nucleolar RNAs, and mRNAs. Finally, this work has led to the development of a software, nanoRMS, that estimates per-site modification stoichiometries by identifying single-molecule reads with altered current intensity and trace profiles. This work demonstrates that Nm and Ψ RNA modifications can be detected in cellular RNAs and that their modification stoichiometry can be quantified by nanopore sequencing of native RNA.

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. As a final part of this thesis, I developed Nano3P-seq, a novel method that relies on nanopore sequencing to simultaneously quantify RNA abundance 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. I 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. I find that, in addition to mRNAs and lncRNAs, polyA tails can be identified in 16S mitochondrial rRNA in both mice and zebrafish. Moreover, I 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.

7. References

1. Carter R, Drouin G. Structural differentiation of the three eukaryotic RNA polymerases. Genomics. 2009;94: 388–396.

2. Barba-Aliaga M, Alepuz P, Pérez-Ortín JE. Eukaryotic RNA Polymerases: The Many Ways to Transcribe a Gene. Front Mol Biosci. 2021;8: 663209.

3. Buratowski S, Hahn S, Guarente L, Sharp PA. Five intermediate complexes in transcription initiation by RNA polymerase II. Cell. 1989;56: 549–561.

4. Flores O, Lu H, Reinberg D. Factors involved in specific transcription by mammalian RNA polymerase II. Identification and characterization of factor IIH. J Biol Chem. 1992;267: 2786–2793.

5. Matsui T, Segall J, Weil PA, Roeder RG. Multiple factors required for accurate initiation of transcription by purified RNA polymerase II. J Biol Chem. 1980;255: 11992–11996.

6. Watkins NJ, Bohnsack MT. The box C/D and H/ACA snoRNPs: key players in the modification, processing and the dynamic folding of ribosomal RNA. Wiley Interdiscip Rev RNA. 2012;3: 397–414.

7. Venema J, Tollervey D. Ribosome synthesis in Saccharomyces cerevisiae. Annu Rev Genet. 1999;33: 261–311.

8. Turowski TW, Tollervey D. Transcription by RNA polymerase III: insights into mechanism and regulation. Biochem Soc Trans. 2016;44: 1367–1375.

9. Hsin J-P, Manley JL. The RNA polymerase II CTD coordinates transcription and RNA processing. Genes Dev. 2012;26: 2119–2137.

10. Cramer P. Eukaryotic Transcription Turns 50. Cell. 2019;179: 808–812.

11. Saldi T, Cortazar MA, Sheridan RM, Bentley DL. Coupling of RNA Polymerase II Transcription Elongation with Pre-mRNA Splicing. J Mol Biol. 2016;428: 2623–2635.

12. Corbett AH. Post-transcriptional regulation of gene expression and human disease. Curr Opin Cell Biol. 2018;52: 96–104.

13. Roundtree IA, Evans ME, Pan T, He C. Dynamic RNA Modifications in Gene Expression Regulation. Cell. 2017;169: 1187–1200.

14. Kaneko T, Suzuki T, Kapushoc ST, Rubio MA, Ghazvini J, Watanabe K, et al. Wobble modification differences and subcellular localization of tRNAs in Leishmania tarentolae: implication for tRNA sorting mechanism. EMBO J. 2003;22: 657–667.

15. Wang X, Zhao BS, Roundtree IA, Lu Z, Han D, Ma H, et al. N(6)-methyladenosine Modulates Messenger RNA Translation Efficiency. Cell. 2015;161: 1388–1399.

16. Novoa EM, Ribas de Pouplana L. Speeding with control: codon usage, tRNAs, and ribosomes. Trends Genet. 2012;28: 574–581.

17. Ke S, Pandya-Jones A, Saito Y, Fak JJ, Vågbø CB, Geula S, et al. m6A mRNA modifications are deposited in nascent pre-mRNA and are not required for splicing but do specify cytoplasmic turnover. Genes Dev. 2017;31: 990–1006.

18. Schaefer M, Pollex T, Hanna K, Tuorto F, Meusburger M, Helm M, et al. RNA methylation by Dnmt2 protects transfer RNAs against stress-induced cleavage. Genes Dev. 2010;24: 1590–1595.

19. Alexandrov A, Chernyakov I, Gu W, Hiley SL, Hughes TR, Grayhack EJ, et al. Rapid tRNA decay can result from lack of nonessential modifications. Mol Cell. 2006;21: 87–96.

20. Shatkin AJ, Manley JL. The ends of the affair: capping and polyadenylation. Nat Struct Biol. 2000;7: 838–842.

21. Shuman S. Structure, mechanism, and evolution of the mRNA capping apparatus. Prog Nucleic Acid Res Mol Biol. 2001;66: 1–40.

22. Ramanathan A, Robb GB, Chan S-H. mRNA capping: biological functions and applications. Nucleic Acids Res. 2016;44: 7511–7526.

23. Pan Q, Shai O, Lee LJ, Frey BJ, Blencowe BJ. Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. Nat Genet. 2008;40: 1413–1415.

24. Deveson IW, Brunck ME, Blackburn J, Tseng E, Hon T, Clark TA, et al. Universal Alternative Splicing of Noncoding Exons. Cell Syst. 2018;6: 245–255.e5.

25. Kastner B, Will CL, Stark H, Lührmann R. Structural Insights into Nuclear pre-mRNA Splicing in Higher Eukaryotes. Cold Spring Harb Perspect Biol. 2019;11. doi:10.1101/cshperspect.a032417

26. Will CL, Lührmann R. Spliceosome structure and function. Cold Spring Harb Perspect Biol. 2011;3. doi:10.1101/cshperspect.a003707

27. Roy SW, Gilbert W. The evolution of spliceosomal introns: patterns, puzzles and progress. Nat Rev Genet. 2006;7: 211–221.

28. Chen W, Feng P-M, Lin H, Chou K-C. iSS-PseDNC: identifying splicing sites using pseudo dinucleotide composition. Biomed Res Int. 2014;2014: 623149.

29. Nicholson AL, Pasquinelli AE. Tales of Detailed Poly(A) Tails. Trends Cell Biol. 2019;29: 191–200.

30. Takagaki Y, Ryner LC, Manley JL. Separation and characterization of a poly(A) polymerase and a cleavage/specificity factor required for pre-mRNA polyadenylation. Cell. 1988. pp. 731–742. doi:10.1016/0092-8674(88)90411-4

31. Keller W, Bienroth S, Lang KM, Christofori G. Cleavage and polyadenylation factor CPF specifically interacts with the pre-mRNA 3′ processing signal AAUAAA. EMBO J. 1991;10: 4241–4249.

32. Carswell S, Alwine JC. Efficiency of utilization of the simian virus 40 late polyadenylation site: effects of upstream sequences. Mol Cell Biol. 1989;9: 4248–4258.

33. Brackenridge S, Proudfoot NJ. Recruitment of a basal polyadenylation factor by the upstream sequence element of the human lamin B2 polyadenylation signal. Mol Cell Biol. 2000;20: 2660–2669.

34. Gil A, Proudfoot NJ. A sequence downstream of AAUAAA is required for rabbit beta-globin mRNA 3'-end formation. Nature. 1984;312: 473–474.

35. Gil A, Proudfoot NJ. Position-dependent sequence elements downstream of AAUAAA are required for efficient rabbit beta-globin mRNA 3' end formation. Cell. 1987;49: 399–406.

36. Shi Y, Manley JL. The end of the message: multiple protein–RNA interactions define the mRNA polyadenylation site. Genes Dev. 2015;29: 889–897.

37. Mandel CR, Kaneko S, Zhang H, Gebauer D, Vethantham V, Manley JL, et al. Polyadenylation factor CPSF-73 is the pre-mRNA 3'-end-processing endonuclease. Nature. 2006;444: 953–956.

38. Kühn U, Wahle E. Structure and function of poly(A) binding proteins. Biochimica et Biophysica Acta (BBA) - Gene Structure and Expression. 2004. pp. 67–84. doi:10.1016/j.bbaexp.2004.03.008

39. Tarun SZ Jr, Wells SE, Deardorff JA, Sachs AB. Translation initiation factor eIF4G mediates in vitro poly(A) tail-dependent translation. Proc Natl Acad Sci U S A. 1997;94: 9046–9051.

40. Kahvejian A, Svitkin YV, Sukarieh R, M'Boutchou M-N, Sonenberg N. Mammalian poly (A)-binding protein is a eukaryotic translation initiation factor, which acts via multiple mechanisms. Genes Dev. 2005;19: 104–113.

41. Eliseeva IA, Lyabin DN, Ovchinnikov LP. Poly(A)-binding proteins: Structure, domain organization, and activity regulation. Biochemistry . 2013;78: 1377–1391.

42. Subtelny AO, Eichhorn SW, Chen GR, Sive H, Bartel DP. Poly(A)-tail profiling reveals an embryonic switch in translational control. Nature. 2014;508: 66–71.

43. Giraldez AJ, Mishima Y, Rihel J, Grocock RJ, Van Dongen S, Inoue K, et al. Zebrafish MiR-430 promotes deadenylation and clearance of maternal mRNAs. Science. 2006;312: 75–79.

44. Weill L, Belloc E, Bava F-A, Méndez R. Translational control by changes in poly (A) tail length: recycling mRNAs. Nat Struct Mol Biol. 2012;19: 577–585.

45. D'Ambrogio A, Nagaoka K, Richter JD. Translational control of cell growth and malignancy by the CPEBs. Nat Rev Cancer. 2013;13: 283–290.

46. Barbieri I, Kouzarides T. Role of RNA modifications in cancer. Nat Rev Cancer. 2020;20: 303–322.

47. Saletore Y, Meyer K, Korlach J, Vilfan ID, Jaffrey S, Mason CE. The birth of the Epitranscriptome: deciphering the function of RNA modifications. Genome Biol. 2012;13: 175.

48. Novoa EM, Mason CE, Mattick JS. Charting the unknown epitranscriptome. Nat Rev Mol Cell Biol. 2017;18: 339–340.

49. Lence T, Akhtar J, Bayer M, Schmid K, Spindler L, Ho CH, et al. m6A modulates neuronal functions and sex determination in Drosophila. Nature. 2016;540: 242–247.

50. Zhang Y, Zhang X, Shi J, Tuorto F, Li X, Liu Y, et al. Dnmt2 mediates intergenerational transmission of paternally acquired metabolic disorders through sperm small non-coding RNAs. Nat Cell Biol. 2018;20: 535–540.

51. Batista PJ, Molinie B, Wang J, Qu K, Zhang J, Li L, et al. m(6)A RNA modification controls cell fate transition in mammalian embryonic stem cells. Cell Stem Cell. 2014;15: 707–719.

52. Jonkhout N, Tran J, Smith MA, Schonrock N, Mattick JS, Novoa EM. The RNA modification landscape in human disease. RNA. 2017;23: 1754–1769.

53. Torres AG, Batlle E, Ribas de Pouplana L. Role of tRNA modifications in human diseases. Trends Mol Med. 2014;20: 306–314.

54. Bednarova A, Hanna M, Durham I, VanCleave T, England A, Chaudhuri A, et al. Lost in Translation: Defects in Transfer RNA Modifications and Neurological Disorders. Front Mol Neurosci. 2017;10: 135.

55. Sarin LP, Leidel SA. Modify or die?--RNA modification defects in metazoans. RNA Biol. 2014;11: 1555–1567.

56. Pereira M, Francisco S, Varanda AS, Santos M, Santos MAS, Soares AR. Impact of tRNA Modifications and tRNA-Modifying Enzymes on Proteostasis and Human Disease. Int J Mol Sci. 2018;19. doi:10.3390/ijms19123738

57. Jia G, Fu Y, Zhao X, Dai Q, Zheng G, Yang Y, et al. N6-methyladenosine in nuclear RNA is a major substrate of the obesity-associated FTO. Nat Chem Biol. 2011;7: 885–887.

58. Mauer J, Luo X, Blanjoie A, Jiao X, Grozhik AV, Patil DP, et al. Reversible methylation of m6Am in the 5’ cap controls mRNA stability. Nature. 2016. doi:10.1038/nature21022

59. Qin Y, Li L, Luo E, Hou J, Yan G, Wang D, et al. Role of m6A RNA methylation in cardiovascular disease (Review). Int J Mol Med. 2020;46: 1958–1972.

60. Esteve-Puig R, Bueno-Costa A, Esteller M. Writers, readers and erasers of RNA modifications in cancer. Cancer Lett. 2020;474: 127–137.

61. Bokar JA, Rath-Shambaugh ME, Ludwiczak R, Narayan P, Rottman F. Characterization and partial purification of mRNA N6-adenosine methyltransferase from HeLa cell nuclei. Internal mRNA methylation requires a multisubunit complex. J Biol Chem. 1994;269: 17697–17704.

62. Westbye MP, Feyzi E, Aas PA, Vågbø CB, Talstad VA, Kavli B, et al. Human AlkB homolog 1 is a mitochondrial protein that demethylates 3-methylcytosine in DNA and RNA. J Biol Chem. 2008;283: 25046–25056.

63. Liu F, Clark W, Luo G, Wang X, Fu Y, Wei J, et al. ALKBH1-Mediated tRNA Demethylation Regulates Translation. Cell. 2016;167: 1897.

64. Monsen VT, Sundheim O, Aas PA, Westbye MP, Sousa MML, Slupphaug G, et al.

Divergent β-hairpins determine double-strand versus single-strand substrate recognition of human AlkB-homologues 2 and 3. Nucleic Acids Res. 2010;38: 6447–6455.

65. Woo H-H, Chambers SK. Human ALKBH3-induced m1A demethylation increases the CSF-1 mRNA stability in breast and ovarian cancer cells. Biochim Biophys Acta Gene Regul Mech. 2019;1862: 35–46.

66. Kasowitz SD, Ma J, Anderson SJ, Leu NA, Xu Y, Gregory BD, et al. Nuclear m6A reader YTHDC1 regulates alternative polyadenylation and splicing during mouse oocyte development. PLoS Genet. 2018;14: e1007412.

67. Xiao W, Adhikari S, Dahal U, Chen Y-S, Hao Y-J, Sun B-F, et al. Nuclear m6A Reader YTHDC1 Regulates mRNA Splicing. Mol Cell. 2016;61: 507–519.

68. Wang X, Lu Z, Gomez A, Hon GC, Yue Y, Han D, et al. N6-methyladenosine-dependent regulation of messenger RNA stability. Nature. 2014;505: 117–120.

69. Huang H, Weng H, Sun W, Qin X, Shi H, Wu H, et al. Recognition of RNA N6-methyladenosine by IGF2BP proteins enhances mRNA stability and translation. Nat Cell Biol. 2018;20: 285–295.

70. Yang X, Yang Y, Sun BF, Chen YS, Xu JW, Lai WY, et al. 5-methylcytosine promotes mRNA export - NSUN2 as the methyltransferase and ALYREF as an m(5)C reader. Cell Res. 2017;27: 606–625.

71. Lian H, Wang Q-H, Zhu C-B, Ma J, Jin W-L. Deciphering the Epitranscriptome in Cancer. Trends Cancer Res. 2018;4: 207–221.

72. Vu LP, Pickering BF, Cheng Y, Zaccara S, Nguyen D, Minuesa G, et al. The N6-methyladenosine (m6A)-forming enzyme METTL3 controls myeloid differentiation of normal hematopoietic and leukemia cells. Nat Med. 2017;23: 1369–1376.

73. Li A, Chen Y-S, Ping X-L, Yang X, Xiao W, Yang Y, et al. Cytoplasmic m6A reader YTHDF3 promotes mRNA translation. Cell Res. 2017;27: 444–447.

74. Tang C, Klukovich R, Peng H, Wang Z. ALKBH5-dependent m6A demethylation controls splicing and stability of long 3′-UTR mRNAs in male germ cells. Proceedings of the. 2018. Available: https://www.pnas.org/content/115/2/E325/

75. Desrosiers R, Friderici K, Rottman F. Identification of methylated nucleosides in messenger RNA from Novikoff hepatoma cells. Proc Natl Acad Sci U S A. 1974;71: 3971–3975.

76. Fazi F, Fatica A. Interplay Between N 6-Methyladenosine (m6A) and Non-coding RNAs in Cell Development and Cancer. Front Cell Dev Biol. 2019;7: 116.

77. Alarcón CR, Lee H, Goodarzi H, Halberg N, Tavazoie SF. N6-methyladenosine marks primary microRNAs for processing. Nature. 2015;519: 482–485.

78. Maden BE. Identification of the locations of the methyl groups in 18 S ribosomal RNA from Xenopus laevis and man. J Mol Biol. 1986;189: 681–699.

79. Maden BE. Locations of methyl groups in 28 S rRNA of Xenopus laevis and man. Clustering in the conserved core of molecule. J Mol Biol. 1988;201: 289–314.

80. Liu J, Yue Y, Han D, Wang X, Fu Y, Zhang L, et al. A METTL3-METTL14 complex mediates mammalian nuclear RNA N6-adenosine methylation. Nat Chem Biol. 2014;10: 93–95.

81. Pendleton KE, Chen B, Liu K, Hunter OV, Xie Y, Tu BP, et al. The U6 snRNA m6A Methyltransferase METTL16 Regulates SAM Synthetase Intron Retention. Cell. 2017;169: 824–835.e14.

82. van Tran N, Ernst FGM, Hawley BR, Zorbas C, Ulryck N, Hackert P, et al. The human 18S rRNA m6A methyltransferase METTL5 is stabilized by TRMT112. Nucleic Acids Res. 2019;47: 7719–7733.

83. Ma H, Wang X, Cai J, Dai Q, Natchiar SK, Lv R, et al. N6-Methyladenosine methyltransferase ZCCHC4 mediates ribosomal RNA methylation. Nat Chem Biol. 2019;15: 88–94.

84. Shi H, Wei J, He C. Where, When, and How: Context-Dependent Functions of RNA Methylation Writers, Readers, and Erasers. Mol Cell. 2019;74: 640–650.

85. Yang Y, Hsu PJ, Chen YS, Yang YG. Dynamic transcriptomic m(6)A decoration: writers, erasers, readers and functions in RNA metabolism. Cell Res. 2018;28: 616–624.

86. Post-transcriptional gene regulation by mRNA modifications. Nature Publishing Group 2016 pp. 1–12. doi:10.1038/nrm.2016.132

87. Csepany T, Lin A, Baldick CJ Jr, Beemon K. Sequence specificity of mRNA N6-adenosine methyltransferase. J Biol Chem. 1990;265: 20117–20122.

88. Kane SE, Beemon K. Precise localization of m6A in Rous sarcoma virus RNA reveals clustering of methylation sites: implications for RNA processing. Mol Cell Biol. 1985;5: 2298–2306.

89. Dominissini D, Moshitch-Moshkovitz S, Schwartz S, Salmon-Divon M, Ungar L, Osenberg S, et al. Topology of the human and mouse m6A RNA methylomes revealed by m6A-seq. Nature. 2012;485: 201–206.

90. Meyer KD, Saletore Y, Zumbo P, Elemento O, Mason CE, Jaffrey SR. Comprehensive analysis of mRNA methylation reveals enrichment in 3' UTRs and near stop codons. Cell. 2012;149: 1635–1646.

91. Chen K, Lu Z, Wang X, Fu Y, Luo G-Z, Liu N, et al. High-resolution N(6) -methyladenosine (m(6) A) map using photo-crosslinking-assisted m(6) A sequencing. Angew Chem Int Ed Engl. 2015;54: 1587–1590.

92. Linder B, Grozhik AV, Olarerin-George AO, Meydan C, Mason CE, Jaffrey SR. Single-nucleotide-resolution mapping of m6A and m6Am throughout the transcriptome. Nat Methods. 2015;12: 767–772.

93. Harcourt EM, Ehrenschwender T, Batista PJ, Chang HY, Kool ET. Identification of a selective polymerase enables detection of N(6)-methyladenosine in RNA. J Am Chem Soc. 2013;135: 19079–19082.

94. Liu N, Parisien M, Dai Q, Zheng G, He C, Pan T. Probing N6-methyladenosine RNA modification status at single nucleotide resolution in mRNA and long noncoding RNA. RNA. 2013;19: 1848–1856.

95. Molinie B, Wang J, Lim KS, Hillebrand R, Lu Z-X, Van Wittenberghe N, et al. m(6)A-LAIC-seq reveals the census and complexity of the m(6)A epitranscriptome. Nat Methods. 2016;13: 692–698.

96. Bohnsack KE, Höbartner C, Bohnsack MT. Eukaryotic 5-methylcytosine (m^5C) RNA Methyltransferases: Mechanisms, Cellular Functions, and Links to Disease. Genes . 2019;10. doi:10.3390/genes10020102

97. Haag S, Warda AS, Kretschmer J, Günnigmann MA, Höbartner C, Bohnsack MT. NSUN6 is a human RNA methyltransferase that catalyzes formation of m5C72 in specific tRNAs. RNA. 2015;21: 1532–1543.

98. Li Q, Li X, Tang H, Jiang B, Dou Y, Gorospe M, et al. NSUN2-Mediated m5C Methylation and METTL3/METTL14-Mediated m6A Methylation Cooperatively Enhance p21 Translation. Journal of Cellular Biochemistry. 2017. pp. 2587–2598. doi:10.1002/jcb.25957

99. Schosserer M, Minois N, Angerer TB, Amring M, Dellago H, Harreither E, et al. Methylation of ribosomal RNA by NSUN5 is a conserved mechanism modulating organismal lifespan. Nat Commun. 2015;6: 6158.

100. Gu X, Liang Z. Transcriptome-Wide Mapping 5-Methylcytosine by m5C RNA Immunoprecipitation Followed by Deep Sequencing in Plant. Methods Mol Biol. 2019;1933: 389–394.

101. Squires JE, Patel HR, Nousch M, Sibbritt T, Humphreys DT, Parker BJ, et al. Widespread occurrence of 5-methylcytosine in human coding and non-coding RNA. Nucleic Acids Res. 2012;40: 5023–5033.

102. Legrand C, Tuorto F, Hartmann M, Liebers R, Jacob D, Helm M, et al. Statistically robust methylation calling for whole-transcriptome bisulfite sequencing reveals distinct methylation patterns for mouse RNAs. Genome Res. 2017;27: 1589–1596.

103. Song C-X, Yi C, He C. Mapping recently identified nucleotide variants in the genome and transcriptome. Nat Biotechnol. 2012;30: 1107–1116.

104. Huang T, Chen W, Liu J, Gu N, Zhang R. Genome-wide identification of mRNA 5-methylcytosine in mammals. Nat Struct Mol Biol. 2019;26: 380–388.

105. Peifer C, Sharma S, Watzinger P, Lamberth S, Kötter P, Entian K-D. Yeast Rrp8p, a novel methyltransferase responsible for m1A 645 base modification of 25S rRNA. Nucleic Acids Res. 2013;41: 1151–1163.

106. Bar-Yaacov D, Frumkin I, Yashiro Y, Chujo T, Ishigami Y, Chemla Y, et al. Mitochondrial 16S rRNA Is Methylated by tRNA Methyltransferase TRMT61B in All Vertebrates. PLoS Biol. 2016;14: e1002557.

107. Shima H, Igarashi K. N 1-methyladenosine (m1A) RNA modification: the key to ribosome control. J Biochem. 2020;167: 535–539.

108. Saikia M, Fu Y, Pavon-Eternod M, He C, Pan T. Genome-wide analysis of N1-methyl-adenosine modification in human tRNAs. RNA. 2010;16: 1317–1327.

109. Safra M, Sas-Chen A, Nir R, Winkler R, Nachshon A, Bar-Yaacov D, et al. The m1A landscape on cytosolic and mitochondrial mRNA at single-base resolution.

Nature. 2017;551: 251–255.

110. Li X, Xiong X, Zhang M, Wang K, Chen Y, Zhou J, et al. Base-Resolution Mapping Reveals Distinct m1A Methylome in Nuclear- and Mitochondrial-Encoded Transcripts. Mol Cell. 2017;68: 993–1005.e9.

111. Schwartz S. m1A within cytoplasmic mRNAs at single nucleotide resolution: a reconciled transcriptome-wide map. RNA. 2018;24: 1427–1436.

112. Dominissini D, Nachtergaele S, Moshitch-Moshkovitz S, Peer E, Kol N, Ben-Haim MS, et al. The dynamic N(1)-methyladenosine methylome in eukaryotic messenger RNA. Nature. 2016;530: 441–446.

113. Alriquet M, Calloni G, Martínez-Limón A, Delli Ponti R, Hanspach G, Hengesbach M, et al. The protective role of m1A during stress-induced granulation. J Mol Cell Biol. 2021;12: 870–880.

114. Seo KW, Kleiner RE. YTHDF2 Recognition of N1-Methyladenosine (m1A)-Modified RNA Is Associated with Transcript Destabilization. ACS Chem Biol. 2020;15: 132–139.

115. Motorin Y, Muller S, Behm-Ansmant I, Branlant C. Identification of modified residues in RNAs by reverse transcription-based methods. Methods Enzymol. 2007;425: 21–53.

116. Hauenschild R, Tserovski L, Schmid K, Thuring K, Winz ML, Sharma S, et al. The reverse transcription signature of N-1-methyladenosine in RNA-Seq is sequence dependent. Nucleic Acids Res. 2015;43: 9950–9964.

117. Motorin Y, Helm M. Methods for RNA Modification Mapping Using Deep Sequencing: Established and New Emerging Technologies. Genes . 2019;10. doi:10.3390/genes10010035

118. Grozhik AV, Olarerin-George AO, Sindelar M, Li X, Gross SS, Jaffrey SR. Antibody cross-reactivity accounts for widespread appearance of m1A in 5'UTRs. Nat Commun. 2019;10: 5126.

119. Cohn WE, Volkin E. Nucleoside-5′-Phosphates from Ribonucleic Acid. Nature. 1951. pp. 483–484. doi:10.1038/167483a0

120. Ge J, Yu Y-T. RNA pseudouridylation: new insights into an old modification. Trends Biochem Sci. 2013;38: 210–218.

121. Schwartz S, Bernstein DA, Mumbach MR, Jovanovic M, Herbst RH, León-Ricardo BX, et al. Transcriptome-wide mapping reveals widespread dynamic-regulated pseudouridylation of ncRNA and mRNA. Cell. 2014;159: 148–162.

122. Carlile TM, Rojas-Duran MF, Zinshteyn B, Shin H, Bartoli KM, Gilbert WV. Pseudouridine profiling reveals regulated mRNA pseudouridylation in yeast and human cells. Nature. 2014;515: 143–146.

123. Carlile TM, Martinez NM, Schaening C, Su A, Bell TA, Zinshteyn B, et al. mRNA structure determines modification by pseudouridine synthase 1. Nat Chem Biol. 2019;15: 966–974.

124. Torchet C, Badis G, Devaux F, Costanzo G, Werner M, Jacquier A. The

complete set of H/ACA snoRNAs that guide rRNA pseudouridylations in Saccharomyces cerevisiae. RNA. 2005;11: 928–938.

125. Hamma T, Ferré-D'Amaré AR. Pseudouridine synthases. Chem Biol. 2006;13: 1125–1135.

126. Lovejoy AF, Riordan DP, Brown PO. Transcriptome-wide mapping of pseudouridines: pseudouridine synthases modify specific mRNAs in S. cerevisiae. PLoS One. 2014;9: e110799.

127. Khoddami V, Yerra A, Mosbruger TL, Fleming AM, Burrows CJ, Cairns BR. Transcriptome-wide profiling of multiple RNA modifications simultaneously at single-base resolution. Proc Natl Acad Sci U S A. 2019;116: 6784–6789.

128. Zhao Y, Dunker W, Yu Y-T, Karijolich J. The Role of Noncoding RNA Pseudouridylation in Nuclear Gene Expression Events. Front Bioeng Biotechnol. 2018;6: 8.

129. Dimitrova DG, Teysset L, Carré C. RNA 2'-O-Methylation (Nm) Modification in Human Diseases. Genes . 2019;10. doi:10.3390/genes10020117

130. Piekna-Przybylska D, Decatur WA, Fournier MJ. The 3D rRNA modification maps database: with interactive tools for ribosome analysis. Nucleic Acids Res. 2008;36: D178–83.

131. Dönmez G, Hartmuth K, Lührmann R. Modified nucleotides at the 5' end of human U2 snRNA are required for spliceosomal E-complex formation. RNA. 2004;10: 1925–1933.

132. Karunatilaka KS, Rueda D. Post-transcriptional modifications modulate conformational dynamics in human U2-U6 snRNA complex. RNA. 2014;20: 16–23.

133. Cavaillé J, Nicoloso M, Bachellerie JP. Targeted ribose methylation of RNA in vivo directed by tailored antisense RNA guides. Nature. 1996;383: 732–735.

134. Kiss-László Z, Henry Y, Bachellerie JP, Caizergues-Ferrer M, Kiss T. Site-specific ribose methylation of preribosomal RNA: a novel function for small nucleolar RNAs. Cell. 1996;85: 1077–1088.

135. Somme J, Van Laer B, Roovers M, Steyaert J, Versées W, Droogmans L. Characterization of two homologous 2'-O-methyltransferases showing different specificities for their tRNA substrates. RNA. 2014;20: 1257–1271.

136. Lee K-W, Bogenhagen DF. Assignment of 2'-O-methyltransferases to modification sites on the mammalian mitochondrial large subunit 16 S ribosomal RNA (rRNA). J Biol Chem. 2014;289: 24936–24942.

137. Lebars I, Legrand P, Aimé A, Pinaud N, Fribourg S, Di Primo C. Exploring TAR–RNA aptamer loop–loop interaction by X-ray crystallography, UV spectroscopy and surface plasmon resonance. Nucleic Acids Res. 2008;36: 7146–7156.

138. Hou Y-M. An important 2'-OH group for an RNA-protein interaction. Nucleic Acids Research. 2001. pp. 976–985. doi:10.1093/nar/29.4.976

139. Kurth HM, Mochizuki K. 2'-O-methylation stabilizes Piwi-associated small RNAs and ensures DNA elimination in Tetrahymena. RNA. 2009. pp. 675–685.

doi:10.1261/rna.1455509

140. Sproat BS, Lamond AI, Beijer B, Neuner P, Ryder U. Highly efficient chemical synthesis of 2′-O-methyloligoribonucleotides and tetrabiotinylated derivatives; novel probes that are resistant to degradation by RNA or DNA specific nucleases. Nucleic Acids Research. 1989. pp. 3373–3386. doi:10.1093/nar/17.9.3373

141. Marchand V, Blanloeil-Oillo F, Helm M, Motorin Y. Illumina-based RiboMethSeq approach for mapping of 2′-O-Me residues in RNA. Nucleic Acids Res. 2016;44: e135–e135.

142. Gumienny R, Jedlinski DJ, Schmidt A, Gypas F, Martin G, Vina-Vilaseca A, et al. High-throughput identification of C/D box snoRNA targets with CLIP and RiboMeth-seq. Nucleic Acids Res. 2017;45: 2341–2353.

143. Krogh N, Birkedal U, Nielsen H. RiboMeth-seq: Profiling of 2'-O-Me in RNA. Methods Mol Biol. 2017;1562: 189–209.

144. Incarnato D, Anselmi F, Morandi E, Neri F, Maldotti M, Rapelli S, et al. High-throughput single-base resolution mapping of RNA 2´-O-methylated residues. Nucleic Acids Res. 2017;45: 1433–1441.

145. Zhu Y, Pirnie SP, Carmichael GG. High-throughput and site-specific identification of 2′. O;

146. Dai Q, Moshitch-Moshkovitz S, Han D, Kol N, Amariglio N, Rechavi G, et al. Nm-seq maps 2'-O-methylation sites in human mRNA with base precision. Nat Methods. 2017;14: 695–698.

147. Hsu PJ, Fei Q, Dai Q, Shi H, Dominissini D, Ma L, et al. Single base resolution mapping of 2'-O-methylation sites in human mRNA and in 3' terminal ends of small RNAs. Methods. 2019;156: 85–90.

148. Blanc V, Davidson NO. C-to-U RNA Editing: Mechanisms Leading to Genetic Diversity *. J Biol Chem. 2003;278: 1395–1398.

149. Keegan L, Khan A, Vukic D, O'Connell M. ADAR RNA editing below the backbone. RNA. 2017;23: 1317–1328.

150. Kung C-P, Maggi LB Jr, Weber JD. The Role of RNA Editing in Cancer Development and Metabolic Disorders. Front Endocrinol . 2018;9: 762.

151. Powell LM, Wallis SC, Pease RJ, Edwards YH, Knott TJ, Scott J. A novel form of tissue-specific RNA processing produces apolipoprotein-B48 in intestine. Cell. 1987;50: 831–840.

152. Teng B, Burant CF, Davidson NO. Molecular cloning of an apolipoprotein B messenger RNA editing protein. Science. 1993;260: 1816–1819.

153. Bass BL, Weintraub H. A developmentally regulated activity that unwinds RNA duplexes. Cell. 1987;48: 607–613.

154. Rebagliati MR, Melton DA. Antisense RNA injections in fertilized frog eggs reveal an RNA duplex unwinding activity. Cell. 1987;48: 599–605.

155. Muramatsu M, Kinoshita K, Fagarasan S, Yamada S, Shinkai Y, Honjo T. Class

switch recombination and hypermutation require activation-induced cytidine deaminase (AID), a potential RNA editing enzyme. Cell. 2000;102: 553–563.

156. Lerner T, Papavasiliou FN, Pecori R. RNA Editors, Cofactors, and mRNA Targets: An Overview of the C-to-U RNA Editing Machinery and Its Implication in Human Disease. Genes . 2018;10. doi:10.3390/genes10010013

157. Yang L, Emerman M, Malik HS, McLaughlin RN Jnr. Retrocopying expands the functional repertoire of APOBEC3 antiviral proteins in primates. Elife. 2020;9. doi:10.7554/eLife.58436

158. Smith HC. RNA binding to APOBEC deaminases; Not simply a substrate for C to U editing. RNA Biol. 2017;14: 1153–1165.

159. Sharma S, Patnaik SK, Taggart RT, Kannisto ED, Enriquez SM, Gollnick P, et al. APOBEC3A cytidine deaminase induces RNA editing in monocytes and macrophages. Nat Commun. 2015;6: 6881.

160. Sharma S, Patnaik SK, Taggart RT, Baysal BE. The double-domain cytidine deaminase APOBEC3G is a cellular site-specific RNA editing enzyme. Sci Rep. 2016;6: 39100.

161. Bass BL, Nishikura K, Keller W, Seeburg PH, Emeson RB, O'Connell MA, et al. A standardized nomenclature for adenosine deaminases that act on RNA. RNA. 1997;3: 947–949.

162. Bass BL, Weintraub H. An unwinding activity that covalently modifies its double-stranded RNA substrate. Cell. 1988;55: 1089–1098.

163. Melcher T, Maas S, Herb A, Sprengel R, Higuchi M, Seeburg PH. RED2, a brain-specific member of the RNA-specific adenosine deaminase family. J Biol Chem. 1996;271: 31795–31798.

164. Wang Y, Chung DH, Monteleone LR, Li J, Chiang Y, Toney MD, et al. RNA binding candidates for human ADAR3 from substrates of a gain of function mutant expressed in neuronal cells. Nucleic Acids Res. 2019;47: 10801–10814.

165. Sommer B, Köhler M, Sprengel R, Seeburg PH. RNA editing in brain controls a determinant of ion flow in glutamate-gated channels. Cell. 1991;67: 11–19.

166. Higuchi M, Maas S, Single FN, Hartner J, Rozov A, Burnashev N, et al. Point mutation in an AMPA receptor gene rescues lethality in mice deficient in the RNA-editing enzyme ADAR2. Nature. 2000;406: 78–81.

167. Mladenova D, Barry G, Konen LM, Pineda SS, Guennewig B, Avesson L, et al. Adar3 Is Involved in Learning and Memory in Mice. Front Neurosci. 2018;12: 243.

168. Bahn JH, Lee J-H, Li G, Greer C, Peng G, Xiao X. Accurate identification of A-to-I RNA editing in human by transcriptome sequencing. Genome Res. 2012;22: 142–150.

169. Wahlstedt H, Daniel C, Enstërö M, Ohman M. Large-scale mRNA sequencing determines global regulation of RNA editing during brain development. Genome Res. 2009;19: 978–986.

170. Licht K, Kapoor U, Amman F, Picardi E, Martin D, Bajad P, et al. A high

resolution A-to-I editing map in the mouse identifies editing events controlled by pre-mRNA splicing. Genome Res. 2019;29: 1453–1463.

171. Daniel C, Lagergren J, Öhman M. RNA editing of non-coding RNA and its role in gene regulation. Biochimie. 2015;117: 22–27.

172. Daniel C, Silberberg G, Behm M, Öhman M. Alu elements shape the primate transcriptome by cis-regulation of RNA editing. Genome Biol. 2014;15: R28.

173. Liddicoat BJ, Piskol R, Chalk AM, Ramaswami G, Higuchi M, Hartner JC, et al. RNA editing by ADAR1 prevents MDA5 sensing of endogenous dsRNA as nonself. Science. 2015;349: 1115–1120.

174. Pestal K, Funk CC, Snyder JM, Price ND, Treuting PM, Stetson DB. Isoforms of RNA-Editing Enzyme ADAR1 Independently Control Nucleic Acid Sensor MDA5-Driven Autoimmunity and Multi-organ Development. Immunity. 2015;43: 933–944.

175. Mannion NM, Greenwood SM, Young R, Cox S, Brindle J, Read D, et al. The RNA-editing enzyme ADAR1 controls innate immune responses to RNA. Cell Rep. 2014;9: 1482–1494.

176. Schumacher JM, Lee K, Edelhoff S, Braun RE. Distribution of Tenr, an RNA-binding protein, in a lattice-like network within the spermatid nucleus in the mouse. Biol Reprod. 1995;52: 1274–1283.

177. Hough RF, Bass BL. Analysis of Xenopus dsRNA adenosine deaminase cDNAs reveals similarities to DNA methyltransferases. RNA. 1997;3: 356–370.

178. Gerber AP, Keller W. RNA editing by base deamination: more enzymes, more targets, new mysteries. Trends Biochem Sci. 2001;26: 376–384.

179. Torres AG, Piñeyro D, Filonava L, Stracker TH, Batlle E, Ribas de Pouplana L. A-to-I editing on tRNAs: biochemical, biological and evolutionary implications. FEBS Lett. 2014;588: 4279–4286.

180. Chen X-Y, Zhang J, Zhu J-S. The role of m6A RNA methylation in human cancer. Mol Cancer. 2019;18: 103.

181. Roundtree IA, Luo G-Z, Zhang Z, Wang X, Zhou T, Cui Y, et al. YTHDC1 mediates nuclear export of N6-methyladenosine methylated mRNAs. Elife. 2017;6. doi:10.7554/eLife.31311

182. Spitale RC, Flynn RA, Zhang QC, Crisalli P, Lee B, Jung JW, et al. Structural imprints in vivo decode RNA regulatory mechanisms. Nature. 2015;519: 486–490.

183. Zhao X, Yang Y, Sun BF, Shi Y, Yang X, Xiao W, et al. FTO-dependent demethylation of N6-methyladenosine regulates mRNA splicing and is required for adipogenesis. Cell Res. 2014;24: 1403–1419.

184. Du H, Zhao Y, He J, Zhang Y, Xi H, Liu M, et al. YTHDF2 destabilizes m(6)A-containing RNA through direct recruitment of the CCR4-NOT deadenylase complex. Nat Commun. 2016;7: 12626.

185. Park OH, Ha H, Lee Y, Boo SH, Kwon DH, Song HK, et al. Endoribonucleolytic Cleavage of m6A-Containing RNAs by RNase P/MRP Complex. Mol Cell. 2019;74: 494–507.e8.

186. Collart MA. The Ccr4-Not complex is a key regulator of eukaryotic gene expression. Wiley Interdiscip Rev RNA. 2016;7: 438–454.

187. Ukleja M, Valpuesta JM, Dziembowski A, Cuellar J. Beyond the known functions of the CCR4-NOT complex in gene expression regulatory mechanisms: New structural insights to unravel CCR4-NOT mRNA processing machinery. Bioessays. 2016;38: 1048–1058.

188. Saramago M, da Costa PJ, Viegas SC, Arraiano CM. The Implication of mRNA Degradation Disorders on Human DISease: Focus on DIS3 and DIS3-Like Enzymes. Adv Exp Med Biol. 2019;1157: 85–98.

189. Schmid M, Jensen TH. The Nuclear RNA Exosome and Its Cofactors. Adv Exp Med Biol. 2019;1203: 113–132.

190. Lu W, Tirumuru N, St Gelais C, Koneru PC, Liu C, Kvaratskhelia M, et al. N6-Methyladenosine-binding proteins suppress HIV-1 infectivity and viral production. J Biol Chem. 2018;293: 12992–13005.

191. Shi H, Wang X, Lu Z, Zhao BS, Ma H, Hsu PJ, et al. YTHDF3 facilitates translation and decay of N6-methyladenosine-modified RNA. Cell Res. 2017;27: 315–328.

192. Rosenberg RN, Sassin J, Zimmerman EA, Carter S. The interrelationship of neurofibromatosis and fibrous dysplasia. Arch Neurol. 1967;17: 174–179.

193. Kretschmer J, Rao H, Hackert P, Sloan KE, Höbartner C, Bohnsack MT. The m6A reader protein YTHDC2 interacts with the small ribosomal subunit and the 5'-3' exoribonuclease XRN1. RNA. 2018;24: 1339–1350.

194. Visvanathan A, Patil V, Arora A, Hegde AS, Arivazhagan A, Santosh V, et al. Essential role of METTL3-mediated m6A modification in glioma stem-like cells maintenance and radioresistance. Oncogene. 2018;37: 522–533.

195. Wu R, Li A, Sun B, Sun J-G, Zhang J, Zhang T, et al. A novel m6A reader Prrc2a controls oligodendroglial specification and myelination. Cell Res. 2019;29: 23–41.

196. Edupuganti RR, Geiger S, Lindeboom RGH, Shi H, Hsu PJ, Lu Z, et al. N6-methyladenosine (m6A) recruits and repels proteins to regulate mRNA homeostasis. Nat Struct Mol Biol. 2017;24: 870–878.

197. Zhang F, Kang Y, Wang M, Li Y, Xu T, Yang W, et al. Fragile X mental retardation protein modulates the stability of its m6A-marked messenger RNA targets. Hum Mol Genet. 2018;27: 3936–3950.

198. Wang Y, Li Y, Toth JI, Petroski MD, Zhang Z, Zhao JC. N6-methyladenosine modification destabilizes developmental regulators in embryonic stem cells. Nat Cell Biol. 2014;16: 191–198.

199. Geula S, Moshitch-Moshkovitz S, Dominissini D, Mansour AA, Kol N, Salmon-Divon M, et al. Stem cells. m6A mRNA methylation facilitates resolution of naïve pluripotency toward differentiation. Science. 2015;347: 1002–1006.

200. Zhao BS, Wang X, Beadell AV, Lu Z, Shi H, Kuuspalu A, et al. m(6)A-dependent maternal mRNA clearance facilitates zebrafish maternal-to-zygotic transition. Nature. 2017;542: 475–478.

201. Haussmann IU, Bodi Z, Sanchez-Moran E, Mongan NP, Archer N, Fray RG, et al. m6A potentiates Sxl alternative pre-mRNA splicing for robust Drosophila sex determination. Nature. 2016;540: 301–304.

202. Zheng G, Dahl JA, Niu Y, Fedorcsak P, Huang C-M, Li CJ, et al. ALKBH5 is a mammalian RNA demethylase that impacts RNA metabolism and mouse fertility. Mol Cell. 2013;49: 18–29.

203. Zhang M, Zhang Y, Ma J, Guo F, Cao Q, Zhang Y, et al. The Demethylase Activity of FTO (Fat Mass and Obesity Associated Protein) Is Required for Preadipocyte Differentiation. PLoS One. 2015;10: e0133788.

204. Brzezicha B, Schmidt M, Makałowska I, Jarmołowski A, Pieńkowska J, Szweykowska-Kulińska Z. Identification of human tRNA: m5C methyltransferase catalysing intron-dependent m5C formation in the first position of the anticodon of the. Nucleic Acids Res. 2006;34: 6034–6043.

205. Schaefer M, Pollex T, Hanna K, Lyko F. RNA cytosine methylation analysis by bisulfite sequencing. Nucleic Acids Res. 2009;37: e12.

206. Amort T, Rieder D, Wille A, Khokhlova-Cubberley D, Riml C, Trixl L, et al. Distinct 5-methylcytosine profiles in poly(A) RNA from mouse embryonic stem cells and brain. Genome Biol. 2017;18: 1.

207. Khan MA, Rafiq MA, Noor A, Hussain S, Flores JV, Rupp V, et al. Mutation in NSUN2, which encodes an RNA methyltransferase, causes autosomal-recessive intellectual disability. Am J Hum Genet. 2012;90: 856–863.

208. Martinez FJ, Lee JH, Lee JE, Blanco S, Nickerson E, Gabriel S, et al. Whole exome sequencing identifies a splicing mutation in NSUN2 as a cause of a Dubowitz-like syndrome. J Med Genet. 2012;49: 380–385.

209. Flores JV, Cordero-Espinoza L, Oeztuerk-Winder F, Andersson-Rolf A, Selmi T, Blanco S, et al. Cytosine-5 RNA Methylation Regulates Neural Stem Cell Differentiation and Motility. Stem Cell Reports. 2017;8: 112–124.

210. Van Haute L, Dietmann S, Kremer L, Hussain S, Pearce SF, Powell CA, et al. Deficient methylation and formylation of mt-tRNA(Met) wobble cytosine in a patient carrying mutations in NSUN3. Nat Commun. 2016;7: 12039.

211. Trixl L, Amort T, Wille A, Zinni M, Ebner S, Hechenberger C, et al. RNA cytosine methyltransferase Nsun3 regulates embryonic stem cell differentiation by promoting mitochondrial activity. Cell Mol Life Sci. 2018;75: 1483–1497.

212. Harris T, Marquez B, Suarez S, Schimenti J. Sperm motility defects and infertility in male mice with a mutation in Nsun7, a member of the Sun domain-containing family of putative RNA methyltransferases. Biol Reprod. 2007;77: 376–382.

213. Davis DR. Stabilization of RNA stacking by pseudouridine. Nucleic Acids Res. 1995;23: 5020–5026.

214. Yarian CS, Basti MM, Cain RJ, Ansari G, Guenther RH, Sochacka E, et al. Structural and functional roles of the N1- and N3-protons of at tRNA's position 39. Nucleic Acids Research. 1999. pp. 3543–3549. doi:10.1093/nar/27.17.3543

215. Durant PC, Davis DR. Stabilization of the anticodon stem-loop of tRNALys,3 by

an A+-C base-pair and by pseudouridine. J Mol Biol. 1999;285: 115–131.

216. Lecointe F, Namy O, Hatin I, Simos G, Rousset J-P, Grosjean H. Lack of pseudouridine 38/39 in the anticodon arm of yeast cytoplasmic tRNA decreases in vivo recoding efficiency. J Biol Chem. 2002;277: 30445–30453.

217. Karijolich J, Yu Y-T. Converting nonsense codons into sense codons by targeted pseudouridylation. Nature. 2011;474: 395–398.

218. Nakamoto MA, Lovejoy AF, Cygan AM, Boothroyd JC. mRNA pseudouridylation affects RNA metabolism in the parasite Toxoplasma gondii. RNA. 2017;23: 1834–1849.

219. Karikó K, Muramatsu H, Welsh FA, Ludwig J, Kato H, Akira S, et al. Incorporation of pseudouridine into mRNA yields superior nonimmunogenic vector with increased translational capacity and biological stability. Mol Ther. 2008;16: 1833–1840.

220. Karikó K, Muramatsu H, Keller JM, Weissman D. Increased erythropoiesis in mice injected with submicrogram quantities of pseudouridine-containing mRNA encoding erythropoietin. Mol Ther. 2012;20: 948–953.

221. Warren L, Manos PD, Ahfeldt T, Loh Y-H, Li H, Lau F, et al. Highly efficient reprogramming to pluripotency and directed differentiation of human cells with synthetic modified mRNA. Cell Stem Cell. 2010;7: 618–630.

222. Festen EAM, Goyette P, Green T, Boucher G, Beauchamp C, Trynka G, et al. A meta-analysis of genome-wide association scans identifies IL18RAP, PTPN2, TAGAP, and PUS10 as shared risk loci for Crohn's disease and celiac disease. PLoS Genet. 2011;7: e1001283.

223. Preumont A, Rzem R, Vertommen D, Van Schaftingen E. HDHD1, which is often deleted in X-linked ichthyosis, encodes a pseudouridine-5'-phosphatase. Biochem J. 2010;431: 237–244.

224. Patton JR, Bykhovskaya Y, Mengesha E, Bertolotto C, Fischel-Ghodsian N. Mitochondrial myopathy and sideroblastic anemia (MLASA): missense mutation in the pseudouridine synthase 1 (PUS1) gene is associated with the loss of tRNA pseudouridylation. J Biol Chem. 2005;280: 19823–19828.

225. Byszewska M, Śmietański M, Purta E, Bujnicki JM. RNA methyltransferases involved in 5' cap biosynthesis. RNA Biol. 2014;11: 1597–1607.

226. Kumar S, Mapa K, Maiti S. Understanding the effect of locked nucleic acid and 2'-O-methyl modification on the hybridization thermodynamics of a miRNA-mRNA pair in the presence and absence of AfPiwi protein. Biochemistry. 2014;53: 1607–1615.

227. Yildirim I, Kierzek E, Kierzek R, Schatz GC. Interplay of LNA and 2'-O-methyl RNA in the structure and thermodynamics of RNA hybrid systems: a molecular dynamics study using the revised AMBER force field and comparison with experimental results. J Phys Chem B. 2014;118: 14177–14187.

228. Inoue H, Hayase Y, Imura A, Iwai S, Miura K, Ohtsuka E. Synthesis and hybridization studies on two complementary nona(2'-O-methyl)ribonucleotides. Nucleic Acids Res. 1987;15: 6131–6148.

229. Kawai G, Yamamoto Y, Kamimura T, Masegi T, Sekine M, Hata T, et al. Conformational rigidity of specific pyrimidine residues in tRNA arises from posttranscriptional modifications that enhance steric interaction between the base and the 2'-hydroxyl group. Biochemistry. 1992;31: 1040–1046.

230. Majlessi M, Nelson NC, Becker MM. Advantages of 2'-O-methyl oligoribonucleotide probes for detecting RNA targets. Nucleic Acids Res. 1998;26: 2224–2229.

231. Tsourkas A, Behlke MA, Bao G. Hybridization of 2'-O-methyl and 2'-deoxy molecular beacons to RNA and DNA targets. Nucleic Acids Res. 2002;30: 5168–5174.

232. Erales J, Marchand V, Panthu B, Gillot S, Belin S, Ghayad SE, et al. Evidence for rRNA 2'-O-methylation plasticity: Control of intrinsic translational capabilities of human ribosomes. Proc Natl Acad Sci U S A. 2017;114: 12934–12939.

233. Marcel V, Ghayad SE, Belin S, Therizols G, Morel A-P, Solano-Gonzàlez E, et al. p53 acts as a safeguard of translational control by regulating fibrillarin and rRNA methylation in cancer. Cancer Cell. 2013;24: 318–330.

234. Wang X, Li Z-T, Yan Y, Lin P, Tang W, Hasler D, et al. LARP7-Mediated U6 snRNA Modification Ensures Splicing Fidelity and Spermatogenesis in Mice. Mol Cell. 2020;77: 999–1013.e6.

235. Grosjean H, Droogmans L, Roovers M, Keith G. Detection of enzymatic activity of transfer RNA modification enzymes using radiolabeled tRNA substrates. Methods Enzymol. 2007;425: 55–101.

236. Köhrer C, Rajbhandary UL. The many applications of acid urea polyacrylamide gel electrophoresis to studies of tRNAs and aminoacyl-tRNA synthetases. Methods. 2008;44: 129–138.

237. Jora M, Lobue PA, Ross RL, Williams B, Addepalli B. Detection of ribonucleoside modifications by liquid chromatography coupled with mass spectrometry. Biochim Biophys Acta Gene Regul Mech. 2019;1862: 280–290.

238. Helm M, Motorin Y. Detecting RNA modifications in the epitranscriptome: predict and validate. Nat Rev Genet. 2017;18: 275–291.

239. Kellner S, Burhenne J, Helm M. Detection of RNA modifications. RNA Biol. 2010;7: 237–247.

240. Wetzel C, Limbach PA. Mass spectrometry of modified RNAs: recent developments. Analyst. 2016;141: 16–23.

241. Mongan NP, Emes RD, Archer N. Detection and analysis of RNA methylation. F1000Res. 2019;8. doi:10.12688/f1000research.17956.1

242. Tserovski L, Marchand V, Hauenschild R, Blanloeil-Oillo F, Helm M, Motorin Y. High-throughput sequencing for 1-methyladenosine (m(1)A) mapping in RNA. Methods. 2016;107: 110–121.

243. Zheng G, Qin Y, Clark WC, Dai Q, Yi C, He C, et al. Efficient and quantitative high-throughput tRNA sequencing. Nat Methods. 2015;12: 835–837.

244. Ryvkin P, Leung YY, Silverman IM, Childress M, Valladares O, Dragomir I, et al. HAMR: high-throughput annotation of modified ribonucleotides. RNA. 2013;19: 1684–1692.

245. David R, Burgess A, Parker B, Li J, Pulsford K, Sibbritt T, et al. Transcriptome-Wide Mapping of RNA 5-Methylcytosine in Arabidopsis mRNAs and Noncoding RNAs. Plant Cell. 2017;29: 445–460.

246. Thalalla Gamage S, Sas-Chen A, Schwartz S, Meier JL. Quantitative nucleotide resolution profiling of RNA cytidine acetylation by ac4C-seq. Nat Protoc. 2021;16: 2286–2307.

247. Schwartz S, Motorin Y. Next-generation sequencing technologies for detection of modified nucleotides in RNAs. RNA Biol. 2017;14: 1124–1137.

248. Anreiter I, Mir Q, Simpson JT, Janga SC, Soller M. New Twists in Detecting mRNA Modification Dynamics. Trends Biotechnol. 2020;0. doi:10.1016/j.tibtech.2020.06.002

249. Schaefer M, Kapoor U, Jantsch MF. Understanding RNA modifications: the promises and technological bottlenecks of the “epitranscriptome.” Open Biology. 2017. p. 170077. doi:10.1098/rsob.170077

250. Lahens NF, Kavakli IH, Zhang R, Hayer K, Black MB, Dueck H, et al. IVT-seq reveals extreme bias in RNA sequencing. Genome Biol. 2014;15: R86.

251. Garalde DR, Snell EA, Jachimowicz D, Sipos B, Lloyd JH, Bruce M, et al. Highly parallel direct RNA sequencing on an array of nanopores. Nat Methods. 2018;15: 201–206.

252. Vilfan ID, Tsai Y-C, Clark TA, Wegener J, Dai Q, Yi C, et al. Analysis of RNA base modification and structural rearrangement by single-molecule real-time detection of reverse transcription. J Nanobiotechnology. 2013;11: 8.

253. Smith AM, Jain M, Mulroney L, Garalde DR, Akeson M. Reading canonical and modified nucleobases in 16S ribosomal RNA using nanopore native RNA sequencing. PLoS One. 2019;14: e0216709.

254. Smith AM, Jain M, Mulroney L, Garalde DR, Akeson M. Reading canonical and modified nucleotides in 16S ribosomal RNA using nanopore direct RNA sequencing. bioRxiv. 2017. doi:10.1101/132274

255. Liu H, Begik O, Lucas MC, Ramirez JM, Mason CE, Wiener D, et al. Accurate detection of m6A RNA modifications in native RNA sequences. Nat Commun. 2019;10: 4079.

256. Begik O, Lucas MC, Liu H, Ramirez JM, Mattick JS, Novoa EM. Integrative analyses of the RNA modification machinery reveal tissue- and cancer-specific signatures. Genome Biol. 2020;21: 97.

257. Parker MT, Knop K, Sherwood AV, Schurch NJ, Mackinnon K, Gould PD, et al. Nanopore direct RNA sequencing maps the complexity of Arabidopsis mRNA processing and m6A modification. eLife. 2020. doi:10.7554/elife.49658

258. Wongsurawat T, Jenjaroenpun P, Wassenaar TM, Wadley TD, Wanchai V, Akel NS, et al. Decoding the Epitranscriptional Landscape from Native RNA Sequences.

bioRxiv. 2018. p. 487819. doi:10.1101/487819

259. Viehweger A, Krautwurst S, Lamkiewicz K, Madhugiri R, Ziebuhr J, Hölzer M, et al. Direct RNA nanopore sequencing of full-length coronavirus genomes provides novel insights into structural variants and enables modification analysis. Genome Res. 2019;29: 1545–1554.

260. Pratanwanich PN, Yao F, Chen Y, Koh CWQ, Wan YK, Hendra C, et al. Identification of differential RNA modifications from nanopore direct RNA sequencing with xPore. Nat Biotechnol. 2021. doi:10.1038/s41587-021-00949-w

261. Sippel AE, Stavrianopoulos JG, Schutz G, Feigelson P. Translational properties of rabbit globin mRNA after specific removal of poly(A) with ribonuclease H. Proc Natl Acad Sci U S A. 1974;71: 4635–4639.

262. Sallés FJ, Richards WG, Strickland S. Assaying the polyadenylation state of mRNAs. Methods. 1999;17: 38–45.

263. Salles FJ, Darrow AL, O'Connell ML. Isolation of novel murine maternal mRNAs regulated by cytoplasmic polyadenylation. Genes . 1992. Available: http://genesdev.cshlp.org/content/6/7/1202.short

264. Wang Z, Gerstein M, Snyder M. RNA-Seq: a revolutionary tool for transcriptomics. Nat Rev Genet. 2009;10: 57–63.

265. Chang H, Lim J, Ha M, Kim VN. TAIL-seq: genome-wide determination of poly(A) tail length and 3' end modifications. Mol Cell. 2014;53: 1044–1052.

266. 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. 2019;16: 879–886.

267. 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. 2019;10: 5292.

268. Oikonomopoulos S, Bayega A, Fahiminiya S, Djambazian H, Berube P, Ragoussis J. Methodologies for Transcript Profiling Using Long-Read Technologies. Front Genet. 2020;11: 606.

269. Ramsköld D, Luo S, Wang Y-C, Li R, Deng Q, Faridani OR, et al. Full-length mRNA-Seq from single-cell levels of RNA and individual circulating tumor cells. Nat Biotechnol. 2012;30: 777–782.

270. Au KF, Sebastiano V, Afshar PT, Durruthy JD, Lee L, Williams BA, et al. Characterization of the human ESC transcriptome by hybrid sequencing. Proc Natl Acad Sci U S A. 2013;110: E4821–30.

271. Ozsolak F, Platt AR, Jones DR, Reifenberger JG, Sass LE, McInerney P, et al. Direct RNA sequencing. Nature. 2009;461: 814–818.

272. Workman RE, Tang AD, Tang PS, Jain M, Tyson JR, Razaghi R, et al. Nanopore native RNA sequencing of a human poly(A) transcriptome. Nat Methods. 2019;16: 1297–1305.

273. Boccaletto P, Machnicka MA, Purta E, Piatkowski P, Baginski B, Wirecki TK, et

al. MODOMICS: a database of RNA modification pathways. 2017 update. Nucleic Acids Res. 2018;46: D303–D307.

274. UniProt Consortium. UniProt: a worldwide hub of protein knowledge. Nucleic Acids Res. 2019;47: D506–D515.

275. Xu L, Liu X, Sheng N, Oo KS, Liang J, Chionh YH, et al. Three distinct 3-methylcytidine (m(3)C) methyltransferases modify tRNA and mRNA in mice and humans. J Biol Chem. 2017;292: 14695–14703.

276. Schöller E, Marks J, Marchand V, Bruckmann A, Powell CA, Reichold M, et al. Balancing of mitochondrial translation through METTL8-mediated m3C modification of mitochondrial tRNAs. Mol Cell. 2021;81: 4810–4825.e12.

277. Vilardo E, Nachbagauer C, Buzet A, Taschner A, Holzmann J, Rossmanith W. A subcomplex of human mitochondrial RNase P is a bifunctional methyltransferase--extensive moonlighting in mitochondrial tRNA biogenesis. Nucleic Acids Res. 2012;40: 11583–11593.

278. GTEx Consortium. Human genomics. The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans. Science. 2015;348: 648–660.

279. Thul PJ, Lindskog C. The human protein atlas: A spatial map of the human proteome. Protein Science. 2018. pp. 233–244. doi:10.1002/pro.3307

280. Li B, Qing T, Zhu J, Wen Z, Yu Y, Fukumura R, et al. A Comprehensive Mouse Transcriptomic BodyMap across 17 Tissues by RNA-seq. Sci Rep. 2017;7: 4200.

281. Guimaraes JC, Zavolan M. Patterns of ribosomal protein expression specify normal and malignant human cells. Genome Biol. 2016;17: 236.

282. Kim M-S, Pinto SM, Getnet D, Nirujogi RS, Manda SS, Chaerkady R, et al. A draft map of the human proteome. Nature. 2014;509: 575–581.

283. Chen Y, Zheng Y, Gao Y, Lin Z, Yang S, Wang T, et al. Single-cell RNA-seq uncovers dynamic processes and critical regulators in mouse spermatogenesis. Cell Res. 2018;28: 879–896.

284. Bettegowda A, Wilkinson MF. Transcription and post-transcriptional regulation of spermatogenesis. Philos Trans R Soc Lond B Biol Sci. 2010;365: 1637–1651.

285. Robles V, Herráez P, Labbé C, Cabrita E, Pšenička M, Valcarce DG, et al. Molecular basis of spermatogenesis and sperm quality. Gen Comp Endocrinol. 2017;245: 5–9.

286. Jiang J, White-Cooper H. Transcriptional activation in Drosophila spermatogenesis involves the mutually dependent function of aly and a novel meiotic arrest gene cookie monster. Development. 2003;130: 563–573.

287. Lin Z, Tong M-H. m6A mRNA modification regulates mammalian spermatogenesis. Biochim Biophys Acta Gene Regul Mech. 2018. doi:10.1016/j.bbagrm.2018.10.016

288. Lin Z, Hsu PJ, Xing X, Fang J, Lu Z, Zou Q, et al. Mettl3-/Mettl14-mediated mRNA N6-methyladenosine modulates murine spermatogenesis. Cell Res. 2017;27: 1216–1230.

289. Green CD, Ma Q, Manske GL, Shami AN, Zheng X, Marini S, et al. A Comprehensive Roadmap of Murine Spermatogenesis Defined by Single-Cell RNA-Seq. Dev Cell. 2018;46: 651–667.e10.

290. Connolly CM, Dearth AT, Braun RE. Disruption of murine Tenr results in teratospermia and male infertility. Dev Biol. 2005;278: 13–21.

291. Xia B, Yan Y, Baron M, Wagner F, Barkley D, Chiodin M, et al. Widespread Transcriptional Scanning in the Testis Modulates Gene Evolution Rates. Cell. 2020;180: 248–262.e21.

292. Jung M, Wells D, Rusch J, Ahmad S, Marchini J, Myers SR, et al. Unified single-cell analysis of testis gene regulation and pathology in five mouse strains. Elife. 2019;8. doi:10.7554/eLife.43966

293. Tuorto F, Liebers R, Musch T, Schaefer M, Hofmann S, Kellner S, et al. RNA cytosine methylation by Dnmt2 and NSun2 promotes tRNA stability and protein synthesis. Nat Struct Mol Biol. 2012;19: 900–905.

294. Lin S, Liu Q, Lelyveld VS, Choe J, Szostak JW, Gregory RI. Mettl1/Wdr4-Mediated m7G tRNA Methylome Is Required for Normal mRNA Translation and Embryonic Stem Cell Self-Renewal and Differentiation. Mol Cell. 2018;71: 244–255.e5.

295. Vogel C, Marcotte EM. Insights into the regulation of protein abundance from proteomic and transcriptomic analyses. Nat Rev Genet. 2012;13: 227–232.

296. Khosronezhad N, Hosseinzadeh Colagar A, Mortazavi SM. The Nsun7 (A11337)-deletion mutation, causes reduction of its protein rate and associated with sperm motility defect in infertile men. J Assist Reprod Genet. 2015;32: 807–815.

297. Hussain S, Tuorto F, Menon S, Blanco S, Cox C, Flores JV, et al. The mouse cytosine-5 RNA methyltransferase NSun2 is a component of the chromatoid body and required for testis differentiation. Mol Cell Biol. 2013;33: 1561–1570.

298. Lim SL, Qu ZP, Kortschak RD, Lawrence DM, Geoghegan J, Hempfling A-L, et al. HENMT1 and piRNA Stability Are Required for Adult Male Germ Cell Transposon Repression and to Define the Spermatogenic Program in the Mouse. PLoS Genet. 2015;11: e1005620.

299. Mathieu C, Guérin JF, Cognat M, Lejeune H, Pinatel MC, Lornage J. Motility and fertilizing capacity of epididymal human spermatozoa in normal and pathological cases. Fertil Steril. 1992;57: 871–876.

300. Wolfson B, Gambone J, Rajfer J. Identification of motile sperm in caput epididymis. Intraoperative observations and clinical correlations. Urology. 1992;40: 335–338.

301. Kotaja N, Bhattacharyya SN, Jaskiewicz L, Kimmins S, Parvinen M, Filipowicz W, et al. The chromatoid body of male germ cells: similarity with processing bodies and presence of Dicer and microRNA pathway components. Proc Natl Acad Sci U S A. 2006;103: 2647–2652.

302. Delaunay S, Frye M. RNA modifications regulating cell fate in cancer. Nat Cell Biol. 2019;21: 552–559.

303. Cui Q, Shi H, Ye P, Li L, Qu Q, Sun G, et al. m6A RNA Methylation Regulates the Self-Renewal and Tumorigenesis of Glioblastoma Stem Cells. Cell Rep. 2017;18: 2622–2634.

304. Paris J, Morgan M, Campos J, Spencer GJ, Shmakova A, Ivanova I, et al. Targeting the RNA m6A Reader YTHDF2 Selectively Compromises Cancer Stem Cells in Acute Myeloid Leukemia. Cell Stem Cell. 2019. doi:10.1016/j.stem.2019.03.021

305. Pinello N, Sun S, Wong JJ-L. Aberrant expression of enzymes regulating m6A mRNA methylation: implication in cancer. Cancer Biol Med. 2018;15: 323–334.

306. Okamoto M, Fujiwara M, Hori M, Okada K, Yazama F, Konishi H, et al. tRNA modifying enzymes, NSUN2 and METTL1, determine sensitivity to 5-fluorouracil in HeLa cells. PLoS Genet. 2014;10: e1004639.

307. Goldman M, Craft B, Hastie M, Repecka K, Kamath A, McDade F, et al. The UCSC Xena platform for public and private cancer genomics data visualization and interpretation. bioRxiv, 326470. 2019.

308. Vagin VV, Sigova A, Li C, Seitz H, Gvozdev V, Zamore PD. A distinct small RNA pathway silences selfish genetic elements in the germline. Science. 2006;313: 320–324.

309. Kirino Y, Mourelatos Z. 2′-O-methyl modification in mouse piRNAs and its methylase. Nucleic Acids Symp Ser . 2007;51: 417–418.

310. Thiaville PC, El Yacoubi B, Köhrer C. Essentiality of threonylcarbamoyladenosine (t6A), a universal tRNA modification, in bacteria. Molecular. 2015. Available: https://onlinelibrary.wiley.com/doi/abs/10.1111/mmi.13209

311. Thiaville PC, Legendre R, Rojas-Benítez D, Baudin-Baillieu A, Hatin I, Chalancon G, et al. Global translational impacts of the loss of the tRNA modification t6A in yeast. Microb Cell Fact. 2016;3: 29–45.

312. Janin M, Ortiz-Barahona V, de Moura MC, Martínez-Cardús A, Llinàs-Arias P, Soler M, et al. Epigenetic loss of RNA-methyltransferase NSUN5 in glioma targets ribosomes to drive a stress adaptive translational program. Acta Neuropathol. 2019. doi:10.1007/s00401-019-02062-4

313. Õunap K, Käsper L, Kurg A, Kurg R. The human WBSCR22 protein is involved in the biogenesis of the 40S ribosomal subunits in mammalian cells. PLoS One. 2013;8: e75686.

314. Huang Y, Su R, Sheng Y, Dong L, Dong Z, Xu H, et al. Small-Molecule Targeting of Oncogenic FTO Demethylase in Acute Myeloid Leukemia. Cancer Cell. 2019;35: 677–691.e10.

315. de Crécy-Lagard V, Boccaletto P, Mangleburg CG, Sharma P, Lowe TM, Leidel SA, et al. Matching tRNA modifications in humans to their known and predicted enzymes. Nucleic Acids Res. 2019. doi:10.1093/nar/gkz011

316. Lobo J, Costa AL, Cantante M, Guimarães R, Lopes P, Antunes L, et al. m6A RNA modification and its writer/reader VIRMA/YTHDF3 in testicular germ cell tumors: a role in seminoma phenotype maintenance. J Transl Med. 2019;17: 79.

317. Katoh K, Rozewicki J, Yamada KD. MAFFT online service: multiple sequence alignment, interactive sequence choice and visualization. Brief Bioinform. 2017. doi:10.1093/bib/bbx108

318. Nguyen L-T, Schmidt HA, von Haeseler A, Minh BQ. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. Mol Biol Evol. 2015;32: 268–274.

319. Rambaut A. FigTree v1. 4. 2012.

320. Uhlén M, Fagerberg L, Hallström BM, Lindskog C, Oksvold P, Mardinoglu A, et al. Proteomics. Tissue-based map of the human proteome. Science. 2015;347: 1260419.

321. Pervouchine DD, Djebali S, Breschi A, Davis CA, Barja PP, Dobin A, et al. Enhanced transcriptome maps from multiple mouse tissues reveal evolutionary constraint in gene expression. Nat Commun. 2015;6: 5903.

322. Yue F, Cheng Y, Breschi A, Vierstra J, Wu W, Ryba T, et al. A comparative encyclopedia of DNA elements in the mouse genome. Nature. 2014;515: 355–364.

323. Brawand D, Soumillon M, Necsulea A, Julien P, Csárdi G, Harrigan P, et al. The evolution of gene expression levels in mammalian organs. Nature. 2011;478: 343–348.

324. Griffin MC, Robinson RA, Trask DK. Validation of tissue microarrays using p53 immunohistochemical studies of squamous cell carcinoma of the larynx. Mod Pathol. 2003;16: 1181–1188.

325. Guschanski K, Warnefors M, Kaessmann H. The evolution of duplicate gene expression in mammalian organs. Genome Res. 2017;27: 1461–1474.

326. Pandolfini L, Barbieri I, Bannister AJ, Hendrick A, Andrews B, Webster N, et al. METTL1 Promotes let-7 MicroRNA Processing via m7G Methylation. Mol Cell. 2019;74: 1278–1290.e9.

327. Kato T, Daigo Y, Hayama S, Ishikawa N, Yamabuki T, Ito T, et al. A novel human tRNA-dihydrouridine synthase involved in pulmonary carcinogenesis. Cancer Res. 2005;65: 5638–5646.

328. Lan Q, Liu PY, Haase J, Bell JL, Hüttelmaier S, Liu T. The Critical Role of RNA m6A Methylation in Cancer. Cancer Res. 2019;79: 1285–1292.

329. Fawcett KA, Barroso I. The genetics of obesity: FTO leads the way. Trends Genet. 2010;26: 266–274.

330. Claussnitzer M, Dankel SN, Kim K-H, Quon G, Meuleman W, Haugen C, et al. FTO Obesity Variant Circuitry and Adipocyte Browning in Humans. N Engl J Med. 2015;373: 895–907.

331. Mauer J, Sindelar M, Despic V, Guez T, Hawley BR, Vasseur J-J, et al. FTO controls reversible m6Am RNA methylation during snRNA biogenesis. Nat Chem Biol. 2019;15: 340–347.

332. Li Z, Weng H, Su R, Weng X, Zuo Z, Li C, et al. FTO Plays an Oncogenic Role in Acute Myeloid Leukemia as a N 6 -Methyladenosine RNA Demethylase. Cancer

Cell. 2017. pp. 127–141. doi:10.1016/j.ccell.2016.11.017

333. Braun DA, Rao J, Mollet G, Schapiro D, Daugeron M-C, Tan W, et al. Mutations in KEOPS-complex genes cause nephrotic syndrome with primary microcephaly. Nat Genet. 2017;49: 1529–1538.

334. Stoiber MH, Quick J, Egan R, Lee JE, Celniker SE, Neely R, et al. De novo Identification of DNA Modifications Enabled by Genome-Guided Nanopore Signal Processing. bioRxiv. 2017.

335. Loose M, Malla S, Stout M. Real-time selective sequencing using nanopore technology. Nat Methods. 2016;13: 751–754.

336. Jain M, Olsen HE, Paten B, Akeson M. The Oxford Nanopore MinION: delivery of nanopore sequencing to the genomics community. Genome Biol. 2016;17: 239.

337. McIntyre ABR, Rizzardi L, Yu AM, Alexander N, Rosen GL, Botkin DJ, et al. Nanopore sequencing in microgravity. NPJ Microgravity. 2016;2: 16035.

338. Teng H, Cao MD, Hall MB, Duarte T, Wang S, Coin LJM. Chiron: translating nanopore raw signal directly into nucleotide sequence using deep learning. Gigascience. 2018;7. doi:10.1093/gigascience/giy037

339. McIntyre ABR, Alexander N, Grigorev K, Bezdan D, Sichtig H, Chiu CY, et al. Single-molecule sequencing detection of N6-methyladenine in microbial reference materials. Nat Commun. In Press 2019.

340. Schwartz S, Agarwala SD, Mumbach MR, Jovanovic M, Mertins P, Shishkin A, et al. High-resolution mapping reveals a conserved, widespread, dynamic mRNA methylation program in yeast meiosis. Cell. 2013;155: 1409–1421.

341. Garcia-Campos MA, Edelheit S, Toth U, Shachar R, Nir R, Lasman L, et al. Deciphering the “m6A code” via quantitative profiling of m6A at single-nucleotide resolution. bioRxiv. 2019. p. 571679. doi:10.1101/571679

342. Agarwala SD, Blitzblau HG, Hochwagen A, Fink GR. RNA methylation by the MIS complex regulates a cell fate decision in yeast. PLoS Genet. 2012;8: e1002732.

343. Schwartz S, Mumbach MR, Jovanovic M, Wang T, Maciag K, Bushkin GG, et al. Perturbation of m6A writers reveals two distinct classes of mRNA methylation at internal and 5’ sites. Cell Rep. 2014;8: 284–296.

344. Kan L, Grozhik AV, Vedanayagam J, Patil DP, Pang N, Lim K-S, et al. The m6A pathway facilitates sex determination in Drosophila. Nat Commun. 2017;8: 15737.

345. Weng Y-L, Wang X, An R, Cassin J, Vissers C, Liu Y, et al. Epitranscriptomic m6A Regulation of Axon Regeneration in the Adult Mammalian Nervous System. Neuron. 2018;97: 313–325.e6.

346. Widagdo J, Zhao Q-Y, Kempen M-J, Tan MC, Ratnu VS, Wei W, et al. Experience-Dependent Accumulation of N6-Methyladenosine in the Prefrontal Cortex Is Associated with Memory Processes in Mice. J Neurosci. 2016;36: 6771–6777.

347. Yoon K-J, Ringeling FR, Vissers C, Jacob F, Pokrass M, Jimenez-Cyrus D, et al.

Temporal Control of Mammalian Cortical Neurogenesis by m6A Methylation. Cell. 2017;171: 877–889.e17.

348. Li Z, Weng H, Su R, Weng X, Zuo Z, Li C, et al. FTO Plays an Oncogenic Role in Acute Myeloid Leukemia as a N6-Methyladenosine RNA Demethylase. Cancer Cell. 2017;31: 127–141.

349. Dai D, Wang H, Zhu L, Jin H, Wang X. N6-methyladenosine links RNA metabolism to cancer progression. Cell Death Dis. 2018;9: 124.

350. Liu Z-X, Li L-M, Sun H-L, Liu S-M. Link Between m6A Modification and Cancers. Front Bioeng Biotechnol. 2018;6: 89.

351. Begik O, Lucas MC, Pryszcz LP, Ramirez JM, Medina R, Milenkovic I, et al. Quantitative profiling of pseudouridylation dynamics in native RNAs with nanopore sequencing. Nat Biotechnol. 2021. doi:10.1038/s41587-021-00915-6

352. Leger A, Amaral PP, Pandolfini L, Capitanchik C. RNA modifications detection by comparative Nanopore direct RNA sequencing. BioRxiv. 2019. Available: https://www.biorxiv.org/content/10.1101/843136v1.abstract

353. Price AM, Hayer KE, McIntyre ABR, Gokhale NS, Della Fera AN, Mason CE, et al. Direct RNA sequencing reveals m6A modifications on adenovirus RNA are necessary for efficient splicing. bioRxiv. 2019. p. 865485. doi:10.1101/865485

354. Lorenz DA, Sathe S, Einstein JM, Yeo GW. Direct RNA sequencing enables m6A detection in endogenous transcript isoforms at base-specific resolution. RNA. 2020;26: 19–28.

355. Pratanwanich PN, Yao F, Chen Y, Koh CWQ, Hendra C, Poon P, et al. Detection of differential RNA modifications from direct RNA sequencing of human cell lines. bioRxiv. 2020. p. 2020.06.18.160010. doi:10.1101/2020.06.18.160010

356. Jack K, Bellodi C, Landry DM, Niederer RO, Meskauskas A, Musalgaonkar S, et al. rRNA pseudouridylation defects affect ribosomal ligand binding and translational fidelity from yeast to human cells. Mol Cell. 2011;44: 660–666.

357. Yoon A, Peng G, Brandenburger Y, Zollo O, Xu W, Rego E, et al. Impaired control of IRES-mediated translation in X-linked dyskeratosis congenita. Science. 2006;312: 902–906.

358. Bellodi C, Krasnykh O, Haynes N, Theodoropoulou M, Peng G, Montanaro L, et al. Loss of Function of the Tumor Suppressor DKC1 Perturbs p27 Translation Control and Contributes to Pituitary Tumorigenesis. Cancer Research. 2010. pp. 6026–6035. doi:10.1158/0008-5472.can-09-4730

359. Wu G, Xiao M, Yang C, Yu Y-T. U2 snRNA is inducibly pseudouridylated at novel sites by Pus7p and snR81 RNP. EMBO J. 2011;30: 79–89.

360. Taoka M, Nobe Y, Hori M, Takeuchi A. A mass spectrometry-based method for comprehensive quantitative determination of post-transcriptional RNA modifications: the complete chemical structure of …. Nucleic acids. 2015. Available: https://academic.oup.com/nar/article-abstract/43/18/e115/2414315

361. Basu A, Das P, Chaudhuri S, Bevilacqua E, Andrews J, Barik S, et al. Requirement of rRNA methylation for 80S ribosome assembly on a cohort of

cellular internal ribosome entry sites. Mol Cell Biol. 2011;31: 4482–4499.

362. Belin S, Beghin A, Solano-Gonzàlez E, Bezin L, Brunet-Manquat S, Textoris J, et al. Dysregulation of Ribosome Biogenesis and Translational Capacity Is Associated with Tumor Progression of Human Breast Cancer Cells. PLoS ONE. 2009. p. e7147. doi:10.1371/journal.pone.0007147

363. Buchhaupt M, Sharma S, Kellner S, Oswald S, Paetzold M, Peifer C, et al. Partial methylation at Am100 in 18S rRNA of baker's yeast reveals ribosome heterogeneity on the level of eukaryotic rRNA modification. PLoS One. 2014;9: e89640.

364. Chen H, Liu Q, Yu D, Natchiar K, Zhou C, Hsu C-H, et al. METTL5, an 18S rRNA-specific m6A methyltransferase, modulates expression of stress response genes. bioRxiv. 2020. p. 2020.04.27.064162. doi:10.1101/2020.04.27.064162

365. Li H. Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics. 2018. pp. 3094–3100. doi:10.1093/bioinformatics/bty191

366. Sović I, Šikić M, Wilm A, Fenlon SN, Chen S, Nagarajan N. Fast and sensitive mapping of nanopore sequencing reads with GraphMap. Nat Commun. 2016;7: 11307.

367. Taoka M, Nobe Y, Yamaki Y, Yamauchi Y, Ishikawa H, Takahashi N, et al. The complete chemical structure of Saccharomyces cerevisiae rRNA: partial pseudouridylation of U2345 in 25S rRNA by snoRNA snR9. Nucleic Acids Res. 2016;44: 8951–8961.

368. Liu Q, Fang L, Yu G, Wang D, Xiao C-L, Wang K. Detection of DNA base modifications by deep recurrent neural network on Oxford Nanopore sequencing data. Nat Commun. 2019;10: 2449.

369. McIntyre ABR, Alexander N, Grigorev K, Bezdan D, Sichtig H, Chiu CY, et al. Single-molecule sequencing detection of N6-methyladenine in microbial reference materials. Nat Commun 10: 579. 2019.

370. De Coster W, Stovner EB, Strazisar M. Methplotlib: analysis of modified nucleotides from nanopore sequencing. Bioinformatics. 2020;36: 3236–3238.

371. Stoiber M, Quick J, Egan R, Lee JE, Celniker S, Neely RK, et al. De novo Identification of DNA Modifications Enabled by Genome-Guided Nanopore Signal Processing. bioRxiv. 2017. p. 094672. doi:10.1101/094672

372. Pintard L, Bujnicki JM, Lapeyre B, Bonnerot C. MRM2 encodes a novel yeast mitochondrial 21S rRNA methyltransferase. EMBO J. 2002;21: 1139–1147.

373. Hebras J, Krogh N, Marty V, Nielsen H, Cavaillé J. Developmental changes of rRNA ribose methylations in the mouse. RNA Biol. 2019; 1–15.

374. Higa-Nakamine S, Suzuki T, Uechi T, Chakraborty A, Nakajima Y, Nakamura M, et al. Loss of ribosomal RNA modification causes developmental defects in zebrafish. Nucleic Acids Res. 2012;40: 391–398.

375. Sahoo T, del Gaudio D, German JR, Shinawi M, Peters SU, Person RE, et al. Prader-Willi phenotype caused by paternal deficiency for the HBII-85 C/D box small nucleolar RNA cluster. Nat Genet. 2008;40: 719–721.

376. Heiss NS, Knight SW, Vulliamy TJ, Klauck SM, Wiemann S, Mason PJ, et al. X-linked dyskeratosis congenita is caused by mutations in a highly conserved gene with putative nucleolar functions. Nat Genet. 1998;19: 32–38.

377. Knight SW, Heiss NS, Vulliamy TJ, Greschner S, Stavrides G, Pai GS, et al. X-Linked Dyskeratosis Congenita Is Predominantly Caused by Missense Mutations in the DKC1 Gene. The American Journal of Human Genetics. 1999. pp. 50–58. doi:10.1086/302446

378. Liao J, Yu L, Mei Y, Guarnera M, Shen J, Li R, et al. Small nucleolar RNA signatures as biomarkers for non-small-cell lung cancer. Mol Cancer. 2010;9: 198.

379. Mei Y-P, Liao J-P, Shen J, Yu L, Liu B-L, Liu L, et al. Small nucleolar RNA 42 acts as an oncogene in lung tumorigenesis. Oncogene. 2012;31: 2794–2804.

380. Bortolin-Cavaille M-L, -L. Bortolin-Cavaille M, Cavaille J. The SNORD115 (H/MBII-52) and SNORD116 (H/MBII-85) gene clusters at the imprinted Prader-Willi locus generate canonical box C/D snoRNAs. Nucleic Acids Research. 2012. pp. 6800–6807. doi:10.1093/nar/gks321

381. Krogh N, Jansson MD, Häfner SJ, Tehler D, Birkedal U, Christensen-Dalsgaard M, et al. Profiling of 2′-O-Me in human rRNA reveals a subset of fractionally modified positions and provides evidence for ribosome heterogeneity. Nucleic Acids Research. 2016. pp. 7884–7895. doi:10.1093/nar/gkw482

382. Birkedal U, Christensen-Dalsgaard M, Krogh N, Sabarinathan R, Gorodkin J, Nielsen H. Profiling of ribose methylations in RNA by high-throughput sequencing. Angew Chem Int Ed Engl. 2015;54: 451–455.

383. Natchiar SK, Myasnikov AG, Kratzat H, Hazemann I, Klaholz BP. Visualization of chemical modifications in the human 80S ribosome structure. Nature. 2017;551: 472–477.

384. Taoka M, Nobe Y, Yamaki Y, Sato K, Ishikawa H, Izumikawa K, et al. Landscape of the complete RNA chemical modifications in the human 80S ribosome. Nucleic Acids Res. 2018;46: 9289–9298.

385. van der Feltz C, DeHaven AC, Hoskins AA. Stress-induced Pseudouridylation Alters the Structural Equilibrium of Yeast U2 snRNA Stem II. J Mol Biol. 2018;430: 524–536.

386. Parker S, Fraczek MG, Wu J, Shamsah S, Manousaki A, Dungrattanalert K, et al. A resource for functional profiling of noncoding RNA in the yeast Saccharomyces cerevisiae. RNA. 2017;23: 1166–1171.

387. Smith MA, Ersavas T, Ferguson JM, Liu H, Lucas MC, Begik O, et al. Barcoding and demultiplexing Oxford Nanopore native RNA sequencing reads with deep residual learning. bioRxiv. 2019. p. 864322. doi:10.1101/864322

388. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 2009. pp. 2078–2079. doi:10.1093/bioinformatics/btp352

389. Robinson JT, Thorvaldsdóttir H, Winckler W, Guttman M, Lander ES, Getz G, et al. Integrative genomics viewer. Nat Biotechnol. 2011;29: 24–26.

390. Loman NJ, Quick J, Simpson JT. A complete bacterial genome assembled de novo using only nanopore sequencing data. Nat Methods. 2015;12: 733–735.

391. Frye M, Harada BT, Behm M, He C. RNA modifications modulate gene expression during development. Science. 2018;361: 1346–1349.

392. Li S, Mason CE. The pivotal regulatory landscape of RNA modifications. Annu Rev Genomics Hum Genet. 2014;15: 127–150.

393. Louloupi A, Ntini E, Conrad T, Orom UAV. Transient N-6-Methyladenosine Transcriptome Sequencing Reveals a Regulatory Role of m6A in Splicing Efficiency. Cell Rep. 2018;23: 3429–3437.

394. Zhou KI, Shi H, Lyu R, Wylder AC, Matuszek Ż, Pan JN, et al. Regulation of Co-transcriptional Pre-mRNA Splicing by m6A through the Low-Complexity Protein hnRNPG. Mol Cell. 2019;76: 70–81.e9.

395. Lee Y, Choe J, Park OH, Kim YK. Molecular Mechanisms Driving mRNA Degradation by m6A Modification. Trends Genet. 2020;36: 177–188.

396. Guo M, Liu X, Zheng X, Huang Y, Chen X. m6A RNA Modification Determines Cell Fate by Regulating mRNA Degradation. Cellular Reprogramming. 2017. pp. 225–231. doi:10.1089/cell.2016.0041

397. Weng H, Huang H, Wu H, Qin X, Zhao BS, Dong L, et al. METTL14 Inhibits Hematopoietic Stem/Progenitor Differentiation and Promotes Leukemogenesis via mRNA m6A Modification. Cell Stem Cell. 2018;22: 191–205.e9.

398. Depledge DP, Srinivas KP, Sadaoka T, Bready D, Mori Y, Placantonakis DG, et al. Direct RNA sequencing on nanopore arrays redefines the transcriptional complexity of a viral pathogen. Nat Commun. 2019;10: 754.

399. Roach NP, Sadowski N, Alessi AF, Timp W, Taylor J, Kim JK. The full-length transcriptome of C. elegans using direct RNA sequencing. Genome Res. 2020;30: 299–312.

400. Workman RE, Tang A, Tang PS, Jain M, Tyson JR, Zuzarte PC, et al. Nanopore native RNA sequencing of a human poly(A) transcriptome. doi:10.1101/459529

401. Kim D, Lee J-Y, Yang J-S, Kim JW, Kim VN, Chang H. The Architecture of SARS-CoV-2 Transcriptome. Cell. 2020;181: 914–921.e10.

402. Begik O, Liu H, Delgado-Tejedor A, Kontur C, Giraldez AJ, Beaudoin J-D, et al. Nano3P-seq: transcriptome-wide analysis of gene expression and tail dynamics using end-capture nanopore sequencing. bioRxiv. 2021. p. 2021.09.22.461331. doi:10.1101/2021.09.22.461331

403. Ulitsky I, Shkumatava A, Jan CH, Subtelny AO, Koppstein D, Bell GW, et al. Extensive alternative polyadenylation during zebrafish development. Genome Res. 2012;22: 2054–2066.

404. Yartseva V, Giraldez AJ. The Maternal-to-Zygotic Transition During Vertebrate Development: A Model for Reprogramming. Curr Top Dev Biol. 2015;113: 191–232.

405. Lee MT, Bonneau AR, Giraldez AJ. Zygotic genome activation during the maternal-to-zygotic transition. Annu Rev Cell Dev Biol. 2014;30: 581–613.

406. Walser CB, Lipshitz HD. Transcript clearance during the maternal-to-zygotic transition. Curr Opin Genet Dev. 2011;21: 431–443.

407. Levy BW, Johnson CB, McCarthy BJ. Diversity of sequences in total and polyadenylated nuclear RNA from Drosophila cells. Nucleic Acids Research. 1976. pp. 1777–1790. doi:10.1093/nar/3.7.1777

408. Hardwick SA, Chen WY, Wong T, Deveson IW, Blackburn J, Andersen SB, et al. Spliced synthetic genes as internal controls in RNA sequencing experiments. Nat Methods. 2016;13: 792–798.

409. Chang H, Yeo J, Kim J-G, Kim H, Lim J, Lee M, et al. Terminal Uridylyltransferases Execute Programmed Clearance of Maternal Transcriptome in Vertebrate Embryos. Mol Cell. 2018;70: 72–82.e7.

410. Bratic A, Clemente P, Calvo-Garrido J, Maffezzini C, Felser A, Wibom R, et al. Mitochondrial Polyadenylation Is a One-Step Process Required for mRNA Integrity and tRNA Maturation. PLoS Genet. 2016;12: e1006028.

411. Krause M, Niazi AM, Labun K, Torres Cleuren YN, Müller FS, Valen E. tailfindr: alignment-free poly(A) length measurement for Oxford Nanopore RNA and DNA sequencing. RNA. 2019;25: 1229–1241.

412. Carninci P, Kasukawa T, Katayama S, Gough J, Frith MC, Maeda N, et al. The transcriptional landscape of the mammalian genome. Science. 2005;309: 1559–1563.

413. Guttman M, Amit I, Garber M, French C, Lin MF, Feldser D, et al. Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals. Nature. 2009;458: 223–227.

414. Marzluff WF, Wagner EJ, Duronio RJ. Metabolism and regulation of canonical histone mRNAs: life without a poly(A) tail. Nature Reviews Genetics. 2008. pp. 843–854. doi:10.1038/nrg2438

415. Battle DJ, Doudna JA. The stem-loop binding protein forms a highly stable and specific complex with the 3' stem-loop of histone mRNAs. RNA. 2001;7: 123–132.

416. Vejnar CE, Messih MA, Takacs CM, Yartseva V. Genome wide analysis of 3′ UTR sequence elements and proteins regulating mRNA stability during maternal-to-zygotic transition in zebrafish. Genome. 2019. Available: https://genome.cshlp.org/content/29/7/1100.short

417. Vejnar CE, Abdel Messih M, Takacs CM, Yartseva V, Oikonomou P, Christiano R, et al. Genome wide analysis of 3' UTR sequence elements and proteins regulating mRNA stability during maternal-to-zygotic transition in zebrafish. Genome Res. 2019;29: 1100–1114.

418. Boo SH, Kim YK. The emerging role of RNA modifications in the regulation of mRNA stability. Exp Mol Med. 2020;52: 400–408.

419. Werner S, Schmidt L, Marchand V, Kemmer T, Falschlunger C, Sednev MV, et al. Machine learning of reverse transcription signatures of variegated polymerases allows mapping and discrimination of methylated purines in limited transcriptomes. Nucleic Acids Res. 2020;48: 3734–3746.

420. Boccaletto P, Bagiński B. MODOMICS: An Operational Guide to the Use of the RNA Modification Pathways Database. Methods Mol Biol. 2021;2284: 481–505.

421. Martelotto L. “Frankenstein” protocol for nuclei isolation from fresh and frozen tissue for snRNAseq v2 (protocols.io.3fkgjkw). protocols.io. ZappyLab, Inc.; 2019. doi:10.17504/protocols.io.3fkgjkw

422. Locati MD, Pagano JFB, Girard G, Ensink WA, van Olst M, van Leeuwen S, et al. Expression of distinct maternal and somatic 5.8S, 18S, and 28S rRNA types during zebrafish development. RNA. 2017;23: 1188–1199.

423. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics. 2010;26: 841–842.

424. Cozzuto L, Liu H, Pryszcz LP, Pulido TH, Delgado-Tejedor A, Ponomarenko J, et al. MasterOfPores: A Workflow for the Analysis of Oxford Nanopore Direct RNA Sequencing Datasets. Front Genet. 2020;11: 211.

425. Shepard PJ, Choi E-A, Lu J, Flanagan LA, Hertel KJ, Shi Y. Complex and dynamic landscape of RNA polyadenylation revealed by PAS-Seq. RNA. 2011. pp. 761–772. doi:10.1261/rna.2581711

426. Harrison PF, Powell DR, Clancy JL, Preiss T, Boag PR, Traven A, et al. PAT-seq: a method to study the integration of 3’-UTR dynamics with gene expression in the eukaryotic transcriptome. RNA. 2015;21: 1502–1510.

427. Lim J, Lee M, Son A, Chang H, Kim VN. mTAIL-seq reveals dynamic poly(A) tail regulation in oocyte-to-embryo development. Genes Dev. 2016;30: 1671–1682.

428. Xu H, Yao J, Wu DC, Lambowitz AM. Improved TGIRT-seq methods for comprehensive transcriptome profiling with decreased adapter dimer formation and bias correction. Sci Rep. 2019;9: 7953.

429. Mohr S, Ghanem E, Smith W, Sheeter D, Qin Y, King O, et al. Thermostable group II intron reverse transcriptase fusion proteins and their use in cDNA synthesis and next-generation RNA sequencing. RNA. 2013;19: 958–970.

8. Appendix