THE UNIVERSITY OF CHICAGO

INVESTIGATION OF MESSENGER RNA MODIFICATIONS IN NANOPORE SEQUENCING DATA BY MACHINE LEARNING METHODS

A DISSERTATION SUBMITTED TO

THE FACULTY OF THE DIVISION OF THE BIOLOGICAL SCIENCES

AND THE PRITZKER SCHOOL OF MEDICINE

IN CANDIDACY FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

GRADUATE PROGRAM IN BIOCHEMISTRY AND MOLECULAR BIOPHYSICS

BY

SIHAO HUANG

CHICAGO, ILLINOIS

JUNE 2024

Table of Contents

List of figures

List of tables

Abbreviations

5moU	5-methoxyuridine
A	adenosine
a^6A	N^6-allyladenosine
ac^4C	N^4-acetylcytidine
AI	artificial intelligence
ALKBH	AlkB homolog
AML	acute myeloid leukemia
AUC	area under curve
C	cytosine
cDNA	complementary DNA
CDS	coding sequence
CLIP	crosslinking immunoprecipitation
CMC	*N*-cyclohexyl-*N*′-(2-morpholinoethyl) carbodiimide methyl-*p*-toluenesulfonate
CNN	convolutional neural network
COVID	coronavirus disease
CPU	central processing unit
CsgG	Curlin sigma S-dependent growth
DEPC	Diethyl pyrocarbonate
DNA	deoxyribonucleic acid
DRS	direct RNA sequencing
DTT	dithiothreitol
ESC	embryonic stem cell
EXT	extremely randomized trees
FNN	feedforward neural network
FTO	fat mass and obesity-associated
G	guanosine
GMM	Gaussian mixture models
GO	gene ontology
HEK	human embryonic kidney
hm^5C	5-hydroxymethylcytosine
HMM	hidden Markov model
hnRNP	heterogeneous nuclear ribonucleoprotein
I	inosine
IFN	interferon
IGFBP	insulin-like growth factor-2 mRNA-binding proteins
ISG	interferon stimulated genes
IVT	in vitro transcription
KNN	K-nearest neighbors,
K-S	Kolmogorov-Smirnov
LC/MS	liquid chromatography/mass spectrometry
lncRNA	long non-coding RNA

LSTM	long short term memory
m^1A	N1-methyladenosine
m^5C	5-methylcytosine
m^6A	N^6-methyladenosine
m^6Am	N^6,2′-*O*-dimethyladenosine
m^7G	N^7-methylguanosine
METTL	methyltransferase-like
MHC	major histocompatibility complex
ML	machine learning
mRNA	messenger RNA
MspA	Mycobacterium smegmatis porin A
NAI-N3	2-methylnicotinic acid imidazolide azide
NanoPsu	nanopore investigation of pseudouridine
NanoSPA	nanopore simultaneous investigation for pseudouridine and m6A
NAT	N-acetyltransferase
NGS	Next generation sequencing
Nm	2′-*O*-methylation
NN	neural network
NSUN	NOP2/Sun domain protein
ONT	Oxford Nanopore Technologies
PCR	polymerase chain reaction
PKR	protein kinase R
PUS	pseudouridine synthase
RF	random forest
RNA	ribonucleic acids
RNN	recurrent neural network
ROC	receiver operating characteristic
rRNA	ribosomal RNA
RT	reverse transcription
RT-qPCR	reverse transcription-quantitative polymerase chain reaction
SAM	S-adenosyl methionine
SHAPE	2′-hydroxyl acylation analyzed by primer extension
snoRNP	small nucleolar ribonucleoproteins
SNP	single nucleotide polymorphisms
SV	structural variations
SVM	support vector machine
T2T	Telomere-to-Telomere
TE	translation efficiency
TGS	third generation sequencing
tRNA	transfer RNA
U	uridine
UTR	untranslated region
UV	ultraviolet
WT	wild type

XGBoost	extreme gradient boosting
YTH	YT521 homology
YTHDC	YTH domain containing
YTHDF	YTH domain family
Ψ	pseudouridine

Acknowledgement

It's really a wonderful journey to pursue my PhD degree at the University of Chicago. There are so many people that help me and affect me during the six years of time. First of all, I would like to thank my thesis adviser Prof. Tao Pan, who kindly accepted my application when I was bad at writing and speaking in English and had little experience living and working in a new country. Tao knows so well what a graduate student needs at different stage of research training. We sat together, shared about ideas, discussed about results, argued about thoughts, reached consensus and made beautiful publications. Tao drove me to the hospital twice when I was badly sick during the COVID outbreak. Tao helps me a lot both on work and life.

I would like to thank my thesis committee members: Prof. Jingyi Fei, Prof. Chuan He and Prof. A. Murat Eren. They are not only great advisers but also fantastic collaborators. They provided essential comments and ideas on my thesis projects and helped solve a lot of problems in my research. I would like to thank Prof. Chuan He for telling me how to make a good presentation during my preliminary examination.

I would like to thank all the members in Pan Lab. I would like to thank Dr. Wen Zhang, Dr. Chris Katanski, Dr. Qing Dai, Dr. Chris Watkins, Dr. Adam Wylder, Yichen Hou and Noah Pena for direct collaborations on my publications and all lab mates for helping with the experiments and data. I enjoy the years working with all of them. I would like to thank Dr. Qing Dai for guiding me when I did my rotation in Pan Lab and recommended me to stay.

I would like to thank the University of Chicago for providing detailed help. I would like to thank the staffs in UChicagoGRAD and MyChoice for helping with internship and job hunting, especially for giving me the chance to participate in the comp bio trek to bay area to learn about the companies and working opportunities there. I would like to thank UChicago

hospital and the student wellness center for helping me when I was sick, especially during the hard period of COVID pandemic.

I would like to thank my family for supporting me. The high-speed Internet enables video chat with my parents every day and easy communication. Although I live alone abroad, it's like that I still live with my family at home. I would like to dedicate this dissertation to the memory of my grandmother, who passed away on September 15th, 2023, due to stage IV glioblastoma. It's a pity that we still could not conquer and cure all types of cancers, and I hope I could do something to make it possible in the future. I would like to thank my girlfriend, Dr. Zhouzerui Liu, for companion during my whole PhD life. Although we don't live together, but I still feel that there is someone that I could talk to, share me thoughts with every day and I never feel lonely.

I would like to thank my friends. I was not very active in exploring the city of Chicago or the whole country. Thanks to my friends, I went to the forests, lakes, markets and events around the Chicago area. I tasted the food from all over the world in Chicago. Together with my friends, I left my footprints in California, Texas, Florida, Massachusetts etc. The experience outside scientific research makes my life not boring. Some of my friends also became my collaborators. It was so unexpected experience that the collaborations were just built during a trip or a ride together. Although our major research directions were different, we had happy times working on the same biological problems.

I would like to thank the colleagues I worked with during my internships at Guardant Health (GH) and Daiichi Sankyo. I learnt a lot of new technical skills during the internships. The ML related skilled learnt at GH directly benefit my publication "tRNA abundance, modification

and fragmentation in nasopharyngeal swabs as biomarkers for COVID-19 severity" right in the first month that I came back to lab from the internship.

I'm definitely not going to thank the next one, COVID-19, but it did profoundly affect and change the life of many people, including me. I'm among one of the PhD students whose PhD life was affected the most by the pandemic. In 2020, when I was in my second year, I was kept at home first. Then unfortunately, I was kept in the hospital for a week. After that, I started to live a healthier life. Due to the pandemic, the whole world changed and could hardly go back to the same as before. The conferences and seminars were virtual, which provided me more chances to attend events of different topics and allowed me to attend without worrying about the travel fees. The jobs and internships could be done remotely, which made my internships possible. The working efficiency raised a lot by meeting, communicate, sharing data and results and writing manuscripts online. Personally, I turned my direction to full dry work after the COVID outbreak, which I think was a great selection. I stopped trying to identify RNA modifications by wet lab strategies and started to apply machine learning methods instead, which resulted in the projects discussed in Chapter 2 to 5. I mourn for all victims in the pandemic. I hope we could win the war against infectious diseases with less cost in the future. I hope I could contribute to it.

This thesis concludes the major work I have done during my graduate program at the University of Chicago.

Abstract

To date, over 170 types of modifications have been identified in RNA, in which around 10 types are discovered in mRNA. RNA modifications play important roles in transcription, mRNA stability, decay, splicing, translation, regulate the expression of genes and affect metabolisms. Thus, it's important to understand the abundance and distribution of RNA modifications in transcriptome, to better understand how these modifications affect the metabolisms and how these modifications are regulated to execute proper functions. Next generation sequencing methods provide a group of strategies to map the transcriptome wide distributions of RNA modifications and has resulted in meaningful biological discoveries. However, only DNA molecules could be directly run by NGS methods and thus all RNA modifications are detected by indirect approaches, depending on mutations, indels, reverse transcription stops, or immunoprecipitation enrichment brought about by the modified sites. In the past decade, the development of Nanopore sequencing enables the direct sequencing of RNA molecules, as well as RNA modifications. In this dissertation, I developed machine learning based pipelines NanoPsu and NanoSPA for mRNA modification identification from nanopore direct RNA sequencing data. NanoPsu identifies pseudouridine modifications from human transcriptome and the correlation of interferon induced gene expression and pseudouridylation is revealed. NanoSPA enables simultaneous mapping of mRNA m^6A and pseudouridine in human transcriptome and reveals the anti-coordination of the two modifications. Both m^6A and pseudouridine are discovered to have positive effect on translation and the effect of pseudouridine is stronger than m^6A. Besides, I and others in the Pan Lab also attempted to develop a pipeline to predict pseudouridine based on single reads and revealed the stoichiometry of pseudouridine and the linkages between multiple modification sites. The study develops

pipelines to facilitate the modification identification from nanopore direct RNA sequencing data and reveals the potential roles of the modifications in viral infection response and translation. The methods could be applied to other species and samples for more biological discoveries. The pipelines are designed for convenient usage of public users and could be easily expanded to more RNA modifications in the future.

Chapter 1. Introduction

RNA modifications play important roles in transcription, mRNA stability, decay, splicing, translation, regulate the expression of genes and affect metabolisms. To learn the biological functions of RNA modification, the essential step is to know the transcriptome wide distribution of the modifications. Thus, high resolution, high coverage, high accuracy, low sample amount demand, low chemical toxicity, fast speed and easy execution transcriptome wide RNA sequencing methods become the goal of researchers. Here, I go through the previous effort on the RNA modification studies and talk about our progress in this field.

1.1 RNA modifications

Ribonucleic acids (RNA) are one of the key macromolecules in central dogma, which serve as a bridge connecting the stable inherited information and the diverse expression and differentiation. The information of the macromolecule is stored in four units, adenosine (A), cytosine (C), guanosine (G) and uridine (U), and the billions of ways to order the four nucleosides in a genome result in thousands of different proteins. Beyond that, RNA modifications add more diversity and regulation possibilities to the dynamics of expression. The first RNA modification was discovered in 1951 and was called “unknown constituents” at that time (Cohn & Volkin, 1951). Since then, over 170 types of RNA modifications have been discovered in the past several decades (Boccaletto et al., 2022). The first demethylase of RNA modifications was discovered in 2011, which was a milestone as it revealed the reversibility and thus dynamics of RNA modifications (Jia et al., 2011). Since then, the study of RNA

modifications, or “epitranscriptomics”, became a hot field and more resources were put into this field.

RNA modification was widely distributed in tRNA, rRNA, mRNA and other non-coding RNA. RNA molecules have much more types of modifications than DNA, which owns around 17 types of modifications (L. Y. Zhao, Song, Liu, Song, & Yi, 2020), reflecting the high dynamics of gene expression and regulation. tRNA has the greatest number of modifications and around 15% to 25% of all tRNA nucleotides are modified in eukaryotic species (El Yacoubi, Bailly, & de Crecy-Lagard, 2012). Each tRNA molecule has ~13 modified nucleotides on average (Pan, 2018). The modifications in human 80S rRNA has been thoroughly identified and quantified (Masato Taoka et al., 2018). In mRNA, there are around 10 types of modifications discovered, including pseudouridine (Ψ), N^6-methyladenosine (m^6A), N^1-methyladenosine (m^1A), N^6,2′-*O*-dimethyladenosine (m^6Am), 5-methylcytosine (m^5C), 5-hydroxymethylcytosine (hm^5C), N^4-acetylcytidine (ac^4C), N^7-methylguanosine (m^7G), inosine (I), 2′-*O*-methylation (Nm) (Roundtree, Evans, Pan, & He, 2017) (**Fig. 1.1a**). The majority of these modification is on the base of the nucleotide, except for Nm which is on the ribose. Chemically, the modifications could be classified as base methylation, base acetylation, base isomerization, backbone methylation and base editing. RNA modification level and distribution differ among species and cell types.

The function of mRNA modifications in splicing, translation, mRNA stability and localization has been gradually revealed in the last decade (Delaunay & Frye, 2019; Sun, Li, Liu, & Yi, 2023; B. S. Zhao, Roundtree, & He, 2017), but there are still major unknowns about them.

There are enzymes to synthesize, bind or remove RNA modifications and they are usually called “writers”, “readers” and “erasers”. Writers and erasers enable the change of RNA

modification level and distribution globally or locally and thus result in regulation effect. Readers bind to the modifications and convey the downstream effect brought about by the modifications.

The thesis mainly focuses on pseudouridine and m^6A in mRNA studies. Below I describe more about the details of these two modifications.

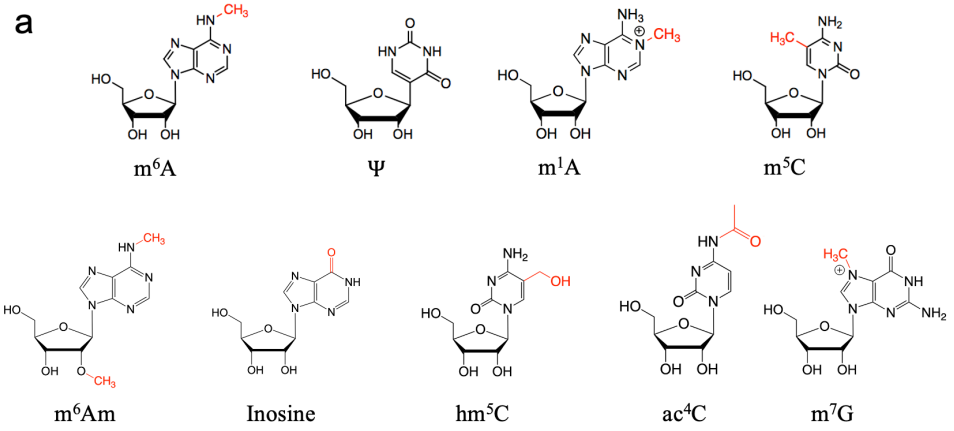

Figure 1.1 RNA modifications
(a) Structure of mRNA modifications

1.1.1 Pseudouridine

Pseudouridine (Ψ) is the most abundant RNA modification (Charette & Gray, 2000; Ge & Yu, 2013; X. Li, Ma, & Yi, 2016) and the second most abundant in mammalian mRNA, following m^6A. It is also widely distributed in non-coding RNA (Schwartz et al., 2014). It was the first RNA modification discovered back in 1951 when researchers had not yet figured out how ribonucleotides were connected with each other to form RNA molecules (Cohn & Volkin, 1951). It was later discovered again in yeast in 1957 and called the "fifth nucleotide" (F. F. Davis & Allen, 1957) and identified as 5-ribosyl uracil (C.-T. Yu & Allen, 1959). The modification was named as pseudouridine with the symbol Ψ by Dr. A. Michelson in the same year (Cohn, 1959).

Pseudouridine is an isomer of uridine, and the difference is that uridine base is connected to the ribose by N1 forming a C-N bond and pseudouridine base is connected to the ribose by C5 which forms a C-C bond. Thus, the connection of pseudouridine base to the backbone is stronger than uridine base. The N1 provides an additional hydrogen and thus enables formation of non-Watson-Crick pairs, making the chemical properties of pseudouridine different from its isomer (Ge & Yu, 2013). Uridine and pseudouridine share the same molecular weight but different mass spectrometric dissociation (Durairaj & Limbach, 2008).

1.1.1.1 Distributions and functions of pseudouridine

Pseudouridine is overall the most abundant RNA modification. It is discovered in human mRNA and most types of non-coding RNA like rRNA, tRNA and snRNA (Ge & Yu, 2013). The distribution of pseudouridine in human rRNA has been thoroughly studied and the stoichiometry of each pseudouridine site has been measured by mass spectrometry (Masato Taoka et al., 2018). There are many pseudouridine sites discovered in human cytoplasmic and mitochondrial tRNA. Pseudouridine levels vary among tissues and cell cycle stages (Brandmayr et al., 2012; Patil et al., 2012) and may have different functions in different conditions.

Conserved pseudouridine sites were reported to stabilize RNA structures when it's in tRNA (Arnez & Steitz, 1994). The pseudouridine at anticodon positions are reported to affect base pairing and thus affect translation efficiency and fidelity (D. R. Davis, Veltri, & Nielsen, 1998; Harrington, Nazarenko, Dix, Thompson, & Uhlenbeck, 1993). It was reported to promote RNA stacking chemically (D. R. Davis, 1995). In U2 snRNA, pseudouridine is reported to be essential for spliceosome branch site recognition and base pairing (Newby & Greenbaum, 2001, 2002). In ribosomal RNA, pseudouridine is reported to affect folding, due to existence of the

extra hydrogen bond (Decatur & Fournier, 2002). It could also affect ribosome subunit association (Sakakibara & Chow, 2012). When pseudouridine is present at specific positions in mRNA, it shows the ability to suppress stop codons *in vivo* and *in vitro* (Karijolich & Yu, 2011). To summarize, pseudouridine plays an important role in the non-coding RNA structurally and functionally.

It is reported that serum starvation, heat shock and H_2O_2 stress could alter the mRNA pseudouridine modifications in human cell lines (Carlile et al., 2014; X. Li et al., 2015). Pseudouridine is reported to affect translation *in vitro*. It is reported to promote translation in rabbit reticulocyte system, while repress translation in the wheat germ system and *E.coli* system (Kariko et al., 2008). The pseudouridine derivative, N^1-methyl-pseudouridine, was incorporated into mRNA vaccines to increase efficacy (Morais, Adachi, & Yu, 2021). The modifications benefit the RNA vaccine by stabilizing against enzyme degradation, enhance RNA lifespan and reduce toxic effect and immunogenicity (Ho, Schiess, Miranda, Weber, & Astakhova, 2024). However, it was also reported that N^1-methyl-pseudouridine could cause ribosomal shifting in mRNA and produce unexpected proteins, which introduced doubt on the safety of mRNA vaccines (Mulroney et al., 2024). It was reported that pseudouridine incorporated by PUS1 into oncogenic mRNA could promote the translation of oncogenes and thus promote hepatocellular carcinoma (Y.-X. Hu et al.). However, to date, functions of individual mRNA pseudouridine sites are still challenging to determine, due to the limitation of high resolution and high accuracy pseudouridine mapping methods, or potential nature of group functioning of pseudouridine.

1.1.1.2 Enzymes

There are thirteen reported pseudouridine synthase (PUS) in human cells (E. K. Borchardt, N. M. Martinez, & W. V. Gilbert, 2020), which makes the studies on pseudouridine complicated and difficult, due to the potential redundancy and overlap of enzyme functions and that each specific pseudouridine modification could be installed by multiple PUS enzymes (Dai et al., 2023). Knockdown or knockout of multiple PUS enzymes are likely to be lethal thus it is not possible to fully block pseudouridine generation in cells. The thirteen synthases could be divided into 6 classes, including TruA, TruB, TruD, RluA, RsuA and PUS10 families, mainly depending on the domains they have (E. K. Borchardt et al., 2020). Among the thirteen writers, only DKC1 could be guided by H/ACA small nucleolar ribonucleoproteins (snoRNP) for pseudouridylation and all the rest work independently for targets and are called stand-alone pseudouridine synthase (Garus & Autexier, 2021; Hamma & Ferré-D'Amaré, 2006; Hur, Stroud, & Finer-Moore, 2006; WATKINS et al., 1998). The pseudouridine synthases have their own preferred regions and substrates and some synthases are only responsible for specific pseudouridine sites, especially in tRNA and rRNA (Spenkuch, Motorin, & Helm, 2014). Among the thirteen enzymes, PUS1, PUS7, TRUB1 and TRUB2 are reported to generate pseudouridine modifications in human mRNA. PUS1, PUS7 and TRUB1 are primarily localized in nucleus, while TRUB1 and TRUB2 are localized in mitochondria. It was reported that TRUB1 is the dominant pseudouridine writer in human mRNA, followed by PUS7 (M. Safra, Nir, Farouq, Vainberg Slutskin, & Schwartz, 2017). The redundancy of mRNA pseudouridine writers make it challenging to study the exact functions of pseudouridine and also the enzymes.

To date, the eraser of pseudouridine remains unknown, probably due to the stable C-C bond between the base and the ribose. How or whether pseudouridine could be reversed like m^6A

dynamically remains unknown. This results in doubt whether pseudouridine is reversible and limits further studies for potential pseudouridine functions biologically. There are also no confirmed pseudouridine reader proteins and Prp5 RNA helicase is the only putative pseudouridine reader from yeast (Wu et al., 2016). Thus how the information from mRNA pseudouridine passes onto the downstream pathways remains poorly understood. Of course, it is possible that pseudouridine does not function in the same ways as other mRNA modifications like m^6A, but new mechanism needs to be proposed by the future works.

1.1.1.3 NGS based Ψ sequencing methods

Next generation sequencing based strategies have been used to identify the distribution of RNA modifications like pseudouridine and m^6A in mRNA in the past decade. Different from mass spectrometry which could only report the overall level of pseudouridine (Addepalli & Limbach, 2011; Taucher, Ganisl, & Breuker, 2011), sequencing based methods could provide the modification position information and probably stoichiometry information of each site. If the RNA molecules are reverse transcribed without any special treatment, then pseudouridine will be paired with A, and m^6A will be paired with T, just the same as unmodified U and A bases. The modifications will not be identified in the cDNA. Thus, all NGS based modification sequencing methods require either chemical or enzyme treatments on the RNA molecules. The logic is straightforward. Any treatment that will make the modified and unmodified bases look different in the final sequencing reads could be considered as potential strategy to identify RNA modifications. There are two ways to induce the different signals between modified and unmodified nucleotides. One is to change the signal of the modified nucleotides, the other is to change the signal of the unmodified nucleotides. The signal change could be mutations,

deletions, RT stops or enrichment folds in pulldown samples. Tens of methods have been developed based on the criteria above in the past decade for several types of mRNA modifications. The various methods have various performance in resolution, transcriptome recall, sensitivity, accuracy and quantifiability. New methods all try to either solve some drawbacks or improve performance of some aspects of older methods.

The first strategy to identify pseudouridine transcriptome wide was raised by three groups independently almost simultaneously, inspired by a primer extension strategy for pseudouridine mapping in 1993 (Bakin & Ofengand, 1993). In Pseudo-seq (Carlile, Rojas-Duran, & Gilbert, 2015; Carlile et al., 2014), Ψ-seq (Schwartz et al., 2014), and PSI-seq (Lovejoy, Riordan, & Brown, 2014), a Ncyclohexyl-*N*′-(2-morpholinoethyl) carbodiimide methyl-*p*-toluenesulfonate (CMC) group is added to the pseudouridine bases which induces reverse transcription stop and thus truncated reads in RNA-seq. However, only high stoichiometry pseudouridine sites could generate strong enough RT stop signals and be detected. The strategy is improved in CeU-seq by adding an azide group to CMC followed by addition of biotin with click chemistry (X. Li et al., 2015). Then the RNA molecules with biotin labeled pseudouridine bases could be enriched and the number of identified pseudouridine sites increases massively from 100-400 sites to more than 2000 sites in human mRNA (X. Li, Ma, et al., 2016). However, the enrichment step makes it hard to quantify pseudouridine levels. The RT stop nature means the pseudouridine sites close to 5’ end of the mRNA molecules is more difficult to be detected then the ones close to 3’ end, thus result in bias for transcriptome wide investigation. Also, the alkaline treatment step can result in severe RNA degradation and the required mRNA input is 5-10 μg, which further limit the application of these methods on biological samples.

Pseudouridine was reported to show deletion signal under bisulfite treatment by RBS-seq in 2019 (Khoddami et al., 2019), which shed on new single base resolution mapping strategies. The ribose ring of pseudouridine was opened to form two types of adduct products and result in a deletion in the reverse transcription product (Everett, 1980; Singhal, 1974). The condition for bisulfite treatment was optimized later in BID-seq to achieve better deletion rate and less background noise (Dai et al., 2023; L.-S. Zhang, C. Ye, et al., 2023). This method also enables quantification of single pseudouridine sites and required input mRNA at as low as ~10 ng. A method using a similar strategy, PRAISE, was also reported in the same year (M. Zhang et al., 2023). PRAISE mixes sulfite with bisulfite at specific ratios to treat pseudouridine and results in higher deletion rate and less C-to-T conversion compared to standard bisulfite treatment.

To note, the identification of pseudouridine transcriptome wide with high accuracy is still challenging to date and the consensus among all existing methods is poor and remains to be further studied in the future (M. Safra et al., 2017). It is also challenging to study the relationship of pseudouridine distributions and functions as it looks like pseudouridine does not function individually and there are no known pseudouridine reader proteins.

1.1.2 N^6-methyladenosine (m^6A)

N^6-methyladenosine (m^6A) is the most abundant mRNA modification in mammalian transcriptome, representing ~ 0.5% of all adenosines (D. Dominissini et al., 2012; K. D. Meyer et al., 2012; I. A. Roundtree et al., 2017). It was first discovered in rat hepatoma cells back in 1974 (Desrosiers, Friderici, & Rottman, 1974), followed by the discovery of its derivative N^6,2′-*O*-dimethyladenosine in 1975 (Wei, Gershowitz, & Moss, 1975). m^6A is discovered to have

preference for DRACH motif (Linder et al., 2015). It has a preference for 3' UTR and regions near stop codon (K. D. Meyer et al., 2012).

1.1.2.1 Enzymes for m^6A

Unlike pseudouridine which has 13 possible writers with no reader or eraser enzymes discovered as of March 2024, m^6A has all three types of related proteins discovered.

The main writer for human m^6A methylation is the METTL3/METTL14 complex. METTL3 was identified in 1997 (Bokar, Shambaugh, Polayes, Matera, & Rottman, 1997) and was demonstrated to work together with METTL14 seventeen years later (J. Liu et al., 2014; Ping et al., 2014; Y. Wang et al., 2014). It was reported that METTL3 is the functioning catalytic component while METTL14 mainly contributes as the structural scaffold (Śledź & Jinek, 2016; P. Wang, Doxtader, & Nam, 2016; X. Wang et al., 2016). The methylation complex also contains a group of cofactors in mammals, including WTAP, FLACC, HAKAI, RBM15 and VIRMA (Balacco & Soller, 2018). The simplicity of writer enzyme composition makes it much easier to deplete m^6A than pseudouridine in cells by knocking out or knocking down METTL3. This is a key factor when designing strategies and validation methods for transcriptome wide m^6A mapping.

The first human m6A demethylase ("eraser") fat mass and obesity-associated (FTO) was discovered in 2011 (Jia et al., 2011), followed by the second one ALKBH5 discovered by the same team shortly after (Zheng et al., 2013). The discovery of m6A erasers marked the thrive of epitranscriptomics studies.

For readers, three classes of m^6A binding proteins have been discovered. The first class, YTH (YT521 homology) domain proteins, recognize specific m^6A base containing RNA

structures directly. YTH domain family (YTHDF1-3) (Shi et al., 2017; X. Wang et al., 2014; X. Wang et al., 2015) and YTH domain containing (YTHDC1-2) (Hsu et al., 2017; Ian A Roundtree et al., 2017; W. Xiao et al., 2016) proteins belong to this class. The second class proteins require the presence of m^6A to alter the local RNA structure, which include heterogeneous nuclear ribonucleoproteins (hnRNPs) (N. Liu et al., 2015; N. Liu et al., 2017). The third class, which includes Insulin-like growth factor-2 mRNA-binding proteins (IGFBP1-3), binds RNA with a folded RNA binding domain and then uses its flanking sequence to recognize m^6A bases (K. I. Zhou & Pan, 2018). The proteins are important for m^6A to be involved in gene expression regulation events and the functions are discussed in the next section.

1.1.2.2 m^6A functions

m^6A was reported to participate in many events in gene expression like splicing, nuclear export and translation (I. A. Roundtree et al., 2017). Most of the m^6A functions involve the writer, reader and eraser proteins. It was reported that m^6A at splice junctions could increase splicing kinetics and m^6A placed in introns are related to slow and alternative splicing events (Louloupi, Ntini, Conrad, & Ørom, 2018). m^6A was also reported to be related to mRNA stability and life span. Transcripts bound by m^6A reader protein YTHDF2 are directed to mRNA decay site rather than to translatable pool (X. Wang et al., 2014). Another m^6A reader protein YTHDC1 was reported to regulate the nuclear export of mRNA in mammalian cells (Ian A Roundtree et al., 2017). It was reported that human YTHDF1 bound to mRNA m^6A could recruit translation initiation factors and promote translation (X. Wang et al., 2015). It was reported that m6A could form clusters, and transcripts with more m^6A clusters had significant lower level of translation (C. Liu et al., 2023).

m^6A levels was regulated in response to stress. In acutely stressed mice, m^6A is altered by glucocorticoid administration (Engel et al., 2018). Depressive disorder could lower blood m^6A levels in patients. METTL3/METTL14 complex recruited to UV damaged DNA could recruit DNA damage repair polymerase κ (Xiang et al., 2017). m^6A could affect development and diseases (Jonkhout et al., 2017). mRNA was shown to maintain mESC at ground states and negatively correlated with gene expression in many developmental regulators (Y. Wang et al., 2014). Knockout of one of the m^6A erasers Alkbh5 in mice impaired their fertility (Zheng et al., 2013). Knockout of m^6A writer Mettl3 would block spermatogonial differentiation and initiation of meiosis in mouse germ cells (Xu et al., 2017). Inhibition of METTL3 could be beneficial to treatment of acute myeloid leukemia (AML) (Yankova et al., 2021).

1.1.2.3 NGS based m^6A sequencing methods

Tens of methods based on different strategies have been developed to map m^6A transcriptome wide in the past decade. Their performance differs in coverage, resolution, quantifiability and input RNA requirement.

The earliest m^6A transcriptome-wide mapping methods based on m^6A specific antibodies. In m^6A-seq (D. Dominissini et al., 2012) and MeRIP-seq (K. D. Meyer et al., 2012), m^6A enriched RNA fragments are pulled down and sequenced and the resulting peaks contain m^6A nucleotides. The width of the peaks determined the sequencing resolution and it's usually 100-200 bases. Later, miCLIP used UV to induce RNA-protein crosslink and following m6A flanking sites mutations during reverse transcription to increase resolution (Linder et al., 2015). m^6A-LAIC-seq switched the fragmentation and immunoprecipitation steps and enabled the detection of overall m^6A levels in each transcript (Molinie et al., 2016). Identification of m^6A by antibodies

results in relatively low-resolution results. Each peak may contain multiple m^{6}A sites and it's hard to coordinate the ~100nt wide features with other single base features in RNA. Also, the immunoprecipitation steps make it difficult to quantify the stoichiometry of single m6A sites. Thus, antibody-free methods are needed for further m6A transcriptome-wide studies.

In 2019, MazF RNase cleavage based methods MAZTER-seq (Garcia-Campos et al., 2019) and m^{6}A -REF-seq (Z. Zhang et al., 2019) were developed, marked the progress on antibody-free m^{6}A sequencing methods. MazF is an RNase that cleaves at the 5' side of an ACA motif and the cleavage is blocked when the first A becomes m^{6}A. Both methods use the strategy to identify m^{6}A sites. To note, the reliability and accuracy of this strategy is based on the 100% cleavage of MazF on non-modified A sites in ACA, otherwise it will not work due to high fraction of false positive results. Although the motif preference of the MazF enzyme limit the detection of m^{6}A sites to ACA motif only, which covers 16% to 25% of all m6A sites, these methods raised the resolution to single base and enabled quantification of the identified m^{6}A sites.

Besides MazF based methods, other chemical assisted m^{6}A sequencing methods have also been published in the same period. m^{6}A -SEAL (Y. Wang, Xiao, Dong, Yu, & Jia, 2020) utilize FTO and dithiothreitol (DTT) to add thiol groups to m^{6}A so that the modification could be enriched by biotin. This method avoids the usage of antibodies but still need pulldown of reads thus could not provide stoichiometry information. DART-seq (K. D. Meyer, 2019) couples m^{6}A reader domain YTH with APOBEC1 which induces C-to-U editing near the m^{6}A sites, but the natural C-to-U editing events in cells could cause false positive results. In m^{6}A -label-seq (Shu et al., 2020), allyl groups are added to m^{6}A by SAM cofactor to induce mutations during reverse transcription. However, the labeling efficiency is not high so that the result could not well cover

the whole transcriptome. Also, the N^6-allyladenosine (a^6A) modifications was added during cell culturing, which detected those positions that have the potential to have m^6A but not really have m^6A in the specific sample. If the reaction preference of a^6A and m^6A are different under the experimental condition, then the sequencing of a^6A could not reflect the actual distribution of m^6A. These methods are good attempts for antibody-free single base resolution m^6A sequencing, while quantification remains a difficult problem.

Recent methods start to seek strategies to quantify single m^6A sites transcriptome wide. In m^6A-SAC-seq, an allyl group is added to m^6A. followed by I_2 treatment and mutation signal in the reverse transcription by a specific reverse transcriptase (L. Hu et al., 2022). This base resolution labeling method yields ~100-fold preference for m^6A over A and has no motif limitation. It is also able to quantify m^6A stoichiometry with as low as 2ng mRNA input. This method changes the signals of the rare positive cases m^6A and maintains the signals of the majority unmodified A, which is a good strategy to avoid false positive identifications. However, the main disadvantage of m^6A-SAC-seq is that the reaction happens on m^6A could also happen on unmodified A, with much less preference, which introduces false positive identifications. Another risk is that if the reaction efficiency is not 100% then some m^6A sites will be missing and the recall will be lower, but such shortcomings is much acceptable compared with high false positive predictions. Thus, when I developed our nanopore data based m^6A prediction models, I used the data from m^6A-SAC-seq as the ground truth. The details will be described in chapter 3.

Different from m^6A-SAC-seq, the following two methods use a complementary labeling strategy. m^6A-SAC-seq changes the signal of m6A sites to have mutations so that m^6A could be distinguished from unmodified A sites. In GLORI, unmodified A sites could be deaminated to inosine (I) while the methyl group in m^6A blocks the reaction (C. Liu et al., 2023). In NGS RNA

sequencing, inosine is read as a G and thus mutation is generated. eTAM-seq uses a similar strategy to deaminate A but not m^6A and results in mutation signals (Y.-L. Xiao et al., 2023). The difference is that eTAM-seq relies on TadA enzyme, but GLORI uses chemicals. Both methods could quantify m^6A fractions at single base resolution. However, mapping modifications by converting the unmapped bases requires very high conversion accuracy, otherwise false positive results will be dominant. For the modification like m^6A occupying ~0.5% of all A bases, at least 99.90% conversion rate for the unmodified bases could result in a false discovery rate (1 - precision) less than 16.67%. Both methods make good progress in finding a high conversion reaction with ~99% conversion rate and we could foresee future work to further raise the conversion rate, although it can be very challenging.

To conclude, the mapping strategies start with antibody pulldown, which generates the very first insight of m^6A distribution, but the resolution is low. Then methods tackling m6A with single base resolution by enzyme or chemical treatment appear, which have limitation in motifs and could not quantify single m^6A levels. Recent methods quantify m^6A transcriptome wide with two different strategies, either mutating m^6A or unmodified A, and achieved m^6A stoichiometry and relative high accuracy.

1.1.3 Other mRNA modifications

Beyond m^6A and pseudouridine, there are several other types of modifications identified in mRNA in the past decade, and their transcriptome wide mapping strategies have been developed, like N^1-methyladenosine (m^1A), 5-methylcytosine (m^5C), N^4-acetylcytidine (ac^4C), N^7-methylguanosine (m^7G), etc. The basic strategies are similar for all modifications, and we could also learn from the cases for other modifications. The modifications could be enriched by

modification specific antibodies. The samples could be treated by either chemicals or enzymes to induce difference between modified and unmodified bases. Here I discuss about some examples of these strategies.

N^1-methyladenosine was first discovered back in 1960s (Dunn, 1961; Hall, 1963) but its appearance in mRNA was determined around half a century later (Dan Dominissini et al., 2016). m^1A contains a methyl group at the N1 position and is positively charged. The additional methyl group disrupts formation of Watson-Crick pairs. Transcriptome wide mapping of m^1A was first revealed in 2016 by two studies. m^1A-seq utilized m^1A specific antibody to enrich RNA containing m^1A and followed by chemical conversion of m^1A to m^6A (Dan Dominissini et al., 2016). m^1A-ID-seq applied both antibody enrichment and base pair disrupt during reverse transcription caused by the methyl group at position N1 (X. Li, Xiong, et al., 2016). m^1A was discovered to appear in ~20% of human transcripts and enriched in 5' UTR and around the start codon. m^1A was discovered to affect translation and could be demethylated by ALKBH3. Later the reverse transcriptase was engineered in m^1A-quant-seq to generate more robust mutation signal during reverse transcription (H. Zhou et al., 2019). The method also enabled base resolution detection and quantification of m^1A sites.

5-methylcytosine modification is most widely spread and studied modification in DNA but it is not broadly studied in RNA molecules (Suzuki & Bird, 2008). Although known to be present in eukaryotic mRNA (Dubin & Taylor, 1975), it was better studied in tRNA than in mRNA, with the discovery of two writers Dnmt2 and Nsun2 (Brzezicha et al., 2006; Goll et al., 2006). One strategy to study m^5C is using modification specific antibodies like m^5C-RIP-seq (Cui et al., 2017). The limitation for antibody-based methods are low resolution and lack of stoichiometry. Like the methods for DNA 5mC, the bisulfite sequencing method for RNA m^5C

was developed in 2012 (Squires et al., 2012). Unmodified C bases are converted to U while m^5C remains to be read as a C. It enabled identification of thousands of mRNA m^5C sites and discovered the enrichment in untranslated regions, while possessing the same problems as the DNA bisulfite sequencing methods. m^5C is not distinguishable from other cytosine modifications like hm^5C and m^3C during bisulfite treatment. Furthermore, bisulfite degrades both DNA and RNA, which makes it hard to achieve long reads and reduces throughput in NGS. For nanopore sequencing, this disadvantage is fatal. To note, this strategy is a negative conversion one, which shares the same high false positive rate problem as GLORI and eTAM-seq mentioned above for m^6A. The experimental conditions for bisulfite sequencing for m^5C was later optimized for better C-to-U conversion and quantification in 2019 (T. Huang, Chen, Liu, Gu, & Zhang, 2019). Ultrafast BS-seq (UBS-seq) was developed in 2024 to reduce RNA damage and improve C-to-U conversion performance by high bisulfite concentration and high reaction temperature (Dai et al., 2024). The required mRNA input could be as low as 10ng, while maintaining the ability to quantify m^5C fractions.

Unlike methylation, acetylation of cytosine is not well studied in the past several decades. N^4-acetylcytidine was first discovered in bacteria tRNA (Stern & LH, 1978) and then identified in eukaryotic tRNA and 18S rRNA (Boccaletto et al., 2022). It was reported that ac^4C has stronger base pairing with G than unmodified C (Kumbhar, Kamble, & Sonawane, 2013). ac^4C was reported to be present in mRNA by mass spectrometry and its transcriptome wide distribution was revealed by acRIP-seq with ac^4C specific antibody (D. Arango et al., 2018). N-acetyltransferase 10 (NAT10) was the only known writer of eukaryotic mRNA ac^4C. ac^4C was discovered to maintain mRNA stability and promote translation when acetylation happens at the wobble position. The sequencing resolution was raised to single base and the stoichiometry is

achieved in ac^{4}C-seq, with C-to-T mutation signal induced by chemical treatment (Sas-Chen et al., 2020; Thalalla Gamage, Sas-Chen, Schwartz, & Meier, 2021). It was reported that ac^{4}C was absent in human and yeast mRNA but could be induced by overexpression of its enzyme complexes. A later paper using acRIP-seq reported the existence of ac^{4}C and revealed that ac^{4}C in 5'UTR affect translation initiation (Daniel Arango et al., 2022). The existence of ac^{4}C in human RNA was supported by another method named FAM-seq in 2023, which incorporate modifications by CoA metabolite fluoroacetyl-CoA (Yan et al., 2023). The major doubt of FAM-seq is similar as m^{6}A-label-seq, that is the incorporation of an alternative chemical may not reflect the transcriptome distribution of the original target modification. Thus it's hard to be used as evidence for existence of ac^{4}C in human transcriptome. Also, as biotin enrichment is required in the following protocol, the method could not quantify ac^{4}C fractions. It looks like the existence of ac4C in human transcriptome still remains to be determined by better methods in the future.

N^7-methylguanosine is widely known to be the 5' cap of mRNA to protect the RNA molecule from degradation as well as affects RNA events like splicing and translation. However, the existence of internal m^{7}G within mRNA was not demonstrated until 2019 by two sequencing methods developed in the same manuscript (L.-S. Zhang et al., 2019). m^{7}G-MeRIP-Seq uses m^{7}G specific antibody to enrich m^{7}G for RNA-seq and thus is not single base resolution. Internal m^{7}G was discovered to prefer GA enriched motifs. In m^{7}G-seq, m^{7}G was turned into an abasic site by chemical treatment and could be labeled by biotin, followed by enrichment and sequencing. The abasic site would result in mutation in reverse transcription. m^{7}G /G ratio was discovered to be 0.02%-0.05% and METTL1 was discovered as the major m^{7}G writer protein.

1.1.4 Mapping multiple modifications in the same sample

To date, the majority of NGS based RNA modification identification methods focus on one specific type of modification. It is challenging to identify multiple modifications in the same sample at the same time, mainly because the pre-treatment steps for different modifications are not compatible with each other. Also, the signal of different modifications could be the same in RNA-seq and make it hard to tell them from each other. However, the demand of mapping multiple modifications in the same samples always exists. Mapping multiple modifications simultaneously helps to understand the potential coordination between modifications and reveal the potential regulation relationship between them. It also helps better understand how different modifications function together in the same pathway or the same spatial location. Studies of multi-omics are hot topics for disease biomarker discovery and diagnosis, but the range of omics usually doesn't contain any modifications or at most DNA 5mC. Little attention has been paid to multiple modification omics and their potential contributions to biomedical research. Thus, the field to study multiple modifications remain to be developed.

Although not prevalent, there are NGS based methods which could deal with multiple RNA modifications at the same time. RBS-seq was designed for mapping of pseudouridine, m^5C and m^1A at base resolution (Khoddami et al., 2019). After bisulfite treatment, m^5C maintains its original read out as C but unmodified C will be converted to U. m^1A is converted to m^6A under bisulfite treatment and thus is read as A, while it results in misincorporation in RT without bisulfite and is read as a T. Pseudouridine is read as deletion signal after bisulfite treatment. DAMM-seq was designed to map m^1A, N^3-methylcytidine (m^3C), N^1-methylguanosine (m^1G), and N^2,N^2-dimethylguanosine ($m^2{}_2G$) simultaneously and quantitatively at base resolution in

mitochondrial RNA and tRNA (L.-S. Zhang, Ju, Jiang, & He, 2023). Its basic idea is to report the misincorporation rate at the modification sites.

The basic logic for NGS based simultaneous detection of multiple modifications is that multiple modifications show different signal changes during the same treatment reactions. The difference could be read out when comparing the reference sequences, the treated reads and the untreated reads. The chemical nature of different RNA modifications limits the development of chemical or enzyme assisted multi-modification mapping methods. Thus, it is a good idea to involve nanopore sequencing and computation assisted strategies for such problems, and it will be discussed in chapter 3.

1.1.5 tRNA and rRNA modifications

tRNA and rRNA are important non-coding RNA in cells. Both are heavily modified by RNA modifications, which largely affect the structures and functions of tRNA and rRNA. The modification fraction in tRNA and rRNA are usually higher compared with mRNA, making it easier to quantify the presence and fraction of modifications by mass spectrometry (Masato Taoka et al., 2018).

The major function of tRNA is to generate the peptide chain based on the information provided by mRNA. Based on the fundamental role of tRNA, recent years of studies have revealed its potential in disease treatment, for correcting the mismatches within mRNA and produce correct protein product (Hou et al., 2023). tRNA is heavily modified and each tRNA is reported to have 13 modified sites on average (Pan, 2018). Most types of identified modifications appear in tRNA. The modification could be addition of methyl, acetyl, amino acid side groups, or isomerization and deaminated nucleotides (Boccaletto et al., 2022). For example, beyond

pseudouridine, a group of pseudouridine derivatives also appear in tRNA, like 2'-*O*-methyl-Ψ (Ψm) and 1-methyl-Ψ ($m^1Ψ$) (Boccaletto et al., 2022).

tRNA modifications have many functions. Mutations in tRNA modification enzymes have been reported to be correlated to many diseases (Jonkhout et al., 2017; Torres, Batlle, & Ribas de Pouplana, 2014). tRNA modifications could be used as biomarkers to predict the severity of diseases like COVID(Katanski et al., 2022). Engineered pseudouridine could be used to raise the read through of premature termination codons (PTC) (Luo et al., 2024), while PTC is reported to be related to many diseases like cystic fibrosis (Cheng et al., 1990) and Hurler syndrome (Ballabio & Gieselmann, 2009). m^5C in tRNA was reported to be related to stability and cleavage (Schaefer et al., 2010; Tuorto et al., 2012).

rRNA is the fundamental component of the translation machine. Human rRNA is heavily modified by pseudouridine and 2'-*O*-methyl, including all Am, Cm, Gm and Um. 231 of these sites were quantified in the previous publications (M. Taoka et al., 2018). There is also a group of pseudouridine derivatives in rRNA, including 3-methyl-Ψ ($m^3Ψ$), 3-(3-amino-3-carboxypropyl)-Ψ ($acap^3Ψ$) and $m^1acap^3Ψ$ (Boccaletto et al., 2022; M. Taoka et al., 2018). ac^4C, m^7G, m^6A, $m^6{}_2A$ are also discovered in human 18S rRNA at specific positions while m^1A, m^6A, m^5C and m^3U appear in 28S.

The nature that rRNA is heavily modified and the modification sites are relatively conserved compared to mRNA is strength for generating known sites. In chapter 2, I used this strategy and got a list of pseudouridine sites from rRNA of multiple species for nanopore data supervised learning model training. It is also beneficial to have all human rRNA pseudouridine site fractions quantified in the previous studies. We used such information to train single read pseudouridine prediction model in chapter 5.

1.2 Nanopore direct RNA sequencing (DRS)

Whole genome and transcriptome sequencing has become possible with the development of next generation sequencing technologies. However, NGS technologies have many limitations. For example, for Illumina sequencing, the difficulty in keeping the same pace for all synthesis event at the same spot results in messy fluorescence signals as the read gets longer and thus the read length is limited to hundreds of nucleotides. To avoid mixture of signals from multiple synthesis events, single molecule sequencing is a possible solution. Although single molecule sequencing doesn't need to worry about messy signals, the major problem becomes how to detect the weak signal from only one molecule. Such technologies are called third generation sequencing technologies (TGS). Two major types of TGS techniques have been widely used in biology studies, PacBio SMRT-seq (Flusberg et al., 2010) and Oxford Nanopore Technologies (ONT) nanopore sequencing. Both methods are single molecule sequencing technologies with the ability to tell apart modification signals from unmodified bases directly. In this dissertation, we mainly focus on ONT nanopore sequencing.

1.2.1 History

Oxford Nanopore Technologies started on the idea of "strand sequencing" in 2009 and the ability of reporting *de novo* base calling of genome sequences became available in 2012 (Brown & Clarke, 2016). The "pore" was first nanopore protein *Mycobacterium smegmatis* porin A (MspA) (Morton et al., 2015) and later an engineered version of Curlin sigma S-dependent growth protein CsgG from *E. coli* (R9.4 or R9.5) (Ayub & Bayley, 2016; Henley, Carson, & Wanunu, 2016; Ip et al., 2015). In the first several years, nanopore is only used to sequencing

DNA molecules. It was shown in 2013 that MspA could be used to identify DNA modifications (Laszlo et al., 2013; Schreiber et al., 2013). 5-methylcytosine, and 5-hydroxymethylcytosine could be distinguished from the unmodified cytosine in synthesized DNA. It was demonstrated that nanopore could be used to identify DNA modifications in human genome samples in 2017 (Simpson et al., 2017). Based on the principle of nanopore sequencing, any macromolecules with appropriate diameter could be sequenced, with some customizations to the hardware and software. It was not until 2018 that nanopore direct RNA sequencing technology was reported by ONT (Garalde et al., 2018). Since then, nanopore has been used to sequence transcriptomes of all kinds of species and reveal different RNA events including RNA modifications.

1.2.2 How nanopore sequencing works

Nanopore sequencing uses engineered pore proteins as sensors to detect the molecules going through it (Banerjee et al., 2010). There are thousands of pores attached to a membrane. A voltage difference is applied to the two sides of the membrane so that constant current flow is formed in the pores. The diameter of the pores is compatible to the width of the macromolecules to be measured. Usually DNA and RNA molecules has overall negative charges, so they have a trend to move to the high electric potential side automatically. To maintain a stable speed of going through the pores, a helicase motor protein sits on top of each pore so that the pace is controlled. When the macromolecule goes through the pore, it will partly block the current flow. The narrowest region of the pore which is valid for reflecting the current blockage is called sensing region. Different nucleotides have diverse size, shape and charge state so that their abilities to block the current flow are different. The difference of blockage could be reflected as the change of the current signal over time. The current signal has two attributes, one is the

strength of the signal, the other is the length of the signal, which is also called dwell time. Usually the sensing region of the nanopore is longer than the length of one nucleotide, which means there are more than one nucleotide in the sensing region at a time, so each current signal recorded is the sum of the contribution of several (usually around 5-6) neighboring nucleotides (Wick, Judd, & Holt, 2019). To achieve the DNA/RNA sequences from the raw current signal, hidden Markov model (HMM) or neural network (NN) algorithms were developed to deduce the nucleotides information. RNA modifications differ from the four common nucleotides in size, shape and charge state, so theoretically modifications could be identified from the unmodified nucleotides, just like A, C, G, U could be identified from each other.

For nanopore direct RNA sequencing (DRS), RNA molecules with poly A tail are collected with a poly T adaptor and the sample preparation kit also use magnetic beads to select RNA molecules longer that 200nt. Thus, nanopore DRS is designed for direct sequencing of mRNA samples. For other types of RNA without poly A tail, specific adaptors need to be designed for library preparation.

1.2.3 Advantages

Compared with next generation sequencing, nanopore sequencing has many advantages. These advantages enable specific applications of nanopore sequencing to add novel knowledge to current discoveries by NGS methods.

Nanopore sequencing run RNA molecules directly, while next generation sequencing (NGS) handles RNA in an indirect manner. To sequence RNA samples by NGS, the RNA molecules need to be reverse transcribed into complementary DNA (cDNA) and then generate double-stranded DNA molecules followed by amplification and then the sequencer could run the

sample. RNA molecules themselves could not be directly run in the sequencer. However, there is no limitation for the type of macromolecule that goes through the nanopore, which means RNA molecules could be sequenced directly without the necessity of reverse transcribed into cDNA. The direct sequencing of RNA molecules enables the detection of any additional information in the RNA molecules, for example RNA modifications. For NGS based methods, all mRNA modification information is lost during reverse transcription (RT) without additional chemical or enzyme treatment, as all nucleotides in cDNA derived from mRNA are the four common bases. To identify RNA modifications by NGS, we have to rely on the footprint left by RNA modifications during reverse transcription, like RT stops, induced mutations, deletions etc. The limitation is obvious. RT stops truncate reads and all nucleotides after the stop are lost. Mutations and deletions could be introduced by factors other than RNA modifications. As a comparison, modification detection by nanopore direct RNA seq doesn't have such limitations. The modified RNA nucleotides go through the pores as they originally are and the changes in signal are recorded directly.

Nanopore sequencing doesn't have a limitation on the length of reads. Theoretically, the length of the read depends on the length of the molecule itself. This means we could achieve full length mRNA, rRNA and long non-coding RNA (lncRNA) reads. Based on this, the alternative splicing events and splicing isoforms could be reported from the data, which expands the scopes of data analysis and biological discoveries. For DNA, linkage of distant SNPs and large structural variations could be reported. To note, in lab practice, it's difficult to obtain 100% full length DNA or RNA molecules as the long chains are easily broken during library preparation, often by pipetting. Despite this, the average length of produced reads by nanopore sequencing is

still longer than NGS methods. For poly A RNA samples, the average could be more than 1000 nucleotides.

Nanopore sequencing doesn't require PCR amplification for sample preparation. PCR amplification could result in bias in expression levels as different transcripts are not amplified the same times during the reactions. This could result in inaccuracy in sequencing results. Also, modification information could be removed during PCR amplification. Nanopore sequencing does not require PCR amplification during sample preparation and thus could reduce bias introduced from PCR process.

In addition, poly A tail length could be directly estimated from nanopore direct RNA sequencing data without requirement of extra pre-treatment of the samples (Rachael E Workman et al., 2019). This is a novel application that could not be completed in the previous methods.

The size of the nanopore sequencer "MinION" is very small and is portable so that it's possible to do the sequencing in real time outside the lab, which has the potential to be applied in medical or field research settings.

1.2.4 Limitations

Although nanopore sequencing has many edges over NGS based RNA sequencing, as it is a new technology in development, there are also many limitations which needs to be paid attention to when applying the method to avoid mistakes. The problems are to be solved in the future development of the technology.

One limitation often ignored by many researchers is that nanopore sequencing is not compatible with those library preparation protocols that could result in fragmented RNA molecules. Fragmentation is not only brought about by mechanical shearing during pipetting, but

also by other key factors in the experimental design. For example, in ribosome profiling (Ribo-seq), the sample is treated by RNase and only the ribosome attached mRNA fragment is kept. This means that all mRNA reads are short fragments which is not suitable for nanopore sequencing. In another example, divalent metal cations have the potential to degrade RNA, which means if the pretreatment steps of the RNA sample require divalent metal cations, then the product RNA is not suitable for nanopore sequencing. In the protocol of m^6A-SEAL (Y. Wang et al., 2020), the first step of FTO oxidation of m^6A requires existence of Fe^{2+}, which degrades mRNA. Thus this protocol is only suitable for NGS based methods but not nanopore sequencing. This drawback deserves attention from the researchers who would like to couple nanopore sequencing with a second protocol for RNA modification studies.

The accuracy of base calling by nanopore sequencing is not as high as NGS. Usually NGS could call bases with >99.9% accuracy, while according to previous reports, the accuracy for nanopore direct RNA sequencing at the beginning of commercialization was only 86% (Jain, Abu-Shumays, Olsen, & Akeson, 2022). Another study reported the error rate for RNA molecules as 7-12% (Wick et al., 2019). Oxford Nanopore Technology (ONT) keeps working on raising the single base accuracy of nanopore sequencing. They updated the base calling software guppy base caller and the accuracy is reported to be more than 91% (Grünberger, Ferreira-Cerca, & Grohmann, 2022; Jain et al., 2022; Rousseau-Gueutin et al., 2020). Another strategy is to apply two tandem nanopores for each single molecule. The molecule goes through the first pore and then the second one so that the current signal of each molecule is recorded twice. In this way, the accuracy is reported to rise to 99% for nanopore DNA sequencing (Sereika et al., 2022); however, direct RNA sequencing technology is still waiting to be optimized in this way. The

increase of accuracy would benefit more reliable studies of transcriptomes by nanopore sequencing and we could foresee that the accuracy will be raised in the future.

Another major limitation for nanopore direct RNA sequencing is high cost and low amount of production. The cost has two aspects. The first aspect is on the sample requirement. Back in 2020, each DRS run requires 500 ng polyA RNA as input. Usually only culturable cell lines could afford the demand and the door is closed for precious samples or clinical samples. According to the website of ONT and previous reports, the RNA input requirement decreased to 50 ng polyA RNA later (Jain et al., 2022). The second aspect is about the cost. Each flow cell, which is not cheap, could only be used for one run and for one RNA sample, and there is no commercial barcoding system. The researchers need to run multiple flow cells in one project or they have to do barcoding on their own. However, barcoding for mRNA may not be a good idea, as each flow cell is supposed to produce only 1 million DRS reads. In our own research, the number of reads produced by each flow cell is not stable and have very high variance, ranges from 100K to over 3 million. This increases the challenge of achieving replicates of samples and getting results from each replicate with similar quality. Of course, as the technology becomes more developed in the future, it's promising that the cost will decrease, and the yield will increase.

Beyond, it was reported that nanopore direct RNA seq has bias based on the length of RNA molecules. In specific ranges, shorter reads are preferred over longer reads and thus it results in the missing of mRNA isoform information from specific genes (Jain et al., 2022). In our practice, we also noticed that samples with too many short RNA molecules (usually < 200nt) would largely decrease the number of read yields of a nanopore run. Thus, it is very challenging to use nanopore direct RNA seq on small RNA samples.

1.2.5 Nanopore sequencing strategies for RNA modifications

As mentioned above, one of the major advantages of nanopore sequencing is that it could run RNA samples directly without the necessity of reverse transcription. As RNA modifications differ from unmodified bases in sizes, shapes and charging states, it is possible to identify RNA modifications directly from nanopore direct RNA sequencing data without any pre-treatment like bisulfite. As the data processing software provided by Oxford Nanopore Technologies have very limited functions in modification detection, researchers usually develop their own computation pipelines to identify RNA modifications directly from raw nanopore sequencing data. Those statistical methods and machine learning methods for nanopore data processing will be discussed in the section 1.3. The machine learning based pipelines for pseudouridine and m^6A identification developed in this dissertation will be discussed in chapter 2 and 3.

Although pre-treatment is not a necessity for nanopore RNA modification detection, we could still learn from NGS based modification identification strategies. The fundamental idea for RNA detection is to have different signals for modified and unmodified nucleotides. In NGS methods, the difference is from mutations, deletions, RT stops or immunoprecipitation enrichment fold. In nanopore sequencing, the current signals of modifications and unmodified bases are usually different. If the difference is too small to be notified by naked eyes, then we enhance the difference either by experimental methods or data processing methods.

It is possible to use chemical or enzyme pre-treatment on the mRNA library to enhance the difference of modified nucleotides and unmodified ones. For example, by adding a big extra ring structure to m^6A but not A, it is possible to generate a big different signal for m^6A in nanopore raw data. However, as one of the major advantages is long read length, if the pre-

treatment of adding a big ring to m^6A will result in fragmented RNA molecules, then the pre-treatment methods may not be appropriate, like the bisulfite treatment for m^5C, CMC treatment for pseudouridine or the protocol of m^6A-SEAL for m^6A mentioned above. Also, the strategy like ribo-seq that protect specific mRNA regions and digest the rest is not suitable for nanopore sequencing, as the protected regions are usually very short. Instead, polysome profiling is currently the only method to study mRNA translation on the ribosome by nanopore sequencing. We used mRNA samples from polysome profiling to study effect of RNA modifications on translation in chapter 4.

Immunoprecipitation with modification specific antibodies or chemicals is widely used for all kinds of RNA modifications in NGS methods. This strategy is beneficial when the modification is very rare in the transcriptome. It could also contribute in nanopore sequencing to raise the ratio of positive cases so that lowering the false discovery rate. However, in spite of its potential benefit, it's challenging to apply immunoprecipitation-based methods in nanopore sequencing as the mRNA molecules are of full length. Better idea is needed for such strategy to be applied in nanopore sequencing.

1.2.6 Other applications of nanopore direct RNA sequencing

Beyond identification of RNA modifications, the advantages of nanopore sequencing also enable its application in many other fields. It has been used to determine the sequence of RNA viruses. It was used to identify the whole transcriptome of SARS-CoV-2, contributing to the fight against the pandemic in time (Kim et al., 2020). It has also been applied to other RNA viruses (Viehweger et al., 2019; Wongsurawat et al., 2019).

Poly A tail length was directly estimated by nanopore direct RNA sequencing data and its correlation with RNA expression levels was revealed (Rachael E Workman et al., 2019).

Nanopore direct RNA sequencing could be used to determine RNA structures. In PORE-cupine, single stranded RNA was labeled by 2-methylnicotinic acid imidazolide azide (NAI-N3) to reveal the secondary structures (Aw et al., 2021). In nanoSHAPE, the researchers used the 2′-hydroxyl acylation analyzed by primer extension (SHAPE) strategy to detect RNA structure as well as 2'-methyl modifications, by labeling the 2'hydroxyl group of exposed RNA nucleotides with chemical labels (Stephenson et al., 2022). In SMS-seq, exposed RNA is labeled by Diethyl pyrocarbonate (DEPC) to depict secondary structures of RNA molecules (Bizuayehu et al., 2022). DEPC could be used to report single stranded adenine sites. To conclude, these methods label exposed RNA with chemicals to deduce the secondary structure of RNA molecules.

Recently the nanopore direct RNA sequencing processing protocol has been optimized to enable quantitative analysis of tRNA in Nano-tRNAseq (Lucas et al., 2024). It overcame the problem that the software discards short reads and raised the coverage of tRNA by over 10 folds.

Nanopore sequencing has been reported to be combined with the library preparation protocol of single cell RNA seq. In SCAN-seq2, two groups of barcodes are added to 5' and 3' ends of each transcript as its cell identity (Liao et al., 2023). Over 5000 cells are sequenced in a pool. To note, the barcodes are added during reverse transcription and finally cDNA is sequenced so it's not direct RNA sequencing.

1.2.7 Nanopore sequencing for other macromolecules

1.2.7.1 DNA

The nature that nanopore sequencing could produce long reads and read modification directly enables its many applications in DNA sequencing. Nanopore DNA sequencing could be used to identify DNA modifications, as the modification bases show different signals from unmodified bases. Hidden Markov models could be used to identify 5mC, 5hmC and 6mA (Rand et al., 2017; Simpson et al., 2017). Deep learning models like DeepSignal and DeepMod could identify 5mC and 6mA (Q. Liu et al., 2019; Ni et al., 2019). Nanopore DNA sequencing could also be used to detect complex structural variations (SV) (Sedlazeck et al., 2018).

Nanopore DNA sequencing could product high quality long reads thus it is a very appropriate tool to study the complicated metagenome sequences (Sereika et al., 2022). It is also used for fast same-day diagnosis for diseases like brain tumors (Euskirchen et al., 2017), which shed light on practical clinical usage.

The long read property enables more detailed mapping of the human genome, especially for regions with high copy number repeats like centromere and telomere. A reference of human GM12878 Utah/Ceph cell line was assembled with nanopore long reads (Jain et al., 2018). In 2022, a series of papers were published for the complete human genome (Aganezov et al., 2022; Altemose et al., 2022; Gershman et al., 2022; Hoyt et al., 2022; Nurk et al., 2022; Vollger et al., 2022). With the help of PacBio and Nanopore long read sequencing technologies, Telomere-to-Telomere (T2T) Consortium managed to complete the last 8% of human genome left by the Human Genome Project. More detailed maps of genetic variations, epigenetics pattern and repeat elements are also revealed. This project reflects the edges of TGS technologies on highly repetitive regions in the genome.

1.2.7.2 Protein

Besides DNA and RNA, other macromolecules like peptides could also go through engineered pores and thus have the potential to be sequenced by nanopore sequencing. Currently, the high throughput sequencing method for protein has not been developed yet but such technology is of high demand for better understanding of the omics of all expression products of a cell and the regulation and metabolisms of them.

Recently, it was reported that an engineered MspA could be used to discriminate all 20 proteinogenic amino acids and 4 amino acid modifications including N^{ω},N'^{ω}-dimethyl-arginine (Me-R), *O*-phosphoserine (P-S), *O*-acetyl-threonine (Ac-T) and N^4-(β-*N*-acetyl-$_D$-glucosaminyl)-asparagine (GlcNAc-N) (K. Wang et al., 2024). A quadratic SVM model was used to classify the signals from 5 features. Another group used α-hemolysin nanopores to achieve similar results (Yun Zhang et al., 2024). However, these methods could only identify the existence of specific types of amino acids in their free state within the samples but could not read out the sequence content as a series. Both teams tried to digest a peptide and then use nanopore to sense the free amino acid excised from the peptide, which were good proof-of-concept attempts but are still far from practical usage.

The challenge for nanopore protein sequencing is obvious. There are only 4 types of common DNA or RNA nucleotides but there are 20 for proteins. It's much harder to develop a method to distinguish 20 classes than 4 classes. Also, there is very little prior knowledge for peptide sequencing, which means researchers could not conveniently design experiments and train models based on previous data, like nanopore RNA models based on NGS data. The existence of post translational modifications of amino acids makes it even more challenging to

know the information of peptides. Secondary structures also place obstacles for sequencing library preparation. To conclude, nanopore sequencing is one of the most promising approaches for peptide sequencing but many challenges remain to be solved before the technology could be put into practical use.

1.2.7.3 Glycan

Recently, a modified MspA nanopore was used to identify different disaccharide isomers and was applied to detect the existence of sucrose in yogurt (S. Zhang et al., 2023). The working mode is similar to amino acid sequencing mentioned above. Here nanopore is more like a sensor to identify the existence of specific molecules in the samples, rather than read out the series of a sequence from a macromolecule. Although these are small molecules, but it shed on promise in detecting glycan side groups in macromolecules like RNA or protein.

1.3 Machine learning for nanopore direct RNA sequencing data analysis

The drastic development of ML/AI technology makes it possible to solve problems in biology field. ML models are good at dealing with large amount of data and extract useful information from them. RNA sequencing data usually contain the expression information of genes, as well as the distribution of RNA modifications. The size of raw data is usually huge and needs to be processed by high performance computers. Thus, it is a good idea to reveal the information of RNA modifications with the help of machine learning methods.

1.3.1 Machine learning strategies

The machine learning strategies could be classified based on whether labeled ground truth is needed for model generation. Supervised learning, unsupervised learning and their combinations are widely used in natural science studies including biology (Greener, Kandathil, Moffat, & Jones, 2022; Mahesh, 2020; Tarca, Carey, Chen, Romero, & Drăghici, 2007). Supervised learning methods use data with labels to train models to complete tasks. The labels are the "ground truth" that you believe to be true to describe the records. Usually, the labels are from prior knowledge of the field. The input for the model training is the features, which describe all kinds of different aspects of each record. The process of model training is basically looking for a best map from the features to the labels. It's something like a function, but the format of the model doesn't necessarily look like a function. Supervised learning methods are usually used when we have the data to show the phenomenon and the consequences, and we want to build a model to describe how the phenomenon and the consequences are linked with each other. For example, if we have RNA expression data of healthy and cancer patients, we could build supervised learning models for the data. We could learn from the data about the features that are more important than others, or in the case, the genes that contribute more to the generation of tumors. Sometimes, these statistically significant features are also biologically important so that we could gain more insight into the specific problems. For example, find biomarkers for cancers. After we build the supervised learning models, we could also apply them to new records without labels and do prediction. This help to evaluate which category the record belongs to, or how much the output value would be. For example, the models connecting RNA expression and cancer diagnosis could be used to evaluate whether the new patients have tumors.

There are two major types of questions supervised learning models could solve, depending on the type of the labels. One is classification, where the labels are usually discrete categories. The other is regression, where the labels are continuous values describing the extent of some variables. Both types are widely used in biology questions. For example, we could build classification models to identify whether the RNA nucleotide is modified or not. We could also build regression models to predict the modification fractions of a site. The choice of model type relies on the type of questions to solve.

Unsupervised learning methods do not rely on pre-defined labels. They usually gather the records which have similar patterns of features and form clusters. We usually want to find new patterns and subclasses from the clustering results of the whole group of data.

Machine learning methods are widely used in identification of RNA modifications from nanopore direct RNA sequencing data. For the current methods, there are two strategies concerning the types of samples used to train the models (Zhong et al., 2023). For “single mode” methods, there is only one sample containing both modified and unmodified sites, and the ground truth is from previous studies, probably from NGS data. For “compare mode” methods, the signals from two samples, one with modifications and the other one without, are compared, and the difference between signals is used to determine whether the site is modified or not.

1.3.2 Single mode methods

There is a group of supervised learning methods for modification identification based on nanopore direct RNA sequencing data. Single mode methods are usually supervised learning methods, which require training materials from nanopore direct RNA sequencing data and

labeled ground truth from previous studies. The existing methods covered a large range of different supervised learning algorithms.

1.3.2.1 Methods for m^6A

Nanom^{6}A generated extreme gradient boosting (XGBoost) models to identify m^6A in mRNA (Gao et al., 2021). It used mean, median, standard deviation and width of raw current signals from centered and flanking sites as features to build the models. The data were from 130 m^6A sites in synthesized RNA from RRACH motifs. It was applied to stem-differentiating xylem of *Populus trichocarpa* and revealed the correlation between poly A tail length and m^6A modifications.

EpiNano used the base calling errors of m^6A and trained support vector machine (SVM) models to reveal m^6A (H. Liu et al., 2019). Its features include both base calling errors like mismatches, indels, change in base quality scores, as well as the current signal intensity and standard deviation. Synthesized RNA molecules with 100% m^6A or 100% unmodified A sites are used to generate the signals. These molecules are de Bruijn sequences for all possible 5mers. The final model was applied to yeast samples to detect m^6A modification sites *in vivo*.

MINES used a random forest (RF) model to identify m^6A (D. A. Lorenz, S. Sathe, J. M. Einstein, & G. W. Yeo, 2020). It limited the range to 6 motifs, AGACT, GAACT, GGACA, GGACC, GGACT and TGACT, which covered over half of all m^6A sites. This could raise the model accuracy for each model and facilitate the feature extraction of flanking sites. Its ground truth was from miCLIP data. It used raw current signals extracted by Tombo as features to train the models. Tombo is a package provided by ONT to extract current signal intensity and dwell time from raw sequencing data. Its problem is that Tombo could not deal with spliced reads.

Tombo itself also has commands for detection of DNA and RNA modifications like 5mC and m^6A. It compares the signals of samples and reference sequences and call modification sites when the difference is big enough, which does not provide high accuracy in modification calling.

DENA was a bidirectional RNN model with LSTM for m^6A identification (Qin et al., 2022). It didn't limit the sequence context of m^6A. It used data from the WT and m^6A deplete Arabidopsis samples to train the m^6A models to avoid problems of synthesized sequences. The ground truth of m^6A sites was achieved by applying a software differr to compare the signal difference of WT and m^6A deplete samples, which heavily relied on the performance of the software. It used feathers directly from the current signal, including mean, median, standard deviation, base quality score and dwell time.

m^6Anet was a multiple instance learning neural network model which considered the mixture of modified and unmodified reads when dealing with the training materials (Hendra et al., 2022). It used m^6A ground truth from m^6ACE-seq and view all m^6A sites as partially modified. m^6ACE-seq was an antibody-based photo-crosslinking method for m^6A identification (Koh, Goh, & Goh, 2019), which might not be the best choice for ground truth resource but it could be due to limitation of better methods at that time. Instead of using average values from all reads covering a specific site, it generated high dimension presentations of the information collected from each read. Theoretically the method could be applied to any m^6A motifs, but the authors limited it to DRACH motifs. It used current intensity, standard deviation and dwell time to generate features.

1.3.2.2 Methods for pseudouridine

NanoRMS was designed for pseudouridine identification and was applied to *Saccharomyces cerevisiae* (O. Begik et al., 2021). It tried both unsupervised (K-means) and supervised (K-nearest neighbors, KNN) learning methods for the quantification of pseudouridine sites. It used current signal intensity and trace as features and was trained on synthesized RNA "curlcake" with either modified or unmodified bases covering all possible 5mers. Mixtures of the reads were used to train quantification models.

Penguin was a machine learning based pipeline for pseudouridine prediction from nanopore direct RNA sequencing data (Hassan, Acevedo, Daulatabad, Mir, & Janga, 2022). The authors tried RF, SVM and NN models. The ground truth of pseudouridine sites was from previous databases and literatures. It used k-mer sequence content, current signal mean, standard deviation and dwell time from raw sequencing data. Most of the features are reasonable, but it may not be proper to include k-mer sequence content as features. Under different biological treatment like WT and writer knock out samples, the same positions within the same sequence content in the two samples could differ largely in modification state, which means the modification state could not be predicted from sequence content. Sequence content information is derived from the reference file but not from the sequencing data during alignment, so it is not part of the attributes of the sequencing data. This method could be used in some conditions but will result in bias when comparing two different samples with the same reference.

There is a group of supervised learning methods extracting features from sequence content solely and predict the most likely pseudouridine sites in the reference sequences (Bi, Jin, & Jia, 2020; Chen, Tang, Ye, Lin, & Chou, 2016; He et al., 2018; Khan, He, Wang, Chen, & Xu, 2020; F. Li et al., 2021; Y. H. Li, Zhang, & Cui, 2015; K. Liu, Chen, & Lin, 2020; Lv, Zhang,

Ding, & Zou, 2020; Song, Chen, et al., 2020; Song, Tang, et al., 2020; Tahir, Tayara, & Chong, 2019). These methods did not analyze raw RNA sequencing data and instead viewed the modifications as static. The basic idea is modification state could be determined by sequence content, regardless of the metabolic state of the samples. To train the models, a list of known pseudouridine sites is downloaded from previous results as true labels so that machine learning models could be trained to identify the cooccurrence probabilities of certain sequence and the modifications. When applying the models, the input are sequence content and output are whether there is likely to be pseudouridine modifications in the sequence. However, as we know, the same sequences could be either modified or unmodified based on the sample treatment conditions and metabolic states, which is not predictable from the sequence content solely. Such methods contribute modestly to solving real world biological questions, while are fair practices to apply machine learning approaches on biology materials and could provide new insights into aspects like feature types.

1.3.2.3 Methods for other modifications

In Dinopore method, a convolutional neural network (CNN) model was developed to identify inosine from nanopore direct RNA sequencing data (Nguyen et al., 2022). The method considered both current signal features and base calling error features like insertions and deletions from multiple sites around the modified sites and applied 43 features in the model. It also generated a regression model to evaluate the stoichiometry.

1.3.3 Compare mode methods

"Compare mode" methods usually use clustering strategies to identify modifications. Gaussian mixture models (GMM) were used to evaluate the number of RNA modifications in the samples from the current signals (Ding, Bailey IV, Jain, Olsen, & Paten, 2020).

Nanocompore used a 2 components Gaussian mixture model to identify RNA modifications (Leger et al., 2021). It compared the current signals from an experimental sample and a low-modification control sample which could either be IVT or writer knock out samples. It used logistic regression to test whether the reads in the GMM clusters significantly belong to two clusters. This method does not involve the process of model training on previous known modification data but require pairs of samples when performing modification prediction. In the paper the method was mainly applied to m^6A, but the same method could also be applied to other modifications, as long as the low-modification sample is available.

xPore made a Gaussian mixture model for modified and unmodified nucleotides and used z-test to quantify the significance of difference (Pratanwanich et al., 2021). The features are the current signals from 5mer events. It could be used to perform on single reads to calculate modification fractions. This method does not limit m^6A within specific motifs and theoretically could be applied to any other RNA modifications.

1.3.4 Factors to consider

The choice of machine learning algorithms mostly relies on the type of input features. For series features like current signals of a group of neighboring sites, CNN or RNN could be considered. For other input consists of independent features, all kinds of machine learning methods were applied. For RNA modification identification from nanopore direct RNA

sequencing data, the choice of machine learning algorithm is not the bottleneck to the problem, as long as the overall direction is correct. The performances of different algorithms are mostly similar, with the same training datasets and features.

One of the key factors is how much training data we have and the resources of the data. For supervised learning methods, the training sites are usually from previous studies, mainly based on the results from NGS methods. Currently, there are NGS methods for only around 10 types of mRNA modifications and thus the studies of nanopore sequencing are also limited to these modifications. Usually, nanopore direct RNA sequencing utilizes poly T adaptors provided by ONT kits to process poly A RNA and all reads starts from the 3' end and thus it has a bias for 3' end. The coverage of mRNA slowly decreases from 3' to 5' end. Meanwhile, the throughput of nanopore DRS is not high and could only reach 1-3 million reads which is hard to cover human transcriptome deeply. The coverage preference of nanopore and NGS data could be different and only the overlapped regions could be used for training, which means a large decrease in the sites that could be used for model training. For synthesized sequences, the coverage could be plenty, but the diversity of the sequence contents is usually not enough. In many cases, we could achieve hundreds to thousands of modification sites from different sequence contents for model training. If the training data is limited and the feature space is small, then it is suggested to use small models instead of neural networks. It is a misunderstanding that a bigger model trained for longer time and more epochs is always a better model. One possible way to generate more data is by data amplification. The reads could be shuffled and assigned to small groups to make more "artificial" sites for training.

Another key factor is the quality of the ground truth or the true labels. The ground truth is usually from NGS methods. However, only very few modifications have single base resolution

NGS methods, and the antibody-based peak results could hardly be used as ground truth, as the nanopore models required the state of each individual sites. Also, the NGS methods could reveal false positive sites and it doesn't necessarily mean finding more modification sites would be better. They could be false positive sites. Nanopore prediction models trained on low quality true labels could hardly be good models. Thus, the selection of high quality NGS true labels is also important for good model training.

It is also to be considered whether the model is friendly to users. Usually, general users want to have easy-to-install packages, smaller storage requirement and faster speed. Thus, it is also a balance between fair performance and too much processing steps and running time. For example, in nanopore sequencing data, the extraction of current signal strength and dwell time is very time and space consuming step. It worth thinking whether there are strategies to make the steps faster and easier.

1.3.5 Other methods for nanopore direct RNA sequencing modification identification

Beyond machine learning methods, there are also other strategies for mRNA modification identification from nanopore direct RNA sequencing data. For non-machine-learning methods, "compare mode" strategies are widely used.

RNA modifications show different current signals from unmodified bases. Thus, if we could sequence a pair of samples, one with the specific modifications and the other with modifications depleted, then we could compare the signals of the two samples and figure out the different part, which is likely to be modifications. Such comparison could be completed without machine learning models. The strategy was applied to *Arabidopsis thaliana* in 2020 (Parker et al., 2020). However, such strategy requires two copies of samples, as well as reliable knock out

method of a type of modification, which may be difficult when the sample type or species is not common. If the positive and the negative samples are from different genetic background, then there could be difference in the reference sequencing like SNPs so that it will result in more errors when calling modifications. Also, nanopore signals have much random noise, so that individual different signals between two samples don't necessary means something biologically but could just be due to random noise.

One way to avoid using of m^6A writer knock out samples is to use *in vitro* transcribed RNA without any m^6A modifications. It was applied to human samples and the m^6A showed different signal from the unmodified A sites in the *in vitro* transcription (IVT) sample (Rachael E Workman et al., 2019).

ELIGOS used a concept percent Error of Specific Bases (%ESB) and compare the errors like mismatches and indels between nanopore direct RNA seq data and either cDNA, IVT RNA or reference sequences (P. Jenjaroenpun et al., 2021). It used Fisher's exact test on the contingency table and thus evaluate whether a site is modified or not. It used information both from the centerer base and the flanking bases. The method was applied to many modifications like m^6A, m^1A, 5-methoxyuridine (5moU), pseudouridine, m^7G, inosine, hm^5C, f5C and m^5C. The performance on most modifications were fair but did not perform well on m^5C. Overall the AUC values were not very high as the method was not specifically designed for a specific modification and there is not training process of models.

In a recent study, the U-to-C mutation error rates of natural RNA and IVT RNA were compared to quantitatively describe the pseudouridine levels (Tavakoli et al., 2023). Two types of pseudouridine hyper modifications were defined, which was a good attempt.

1.4 Overview of this thesis

In this thesis, we are going to discuss about the development of two machine learning based pipeline for transcriptome wide mapping of pseudouridine and m^6A modifications from nanopore direct RNA sequencing data. Chapter 2 will mainly talk about the development of pseudouridine prediction pipeline NanoPsu and its application on studying pseudouridine changer under interferon treatment. Chapter 3 will talk about the development of a new m^6A prediction model based on nanopore direct RNA sequencing data and the simultaneous m^6A and pseudouridine prediction pipeline NanoSPA. Chapter 4 will talk about applying NanoSPA to investigate coordination of m^6A and pseudouridine in human transcriptome and their overall effect on translation efficiency. Chapter 5 will talk about the development of a single read pseudouridine prediction model and its application in evaluating stoichiometry and multi-site linkages. Chapter 6 will conclude chapter 2 to 5 and discuss about perspectives about potential future directions.

Chapter 2. Interferon inducible pseudouridine modification in human mRNA by quantitative nanopore profiling

Acknowledgement: This chapter and Chapter 5 are derived from an article published in Genome Biology (S. Huang et al., 2021). The authors in that article were: Sihao Huang, Wen Zhang, Christopher D. Katanski, Devin Dersh, Qing Dai, Karen Lolans, Jonathan Yewdell, A. Murat Eren and Tao Pan. Author contributions: S.H. performed all nanopore experiments and analyzed all nanopore data with guidance from A.M.E. W.Z. and C.D.K. designed the validation by RT-qPCR method. W.Z. performed the validation experiment. C.D.K. analyzed Illumina sequencing data, and helped with nanopore data analysis. D.D. and J.Y. designed interferon experiment, D.D. performed interferon experiment. W.Z. and Q.D. built the Illumina sequencing libraries. K.L. isolated total RNA from stools. S.H. and T.P. conceived the project, designed the experiments, and wrote the paper.

We thank Jordan Brown and Dr. Heng-Chi Lee for the C. elegans total RNA. This work was supported by the grant from the NIH (RM1 HG008935 to T.P.). D.D. and J.W.Y. are supported by the Division of Intramural Research, NIAID. A.M.E. and K.L. were supported by an NIH NIDDK grant (RC2 DK122394).

2.1 Introduction

Pseudouridine (Ψ) is the second most abundant mRNA modification in the mammalian transcriptome as measured by quantitative mass spectrometry (X. Li et al., 2015) and may exert many cellular functions. For example, Ψ incorporation in synthetic, transfected reporter mRNA increases translation (Kariko et al., 2008) through decreased activation of the RNA-dependent

protein kinase (PKR) (B. R. Anderson et al., 2010). The innate immune evading property of Ψ (and its methylated derivative N^1-methyl-Ψ) in mRNA is essential to the remarkable immunogenicity of successful COVID-19 mRNA vaccines (Jackson et al., 2020).

Functional exploration and mechanistic investigation of mRNA Ψ modification requires appropriate mapping methods. Illumina sequencing of Ψ in mRNA relies on chemical RNA treatments that induce stop, mutation or deletion signatures in cDNA synthesis (Carlile et al., 2014; Khoddami et al., 2019; X. Li et al., 2015; Schwartz et al., 2014; K. I. Zhou et al., 2018). Many computational methods have been developed to map mRNA Ψ sites (Bi et al., 2020; Chen et al., 2016; Hassan et al., 2022; He et al., 2018; Khan et al., 2020; F. Li et al., 2021; Y. H. Li et al., 2015; K. Liu et al., 2020; Lv et al., 2020; Song, Chen, et al., 2020; Tahir et al., 2019). However, mRNA Ψ mapping is inconsistent among these studies, in part due to the high false positives and negatives generated by the chemical treatments. The read-length limitation of Illumina sequencing also narrows the possibility to examine Ψ usage in mRNA splice isoforms and the linkage of multiple Ψ sites in single molecules.

The emergence of nanopore sequencing enables direct interrogation of RNA modifications (Garalde et al., 2018; Huanle Liu et al., 2019; Rachael E Workman et al., 2019) . Additionally, nanopore sequencing can extend to the full length of the mRNA (Drexler, Choquet, & Churchman, 2020), revealing all modified sites in single RNA isoforms (Daniel A Lorenz, Shashank Sathe, Jaclyn M Einstein, & Gene W Yeo, 2020). Both signal strength and dwell time have been used to identify Ψ (Aaron M Fleming, Nicole J Mathewson, Shereen A Howpay Manage, & Cynthia J Burrows, 2021). Previously, a nanopore direct RNA sequencing method, nanoRMS was developed by Novoa and co-workers that employs characteristic base calling "error" features in the nanopore data for Ψ mapping (Oguzhan Begik et al., 2021). NanoRMS

identified new Ψ sites in mitochondrial rRNA, small nuclear RNA, small nucleolar RNA, and mRNA under normal and stress conditions in yeast and further, predicted stoichiometry via supervised learning. Although nanoRMS prediction of Ψ site incorporation using a threshold for base mismatch frequency is straightforward, it is unclear whether this approach can be applied to the mammalian transcriptomes, which are much larger than yeast, can contain introns, and occur in multiple isoforms. For example, the standard Tombo software for nanopore data analysis is ineffective with spliced reads. In this chapter, we developed a new machine learning based pipeline for transcriptome wide pseudouridine identification from nanopore direct RNA sequencing data. We named the pipeline **Nano**pore investigation of **Pseu**douridine or "NanoPsu".

2.2 Results

2.2.1 Nanopore Ψ prediction model

The core component of the computation pipeline is the machine learning model for pseudouridine prediction. We aim at developing a "single mode" supervised learning model. Thus, training materials from known pseudouridine sites are required. Ribosome RNA is enriched of pseudouridine sites with high stoichiometry and is good material for training. To maximize our ability to obtain nanopore training data from as many distinct Ψ sites as possible, we generated a mixture of rRNAs from human, yeast, *C. elegans*, Drosophila, and from human fecal bacteria (**Fig. 2.1a**). We Illumina sequenced half of the mixture after fragmentation, using the bisulfite reaction (Khoddami et al., 2019) to map rRNA Ψ sites, providing a total of 2,142 Ψ sites (**Table 2.1**). In Illumina sequencing of the bisulfite method, Ψ sites are found by RT deletions which enables identification and quantitative assessment of closely spaced rRNA Ψ

sites; these sites are more difficult to assess using the more commonly used carbodiimide method that identifies Ψ sites by RT stops. To note, we do not achieve stoichiometry information from the Illumina sequencing results and the rRNA pseudouridine sites are not supposed to be all 100% modified. Thus the quantification information is not available for nanopore model training and the nanopore model is a classification model rather than a regression model.

Species	Number of Ψ
Homo sapiens	87
Saccharomyces cerevisiae	43
Drosophila melanogaster	91
Caenorhabditis elegans	56
microbiome	1865

Table 2.1 Number of rRNA Ψ sites identified by Illumina sequencing

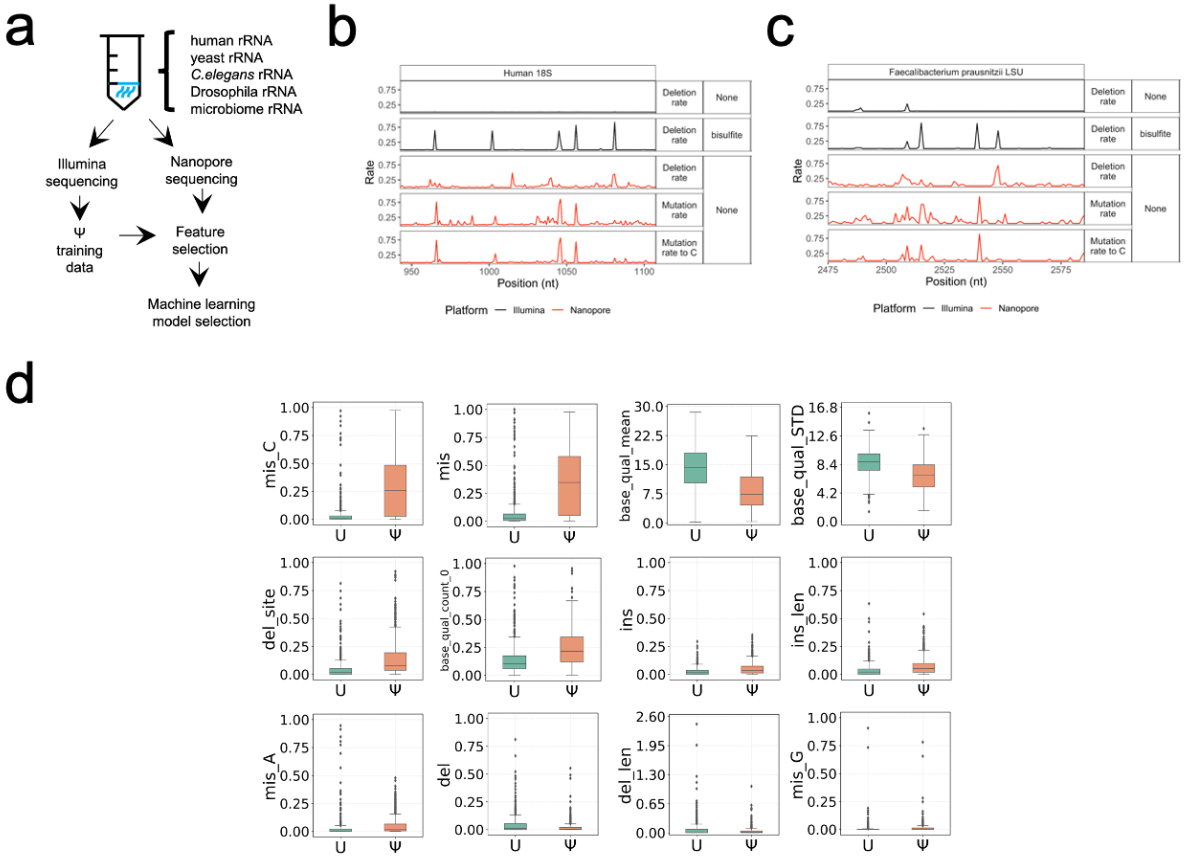

Figure 2.1 Ψ prediction model training using model organisms and microbiome rRNA Ψ modification

(a) Overview of the experiments to generate the Ψ prediction model by nanopore sequencing. (b) Features of a region in human 18S rRNA from Illumina sequencing and nanopore sequencing. (c) Features of a region in a microbial rRNA from Illumina sequencing and nanopore sequencing. (d) Box and Whisker plots with 1.5 times interquartile range of the 12 feature candidates of U and Ψ sites derived from nanopore sequencing. Ins, insertion rate after the base. Ins_len, insertion length mean. Del, deletion rate after the base. Del_len, deletion length mean. Del_site, deleted site ratio (the site is in a deletion). Mis, overall mismatching ratio. Mis_A, mutation to A ratio. Mis_C, mutation to C ratio. Mis_G, mutation to G ratio. Base_qual_mean, average base quality score. Base_qual_STD, base quality score standard deviation. Base_qual_count_0, ratio of bases with a quality score 0 at a site.

Sequencing the second half of the rRNA sample via nanopore direct RNA sequencing, we found that 640 of these Ψ sites passed our filter of 20 read coverage for further analysis (**Table 2.2**). The lower number of Ψ sites in nanopore sequencing was in part derived from the 3' bias of the nanopore sequencing library design where all reads start from the 3' end of the rRNA. These

640 sites were combined with 689 randomly chosen unmodified U sites as the training material (**Table 2.3**). The pseudouridine and unmodified U sites covered 236 of the 256 NN(Ψ/U)NN possible 5mer motifs. These sites are from multiple species and thus the trained model could also be applied to multiple species.

Species	Read count
Homo sapiens	16097
Saccharomyces cerevisiae	19877
Drosophila melanogaster	60062
Caenorhabditis elegans	12135
microbiome	52088

Table 2.2 Number of reads in the model organisms or in microbiome in nanopore sequencing

Species	U	Ψ
Homo sapiens	22	87
Saccharomyces cerevisiae	27	43
Drosophila melanogaster	36	91
Caenorhabditis elegans	18	55
microbiome	586	364
Total	689	640

Table 2.3 Number of U and Ψ sites used in the nanopore Ψ prediction model training determined by Illumina sequencing in the model organisms or in microbiome

High quality features are essential for model performance. NanoRMS (Oguzhan Begik et al., 2021) found that Ψ had negligible effect on nanopore current signals, which means it is hard to directly identify Ψ from the current squiggles like m^6A (Piroon Jenjaroenpun et al., 2021; Daniel A Lorenz et al., 2020; Rachael E Workman et al., 2019). However, distinct features could be found for Ψ identification. For instance, like NanoRMS, we found that apparent mutation to C is a prominent signature for Ψ modification, with apparent deletion also significant for some Ψ sites (**Fig. 2.1b, c**). In total, we examined 12 features of base calling errors for the targeted sites,

including insertion, insertion length, deletion after the site, deletion length after the site, whether the site is deleted or not, total mismatch, mutation to A ratio, mutation to C ratio, mutation to G ratio, base quality score mean, base quality score standard deviation and count of reads with base quality score 0 at the site, and found that Ψ sites tend to have lower base quality mean values and standard deviation (**Fig. 2.1d**). All the 12 features are extracted from the sequencing data independent from sequence content or coverage. The information from flanking sites is not considered in this model, as pseudouridine is not highly enriched in any specific motifs like m^6A in DRACH. The neighboring effects of different types of flanking bases on the targeted sites are mixed in a pool and averaged out with each other. The ternary plot of mutation signatures confirmed Ψ sites having a strong preference to be read as a C but not A or G (**Fig. 2.2a**). The significance of these features was shown in the correlation heatmap of all features and the modification states (**Fig. 2.2b**). We made extremely randomized trees (EXT) models to carry out Ψ probability prediction for each U site. To decide the combination of features included in the model, we added one feature at a time in the order of their correlation strength with the modification label. This revealed that the performance of Ψ calling maximized when all 12 features were included (**Fig. 2.2c**). Using the optimized parameters of our EXT model, its performance was evaluated by the testing set with an area under curve (AUC) of 0.9383 (**Fig. 2.2d**) and the predicted results highly overlap with the true labels (**Fig. 2.2e**). We choose to keep the predicted "pseudouridine probability" as a score to evaluate the likelihood of a site to be pseudouridine instead of calling either pseudouridine or unmodified U sites based on a threshold. In this way, less information is lost, and it is easier to do statistical analysis for the distributions of the scores. It could be imagined that a score of 0.01 and 0.48 definitely means something different but they will both be called unmodified and they will be viewed as the same in the

downstream analysis if we choose to set a threshold. To note, pseudouridine probability is a score and does not represent the modification fraction, although they are positively correlated with each other.

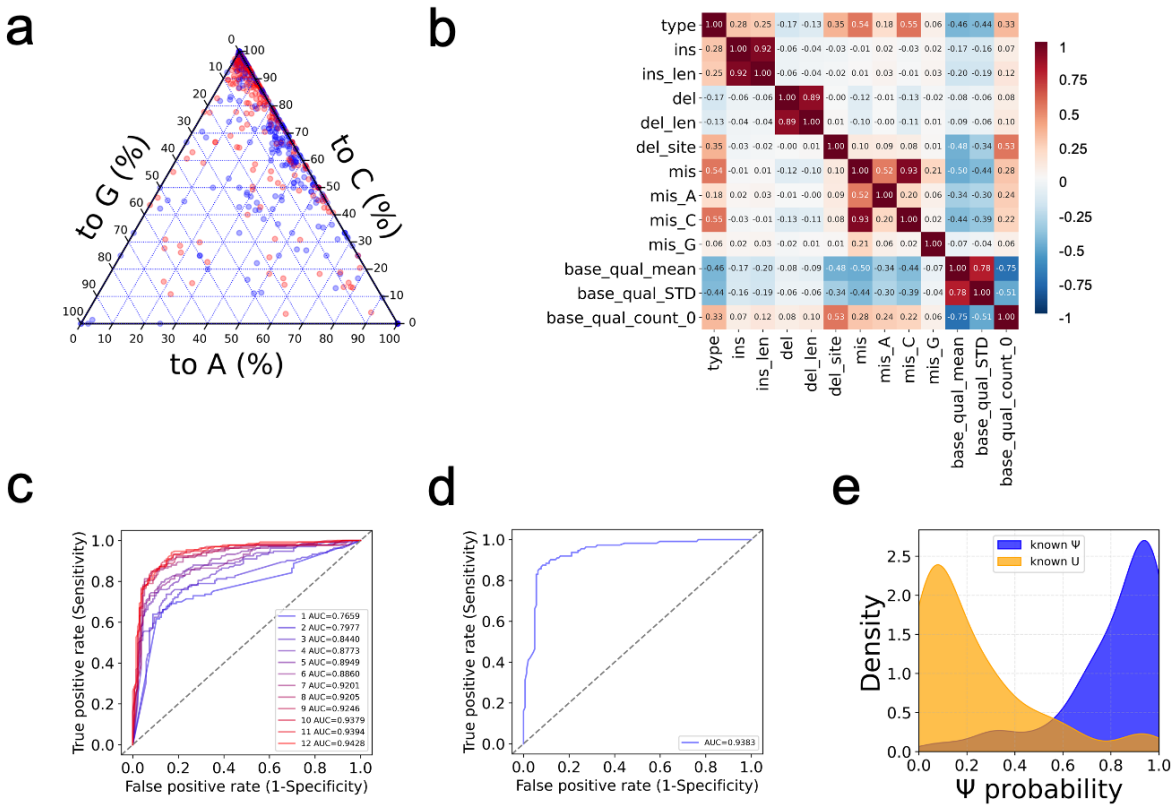

Figure 2.2 Ψ prediction features and model training process
(a) Mutation preference for the Ψ sites in all rRNAs in a ternary plot. Red, Ψ sites in model organisms. Blue, Ψ sites in the microbiome. (b) Correlation matrix of modification state (Ψ=1, U=0) and the 12 feature candidates. The value of correlation coefficient is indicated in each box. Same labels as panel d. Label type, modification state. (c) ROC (receiver operating characteristic) curves of EXT models with different numbers of features included. The number of features and AUC (aera under curve) values of each model are indicated by the legend. The features are added to the model in the order of their correlation with the modification state indicated in panel f. For example, 1 feature means “mis_C”, 2 features means “mis_C” and ”mis”, and so on. (d) ROC curve of the testing set predicted by the optimized EXT model. The AUC value is indicated in the graph. (e) Density plot of predicted Ψ modification probabilities of U and Ψ sites of bulk read prediction in the testing set.

2.2.2 Apply NanoPsu to study effect of interferon treatment

Interferons (IFNs), cytokines produced by nearly all cell types during viral and other microbial infections, play crucial roles regulating immune response (Lee & Ashkar, 2018). mRNA vaccines incorporate Ψ or $m^1Ψ$ to evade host cell foreign RNA sensing and enhance mRNA translation. However, it is unclear whether endogenous mRNAs also use the same strategy through Ψ modification. IFNs can induce the expression of more than a thousand interferon stimulated gene (ISG) transcripts. ISGs includes protein kinase R (PKR) which phosphorylates eIF2α to reduce global translation. It is well established that Ψ-modified reporter mRNA activates PKR much less than the same unmodified mRNA, and is translated at much higher levels (Bart R Anderson et al., 2010). We therefore hypothesize that ISG transcripts may have elevated levels of Ψ modification to enhance translation in the presence of PKR.

We tested this hypothesis by treating cells with either IFN-γ or IFN-β followed by nanopore direct RNA sequencing (**Fig. 2.3a**). IFN treatments worked well as determined by upregulation of surface MHC class I (**Fig. 2.3b**). The mRNA expression levels of the biological replicates were highly correlated (**Fig. 2.3c**). For improved coverage we combined the nanopore data from the biological replicates for downstream analysis (**Table 2.4**). We found strongly up-regulated mRNA transcripts upon IFN treatment that belong to the ISG genes with the expected gene ontology of interferon signaling pathway and viral defense (**Fig. 2.3d, e**). These results indicate the feasibility of using nanopore sequencing to study the interferon response transcriptome.

Sample name	Reads aligned A	Reads aligned B	Total reads	After down sampling
Untreated	1507124	939300	2446424	2446424
IFN β^+	1861242	1581181	3442423	2445584
IFN γ^+	855278	2195506	3050784	2444761

Table 2.4 Read count of each run and read count of the combined samples before and after down sampling

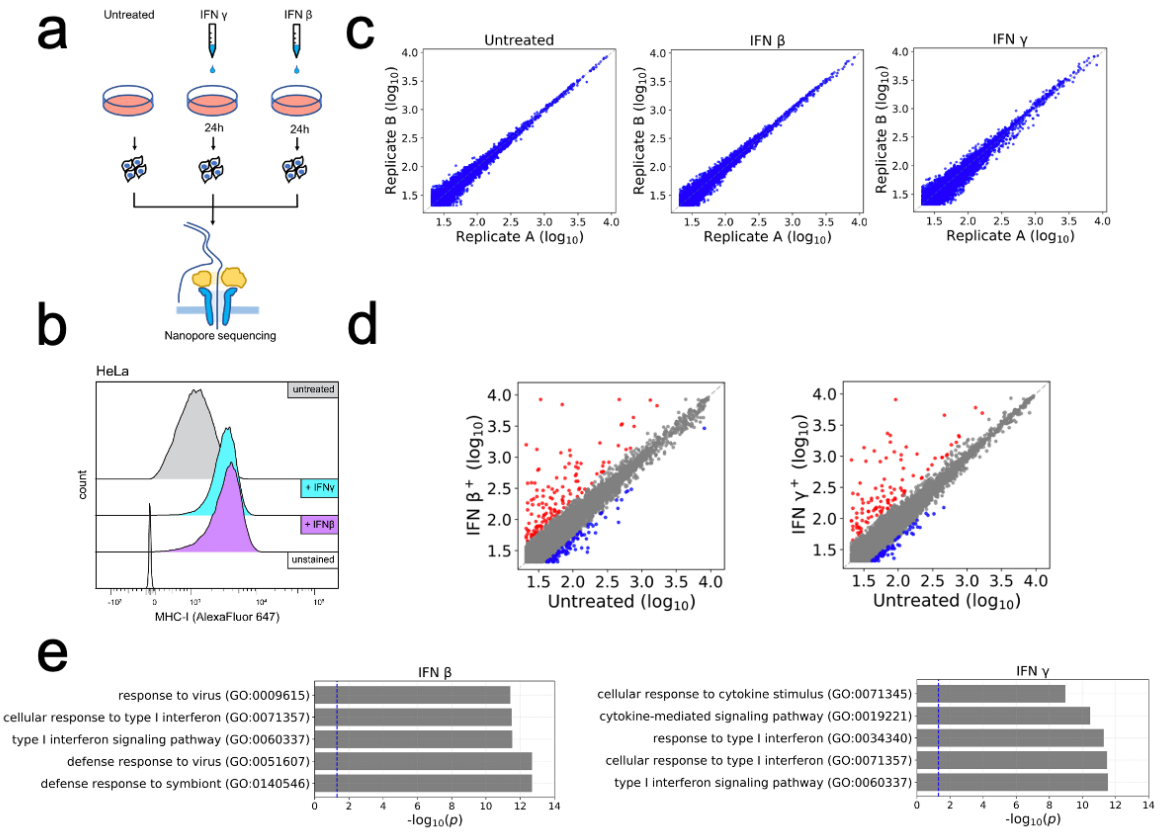

Figure 2.3 Interferon treatment experiment overview
(a) Scheme of the experiment. HeLa cells were treated with either IFN β or IFN γ for 24h or not treated. RNA was extracted and used for nanopore sequencing. (b) Flow cytometry results of the three samples showing the expression of class I MHC molecules upon interferon treatment. (c) Expression levels of biological replicates of each sample. Correlation coefficient for untreated, IFN β and IFN γ samples are 0.9832, 0.9840 and 0.9711 (Pearson's r) respectively. (d) Log10 expression levels of genes in untreated sample and IFN β treated (left) or IFN γ treated (right) samples. Expression level is calculated as the peak height of the piled reads. Red, genes with an increase of > 2 fold in expression. Blue, genes with a decrease of > 2 fold in expression. (e) GO terms of genes with >10 fold increase in expression levels in IFN β (top) or IFN γ (bottom) sample compared with untreated sample.

We used NanoPsu to predict pseudouridine probabilities of all sites in the IFN treated and untreated samples. In total, ~2.6 million U sites were analyzed in each transcriptome (**Fig. 2.4a**). We found a "RAΨU" motif and the previous revealed (Modi Safra, Nir, Farouq, Slutskin, & Schwartz, 2017) "GUΨC" motif among top Ψ sites in the untreated sample (**Fig. 2.4b**). The Ψ sites belonging to "median" or higher groups in the previous study (Modi Safra et al., 2017)

showed significantly higher predicted Ψ probabilities than other U sites in the untreated sample (**Fig. 2.4c**), indicating that our method has consensus with the previous published method on pseudouridine prediction. For the 500 sites with the highest pseudouridine probability, the three samples shared some but also had distinct sites (**Fig. 2.4d**). However, IFN treated samples had a wider range of GO terms than the untreated sample (**Fig. 2.4e**), suggesting that Ψ modification becomes more widespread to transcripts belonging to more diverse cellular processes.

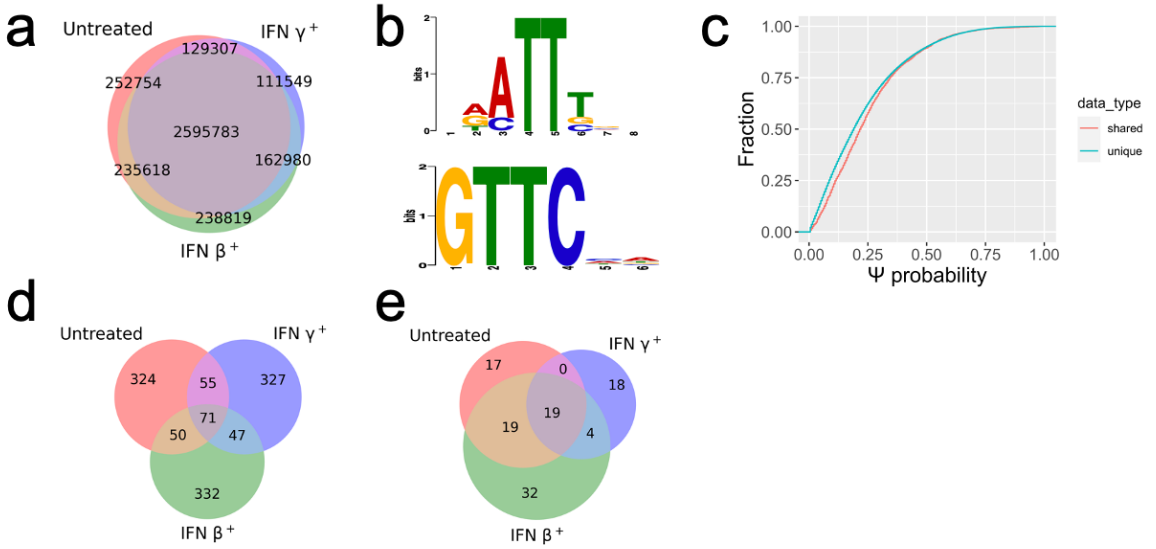

Figure 2.4 Ψ modification overview in samples
(a) Venn diagram of U sites with over 20 read coverage in three samples used for Ψ analysis. (b) Sequence logo of dominant motifs of 10000 sites with highest Ψ modification probabilities (top panel) and sites within GUΨC motif (lower panel) in untreated sample. (c) Cumulative distribution curves of Ψ probabilities of sites predicted as at least "median" level Ψ in Safra et al., 2017 ("Ψ") and other U sites ("U"), p <2.2e-16. (d) Venn diagram of the 500 U sites with the highest Ψ modification probabilities in each sample. (e) Venn diagram of the GO terms of the genes containing the 500 U sites with the highest Ψ probabilities in each sample.

Going beyond the top 500 probable Ψ sites, globally the upregulated gene transcripts had higher average pseudouridine probabilities for IFN treated samples over untreated samples (**Fig. 2.5a**). A higher magnitude of increase in expression level has the preference for a higher level of pseudouridine probability increase (**Fig. 2.5b, c**). Increased average Ψ probability in a mRNA

transcript could be attributed to increased number of Ψ sites and/or increased modification fraction of modified sites. The top 50 genes with highest increase in Ψ probability were related to the interferon pathway and anti-viral response (**Fig. 2.5d**), they included 88.5% of all genes with >10-fold increase and 60.9% of all genes with >5-fold increase in mRNA expression (**Fig. 2.5e**). These results are consistent with increased Ψ modification in the transcriptome upon interferon treatment enhancing ISG function.

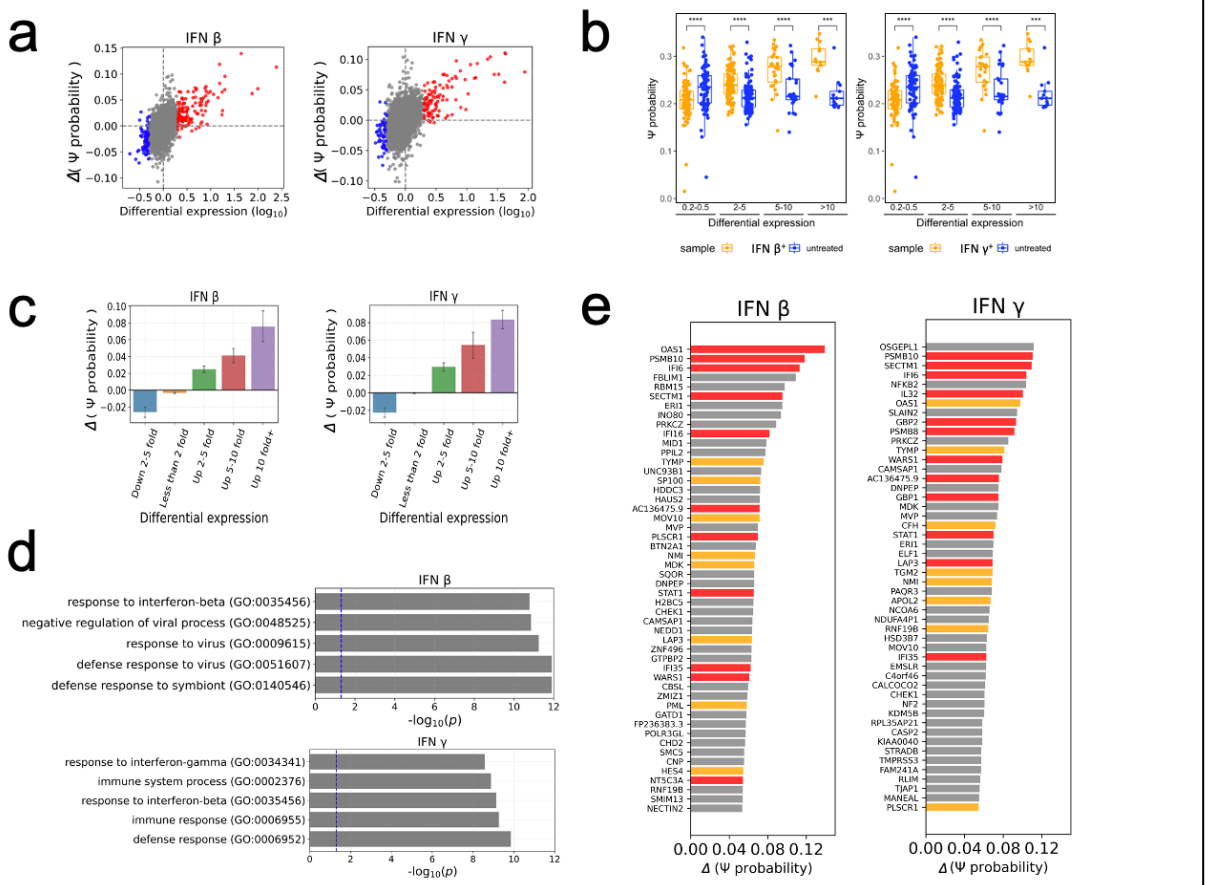

Figure 2.5 Interferon treatment elicits more Ψ modification in mRNA
(a) Scatter plot showing the mean modification probability change versus log10 expression fold change of each gene between untreated and IFN β treated (left) or IFN γ treated (right) sample. Red, genes with an increase of > 2 fold in expression. Blue, genes with a decrease of > 2 fold in expression. (b) Mean Ψ modification probability of genes assigned to groups based on expression fold change between untreated and IFN β treated (left) or IFN γ treated (right) samples. ***, p<10-3, ****, p<10-4. (c) Difference of mean Ψ modification probabilities of genes between IFN β (left) or IFN γ (right) sample and untreated sample grouped by expression fold change. (d) GO analysis of the 50 genes with highest mean Ψ probability change between untreated and IFN β treated (top) or IFN γ treated (bottom) samples. Blue vertical line indicates p=0.05. (e) Mean Ψ probability change of the highest 50 genes between untreated and IFN β treated (left) or IFN γ treated (right) samples. Genes with a significant increase in expression levels are marked in red (>10 fold) or orange (5-10 fold).

2.2.3 Validation of increased Ψ by RT-qPCR

To investigate if the results above predicted by NanoPsu is convincing, we use a RT-qPCR method to validate the increased Ψ level in the ISG transcripts. Our method takes advantage of the standard Ψ detection method using N-cyclohexyl-N′-(2-morpholinoethyl) carbodiimide (CMC). The Ψ-CMC adduct introduces a RT stop in cDNA synthesis which reduces the amount of cDNA product compared to the control reaction without CMC. The differential amount of the cDNA product can then be precisely measured using real-time qPCR (**Fig. 2.6a**). We first showed that the actin mRNA did not change its abundance nor its Ψ level, making it an appropriate internal control for comparing the Ψ levels of the ISG transcripts (**Fig. 2.6b, c**, left panels). Ψ level increase in the ISG15 mRNA upon interferon treatment was validated upon normalization of its expression level and to the actin mRNA within the same sample (**Fig. 2.6b, c**, right panels). The RT-qPCR results validate the NanoPsu predicted increase in average pseudouridine probabilities in ISG transcripts.

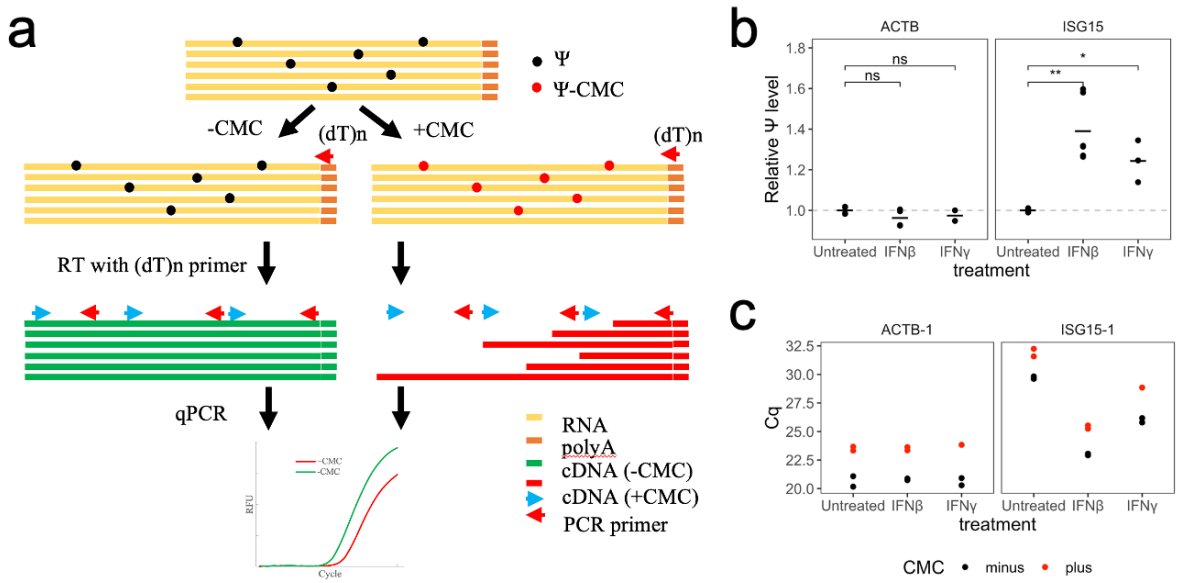

Figure 2.6 Validation of Ψ modification increase in ISG transcripts
(a) Scheme of CMC-mediated RT-qPCR (CRP) validation of Ψ level in specific mRNA transcripts. Total RNA is first treated with and without CMC. Ψs in RNA form Ψ-CMC adducts which cause RT stops. cDNA synthesis uses T22VN primer targeting the 3' polyA tail. Using –CMC RNA as RT template generates longer cDNA products than using +CMC treated RNA. The relative cDNA amount difference is quantified using qPCR with multiple sets of PCR primers targeting different regions of the cDNA. Cq or CT values are obtained for data analysis and comparative analysis of Ψ levels. (b) Relative Ψ level of mRNA transcript of ACTB (left panel, data from set 1 and set 2 primers) and ISG15 (right panel, data from set 1, set 2, and set 3 primers) in the untreated and interferon treated samples measured by RT-qPCR. * $p < 0.05$; ** $p < 0.01$. (c) The Cq values from the CRP validation experiment. Data points are from the biological replicate of the untreated and interferon treated samples. ActB transcript does not show a difference upon CMC treatment and in all samples; this transcript is used as the control for other mRNA transcripts in the same sample. ISG15 is an interferon stimulated gene transcript. "-1" refers to the specific qPCR primer set (see Methods).

2.3 Discussion

In summary, we generated a supervised-learning-based protocol to predict Ψ modification in the human transcriptome and analyzed Ψ on single reads which allows for the evaluation of stoichiometry and linkage between distal Ψ sites in the same mRNA molecule. Human genome contains 13 confirmed and putative Ψ installation enzymes (Erin K Borchardt, Nicole M

Martinez, & Wendy V Gilbert, 2020), suggesting that Ψ installation is a highly robust and dynamic process in human cells. How these enzymes coordinate or antagonize their activities remains to be determined. We found a biological response of Ψ modification change in endogenous mRNA upon IFN treatment which is consistent with Ψ playing a role in IFN signaling pathway and viral defense.

There are also limitations of the Ψ model. First of all, although the performance is satisfying for model training, validation and testing, there is still room for improvement for more practical usage in biological samples. The model was trained on a balanced dataset, while in the real world samples, the ratio of Ψ is much smaller that unmodified U, which means good prediction with low false positive rate requires extremely high model accuracy.

Second, the training material is not perfect. The reported Ψ sites in the previous NGS based papers didn't overlap a lot, while supervised learning models rely heavily on high quality labeled date. Also, Ψ sites distribution may be different for different samples so the sites modified in the previous papers may not be modified in our own samples, even if they are from the same cell line. Instead, we use the same sample for known Ψ sites calling and nanopore model training, which make sure that the Ψ modification state of each site is the same for our Illumina data and nanopore data. However, this relies on the high accuracy of called Illumina pseudouridine sites by BID-seq, which could never reach 100%. Thus, the true labels of the training data are not 100% correct. This also reflect the complexity of machine learning on biological problems or natural science problems. Unlike tasks like distinguishing cat pictures from dog pictures where we can 100% correctly label all the training data, it's hard for most biological questions to have 100% correct materials for training and thus the models may work well during training but not in practical use.

Third, the information from nanopore sequencing data is not fully used and understood. Ψ does not have a strong preference for any specific motif, probably due to the existence of thirteen synthases which may prefer different motif patterns. In this condition, although we know that the signals of flanking sites are also affected by the modification site, it's hard to use the information from flanking sites, as they are from all possible types of base combinations. One possible solution is to train a model for each motif, but unlike m^6A, pseudouridine is not enriched in any motif and its ratio in any single motif is very low. Also it's hard to find enough known pseudouridine sites for each motif, if we divide all known sites by 256 for all possible 5mer motifs.

Using pseudouridine sited from rRNA makes it easy to achieve highly modified high density pseudouridine sites, which facilitate the model training. However, there is also risks using rRNA as training materials. First, rRNA is not only heavily modified by pseudouridine but also by other modifications like 2'-methyl. The signals of other modifications could interfere with the signals from pseudouridine and thus result in bias in the models. Also, the sequence content, modification density and average modification stoichiometry of rRNA is different from mRNA, which may result in systematic bias for mRNA applications. However, this problem is not for rRNA training materials solely and those models trained on synthesized RNA sequences will also have the same problem. Probably only making a "compare mode" model with WT and pseudouridine depleted cell samples will not have such problem but as there are too many pseudouridine synthases in cells it will be extremely difficult to generate pseudouridine depleted samples.

2.4 Methods

2.4.1 Stool sample collection and total RNA extraction

Stool specimens were self-collected by 1 female volunteer using a commercial "toilet hat" stool specimen collection kit (Fisherbrand Commode Specimen Collection System; Thermo Fisher Scientific, 02-544-208). Specimens were immediately transported to the laboratory (<1-hr) and thoroughly homogenized. 100 mg stool was transferred into a cryovial using a sterile spatula and 700 µl RNAlater Stabilization solution was added. Specimens were stored at -80 °C until extraction.

RNA later was first removed from stool sample by centrifugation at 17,200 rcf for 10 minutes at 4 °C. Pelleted material was lysed in 400 µL of 0.3M NaOAc/HOAc,10mM EDTA, pH 4.8 with an equal volume of acetate-saturated phenol:chloroform pH 4.5 (Invitrogen, AM9722). After addition of 1.0 mm glass lysing beads (Bio-Spec Products, 11079110) in a 1:1 ratio (bead:sample weight), samples were placed in a reciprocating bead beater (Mini-Beadbeater-16, Bio-Spec Products) for two 1-min intervals on maximum intensity. Samples were centrifuged at 17,200 rcf for 15 minutes at 4 °C before re-extraction and isopropanol precipitation of total RNA. Pellets were washed with 75% ethanol before resuspension in an acid-buffered elution buffer (10mM NaOAc, 1mM EDTA, pH 4.8).

2.4.2 rRNA mixture sample preparation

A mixture of human HEK293T, yeast BY4741 strain, Drosophila S2 cells, and C. elegans whole animal and stool microbiome total RNA was made by mixing 1 µg RNA from each model organism sample and 8 µg total RNA from a stool microbiome sample. ZYMO RNA Clean & Concentrator-5 (R1013) kit was used on this mixture to remove all small RNAs <200nt. The

final sample was eluted with 20 µl RNase-Free H2O. The mixture was split into two halves. One half was used for Illumina sequencing (see below). For nanopore sequencing, the other half was polyadenylated by yeast Poly(A) Polymerase (ThermoFisher 74225Z25KU) by incubation with 0.48 mM ATP, 20 U/µL Poly(A) Polymerase and 1x Poly(A) Polymerase Reaction Buffer at 37 °C for 15 min. The product was size selected using ZYMO RNA Clean & Concentrator-5 (R1013) kit and RNA molecules >200nt were retained. The sample was eluted with 20 µL RNase-Free H2O. Then ~500 ng of this rRNA mixture was used for nanopore direct RNA seq library preparation and nanopore direct RNA sequencing described below.

2.4.3 rRNA mixture Illumina sequencing and mapping

For Illumina sequencing, bisulfite treatment was performed as described previously(Khoddami et al., 2019). Ψ modification was identified through the deletion at the Ψ site in the sequencing data. Raw reads were demultiplexed via a 4nt barcode on read 2 using je suite (Girardot, Scholtalbers, Sauer, Su, & Furlong, 2016) with the following parameters: je demultiplex F1=#read1 F2=$read2 BF=$barcode_key BPOS=BOTH BM=READ_2 LEN=6:4 O=$output. Only read 2 from paired-end reads were mapped with bowtie2 (version: bowtie2-2.3.3.1-linux-x86_64) (Langmead & Salzberg, 2012) using the following parameters: bowtie2 -x $reference -U $read2 -S $ouput -q -p 10 --local --no-unal. Reads were mapped to either a set of rRNA from model organisms, or a set of bacterial rRNA reads: rfam family RF02541 (bacterial large subunit) and RF00177 (bacterial small subunit). SAM files from bacterial rRNAs were processed with a custom python script to count the total number of reads mapping to each sequence. Only sequences with >1000 reads were processed further. Model organism rRNA sequences from human (NCBI: NR_003286.4, NR_003287.4), yeast (RNACentral:

URS00005F2C2D_559292, URS000061F377_4932), *C. elegans* (RNACentral: URS00005A42AA_6239, URS00008C9AB9_6239), and Drosophila (RNACentral: URS000030AF9A_7227, URS000008C6A9_7227) to form a reference genome for bowtie mapping. Bowtie2 output "sam" files were converted to sorted bam files with samtools (Heng Li et al., 2009). IGV was used to calculate deletion rates with the following parameters: igvtools (Robinson et al., 2011) count -z 5 -w 1 -e 250 --bases $input $output $reference. Custom python scripts were used to reformat the "wig" file.

2.4.4 Nanopore direct RNA seq library preparation and sequencing

The library preparation followed the protocol of Direct RNA Sequencing Kit (SQK-RNA002) provided by Oxford Nanopore Technology. Briefly, ~500 ng of Poly(A)+ RNA sample was used for each run. Each single run contained one biological replicate of one sample. The RT Adaptor (RTA) was ligated to the 3' end of Poly(A)+ RNA by T4 DNA ligase (NEB M0202S) and then reverse transcribed by SuperScript III Reverse Transcriptase (ThermoFisher 12574018). The RNA was purified by 1.8x RNAClean XP beads (72 μL) (Beckman Coulter A63987) and then the RNA Adaptor (RMX) was ligated to the 3' end of Poly(A)+ RNA using T4 DNA ligase (NEB M0202S) and then the RNA was purified with 1x RNAClean XP beads (40 μL). The sample was eluted with 21 μl Elution Buffer. Then the sample was loaded onto a R9.4.1 flow cell (FLO-MIN106D) in a MinION sequencer. Each flow cell was sequenced for 72 hours.

2.4.5 Nanopore data pre-processing

All raw fast5 files generated during sequencing were uploaded to Midway2 cluster for the following steps. Reads were base called by guppy base caller (version 3.2.2+9fe0a78) with

min_qscore 7. The reads were aligned to by minimap2 (version 2.18-r1015) (H. Li, 2018) with parameters -ax splice -uf -k14. The rRNA mixture reads are aligned to the same reference as the rRNA Illumina seq data described above. The human mRNA reads are aligned to the hg38 human genome reference (GRCh38.p13). The mapped reads were piled up to the reference chromosomes by samtools (v1.11). The "error" features were extracted from the mpileup files by customized python scripts (https://github.com/sihaohuanguc/Nanopore_psU).

2.4.6 Model training

For nanopore seq data of rRNA, all sites mapped to "T" in the reference with >20 coverage made up the data pool. 640 Ψ sites revealed by Illumina sequencing and 689 randomly selected U sites from the data pool made up the model training dataset. The dataset was divided into 60% training set, 20% validation set and 20% testing set. The Ψ modification prediction models were generated by training set and validated with the validation set by extremely randomized trees (EXT) models with 1-12 features and customized parameters. Then the models were applied to predict Ψ modification probabilities of the testing set and evaluated by AUC of ROC (Receiver Operating Characteristic) curves derived from the predicted probabilities of the testing set. The final model used EXT algorithm (n_estimators=200, criterion="gini", max_depth=None, min_samples_split=2) with 12 features.

2.4.7 HeLa cell culture and interferon treatment

HeLa cells (ATCC, authenticated and tested for mycoplasma contamination) were cultured in the presence of 500 U/mL human interferon gamma (IFN γ, Peprotech), 500 U/mL human interferon beta (IFN β, Peprotech), or left untreated, with biological duplicates for each.

Cells were incubated for 24 hours, and an aliquot of each was processed for flow cytometry. Cells were washed into a flow cytometry staining buffer (FBS-containing RPMI and Hanks' Balanced Salt Solution) containing the anti-pan-MHC-I antibody W6/32 (BioXcell) conjugated with AlexaFluor 647 (Invitrogen). Cells were then washed 3x and analyzed by a Fortessa X-20 (BD Biosciences) to determine upregulation of MHC class I. The rest of the cells were used for RNA extraction via the RNeasy Mini kit (Qiagen) following the manufacturer's protocol. RNA was eluted in pure water and quantified by Nanodrop (Thermo). PolyA+ RNA from 50 μg total RNA of each sample was extracted by Promega PolyATtract® mRNA Isolation Systems Z5310. Each sample was eluted with 15 μL H2O.

2.4.8 Prediction of Ψ in HeLa samples

The raw data of two replicates for the untreated, IFN γ treated and IFN β treated samples were merged after aligned to the hg38 human genome reference (GRCh38.p13). The merged samples were down sampled so that they have almost the same number of reads and are directly comparable. The Ψ modification probabilities of all sites mapped to "T" in the reference with over 20 coverage were evaluated by the EXT model generated with the rRNA mixture sample. The coverage independence of Ψ probability was examined by down sampling all sites of the samples to similar coverages (expectation = 30) using different random seeds. We found that the change in mean Ψ probability of the transcripts maintained the same after down sampling. The coverage completeness of the transcripts was checked by counting the U sites predicted in the samples (Quinlan & Hall, 2010). For the untreated sample, the U sites within 5'UTR, CDS and 3'UTR represented 2.43%, 42.84% and 54.73% of all U sites, respectively. The gene information was provided by the comprehensive gene annotation file (gencode.v37.annotation.gff3) in the

GENCODE database (https://www.gencodegenes.org) (Adam Frankish et al., 2021). The gene ontology (GO) analysis was performed using the Gene Ontology Resource (http://geneontology.org) (Ashburner et al., 2000; "The Gene Ontology resource: enriching a GOld mine," 2021). The sequence logo plots were generated by MEME (https://meme-suite.org/meme/tools/meme) (Bailey, Johnson, Grant, & Noble, 2015).

2.4.9 CMC-mediated RT-qPCR (CRP) validation of Ψ level in mRNA transcripts

2.4.9.1 Primer design

qPCR primers were designed using NCBI Primer-BLAST tool (https://www.ncbi.nlm.nih.gov/tools/primer-blast/). 2-3 sets of primers were selected to cover the 3' end, middle and 5' end region of the whole transcript. qPCR was performed with TaqMan style fluorescent probes. Probes for each PCR primer pair were designed using IDT PrimerQuest tool (https://www.idtdna.com/pages/tools/primerquest) and examined using NCBI nucleotide BLAST. Primers and probes were purchased from IDT. Actin (NM_001101.5) and ISG15 (NM_005101.4) transcripts were selected for Ψ validation. Below is the list of the sequences of qPCR primers and probes.

ISG15 primer1-Forward: GTGGACAAATGCGACGAACC

ISG15 primer1-Reverse: ATTTCCGGCCCTTGATCCTG

ISG15 probe1: 5'- /56-FAM/TCC TGG TGA /ZEN/GGA ATA ACA AGG GCC /3IABkFQ/ -3'

ISG15 primer2-Forward: GCGCAGATCACCCAGAAGAT

ISG15 primer2-Reverse: GTTCGTCGCATTTGTCCACC

ISG15 probe2: 5'- /56-FAM/TTC CAG CAG /ZEN/CGT CTG GCT GT /3IABkFQ/ -3'

ISG15 primer3-Forward: CAGCGAACTCATCTTTGCCAG

ISG15 primer3-Reverse: GACACCTGGAATTCGTTGCC

ISG15 probe3: 5'- /56-FAM/TGG GAC CTG /ZEN/ACG GTG AAG ATG C/3IABkFQ/ -3'

ACTB primer1-Forward: ACAGGAAGTCCCTTGCCATC

ACTB primer1-Reverse: CAGTGTACAGGTAAGCCCTGG

ACTB probe1: 5'- /56-FAM/ACA CGA AAG /ZEN/CAATGCTATCACCTCCC/31ABkFQ/ -3'

ACTB primer2-Forward: AGATGTGGATCAGCAAGCAGG

ACTB primer2-Reverse: GGGGGATGCTCGCTCCA

ACTB probe2: 5'- /56-FAM/TCG TCC ACC /ZEN/GCA AAT GCT TCT AGG /31ABkFQ/ -3'

2.4.9.2 CMC-mediated RT-qPCR (CRP) experiment

CMC [*N*-cyclohexyl-*N'*-(2-morpholinoethyl) carbodiimide] treatment was done as previously described (W. Zhang, Eckwahl, Zhou, & Pan, 2019). 1.5 µg of untreated, IFNβ treated, and IFNγ treated total RNA in 12 µl was mixed with 24 µl TEU buffer (50 mM Tris-HCl (pH 8.3), 4 mM EDTA, 7 M urea) in microcentrifuge tubes. 4 µl freshly made 1 M CMC (Sigma, C1011) in TEU buffer or 4 µl TEU buffer was added to each sample for +CMC or -CMC treatment, respectively. The sample mixture in 40 µl 0.7× TEU was incubated at 37 ºC for 1 hour. The mixture was diluted to 200 µl with 160 µl of 50 mM KOAc (pH 7), 200 mMKCl. 1 µl 5 µg/µl glycogen and 550 µl ethanol were added to the mixture to precipitate RNA at -80 ºC for >2 hours. The mixture was then centrifuged at highest speed (17000× g) for 30 min. The RNA precipitate was mixed with 500 µl 75% ethanol and kept at -80 ºC for >2 hours followed by centrifugation at 17000× g for 30 min. The washing step was repeated once. The RNA precipitate was mixed with 50 µl of 50 mM Na_2CO_3, 2 mM EDTA (pH 10.4), and incubated at 37 ºC for 6 hours to remove CMC-U/CMC-G adducts. The RNA was purified using Zymo RNA Clean and

Concentrator column (Zymo, R1014) with in-column DNase treatment by following the manufacturer's manual. The RNA was eluted in 11 µl sterile H_2O. The concentration of the ±CMC treated RNA was measured using Nanodrop and equal amount (~300 ng) of total RNA was used for RT-qPCR experiment.

Eleven µl of 300 ng ±CMC treated total RNA from Untreated/IFNβ/IFNγ samples were mixed with 1 µl 50 µM 5'T22VN (V=A,C,G, N=A,C,G,T) primer (IDT) and 1 µl 10 mM dNTP mix. The mixtures were incubated at 65 ºC in thermal cycler for 5 mins followed by incubation at room temperature for 3 min. The PCR tubes were kept on ice until the addition of the SuperScript IV RT mix. 7 µl RT mix was prepared for each sample by combining 4 µl 5× SSIV Buffer, 1 µl 100 mM DTT, 1 µl RNaseOUT RNase inhibitor, and 1 µl SSIV reverse transcriptase. 7 µl RT mix was added to each PCR tube. The tubes were incubated at 55 ºC in thermal cycler for 1.5 hours. The PCR tubes were then incubated at 80 ºC for 10 min followed by incubation on ice immediately to deactivate RT. 45 µl sterile H2O was added to each tube to dilute the RT mixture to 65 µl, and 2 µl was used for qPCR reaction.

qPCR reaction was performed in 10 µl consisting of 5 µl 2× PrimeTime Gene Expression Master Mix (IDT, 1055772), 2 µl RT mix, and 3 µl primer and probe mix. 3 µl primer and probe mix (1.5 µM each PCR primer and 0.6 µM probe) were first added into each well of 384-well plate or 96-well plate. RT mix of each sample and 2× PrimeTime Gene Expression Master Mix were mixed at 2:5 ratio to make master mix based on the number of qPCR reactions for each sample. 7 µl of the template and PrimeTime master mix were then added to each well. The plate was spun on a swing bucket plate centrifuge at 3000 RPM for 2 min. qPCR reaction was performed on Bio-Rad CFX384 or CFX96 qPCR machine for 40 cycles. Cq/CT values was obtained for follow-up data analysis.

Relative Ψ levels for ISG15 transcript was calculated using ACTB-1 as internal reference. First we obtained ΔCq(-) = Cq(ISG15,-CMC) - Cq(ACTB,-CMC), and ΔCq(+) = Cq(ISG15,+CMC) - Cq(ACTB,+CMC); then we obtained ΔΔCq(ISG15) = ΔCq(+) - ΔCq(-). The relative Ψ level is represented as 2^ ΔΔCq(ISG15).

Chapter 3. Nanopore sequencing protocol for simultaneous transcriptome wide m^{6}A and pseudouridine profiling

Acknowledgement: Chapters 3 and 4 are derived from an article published in Nature Biotechnology (S. Huang, Wylder, & Pan, 2024). The authors in that article were: Sihao Huang, Adam C. Wylder and Tao Pan. Author contributions: S.H. performed all computational work including NanoSPA pipeline development and sequencing data analysis. A.C.W. performed all experimental work including si-knockdowns, polysome profiling, and nanopore sequencing. S.H. and T.P. conceived the project, S.H., A.C.W. and T.P. designed the experiments and wrote the paper.

We thank Drs. Lisheng Zhang and Chuan He for providing analyzed m^{6}A-SAC-seq data prior to publication and David Pan for contribution to coding. This work was supported by NIH (RM1 HG008935 to T.P.).

3.1 Introduction

N^6-methyladenosine (m^{6}A) and pseudouridine (Ψ) are the top two most abundant internal mammalian mRNA modifications according to the total m^{6}A or Ψ content in total mRNA measured by mass spectrometry (I. A. Roundtree et al., 2017). m^{6}A is the most extensively studied mRNA modification; it participates in many cellular processes including mRNA stability, splicing, export, localization, and translation (Frye, Harada, Behm, & He, 2018; I. A. Roundtree et al., 2017). The best known Ψ function is innate immune avoidance when in delivered mRNAs, as shown in the successful COVID-19 mRNA vaccines (Jackson et al., 2020); Ψ has also been

shown to affect splicing and translation (B. R. Anderson et al., 2010; Eyler et al., 2019; Kariko et al., 2008; Martinez et al., 2022).

A major gap in the biological studies of m^6A and Ψ is how they enhance or antagonize each other in the same mRNA transcript. The investigation of coordinated m^6A and Ψ function requires mapping methods that can simultaneously report m^6A and Ψ in the same sequencing library. So far, Illumina sequencing of m^6A and Ψ has always been performed separately and independently. Nanopore sequencing has also been employed to map either m^6A or Ψ in numerous studies and pipelines (O. Begik et al., 2021; A. M. Fleming, N. J. Mathewson, S. A. Howpay Manage, & C. J. Burrows, 2021; Gao et al., 2021; Hassan et al., 2022; Hendra et al., 2022; S. Huang et al., 2021; P. Jenjaroenpun et al., 2021; Leger et al., 2021; F. Li et al., 2021; H. Liu et al., 2019; H. Liu, Begik, & Novoa, 2021; R. Liu et al., 2022; D. A. Lorenz et al., 2020; Parker et al., 2020; Piechotta, Naarmann-de Vries, Wang, Altmuller, & Dieterich, 2022; Pratanwanich et al., 2021; Price et al., 2020; Qin et al., 2022; Ramasamy, Mishra, et al., 2022; Ramasamy, Sahayasheela, et al., 2022; Stoiber et al., 2016; Tavakoli et al., 2023; R. E. Workman et al., 2019; F. Yu et al., 2023; Y. Zhang, Huang, Wei, & Chen, 2022), but these studies also considered m^6A or Ψ separately. Therefore, no prior studies investigated the potential crosstalk between m^6A and Ψ in the mRNA transcriptome. In this chapter, we develop a computation pipeline named **Nano**pore **S**imultaneous investigation for **P**seudouridine and m^6**A** (NanoSPA) that analyzes m^6A and Ψ modifications in the same nanopore direct RNA sequencing dataset. We apply NanoSPA to both the human transcriptome with or without knocking down the m^6A writer METTL3 or one of the thirteen Ψ writers TRUB1 to reveal their co-dependence, and to polysome associated mRNA samples to reveal their effects and co-dependence on translation.

3.2 Results

3.2.1 A fused workflow for simultaneous m^{6}A and pseudouridine identification

To investigate m^{6}A and Ψ at the same time, we aimed at designing a fused workflow (**Fig. 3.1a**) for base calling, alignment and feature extraction for these two modifications together, which saves processing steps and storage for different types of features from two independent protocols. It is obvious that for the same nanopore data, the protocol for base calling and alignment could be the same for different models. The feasibility of a fused workflow relies on the usage of same feature space for m^{6}A and pseudouridine prediction. For Ψ prediction, we decided to use our previously published model of NanoPsu (S. Huang et al., 2021). The challenge here was to make models for m^{6}A that could use the same set of features as NanoPsu.

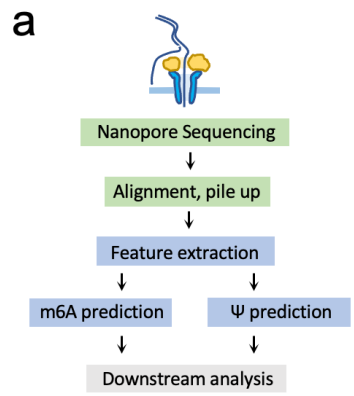

Figure 3.1 NanoSPA method pipeline
(a) Workflow of the NanoSPA pipeline.

We checked the correlation of the 12 features used in NanoPsu with the modification state (m^{6}A=1, A=0) (**Fig. 3.2a**). Both values of centered and flanking sites of the top 8 m^{6}A enriched 5mer motifs are checked. To note, for each motif, some features are useless. For example, in "AGACA" motif, the feature "mutation to A at position 1" is always 0, as A means no mutation at this position in this motif. We chose to keep these useless features to simplify the

process of feature extraction for all motifs and due to the fact that they are harmless to the models. Overall, we found that there are some significant features for each motif but the most significant ones for each motif are different. For example, “mutation to G at position 1” is useful for AGACT but not for TGACT, which means that A bases 2nt 5’ to m^6A are likely to mutate to G in the nanopore signal but T bases at the same relative position don’t. Based on the difference in feature preference, we decided to keep all features for model training to simplify the feature extraction process for different motifs. The correlation map indicated that it’s possible to use this set of features to make models for m^6A prediction.

Of course, there is doubt whether it’s valuable to use the same set of features for m^6A and pseudouridine models. The comparison of the fused workflow with other published nanopore direct RNA sequencing based workflows is shown below in section 3.2.4.

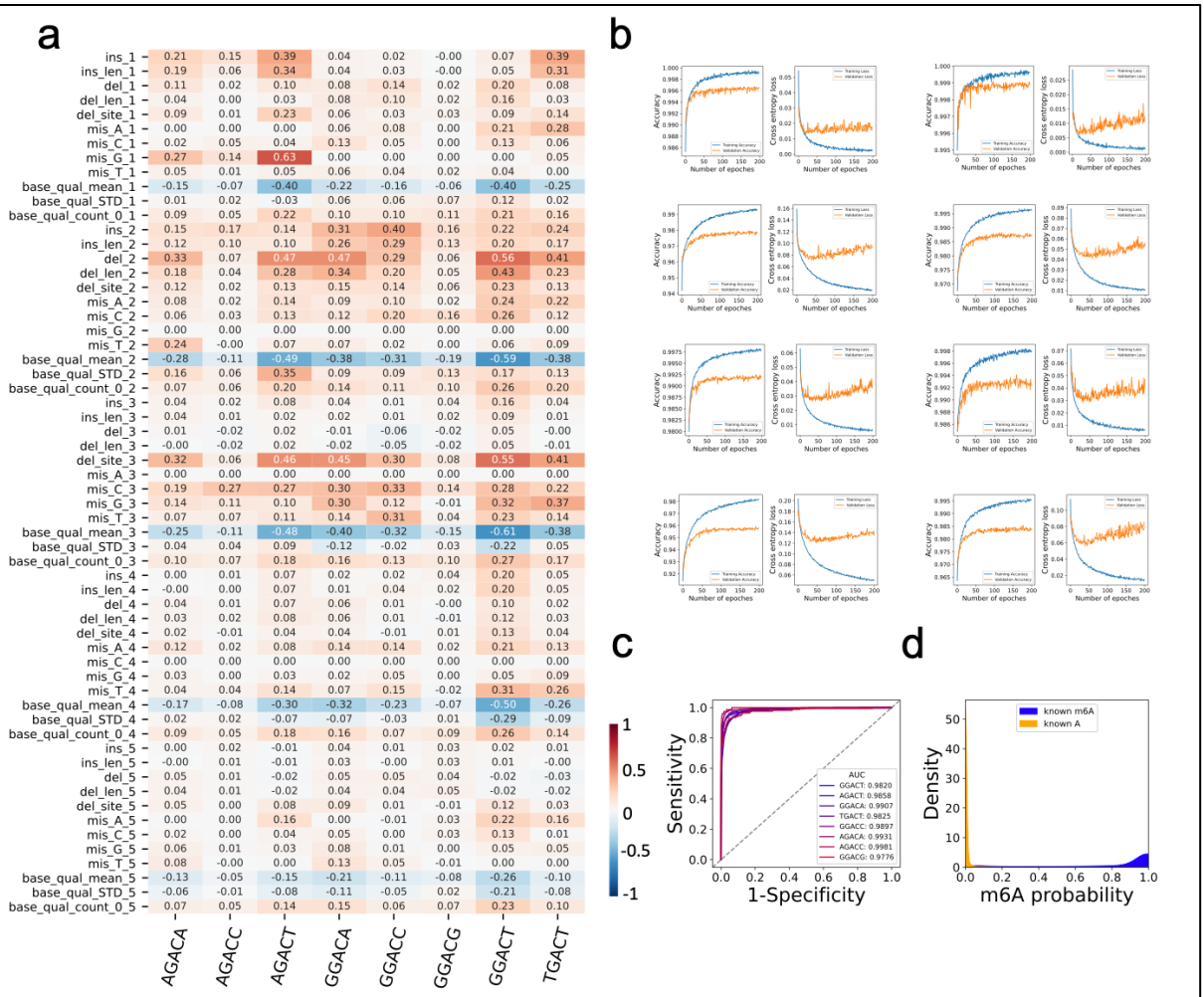

Figure 3.2 m6A prediction model features and training process in NanoSPA

(a) Correlation between features of the five sites in the top 8 m^{6}A motifs and modification state (m^{6}A=1, A=0). The suffix 1-5 indicate the position of the base in each motif. Ins, insertion rate after the base. Ins_len, insertion length mean. Del, deletion rate after the base. Del_len, deletion length mean. Del_site, deleted site ratio (the site is in a deletion). Mis, overall mismatching ratio. Mis_C, mutation to C ratio. Mis_G, mutation to G ratio. Mis_T, mutation to T ratio. Base_qual_mean, average base quality score. Base_qual_STD, base quality score standard deviation. Base_qual_count_0, ratio of bases with a quality score 0 at a site. (b) Training process of the models for the 8 m^{6}A motifs. Every two panels are for one motif. Left panel, training and validation accuracy. Right panel, training and validation loss. The order of the 8 motifs is AGACA, AGACC, AGACT, GGACA, GGACC, GGACG, GGACT, TGACT from left to right first and then from up to bottom. (c) ROC curves of the 8 final FNN models on testing sets for the 8 m^{6}A motifs that covers >90% of all m^{6}A sites in our training data (Table S1). AUC is indicated for each motif. (d) Sum of the prediction results of the annotated m^{6}A and A sites in the testing sets for the 8 m^{6}A motifs.

3.2.2 m^6A model development

Since we are going to use the pseudouridine prediction model in NanoPsu for pseudouridine identification, the only thing we need to build is a new m^6A prediction model. The reasons that we do not use a m^6A model from a published method is shown below in section 3.2.4, where we compare the performance of our model with other methods. To make the output of two modifications in the same format, we decided to make a "single mode" supervised learning model for m^6A. The model will be a classification model which does not provide prediction for modification fraction, the same as the pseudouridine model. For this purpose, high quality training materials for m^6A are crucial.

We took advantage of the recently published m^6A-SAC-seq data by Illumina sequencing, which mapped m^6A at single-base resolution and with modification stoichiometry transcriptome-wide (L. Hu et al., 2022). The reason why we use data from this method is discussed in section 1.1.2.3, where m^6A-SAC-seq is introduced. To maximize accuracy, we employed information from the modified or unmodified A nucleotide, as well as two flanking nucleotides on either side of A, since m^6A modification has a preference for the motifs of DRACH (D = A,G,U, R = A,G, H = A,C,U, **Fig. 3.2a**). Models considering more useful information tend to have better performance. It is feasible to consider the flanking sites as m^6A is enriched in specific motifs so that we could gain enough known sites for training in these motifs, without the necessity to pool the sites from different sequence content to increase the amount of training materials.

To strike a balance between covering as many m^6A sites as possible and ensuring sufficient number of mapped m^6A sites in each motif for training, we included the top 8 motifs quantified by m^6A-SAC-seq which covered 90.46% of all m^6A sites (**Table 3.1**). Although data

of 248 A-centered 5mer motifs will be discarded in advance, this only decreased the recall by less than 10%. In the application for real biological samples, it will only miss a small group of m^6A sites. Also, m^6A in the other 248 motifs are low in frequency, which means there will hardly be enough known m^6A sites in these motifs for model training. To note, although we limit the range of applied sites of each model by sequence content, we do not use sequence content as features in any models. All the values of the features are solely from processed nanopore direct RNA sequencing data, but not prior knowledge from databases or reference sequences. Thus, these models are not static regarding the sequence content of reads and could be applied to cells of different biological or physiological conditions.

Index	Motif	Count	Fraction	Cumulative sum
1	GG-CT	9853	0.2292	0.2292
2	AG-CT	8897	0.2070	0.4362
3	GG-CA	6014	0.1399	0.5762
4	TG-CT	5568	0.1296	0.7057
5	GG-CC	3487	0.0811	0.7868
6	AG-CA	2659	0.0619	0.8487
7	AG-CC	1304	0.0303	0.8790
8	GG-CG	1099	0.0256	0.9046

Table 3.1 Top 8 motifs in HeLa cells from m^6A-SAC-seq

Using the HeLa m^6A-SAC-seq data together with our nanopore direct mRNA sequencing data, we selected high confidence, 100% modified m^6A sites for model training (**Table 3.2**, also see section 3.4.4). High confidence m^6A sites are more likely to be also m^6A sites in our training sample and thus facilitate the training process. For each motif, we decided to train a feedforward neural network (FNN) model. In order to amplify the number of data points used for neural network training, we used a strategy to do data augmentation (**Table 3.3**). The reads covering a site were shuffled and samples of 16 random reads as a group was generated. These 16 reads

were viewed as covering the "artificial" m^6A or A site. From each site, 20 groups of random reads are selected to generate 20 artificial sites. In this way, the number of data points used for training increased by 20 times and the magnitude of data points for each model increased from 10^3 to 10^4, which made it appropriate to use neural network models. Also, when generating the artificial sites, no matter what the coverage of the original site is, only 20 sites of 16 reads will be generated, so that the possible effect of the coverage on the final models are thoroughly removed.

Motif	m^6A in m^6A-SAC-seq	High confidence m^6A in m^6A-SAC-seq	Total A and m^6A in Nanopore sample	A for training	m^6A for training
GG-CT	9853	6360	4600	3457	600
AG-CT	8897	5912	3673	2873	410
GG-CA	6014	3267	5893	5158	229
TG-CT	5568	3239	3970	3446	184
GG-CC	3487	1719	6060	5533	138
AG-CA	2659	1318	4688	4344	74
AG-CC	1304	549	4945	4646	18
GG-CG	1099	511	3098	2894	41

Table 3.2 Number of A and m^6A sites used to train the models

Motif	A train	m6A train	A test	m6A test
GG-CT	53468	9519	13312	2461
AG-CT	43371	6669	10889	1671
GG-CA	80446	3687	20094	893
TG-CT	54708	2936	13672	744
GG-CC	86585	2201	21595	559
AG-CA	65662	1199	16458	281
AG-CC	71043	283	17757	77
GG-CG	44956	651	11264	169

Table 3.3 Number of A and m6A sites in training/validation and testing sets after data augmentation and splitting

We optimized 8 feedforward neural network (FNN) models for the top 8 m^6A motifs (**Fig. 3.2b, Table 3.4**) and the models with the lowest validation loss were the final models. The structures of the 8 neural networks are the same to simplify the pipeline. Also, it was discovered that the structure of the neural networks was not the key factor that affected the model performance. The AUC values ranged from 0.9776 to 0.9931 for the 8 motifs separately on testing sets (**Fig. 3.2c**) and most known m^6A and A sites are predicted correctly (**Fig. 3.2d**). For m^6A, we found that most of the predicted probabilities are very close to either 0 or 1 so there is no necessity to keep the value of probabilities as the final output. Instead, we use 0.5 as the threshold to call m^6A and A sites.

Motif	Minimal validation loss	Epoch	Validation accuracy
GGACT	0.1208	47	0.9553
AGACT	0.0709	44	0.9756
GGACA	0.0408	57	0.9868
TGACT	0.0561	40	0.9829
GGACC	0.0264	34	0.9911
AGACA	0.0132	63	0.9963
AGACC	0.0059	19	0.9989
GGACG	0.0284	32	0.9926

Table 3.4 Number of epochs, minimal validation loss and validation accuracy of the final models for the 8 motifs

3.2.3 Validation of m^6A model by NGS methods

To further validate the performance of NanoSPA on m^6A calling, we performed NanoSPA on a published wild type nanopore human transcriptome sample (S. Huang et al., 2021) and compared the m^6A sites called by NanoSPA with the m^6A sites called by three recently published single-nucleotide resolution m^6A mapping methods, m^6A-SAC-seq (L. Hu et al., 2022), GLORI

(C. Liu et al., 2023), and eTAM-seq (Y.-L. Xiao et al., 2023). Among the 7841 m^6A sites called by NanoSPA, 4133 (52.71%) were called by all three, 3397 (43.32%) were called by one or two, and only 311 (3.97%) were not called by any of the three m^6A sequencing methods (**Fig. 3.3a**). For comparison, the fraction of m^6A sites called only by one of the four methods was 15.28% for m^6A-SAC-seq, 80.29% for GLORI, and 13.13% for eTAM-seq (**Fig. 3.3a**). There is no gold standard for human mRNA m^6A sites, so it is unclear whether these sites called only by one method are false positive or true positive results. These wide discrepancies of m^6A site calling render a precise assessment of the m^6A methylome very difficult. At the same time, high overlaps (>96%) of NanoSPA called m^6A sites with three orthogonal methods demonstrate the reliability of NanoSPA within current scope of knowledge of RNA epitranscriptomics and provide high confidence of using NanoSPA for downstream analysis.

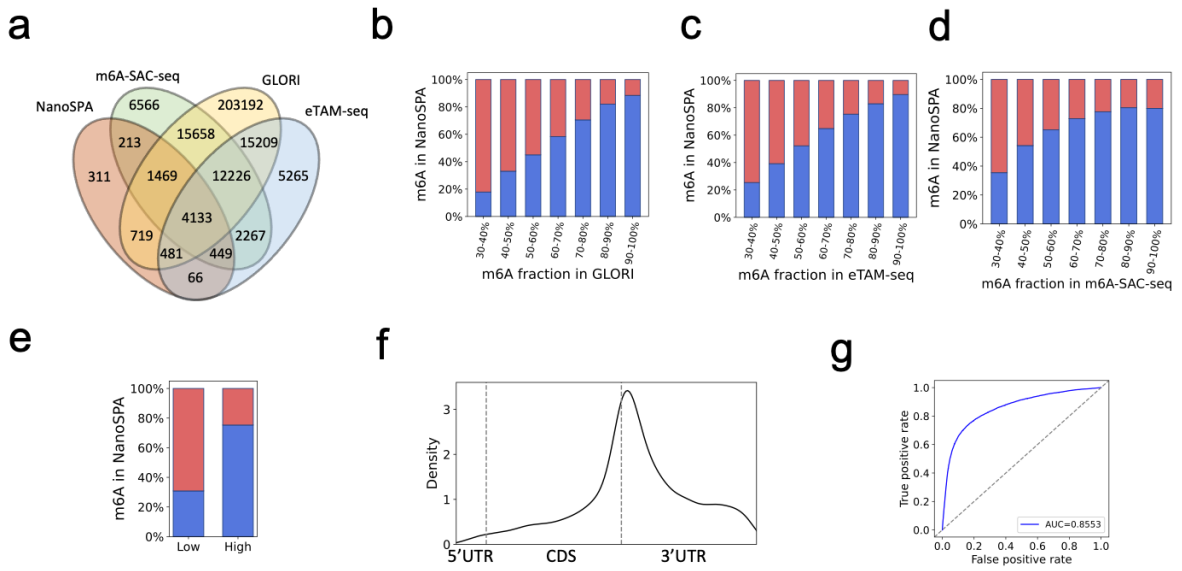

Figure 3.3 Validation of m6A prediction model in NanoSPA by NGS and nanopore based methods

(a) Venn diagram of called m^{6}A sites in NanoSPA, m^{6}A-SAC-seq, GLORI and eTAM-seq. (b) m^{6}A sites from GLORI data are grouped by m^{6}A fraction and the percentage of those m^{6}A sites revealed by NanoSPA are shown. Blue, m^{6}A sites called by NanoSPA; red, m^{6}A sites not called by NanoSPA. (c) m^{6}A sites from eTAM-seq data are grouped by m^{6}A fraction and the percentage of those m^{6}A sites revealed by NanoSPA are shown. Blue, m^{6}A sites called by NanoSPA; red, m^{6}A sites not called by NanoSPA. (d) m^{6}A sites from m^{6}A-SAC-seq data are grouped by m^{6}A fraction and the percentage of those m^{6}A sites called by NanoSPA are shown. Blue, m^{6}A sites called by NanoSPA; red, m^{6}A sites not called by NanoSPA. (e) m^{6}A sites from m^{6}A-SAC-seq data are grouped by confidence (high = more than 20% mutation rates in m^{6}A-SAC-seq data in both replicates, low = others) and the percentage of those m^{6}A sites called by NanoSPA are shown. Blue, m^{6}A sites called by NanoSPA; red, m^{6}A sites not called by NanoSPA. (f) Meta gene distribution of the m^{6}A sites called by NanoSPA. (g) The ROC curve of NanoSPA on a published mESC sample (Zhong et al., 2023), based on the ground truth of miCLIP2 called m6A sites. The AUC is 0.8553. The AUC values of other nanopore direct RNA sequencing methods on m6A prediction under the same setting from fig S2a panel 2 of Zhong et al., 2023 are: Tombo (0.69), MINES (0.64), Nanom6A (0.75), m6Anet (0.89), ELIGOS (0.80) and Epinano (0.77). Compared to m6Anet, NanoSPA took less time and storage to run.

Additional supports for the accuracy of our NanoSPA method included the higher m^{6}A fraction called by m^{6}A-SAC-seq, GLORI, and eTAM-seq, the higher probability for m^{6}A calling by NanoSPA (**Fig. 3.3b-d**), and high-confidence m^{6}A sites called by m^{6}A-SAC-seq were also

better called by NanoSPA (**Fig. 3.3e**). Our predicted m^{6}A sites had a metagene profile consistent with the known m^{6}A enrichment around the stop codon (**Fig. 3.3f**).

3.2.4 Validation of m^{6}A model by nanopore methods

Besides comparison with NGS based m^{6}A calling methods, it is also valuable to show the performance of NanoSPA by directly comparing with other m^{6}A prediction methods based on nanopore direct RNA sequencing data. To compare to other nanopore direct RNA sequencing based m^{6}A detection methods, we run NanoSPA on a previously published mouse ESC sample and call m^{6}A sites based on the same ground truth from miCLIP2(Körtel et al., 2021). NanoSPA has a better AUC over Tombo (0.69), MINES (0.64), Nanom^{6}A (0.75), ELIGOS (0.80) and Epinano (0.77) (Zhong et al., 2023) (**Fig. 3.3g**, Fig S2a panel 2 of Zhong et al., 2023) which mean the m^{6}A model itself in NanoSPA is better than the other models.

Although m^{6}Anet has higher AUC than NanoSPA, the time and storage demand of NanoSPA is much smaller than m^{6}Anet. It took 1.75 hours to run the whole protocol of NanoSPA on 1.0 million base called reads with 16 CPUs with 29 GB of generated intermediate and final data. To note, this includes the pipelines for both m^{6}A and pseudouridine prediction. As a comparison, one of the multiple steps of m^{6}Anet, “nanopolish eventalign”, took 13.3 hours to run on the same data with the same computation resources and generated a 174 GB “eventalign.txt” file.

The fused workflow for m^{6}A and pseudouridine has advantages over other methods. Since m^{6}Anet itself takes more time and storage than NanoSPA, combining m^{6}Anet with any current available pseudouridine detection pipeline for simultaneous m^{6}A and pseudouridine investigation will take even longer processing time and more storage.

The prediction modules are removable and extendable. Since the features for all bases are extracted, the pipeline could easily be extended to other modifications like m^7G, m^1A or m^5C and the only thing needed is a new model for these modifications. The prediction steps are time saving. If the protocol has been run on pseudouridine and the intermediate files are stored, the extra time for running m^6A prediction on 2.6 million reads is just 6 minutes, which is extremely time saving to run predictions on new modifications on previously processed old samples.

Importantly, a lower number of processing steps can be crucial for the application by biological scientists who have less computational background. In the experience of the field, biological scientists using several pipelines are commonly stuck when there are too many packages to install (e.g. one for m^6A, and a different one for Ψ), require too many intermediate steps, and the packages may even conflict with each other.

3.3 Discussion

In summary, we developed a machine learning pipeline NanoSPA for nanopore sequencing to analyze m^6A and Ψ simultaneously. The new m^6A model outperforms most published m^6A models for nanopore direct RNA sequencing and is fused with the pseudouridine model for less time and storage cost. It tries to reach a balance between covering as most known m^6A sites as possible and maintain high accuracy for overall prediction. Of course, the number of motifs included could be discussed. If more motifs are included, the pipeline will be able to cover more m^6A sites, but the model performance will decrease. If fewer motifs are included, the model performance will be better with the risk of losing more m^6A sites. There are no standard answers for such balance questions.

The strategy used in this chapter for simultaneous detection of m^6A and pseudouridine could also be extended and applied to other modifications. Since the features of all 4 bases are extracted, building models for other modifications such as m^7G, m^1A or m^5C could be carried out using the same set of extracted features. All the prediction models could be fused into one single pipeline and the predictions could be done on the same sample without the requirement of running the experiments and data processing of the same sample multiple times. Also, as enzymatic or chemical pre-treatment is not required during library preparation, those historical samples run before the methods are developed could also be reanalyzed by the new protocols or new models in the future. Of course, there are also challenges to extend the models to more modifications. The main challenge is to find out reliable data for training. The modification less prevalent than m^6A and pseudouridine usually have less available NGS based results and thus the availability of high-quality known modification sites could be questionable. One possible solution is to use synthesized sequences with the specific modifications, but usually the synthesized sequences are over-modified with too high density of modification sites and the sequence content could not reflect the real biological conditions. Also, there will be fewer datasets available for validation and comparison to convince the users.

Same as the pseudouridine model in NanoPsu, the m^6A model in NanoSPA is also not perfect. First of all, it only covers the 8 most prevalent motifs in human cells. Although the 8 motifs could cover over 90% of overall m^6A sites in human and the situation is similar for other species like mouse, but it will not perform well if it is applied to those species that have a different preference for types of m^6A motifs. To solve this, a model that equally consider all motifs is needed. The challenge here would be to obtain training dataset, as m^6A is rare in many motifs in the model species that we use, and it will be hard to achieve known m^6A sites from

these motifs. The problem for synthesized sequenced have been mentioned in the last paragraph and thus is also not a good choice.

Second, the m^6A model is a supervised learning model, which rely on the ground truth from known m^6A sites from NGS methods. Here we use the data from m^6A -SAC-seq, which is currently one of the most accurate datasets available. However, as m^6A is highly dynamic, the distributions of m^6A may vary among different samples, even if they are all WT samples from the same human cell line. We tried to maximize the reliability by only using the highly confident m^6A sites as the modification sites for training. The improvement in NGS m^6A detection accuracy could benefit the accuracy in nanopore models in the future, with the impossibility to confirm every single called m^6A site by low throughput experimental methods.

Third, the identification of m^6A is not quantitative in this model. Including stoichiometry prediction will lower the reliability of the model. Of course, it is possible to train a model for m^6A stoichiometry evaluation without the necessity of any extra data. The training dataset from m^6A-SAC-seq includes stoichiometry information and it's straightforward to transform a classification model into a regression model, which provides the stoichiometry.

It is to be emphasized that the development of nanopore based m^6A detection methods will not replace NGS based methods. The major goal for nanopore methods is not to have higher accuracy than any NGS based methods. It is to combine the identification of RNA modifications with the advantages of nanopore sequencing like long read length to facilitate the studies of relationship of modifications and splicing, poly A tail length, etc.

3.4 Methods

3.4.1 WT cell sample culture

HeLa cells (ATCC) used for model training were cultured in Dulbecco's Modified Eagle's Medium (DMEM) with high glucose and L-glutamine, without sodium pyruvate (HyClone, SH30022.01) with 10% FBS and 1% Pen–Strep (Penicillin–Streptomycin). ~3.5 x 10^6 cells were seeded into three 150 mm plates with the same media without 1% Pen–Strep (Penicillin–Streptomycin). The plates were mixed by gently rocking and incubated in a 37°C with 5% CO_2 for 72 hours.

The three 150 mm plates of ~80% confluent HEK293T cells were treated with 100 µg/mL cycloheximide (CHX AC3574200500, Fisher Scientific) for 7 minutes at 37°C. Media was removed from the plates and the cells were washed twice with 10 mL of ice-cold 1x PBS containing 100 µg/mL CHX prior to being scraped and collected in 5 mL 1x ice-cold PBS. Cells were pelleted by centrifugation at 500 g for 5 minutes, then the three plates were combined in 0.8 mL Lysis Buffer (20 mM HEPES, pH 7.6, 100 mM KCl, 5 mM $MgCl_2$, 1% Triton X-100, 100 µg/mL CHX supplemented with fresh 1x Roche protease inhibitor and 1% Superase inhibitor) and lysed by rotating at 4°C for 20 minutes. Cell debris was pelleted by 15-minute centrifugation at 16,000 g. To this lysate, 4 µL T4 Turbo DNAse (Invitrogen, AM2238) was added, and the mixture was incubated at room temperature for 15 minutes. 0.9 mL TRIzol™ Reagent was added to the tube. The sample was incubated for 5 minutes at room temperature prior to addition of 0.18 mL chloroform and an additional 3-minute incubation. The aqueous layer was separated by centrifugation for 15 minutes at 12,000 g and 4°C and then added to a new tube. The sample was frozen overnight at -80°C following addition of 0.45 mL isopropanol. RNA was precipitated by centrifugation for 15-minute at 12,000 g and 4°C. The supernatant was removed, and the RNA

pellet resuspended and washed in 0.9 mL 75% ethanol before being centrifuged for 5 minutes at 7,500 g and 4°C. The supernatant was removed and the pellet air-dried for ten minutes prior to resuspension in 10 µL dH_2O.

The RNA sample was made to 150 µL in dH_2O and cleaned by adding 280 µL Beckman RNAClean XP beads. The sample was mixed by pipetting up and down, incubated at room temperature for 5 minutes, pelleted on a magnetic rack for 5 minutes, washed with 1000 µL 70% EtOH three times, and air-dried for ten minutes before eluting into 150 µL dH_2O. Poly(A)-selection of cleaned RNA sample was then done using Promega PolyATract mRNA Isolation System IV per the manufacturer's protocol. Briefly, the sample was made to 500 µL and incubated at 65°C for 10 minutes before addition of 3 µL Biotin-dT and 1x SSC. Once cooled, the sample was added to 100 µL of washed poly(A) magnetic beads in 0.5x SSC. Following a 10-minute incubation at room temperature, RNA-bound beads were captured using a magnetic rack, washed in 0.1X SSC, and eluted in a total of 250 µL dH_2O.

PolyA+ RNA was concentrated using Zymo Oligo Clean & Concentrator columns per manufacturer's protocol prior to Nanopore sequencing. Briefly, 500 µL Oligo Binding Buffer and 2 mL absolute ethanol were added to 250 µL PolyA+ RNA and mixed by pipetting. The sample was transferred to a Zymo-Spin™ IC Column in a collection tube and centrifuged. The sample was washed using 750 µL DNA Wash buffer prior to elution in 11 µL nuclease-free dH_2O. For quality control, polyA+ RNA was submitted to the University of Chicago Genomics facility and analyzed using an Agilent 2100 Bioanalyzer system. 5 µL of RNA containing ~1 ng/µL was analyzed using the RNA Pico/High Sensitivity Assay (input sensitivity of 0.05-5 ng/µL) to confirm RNA integrity.

3.4.2 Nanopore direct RNA sequencing

The library preparation of direct RNA seq samples followed instructions from Oxford Nanopore Technology for Direct RNA Sequencing Kit (SQK-RNA002). Concisely, 500 ng of Poly(A)$^+$ RNA sample was used to perform a run. The RT Adaptor (RTA) was ligated to the 3' end of Poly(A)$^+$ RNA using T4 DNA ligase (NEB M0202S), followed by reverse transcription by SuperScript III Reverse Transcriptase (ThermoFisher 12574018). The RT product was then purified by 1.8x RNAClean XP beads (72 μL) (Beckman Coulter A63987). A second RNA Adaptor (RMX) was then attached to the 3' end of Poly(A)$^+$ RNA by T4 DNA ligase (NEB M0202S). The RNA product was purified with 1x RNAClean XP beads (40 μL) and eluted with 21 μl Elution Buffer. The sample was loaded onto a R9.4.1 flow cell (FLO-MIN106D) and then run on the MinION sequencer for 72 hours.

3.4.3 Nanopore data pre-processing

Raw sequencing data files were uploaded to UChicago midway2 cluster for pre-processing. Base calling was performed by guppy base caller (version 3.2.2+9fe0a78). Then, the reads were aligned to human genome (GRCh38.p13) by minimap2 (H. Li, 2018) (version 2.18-r1015) with parameters -ax splice -uf -k14. The mapped reads were piled up by samtools (H. Li et al., 2009) (v1.11) and features for modifications prediction were extracted by customized python scripts (https://github.com/sihaohuanguc/NanoSPA).

3.4.4 Model training for m^{6}A prediction

The wild-type HeLa cell nanopore seq sample was used to train the model for m^{6}A prediction. The data was pre-processed as described above. Based on data achieved from the

m^{6}A-SAC-seq study (L. Hu et al., 2022), we screened for all high confidence m^{6}A sites in the 8 motifs. High confidence was defined as >20% induced mutation rate by reverse transcriptase in the mapped cDNA reads in both replicates in m^{6}A-SAC-seq. Those screened high confidence m^{6}A from m6A-SAC-seq which were also covered by > 20 reads in our nanopore sample were used as m^{6}A sites in the training process (**Table 3.2**). At the same time, those A sites covered by > 20 reads in our nanopore sample not in the list of m^{6}A sites (including all high and low confidence m^{6}A sites) were used as the unmodified A sites for training (**Table 3.2**). Data augmentation was performed by shuffling the reads of a site and sampling 16 random reads as a group with 20 groups for each site. Then the dataset for each motif was randomly split into 80% training and validation set and 20% testing set (**Table 3.3**). Sixty features (see **Fig. 3.2a**) for the centered, -2, -1, +1, and +2 sites were collected for all generated sites in the 8 motifs. For each motif, feedforward neural network (FNN) models were trained on training set and evaluated by validation loss (cross entropy). The FNN models had two hidden layers (128 and 64 nodes, activation function ReLU, drop rate 0.1 and 0.2 after each layer respectively), with learning rate 0.001 and maximum epochs 200. The models with the lowest validation loss were stored (**Table 3.4**) as the final models. The final performance for the 8 models was evaluated on the testing sets (not used in training or validation) by AUC (area under curve) of ROC (Receiver Operating Characteristic) (Dean & Monga‘TensorFlow, 2015; Harris et al., 2020; Hunter, 2007; McKinney, 2010; Pedregosa et al., 2011).

3.4.5 Validation for m^{6}A models

To further validate the m^{6}A prediction models, NanoSPA was performed on published HeLa samples (S. Huang et al., 2021) (GSM5467024, GSM5467025). The reads from the two

replicates were combined before running the pipeline. The A sites with an output value >0.5 in the FNN models were called as m^6A sites. The m^6A-SAC-seq HeLa m^6A sites were provided by the authors of this publication. The GLORI study did not have HeLa samples, so HEK293T samples were used instead; the union of the m^6A sites in the two replications (GSM6432590, GSM6432591) was used for comparison. For eTAM-seq, the deep version of replicate 1 HeLa sample (GSE211303) was used for comparison. Venn diagram of the intersections among the four methods was shown in Figure 1d.

The stoichiometry of m^6A modifications was also obtained from the same samples of the three published methods. For GLORI, for sites present in both replicates, the m^6A fraction values were averaged. The m^6A sites were assigned into groups according to their reported modification fractions.

To compare NanoSPA with other nanopore direct RNA sequencing methods, we followed the protocol of a previously published paper (Zhong et al., 2023) and run the same mESC sample (GSM5841801) by NanoSPA and used the same set of miCLIP2 called m^6A sites (Körtel et al., 2021) (GSE163500) as ground truth to draw the ROC curve and calculate AUC.

To compare the time and storage demand of NanoSPA and m^6Anet, we run both protocols on the same 1.0 million base called reads with 16 CPUs on the same computation node. It took 1.75 hours to run the whole protocol of NanoSPA with 29 GB of generated intermediate and final data. As a comparison, one of the multiple steps of m^6Anet, “nanopolish eventalign”, took 13.3 hours to run and generated a 174 GB “eventalign.txt” file.

Chapter 4. Simultaneous mRNA m^6A and pseudouridine nanopore profiling reveals coordination in translation

4.1 Introduction

Translation is the process of producing proteins for metabolism and regulation. RNA modifications are involved in the translation process. Ψ has been shown to affect splicing and translation (B. R. Anderson et al., 2010; Eyler et al., 2019; Kariko et al., 2008; Martinez et al., 2022). The tRNA pseudouridine at anticodon positions were reported to affect base pairing, translation efficiency and fidelity (D. R. Davis et al., 1998; Harrington et al., 1993). Since anticodons are directly base paired with mRNA, it is likely that pseudouridine on mRNA will also affect base pairing efficiency. Effect of pseudouridine in mRNA on translation was reported *in vitro*. It was reported to promote translation in rabbit reticulocyte system, while negatively affect translation in the wheat germ system and *E.coli* system (Kariko et al., 2008). The previous results show totally opposite views and there is still no consensus on whether and how pseudouridine affect translation.

m^6A was reported to be involved a wide range of events in gene expression process like splicing, mRNA decay, nuclear export and translation (Louloupi et al., 2018; I. A. Roundtree et al., 2017; X. Wang et al., 2014). m^6A is involved in translation from multiple aspects. Transcripts bound by m^6A reader protein YTHDF2 are directed to mRNA decay site rather than to translatable pool and thus protein production is controlled (X. Wang et al., 2014). It was reported that human YTHDF1 bound to mRNA m^6A could recruit translation initiation factors and

promote translation (X. Wang et al., 2015). It was reported that m^6A could form clusters, and transcripts with more m^6A clusters had significant lower level of translation (C. Liu et al., 2023).

The NanoSPA pipeline developed in Chapter 3 enables us to identify m^6A and pseudouridine simultaneously transcriptome wide from nanopore direct RNA sequencing data. Thus, in this chapter, we apply the method on input and polysome associated mRNA to study the effect of m^6A and pseudouridine on translation. We also investigate the change of results when m^6A writer METTL3 or one of the pseudouridine writers TRUB1 is knocked down.

4.2 Results

4.2.1 m^6A and Ψ in the siCTRL sample

To investigate the relationship between m^6A and Ψ, our experimental design (**Fig. 4.1a, b**) used siRNA knockdown by negative control siRNA (siCTRL), against the core m^6A writer METTL3, and against one of the major Ψ writers for mRNA in cultured human cell lines, TRUB1 (E. K. Borchardt et al., 2020; Dai et al., 2023; M. Safra et al., 2017). We performed polyA-selection and ran nanopore direct RNA sequencing of biological replicates, which yielded good mapping coverages (**Fig. 4.1c**). Applying NanoSPA on the siCTRL samples, we found that transcript groups with more Ψ had fewer m^6A sites (**Fig. 4.2a, b**). Conversely, transcript groups with more m^6A had less Ψ modification (**Fig. 4.2c, d**). These results suggest that m^6A and Ψ are less likely to co-occur on the same transcripts.

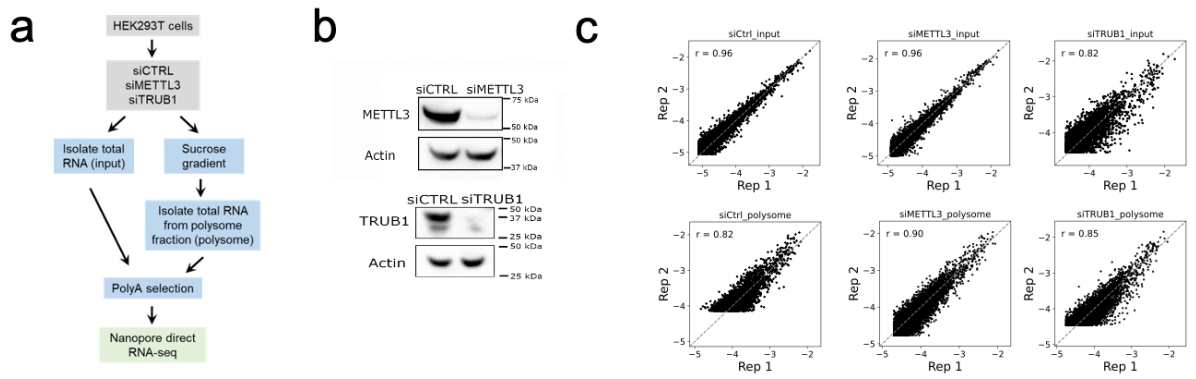

Figure 4.1 Experimental design and sequencing data overview
(a) Experimental design. A total of 12 nanopore data sets were generated: siCTRL, siMETTL3, siTRUB1; input and polysome, two biological replicates. (b) Western blots showing knockdown of the METTL3 or the TRUB1 protein. Actin is the loading control. Full scans are attached in Fig. S4.1. The same experiment was performed in biological replicates. (c) RNA expression of biological replicates of each condition with Pearson's correlation coefficient r values indicated.

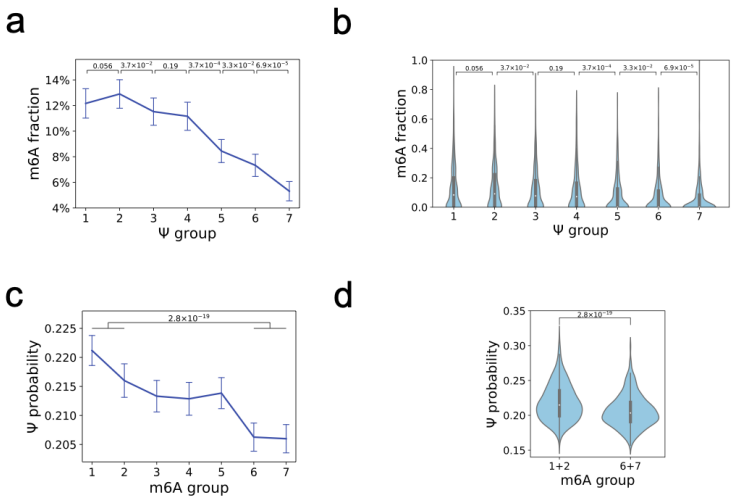

Figure 4.2 Experimental results of the siCTRL sample

(a) Mean m^{6}A fraction of transcripts grouped by mean Ψ probability in the siCTRL input sample. Transcripts with ≥20 U sites and ≥7 A sites are distributed into 7 groups evenly (N=670 each for groups 1-6, N=667 for group 7) based on mean Ψ probability of the sites in each transcript; group 1 has the lowest and group 7 the highest mean Ψ probability. (b) Violin plot of Fig. 4.2a (N=670 each for groups 1-6, N=667 for group 7). (c) Mean Ψ probability of transcripts grouped by m^{6}A fractions in the siCTRL input sample. Transcripts with ≥20 U sites and ≥7 A sites containing at least one m^{6}A site are distributed into 7 groups evenly (N=346 each for groups 1-6, N=339 for group 7) based on mean m^{6}A fraction of the sites in each transcript; group 1 has the lowest and group 7 the highest m^{6}A fraction. (d) Violin plot of Fig. 4.2c. Groups 1 and 2 (N=692), or groups 6 and 7 (N=685) are combined. For **a-d**, p values are determined by two-sided Mann-Whitney U test, error bar represents 95% confidence interval (CI). In violin plots (**b**, **d**), the center line in the inner box plots represents the median, the lower and upper hinges represent the first and third quartiles, and the whiskers represent ±1.5x interquartile range.

4.2.2 m^{6}A and Ψ in the knock down samples

To further evaluate m^{6}A and Ψ relationship, we analyzed the changes of m^{6}A and Ψ upon writer knockdowns. As expected, METTL3 knockdown reduced m^{6}A in all transcript groups regardless of their Ψ status (**Fig. 4.3a**), whereas TRUB1 knockdown reduced Ψ modification globally (**Fig. 4.3b**). We compared the mRNA expression of siCTRL and writer knockdown samples (**Fig. 4.3c**) and found that m^{6}A or Ψ writer knockdown primarily affected genes

involved in metabolic processes (**Fig. 4.3d, e**). We observed an appreciable increase of Ψ upon METTL3 knockdown in transcripts that contained more m^6A sites (groups 6 and 7), but little change in transcripts containing few m^6A sites (groups 1 and 2), consistent with m^6A inhibiting Ψ modification (**Fig. 4.3f, g**). TRUB1 knockdown decreased m^6A modification (**Fig. 4.3h, i**), which was unexpected from the opposing co-occurrence of m^6A and Ψ in the control sample (**Fig. 4.2a, c**). A plausible explanation is that TRUB1 installed Ψ sites promote m^6A, whereas Ψ sites installed by the other 12 human Ψ writers inhibit m^6A installation. Such intricate m^6A and Ψ modification dynamics would be an exciting avenue to pursue for future studies.

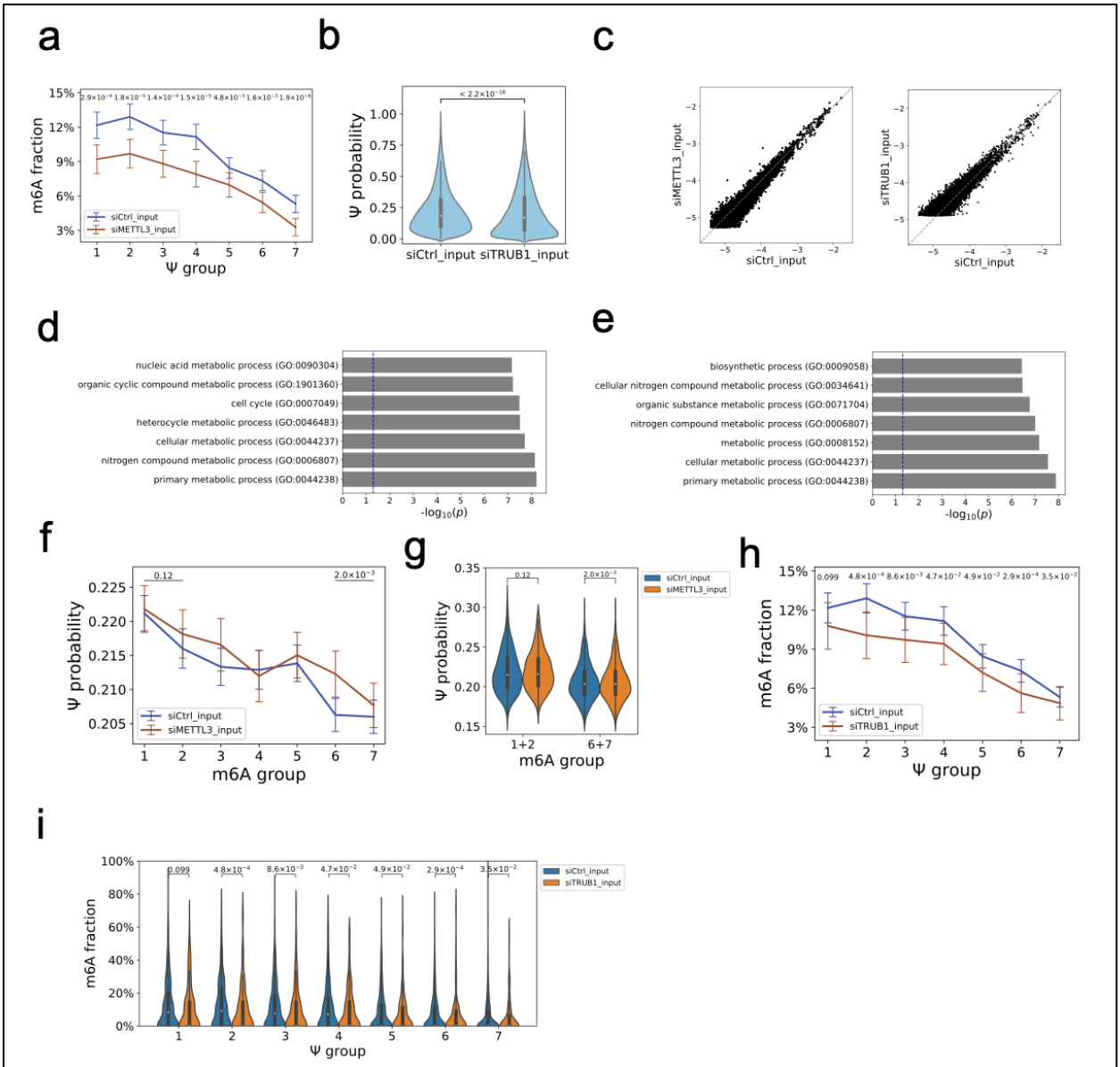

Figure 4.3 Experimental results of KD samples

(a) Mean m^{6}A fractions of transcripts grouped by mean Ψ probability in the siCTRL and siMETTL3 input samples. Transcripts with ≥20 U sites and ≥7 A sites are distributed into 7 groups evenly (N=670/444 each for groups 1-6, N=667/437 each for group 7 in siCTRL/siMETTL3 sample respectively) based on mean Ψ probability of the sites in each transcript; group 1 has the lowest and group 7 the highest mean Ψ probability. Error bar represents 95% confidence interval (CI). (b) Comparison of Ψ probability distribution of siCTRL and siTRUB1 (N=1079233 U sites). (c) Comparison of RNA expression of siCTRL versus siMETTL3 or siTRUB1. (d) Biological process gene ontology (GO) of top 200 highest expressed transcripts in siCTRL over siMETTL3. (e) Biological process GO of 200 highest expressed transcripts in siCTRL over siTRUB1. (f) Mean Ψ probability of transcripts grouped by m^{6}A fractions in the siCTRL and siMETTL3 input samples. Transcripts with ≥20 U sites, ≥7 A sites containing at least one m^{6}A site are distributed into 7 groups evenly (N=346/182

(Figure 4.3 continued) each for groups 1-6, N=339/176 for group 7 in siCTRL/siMETTL3 sample respectively) based on mean m^6A fraction of the sites in each transcript; group 1 has the lowest and group 7 the highest m^6A fraction. (g) Violin plot of Fig. 4.3f. Groups 1 and 2 (N=692/364 for siCTRL/siMETTL3 sample respectively), or groups 6 and 7 (N=685/358 for siCTRL/siMETTL3 sample respectively) are combined. (h) Mean m^6A fractions of transcripts grouped by mean Ψ probability in the siCTRL and siTRUB1 input samples. Transcripts with ≥20 U sites and ≥7 A sites are distributed into 7 groups evenly (N=670/249 each for groups 1-6, N=667/248 each for group 7 in siCTRL/siTRUB1 sample respectively) based on mean Ψ probability of the sites in each transcript; group 1 has the lowest and group 7 the highest mean Ψ probability. (i) Violin plot of Fig. 4.3h (N=670/249 each for groups 1-6, N=667/248 each for group 7 in siCTRL/siTRUB1 sample respectively). For a-b, f-i, p values are determined by two-sided Mann-Whitney U test, error bar represents 95% confidence interval (CI). In violin plots (b, g, i), the center line in the inner box plots represents the median, the lower and upper hinges represent the first and third quartiles, and the whiskers represent ±1.5x interquartile range.

4.2.3 effect of m^6A and Ψ on translation

To investigate the effect of m^6A and Ψ on translation, we performed polysome profiling without and with knockdown of METTL3 or TRUB1 as above (**Fig. 4.4a, 4.1a, b**). Among the two commonly used methods to study translation, ribo-seq and polysome profiling, only polysome profiling retains intact mRNA bound to the ribosome and is useful for nanopore sequencing. Both METTL3 or TRUB1 knockdown reduced the polysomes over the 80S monosome, even more so for the TRUB1 knockdown, implicating a significant change in translation properties upon the loss of this Ψ writer. For the siCTRL sample (**Fig. 4.4b**), gene ontology analysis showed that transcripts with the greatest reduction in the polysome belonged to genes involved in protein synthesis (**Fig. 4.4c**), and transcripts with the greatest increase in the polysome belonged to cellular organelles **(Fig. 4.4d**). The overall m^6A level in the polysome was about the same as the input (**Fig. 4.4e, f**), as was the overall Ψ level in the polysome (**Fig. 4.4g, h**). The ratio of polysome over input mRNA for each transcript is termed translation efficiency (TE) (Ingolia, Hussmann, & Weissman, 2019). TE includes multiple properties in translation

such as initiation and elongation rates. Although m^6A and Ψ in coding region slow elongation, in UTRs they could help recruit ribosome to mRNA which can increase initiation (Choi et al., 2016; Eyler et al., 2019; Kate D Meyer et al., 2015; Svitkin et al., 2017; X. Wang et al., 2015). By convention, transcripts with higher TE are considered to be positively regulated in translation. We found that the transcript groups with more m^6A in both input and polysome had a higher average TE than those with less m^6A (**Fig. 4.4i**). Transcripts in several TE groups also showed higher m^6A levels on the polysome (**Fig. 4.4j**), indicating that m^6A in the transcripts on the polysome can benefit translation. Transcripts with more Ψ in polysome over input also had higher TE (**Fig. 4.4k**), indicating that Ψ on the polysome can also benefit translation.

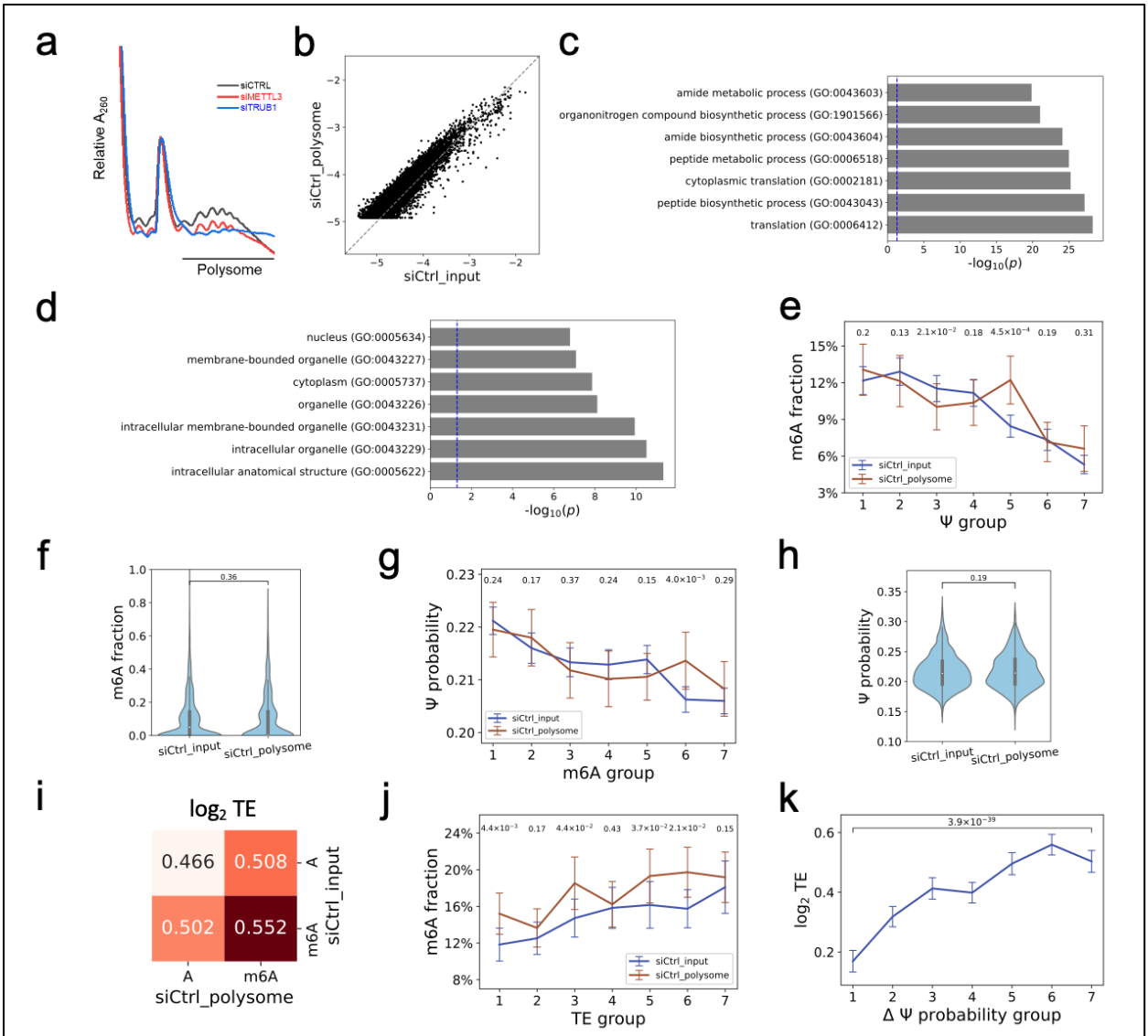

Figure 4.4 Effect of m^{6}A and Ψ in translation in siCTRL samples

(a) Polysome profiles of siCTRL, siMETTL3, and siTRUB1 samples. Knockdown of the m^{6}A writer or a major Ψ writer resulted in global decrease in translation. The "polysome" line shows the combined fractions used for RNA extraction and nanopore sequencing. (b) Comparison of siCTRL mRNA expression between input and polysome. (c) Biological process gene ontology (GO) analysis of top 200 highest expressed transcripts in siCTRL input over polysome. (d) Cellular Component GO of top 200 highest expressed transcripts in siCTRL polysome over input. (e) Mean m^{6}A fractions of transcripts grouped by mean Ψ probability in the siCTRL input and polysome samples. Transcripts with ≥20 U sites and ≥7 A sites are distributed into 7 groups evenly (N=670/210 each for groups 1-6, N=667/204 each for group 7 in siCTRL input/polysome sample respectively) based on mean Ψ probability of the sites in each transcript; group 1 has the lowest and group 7 the highest mean Ψ probability. (f) Distribution of mean m^{6}A fraction of genes of siCTRL input (N=4687) and polysome (N=1464). (g) Mean Ψ probability of transcripts grouped by m^{6}A fractions in the

(Figure 4.4 continued) siCTRL input and polysome samples. Transcripts with ≥20 U sites, ≥7 A sites containing at least one m^6A site are distributed into 7 groups evenly (N=346/100 each for groups 1-6, N=339/96 for group 7 in siCTRL input/polysome sample respectively) based on mean m^6A fraction of the sites in each transcript; group 1 has the lowest and group 7 the highest m^6A fraction. (h) Distribution of mean Ψ probability of genes of siCTRL input (N=4687) and polysome (N=1464). (i) Modification state combinations in input and polysome for all A sites in siCTRL input and polysome and their corresponding mean translation efficiency (TE). A: unmodified, m^6A: modified. (j) Mean m^6A fractions of transcripts grouped by TE in the siCTRL sample. Transcripts with ≥7 A sites containing at least one m^6A site in input or polysome are distributed into 7 groups evenly (N=124 for groups 1-6, N=120 for group 7 in siCTRL input/polysome sample) based on TE of each transcript; group 1 has the lowest and group 7 the highest TE. (k) TE of siCTRL transcripts grouped by differential mean Ψ probability. Transcripts with ≥20 U sites in both input and polysome samples are distributed into 7 groups evenly (N=883, 877, 880, 880, 877, 883, 876) based on Ψ probability difference between polysome and input; group 1 has the lowest and group 7 the highest TE. For **e-h**, **j-k**, p values are determined by two-sided Mann-Whitney U test, error bar represents 95% confidence interval (CI). In violin plots (**f**, **h**), the center line in the inner box plots represents the median, the lower and upper hinges represent the first and third quartiles, and the whiskers represent ±1.5x interquartile range.

4.2.4 m^6A and Ψ effect on translation in knock down samples

We further investigated the m^6A and Ψ writer knockdown effects on polysome profiling. Even though METTL3 knockdown showed a decrease in the input m^6A levels (**Fig. 4.3a**), m^6A levels in polysome transcripts persisted at similar levels as in the siCTRL (**Fig. 4.5a**). TE changes upon METTL3 knockdown were not significant and transcripts with more m^6A still had higher TE than those with less m^6A (**Fig. 4.5b**), which mirrored the m^6A level changes on the polysome. Ψ levels slightly increased in the polysome transcripts in siMETTL3 relative to siCTRL (**Fig. 4.5c**), which could be related to increased Ψ in the input mRNA upon METTL3 knockdown. This increase was accompanied by a slight increase in TE among transcript Ψ probability groups (**Fig. 4.5d**).

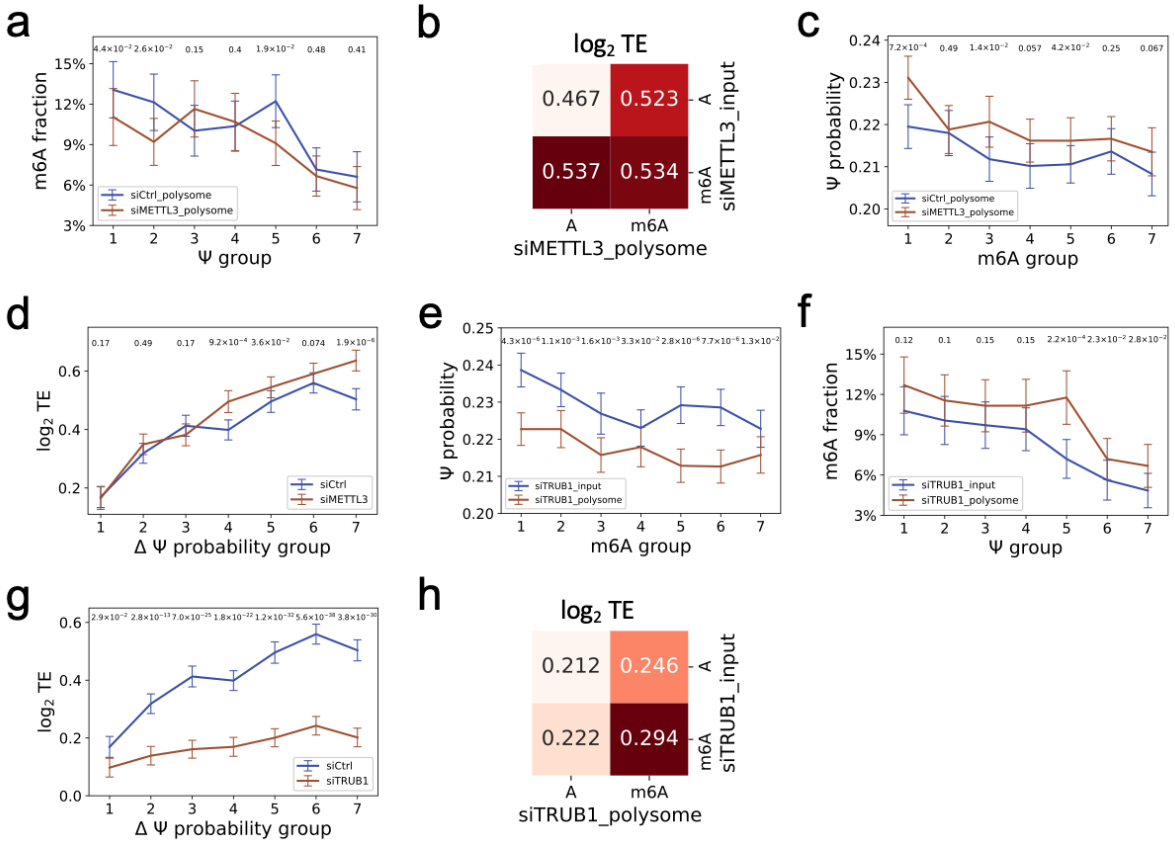

Figure 4.5 Effect of m^{6}A and Ψ in translation in KD samples

(a) Mean m^{6}A fractions of transcripts grouped by mean Ψ probability in the siCTRL and siMETTL3 polysome samples. Transcripts with ≥20 U sites and ≥7 A sites are distributed into 7 groups evenly (N=210/200 each for groups 1-6, N=204/195 each for group 7 in siCTRL/siMETTL3 polysome sample respectively) based on mean Ψ probability of the sites in each transcript; group 1 has the lowest and group 7 the highest mean Ψ probability. (b) Modification state combinations in input and polysome for all A sites in siMETTL3 sample and their corresponding mean translation efficiency (TE). A: unmodified, m^{6}A: modified. (c) Mean Ψ probability of transcripts grouped by m^{6}A fractions in the siCTRL and siMETTL3 polysome samples. Transcripts with ≥20 U sites, ≥7 A sites containing at least one m^{6}A sites are distributed into 7 groups evenly (N=100/89 each for groups 1-6, N=96/84 for group 7 in siCTRL and siMETTL3 polysome sample respectively) based on mean m^{6}A fraction of the sites in each transcript; group 1 has the lowest and group 7 the highest m^{6}A fraction. (d) TE of siCTRL and siMETTL3 transcripts grouped by differential mean Ψ probability. Transcripts with ≥20 U sites in both input and polysome samples are distributed into 7 groups evenly (N=883, 877, 880, 880, 877, 883, 876 for siCTRL and N=884, 887, 881, 881, 881, 881, 875 for siMETTL3) based on Ψ probability difference between polysome and input, with group 1 having the lowest and group 7 the highest TE. (e) Mean Ψ probability of transcripts grouped by m^{6}A fractions in the siTRUB1 input and polysome samples. Transcripts with ≥20 U sites, ≥7 A sites containing at least one m^{6}A site are distributed into 7 groups evenly (N=106/109 each for groups 1-6, N=101/106 for group 7 in siTRUB1 input/polysome sample respectively) based on mean m^{6}A fraction of the sites in each transcript; group 1 has the lowest and group 7

(Figure 4.5 continued) the highest m^6A fraction. (f) Mean m^6A fractions of transcripts grouped by mean Ψ probability in the siTRUB1 input and polysome samples. Transcripts with ≥20 U sites and ≥7 A sites are distributed into 7 groups evenly (N=249/226 each for groups 1-6, N=248/224 each for group 7 in siTRUB1 input/polysome sample respectively) based on mean Ψ probability of the sites in each transcript; group 1 has the lowest and group 7 the highest mean Ψ probability. (g) TE of siCTRL and siTRUB1 transcripts grouped by differential mean Ψ probability. Transcripts with ≥20 U sites in both input and polysome samples are distributed into 7 groups evenly (N=883, 877, 880, 880, 877, 883, 876 for siCTRL and N=794, 795, 791, 794, 794, 797, 786 for siTRUB1) based on Ψ probability difference between polysome and input; group 1 has the lowest and group 7 the highest TE. (h) Modification state combinations in input and polysome for all A sites in siTRUB1 samples and their corresponding mean translation efficiency (TE). A: unmodified, m^6A: modified. For **a, c-g**, p values are determined by two-sided Mann-Whitney U test, error bar represents 95% confidence interval (CI).

As for TRUB1 knockdown, which reduced Ψ levels in the input (**Fig. 4.3b**), Ψ levels of the polysome transcripts were down regardless of the m^6A group (**Fig. 4.5e**), but m^6A level remained similar or became slightly higher in polysome transcripts over input (**Fig. 4.5f**). TE value decreases upon TRUB1 knockdown were very significant in magnitude for almost all Ψ groups (**Fig. 4.5g**) regardless of the m^6A modification state (**Fig. 4.5h**), consistent with the large Ψ level decrease on the polysome. These results indicate that polysome accumulation of transcript m^6A and Ψ is synergistic, but with a hierarchical relationship of Ψ exerting a larger TE effect over m^6A.

4.3 Discussion

We identified an opposing effect of m^6A and Ψ in total mRNA input, but a synergistic effect of m^6A and Ψ for polysome-associated mRNA. Furthermore, m^6A and Ψ have a hierarchical effect on promoting translation efficiency in which Ψ takes precedent over m^6A. However, we should also be cautious about the phenomenon shown in this chapter, as this is so far the only research to check the relationship of pseudouridine and m^6A and their combined

effect on translation. All the observations are not yet further validated by wet lab experiments or convinced in other cell lines. It is possible that some of the conclusions are just occasionally happening in these samples. It is hopeful that in the future, there will be NGS based or other nanopore based simultaneous mapping method for pseudouridine and m^6A, which could be used to confirm or deny the observations of this study.

Among all the results, we could make some conclusions that happens all the time. The first major conclusion is that the anti-correlation relationship of m^6A and pseudouridine happens in all 6 samples, no matter whether it's input or polysome, or whether it is ctrl or knockdown sample. The second on is that both m^6A and pseudouridine have positive effect on translation efficiency, which also happens in all conditions. The conclusions concerning the knockdown of the two writers are less convincing or significant. Knock down of METTL3 results in significantly decrease in m^6A level, as METTL3 is the key component of the only m^6A writer complex. However, in some results, the knock down of TRUB1 results in modest effect. This is probably due to the existence of 12 other pseudouridine synthase and at least 3 of them (PUS1, PUS7, TRUB2) are proved to also work on human mRNA. The redundancy in pseudouridine synthase probably make the change brought about by TRUB1 knock down rescued by other writers. In the meantime, it is challenging to knock down or knockout all the pseudouridine synthases at the same time. Thus, the studies on pseudouridine depletion state in cells remains difficult to realize.

Most of the observations are based on the average values from each transcript but not from single sites. We did not find significant conclusions for the relative positions of m^6A and pseudouridine sites. Biologically, it could mean that either or both m^6A and/or pseudouridine do not function as individuals and only accumulation of modifications within a range could result in

significant biological consequence. In the paper of GLORI, the researchers find that m^6A are not evenly distributed along the transcripts and tend to form clusters, which may indicate the similar conclusion (C. Liu et al., 2023). Technically, the base calling accuracy of nanopore direct RNA sequencing data and the accuracy of our models for m^6A and pseudouridine prediction are not perfect, which may result in errors on the prediction of individual sites. However, as long as the errors are random errors without systematic bias, average results along the transcripts could largely reduce such random errors and show more accurate conclusions.

4.4 Methods

4.4.1 Cell culture and siRNA knockdown

Human embryonic kidney (HEK) HEK293T/17 cells (CRL11268) were cultured in Dulbecco's Modified Eagle's Medium (DMEM) with high glucose and L-glutamine, without sodium pyruvate (HyClone, SH30022.01) and with 10% FBS in a 37°C incubator at 5% CO_2 to seed for reverse transfection by RNAiMax (Sigma 13778150). For each knockdown condition, 75 µL Lipofectamine RNAiMax was added to three separate 150 mm plates containing 300 pmol siRNA (siCTRL, Sigma-Aldrich SIC001-10NMOL MISSION® siRNA Universal Negative Control #1, proprietary sequence; siMETTL3, Sigma-Aldrich PDSIRNA5D SASI_HS_00044317, duplex of GAUCCUAGAGCUAUUAAAU[dT][dT] and AUUUAAUAGCUCUAGGAUC[dT][dT]; siTRUB1, Sigma-Aldrich PDSIRNA5D SASI_Hs02_0036419, duplex of GAGUUCUGGUUGUUGGAAU[dT][dT] and AUUCCAACAACCAGAACUC[dT][dT]) in 5 mL Opti-MEM™ I Reduced Serum Medium (31985070) and incubated at 37°C with 5% CO_2 for 20 minutes. HEK293T cells grown to 80% confluency were washed and detached in 1x Phosphate Buffer Saline (PBS) before pelleting by

centrifugation at 500 g for 3 minutes. The cell pellet was resuspended in media and cells were counted using an Invitrogen™ Countess™ 3 FL Automated Cell Counter, for which a 10 µL aliquot of cells was mixed with 10 µL trypan blue and loaded into a chamber slide. For control and METTL3 knockdowns, ~3.5 x 10^6 cells were seeded into three 150 mm plates containing the appropriate siRNA-lipofectamine mixture. As these conditions were not viable for TRUB1 knockdown cells, ~5 x 10^6 cells were seeded instead. The plates were mixed by gently rocking and incubated in a 37°C with 5% CO_2 for 72 hours.

4.4.2 Polysome Profiling

Polysome profiling procedures were adapted from a previous publication (X. Wang et al., 2014). For each knockdown condition, three 150 mm plates of ~80% confluent HEK293T cells were treated with 100 µg/mL cycloheximide (CHX AC3574200500, Fisher Scientific) for 7 minutes at 37°C. Media was removed from the plates and the cells were washed twice with 10 mL of ice-cold 1x PBS containing 100 µg/mL CHX prior to being scraped and collected in 5 mL 1x ice-cold PBS. Cells were pelleted by centrifugation at 500 g for 5 minutes, then the three plates for each knockdown conditions were combined in 0.8 mL Lysis Buffer (20 mM HEPES, pH 7.6, 100 mM KCl, 5 mM $MgCl_2$, 1% Triton X-100, 100 µg/mL CHX supplemented with fresh 1x Roche protease inhibitor and 1% Superase inhibitor) and lysed by rotating at 4°C for 20 minutes. Cell debris was pelleted by 15-minute centrifugation at 16,000 g. To this lysate, 4 µL T4 Turbo DNAse (Invitrogen, AM2238) was added, and the mixture was incubated at room temperature for 15 minutes. 180 µL of this lysate was saved as “Input” for RNA downstream nanopore sequencing, and 20 µL was saved for western blot (see below).

5-50% sucrose gradient (20 mM HEPES, pH 7.6, 100mM KCl, 5 mM $MgCl_2$, 100 µg/mL CHX supplemented with fresh 1x Roche protease inhibitor and 1% SUPERase Inhibitor) was prepared in SETON 7042 tubes using a Biocomp Gradient Station. 600 µL sucrose gradient was removed from the top of each balanced sucrose gradient tube and then replaced by gently pipetting 600 µL of the respective knockdown cell lysate on top the gradient. Sucrose gradients were centrifuged for 3 hours at 1.41 x 10^5 g in an Optima L-100XP centrifuge using a Beckman SW28 rotor. Sucrose gradient fractions were collected and absorbances continuously measured using a Biocomp gradient station.

For each knockdown replicate, the 30 generated fractions were split in half and 0.9 mL TRIzol™ Reagent was added to each tube. Samples were incubated for 5 minutes at room temperature prior to addition of 0.18 mL chloroform and an additional 3-minute incubation. The aqueous layer was separated by centrifugation for 15 minutes at 12,000 g and 4°C and then added to a new tube. The sample was frozen overnight at -80°C following addition of 0.45 mL isopropanol. RNA was precipitated by centrifugation for 15-minute at 12,000 g and 4°C. The supernatant was removed, and the RNA pellet resuspended and washed in 0.9 mL 75% ethanol before being centrifuged for 5 minutes at 7,500 g and 4°C. The supernatant was removed and the pellet air-dried for ten minutes prior to resuspension in 10 µL dH_2O. The two tubes for each fraction were combined prior to combining all fractions disome and after.

Input and polysome RNA samples were made to 150 µL in dH_2O and cleaned by adding 280 µL Beckman RNAClean XP beads. Samples were mixed by pipetting up and down, incubated at room temperature for 5 minutes, pelleted on a magnetic rack for 5 minutes, washed with 1000 µL 70% EtOH three times, and air-dried for ten minutes before eluting into 150 µL dH_2O. Poly(A)-selection of cleaned RNA samples was then done using Promega PolyATract

mRNA Isolation System IV per the manufacturer's protocol. Briefly, RNA samples were made to 500 µL and incubated at 65°C for 10 minutes before addition of 3 µL Biotin-dT and 1x SSC. Once cooled, this sample was added to 100 µL of washed poly(A) magnetic beads in 0.5x SSC. Following a 10-minute incubation at room temperature, RNA-bound beads were captured using a magnetic rack, washed in 0.1X SSC, and eluted in a total of 250 µL dH_2O.

PolyA+ RNA was concentrated using Zymo Oligo Clean & Concentrator columns per manufacturer's protocol prior to Nanopore sequencing. Briefly, 500 µL Oligo Binding Buffer and 2 mL absolute ethanol were added to 250 µL PolyA+ RNA and mixed by pipetting. The sample was transferred to a Zymo-Spin™ IC Column in a collection tube and centrifuged. The sample was washed using 750 µL DNA Wash buffer prior to elution in 11 µL nuclease-free dH_2O. For quality control, polyA+ RNA was submitted to the University of Chicago Genomics facility and analyzed using an Agilent 2100 Bioanalyzer system. 5 µL of RNA containing ~1 ng/µL was analyzed using the RNA Pico/High Sensitivity Assay (input sensitivity of 0.05-5 ng/µL) to confirm RNA integrity.

4.4.3 Western Blot

Samples were prepared by adding 1x LDS and 100 mM DTT before boiling at 95°C for 5 minutes. Samples were loaded onto 12-well 4–12% polyacrylamide Bis-Tris gels (NP03322, Invitrogen) and ran at 150V for 1 hour. The gels were then transferred to polyvinylidene fluoride membranes (IPVH00010, Millipore). The membranes were blocked overnight in 10% w/v milk (1706404, Bio-Rad). The blots were probed with 1/1000 v/v anti-actin (clone C4 MAB1501), 1/1000 v/v anti-METTL3 (ab195352), or 1/500 v/v anti-TRUB1 (1250-1-AP) in 5% w/v milk (1706404, Bio-Rad) followed by 1/10000 v/v sheep anti-mouse IgG (NA931V, Cytiva) or

1/10000 v/v donkey anti-rabbit IgG conjugated to horseradish peroxidase (NA934V, Cytiva) in 5% w/v milk (1706404, Bio-Rad). The blots were then visualized with ECL Prime Western Blotting Detection Reagents (RPN2232, Amersham) using a BioRad ChemiDoc MP.

4.4.4 Nanopore direct RNA sequencing

The library preparation of direct RNA seq samples followed instructions from Oxford Nanopore Technology for Direct RNA Sequencing Kit (SQK-RNA002). Concisely, 500 ng of Poly(A)$^+$ RNA sample was used to perform a run. The RT Adaptor (RTA) was ligated to the 3' end of Poly(A)$^+$ RNA using T4 DNA ligase (NEB M0202S), followed by reverse transcription by SuperScript III Reverse Transcriptase (ThermoFisher 12574018). The RT product was then purified by 1.8x RNAClean XP beads (72 μL) (Beckman Coulter A63987). A second RNA Adaptor (RMX) was then attached to the 3' end of Poly(A)$^+$ RNA by T4 DNA ligase (NEB M0202S). The RNA product was purified with 1x RNAClean XP beads (40 μL) and eluted with 21 μl Elution Buffer. The sample was loaded onto a R9.4.1 flow cell (FLO-MIN106D) and then run on the MinION sequencer for 72 hours.

4.4.5 Nanopore data pre-processing

Raw sequencing data files were uploaded to UChicago midway2 cluster for pre-processing. Base calling was performed by guppy base caller (version 3.2.2+9fe0a78). Then, the reads were aligned to human genome (GRCh38.p13) by minimap2(H. Li, 2018) (version 2.18-r1015) with parameters -ax splice -uf -k14. The mapped reads were piled up by samtools (H. Li et al., 2009) (v1.11) and features for modifications prediction were extracted by customized python scripts (https://github.com/sihaohuanguc/NanoSPA).

4.4.6 HEK293T cell data processing

The HEK293T samples were sequenced and pre-processed as described above. The mapped reads of two biological replicates were combined to increase the data analysis throughput. We obtained 1.1-3.6 million mapped reads for the input, and 1.2-1.6 million mapped reads for the polysome samples. For Ψ and m^6A prediction, features of all U sites with >20 coverage and all A sites within the 8 motifs with >20 coverage in all 5 nucleotides in each motif were used for prediction by NanoSPA. The expression counts of transcripts were calculated as the maximum peak height of the reads piled at the transcript regions. Relative expression level of a transcript was calculated as the expression count of a transcript divided by the sum of expression counts of the sample. Transcripts with <15 coverage were filtered. Translation efficiency (TE) of a transcript was calculated as its level in polysome sample divided by its level in the input. The gene information was provided by the comprehensive gene annotation file (gencode.v41.annotation.gff3) in the GENCODE database (https://www.gencodegenes.org) (A. Frankish et al., 2021). Gene ontology (GO) analysis was performed using the Gene Ontology Resource (http://geneontology.org) (Ashburner et al., 2000; "The Gene Ontology resource: enriching a GOld mine," 2021). All p values were calculated by Two-sided Mann-Whitney U test unless noticed otherwise.

For transcripts grouped based on Ψ probability and the y-axis being "m^6A fraction", samples were screened for transcripts containing ≥ 20 U sites and ≥ 7 A sites in the 8 motifs. Then, transcripts were sorted and divided into 7 groups with even group sizes, with transcripts in group 1 possessing lowest and transcripts in group 7 highest mean Ψ probability.

The m^6A fraction of a transcript was defined as the number of m^6A sites in the transcript divided by the sum of number of A and m^6A sites in the transcript. For transcripts grouped based on m^6A fraction and the y-axis being "Ψ probability", samples were screened for transcripts with ≥ 20 U sites and ≥ 7 A sites in the 8 motifs in which at least one A was modified to m^6A. Then, transcripts were sorted and divided into 7 groups with even group sizes, with transcripts in group 1 possessing lowest and transcripts in group 7 highest m^6A fraction.

For TE calculation, the transcript must have ≥ 15 coverage in both input and polysome. For transcripts grouped based on TE and the y-axis being "m^6A fraction", samples were screened for transcripts containing ≥ 7 A sites in the 8 motifs in which at least one A was modified to m^6A in input or polysome. For transcripts grouped based on differential mean Ψ probability and the y-axis being TE, samples were screened for transcripts with ≥ 20 U sites in both input and polysome, the differential mean Ψ probability was defined as the difference of mean Ψ probability between polysome and input. Then, transcripts were sorted and divided into 7 groups with even group sizes, with transcripts in group 1 possessing lowest and transcripts in group 7 highest differential mean Ψ probability. For the heatmaps of TE comparison for m^6A and A sites, each A site in both input and polysome was assigned to four groups by its modification state in input and polysome: "A in both input and polysome", "A in input and m^6A in polysome", "m^6A in input and A in polysome", and "m^6A in both input and polysome". Then, the mean TE of all the transcripts corresponding to the A sites in each group were shown in the heatmaps.

Chapter 5. Single read analysis reveals stoichiometry and co-occurrence of pseudouridine

5.1 Introduction

RNA modification fractions could be told from some previous methods by calibration curves (L. Hu et al., 2022). However, the modification state at a site of each RNA molecule is still unknown. When we have two partially modified sites, we do not know whether the RNA molecules tend to have either both modified sites or both unmodified sites, or the molecules with modification on one site tend to be unmodified on the other. The bulk read analysis collect information from all reads and calculate averaged values as features, for example mutation rates. During this process, the information in each single molecule is also averaged out and omitted. For nanopore sequencing data, it's possible to analyze the information of each nucleotide in the full transcript and thus it's possible to know the modification state for each read of two distant sites on the same transcript. Previously, the nanoRMS protocol collects features from single reads, but the single read features were averaged before Ψ prediction, erasing single molecule Ψ site incorporation information (O. Begik et al., 2021).

In this chapter, we are going to show the development of a single read pseudouridine prediction model based on nanopore sequencing data and the application of the model on prediction modification stoichiometry and multi-site linkages.

5.2 Results

5.2.1 Development of the model

We performed single read analysis for quantitative Ψ stoichiometry prediction and investigation of linking modification states of Ψ sites in single molecules of a mRNA transcript. To realize single read analysis, we need to develop a model which take features from each U sites from a read and do prediction. The first problem to solve is to find out reads of pseudouridine and uridine. In the samples, usually the modified sites are only partially modified, which means in all the reads that cover this site, only some of the read provide signals of pseudouridine and the rest provide signals of unmodified U. Such modified sites could not provide high quality training material. Only in 100% modified sites, every read covering this site provide a signal of the modified base. Our training set contained the data points from previously reported (Masato Taoka et al., 2018), 100% modified human rRNA Ψ sites and randomly selected unmodified human rRNA U sites.

We used the same set of features and the same EXT algorithm as NanoPsu to train the model for single read pseudouridine prediction, while we replaced all the “rate” features (like “mismatch ratio”) with “indicator” features (like “mismatch-or-not”). It is much harder to generate good single read prediction model as the information that could be used is much less for each prediction. The model still predicted most known pseudouridine and unmodified U data points correctly in the testing set (**Fig. 5.1a**), although the peaks are wider than the bulk prediction model. The AUC value for the prediction of testing set was 0.8269.

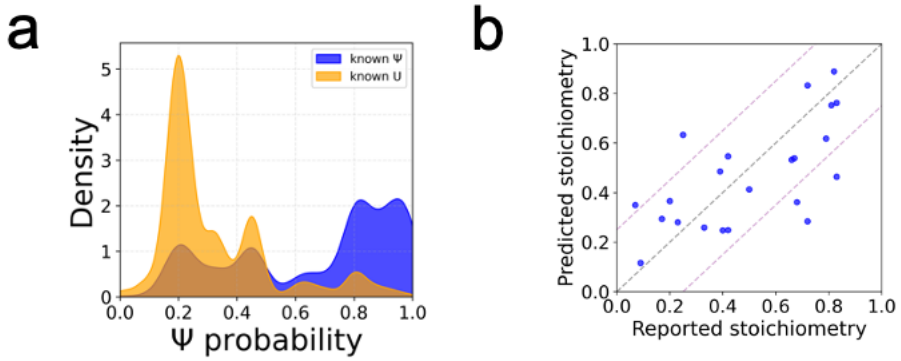

Figure 5.1 Ψ single read prediction model training and stoichiometry calculation
(a) Density plot of predicted Ψ modification probabilities of U and Ψ data points of single read prediction in the rRNA testing set. (b) Single read prediction results for the partially modified Ψ sites in human rRNA. The stoichiometry predicted by our method is compared with the stoichiometry reported previously by quantitative LC/MS. The correlation coefficient is 0.6566 (Pearson's r).

The size of data is much larger for single read prediction, as each U site in each read will result in a prediction. When the work was published in 2021, the algorithm could only be applied to specific transcript by extraction the reads covering the sepcific transcript in advance and then do feature extraction and prediction. The protocol could not be completed in finite time on the whole human transcriptome. Later the algorithm was optimized, and it could be applied to the whole human transcriptome and be completed in several hours.

5.2.2 Prediction of stoichiometry

Since we have the ability to predict the modification state of each read covering a specific site, we could calculate the ratio of data points predicted as modified at a specific site and view it as the stoichiometry at this site. We tested the Ψ stoichiometry prediction from single reads on 22 partially modified Ψ sites (5%-85%) in human rRNA. These sites are partially modified and are not involved in the training process of the single read pseudouridine prediction model above. We found that the predicted stoichiometry is correlated with the previous reported stoichiometry

obtained by LC/MS (Masato Taoka et al., 2018) (**Fig. 5.1b**). This indicates that single read analysis could provide a new strategy for quantification of modification fractions.

5.2.3 Linkage among sites

A new application of the single read prediction method is the ability to perform single read analysis that links occurrence of multiple Ψ sites in individual mRNA transcripts. We examined whether pairs of Ψ site modifications are linked either positively or negatively, meaning whether the modification state of site 2 is affected by the modification state of site 1 and vice versa. We selected 31 positions in the B2M transcript (which encodes the common small subunit of MHC class I molecules) for investigation.

We show two examples of site pairs in **Fig. 5.2a**. In the example on the left, when the first site is modified (dark blue), the second site tends to be unmodified. When the second site is modified, the first site tends to be unmodified. In these two sites, pseudouridine tends to avoid appearing at the same time and these two sites are viewed as negatively linked. In the example on the right, the distributions of the modification state of the second site are the same when the first is either modified or unmodified. This is an example of independent pair, which means the modification states of the two sites are independent from each other. We could use cumulative distribution curves to describe and visualize the different types of linkage (**Fig. 5.2b**). The reads are assigned into two groups based on their modification states at the first position and the cumulative distribution curves of the pseudouridine probabilities of the second site are shown. When the two sites are either positively or negatively linked, the two curves will be separated from each other (**Fig. 5.2b**, left panel). When the two sites are independent from each other, the

two curves will overlap with each other (**Fig. 5.2b**, right panel). Then the patterns could be quantified by two sample Kolmogorov-Smirnov test.

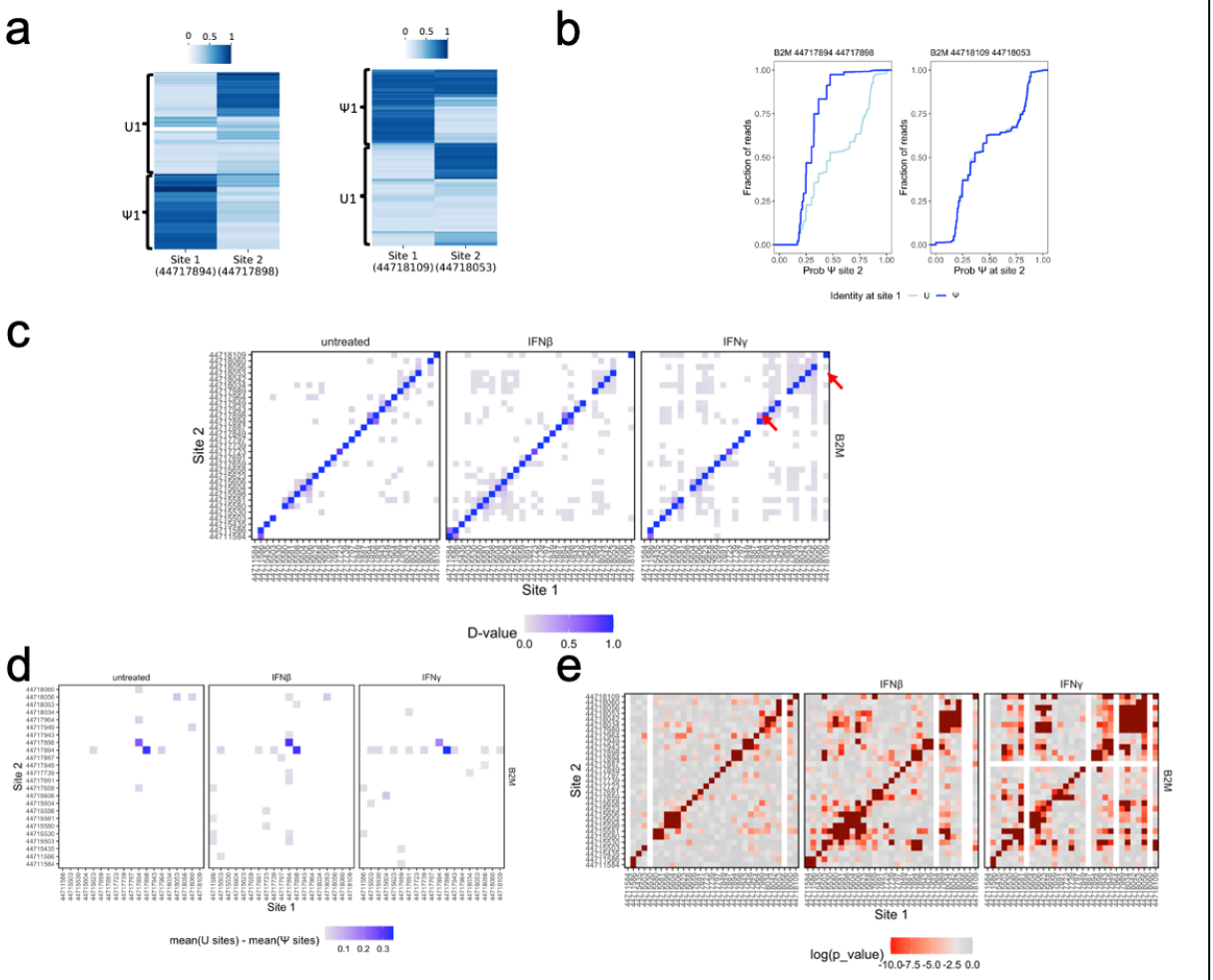

Figure 5.2 Linkage analysis of multiple Ψ sites

(a) Clustering heatmap showing the Ψ probability of two pairs of sites in single reads of the B2M transcript in the IFN γ treated sample. Each row represents a read. Site numbers are defined as the chromosomal locations in the hg38 nomenclature. These two pairs show either negative linkage (left) or no linkage (right). (b) Reads in Fig. 5.2a are assigned to "Ψ" and "U" groups based on the posterior probabilities of site 1 in Gaussian mixture model (k=2). The cumulative distribution curves of Ψ probabilities of site 2 are drawn for reads in "Ψ" or "U" groups or for all reads. The curves for "Ψ" and "U" groups undergo two sample Kolmogorov-Smirnov test; p values are $<2.2\text{x}10^{-16}$ (left) and 0.7684 (right). (c) Value of the maximum distance D (p<0.01) in the two sample Kolmogorov-Smirnov test for selected pair of sites in the B2M transcript in IFN treated and untreated samples. X axis represents site 1 and Y axis represents site 2. Site numbers are defined as the chromosomal locations in the hg38 nomenclature. The two pair of sites shown in Fig. 5.2a and 5.2b are indicated by red arrows. (d) Maximum distance D value in the two sample Kolmogorov-Smirnov test for pair of sites when D >0 and p<0.01 in the B2M transcript in IFN treated and untreated samples. X axis represents site 1 and Y axis represents site 2. (e) P value in the two sample Kolmogorov-Smirnov test for selected pairs of sites in the B2M transcript in the untreated and IFN treated samples.

In most cases, the maximum distance *D* value from two sample K-S test was small (**Fig. 5.2c**), which is consistent with the presence of Ψ at site 1 being independent of Ψ at site 2, these two sites are not linked. A few pairs of sites had high *D* values, but most of those were immediately adjacent Ψ sites. The diagonal has the highest *D* values as every site is perfectly positively linked with itself. The pairs of Ψ sites with negative linkage tend to avoid each other in the same mRNA molecule (**Fig. 5.2d**). This result indicates that the modification of Ψ at two sites in single molecule transcripts is negatively related for some, and completely independent for others. Upon IFN treatment, the linkage between some sites in the B2M transcript became more prominent (**Fig. 5.2e**), suggesting that IFN-induced Ψ installation has stronger co-dependency.

5.3 Discussion

In this chapter we developed a pseudouridine prediction model based on single reads. As the information is much less to predict based on single reads, the performance of the model is not as good as bulk prediction models. Also, as the size of data massively increases, it is much more challenging to process the data on the whole human transcriptome and do the downstream analysis. Thus, example analysis on specific transcripts is provided.

The two major applications of single read pseudouridine prediction is the evaluation of stoichiometry and the investigation of linkages between different sites. Although single read analysis could evaluate stoichiometry, but its efficiency is relatively lower than bulk read regression models and thus it's hard to make it into practical use. The ability to describe site-site linkage is useful, which may reflect the co-regulation of distant sites. These sites maybe distant in the one-dimensional sequence but could be close to each other in the secondary structure and

thus could be biologically meaningful. The application of such analysis need to be further revealed in the future experiments.

5.4 Methods

5.4.1 Single read Ψ prediction model training

The 100% modified human rRNA sites were reported in a previously work measured by quantitative LC/MS (Masato Taoka et al., 2018). A basic assumption was that all reads in our human rRNA sample would have Ψ at the reported 100% modified sites and U at the reported completely unmodified sites. The dataset for training contained 25 100% Ψ sites with 49,437 data points and 26 randomly selected U sites with 50,922 data points. The dataset was divided into 60% training set, 20% validation set and 20% testing set. Features were extracted from each base in each read. The features describing the ratios in bulk prediction model were replaced with features indicating the mismatching and indel states of the base. The Ψ modification prediction models were generated by training set and validated with the validation set using the EXT algorithm (n_estimators=200, criterion="gini", max_depth=None, min_samples_split=2) with 10 features, which are insertion_ot_not, insertion_length, deletion_or_not, deletion_length, deleted_site_or_not, mismatch_or_not, mutate_to_A, mutate_to_C, mutate_to_G, base quality score. The AUC value for the prediction of testing set was 0.8269. To further evaluate the model, Ψ modification probabilities of data points from 22 previously reported (Masato Taoka et al., 2018), partially modified human rRNA Ψ sites (modification fraction from 5% to 85%) were predicted. The base was viewed as Ψ when the probability was larger than 0.5 and as U when the probability was less than 0.5. The stoichiometry of each site was calculated as the number of predicted Ψ bases divided by the coverage of the site.

5.4.2 Single read Ψ analysis in HeLa samples

The Ψ probabilities of all U residues in selected genes were predicted with the protocol above. To investigate the linkage of multiple Ψ on single reads, each read was indexed so that the U data points with the same read index were from the same read. Ψ probabilities of residues of a certain site were fitted by Gaussian mixture model (GMM) with 2 components. The sites with abs(μ1-μ2)>0.5 and λ1, λ2>0.05 were selected for following analysis. When doing pair wise linkage analysis, the reads were assigned into "Ψ" and "U" groups when it had >95% posterior probability for one population in the GMM for site 1. To evaluate whether there was a difference in the Ψ probabilities distribution of site 2 upon the presence or absence of Ψ at site 1, two sample Kolmogorov-Smirnov test was performed on the Ψ probabilities cumulative distribution curves of site 2 in the "Ψ" and "U" groups with an output of the maximum distance D value and p value. The R library to do two sample Kolmogorov-Smirnov test was from GitHub (https://rdrr.io/github/happyrabbit/DataScienceR/man/pairwise_ks_test.html).

Chapter 6. Conclusions and Perspectives

6.1 Better mapping methods for RNA modifications

RNA sequencing methods for RNA modification identification requires special designs and remains challenging. On one hand, there are mapping methods available for prevalent modifications like m^6A and pseudouridine, but none is perfect. As for identification resolution, accuracy, quantifiability, low required input amount, multi-sites coordination, single cell capacity, all methods manage to achieve some by compromising the others. New methods for these modifications come out with the improvement of one or several aspects. On the other hand, more modifications are detected in RNA, especially mRNA, by methods like mass spectrometry, but whole transcriptome distribution and modification fractions of single sites are unknown. Both conditions require novel strategies to map the modifications in transcriptome. There is plenty of demand for more and better RNA modification mapping methods, which will facilitate the future studies of epitranscriptomics.

Ideally, for NGS based methods, the best strategy for modification identification is to change the signals of the modified nucleotides and maintain the signals of the unmodified ones. This requires a reaction that could happen to the modified but not the unmodified nucleotides. The ideal situation is that the reaction on the unmodified nucleotides is $<10^{-4}$(X) preference compared with modified nucleotides. This is due to the fact that the expected number of unmodified nucleotides is usually 200-5000 times of the modified ones. It is less recommended to change the signals of unmodified nucleotides and maintain the signals of modified ones, because this requires the yield of reaction on unmodified nucleotides as close as possible to 100%(1-X), otherwise the small fraction of unreacted unmodified nucleotides will dominate the

positive predictions. For both directions, the purpose is to make the "X" as small as possible. Of course, there are ideas to jump out of the frame. For example, if both the signals of modified and unmodified nucleotides could be changed and their outputs are different and only those nucleotides with signal changes are counted, then there is no need to worry about the magnitude of X. It is hopeful that such strategies could be developed in the future for high accuracy RNA modification identification.

There is still much room for both NGS and TGS based strategies to generate better modification identification methods. It is also possible to have totally novel concepts in sequencing technologies and then apply them to modification identification and quantitation. Each technology has its own advantages and disadvantages. Currently there is no trend for TGS to replace NGS as both TGS and NGS have unique advantages and disadvantages. It's the same case for modification detection. NGS based methods and TGS based methods are more like complement to each other to satisfy high accuracy and long read length. If the advantages of NGS and TGS could be combined to study RNA modifications in the future studies, there will be more detailed and concrete discoveries.

6.2 Nanopore sequencing of more RNA modifications

As we mentioned in the results, the NanoSPA pipeline could be extended to more modifications using the same feature space. For modifications beyond m^6A and pseudouridine, there are few publications working on their whole human transcriptome mapping. Technically, there are some problems for developing nanopore models for other modifications. As there is less prior knowledge on other modifications, it will be more challenging to train good models for nanopore sequencing data. Also, there are fewer cross validations from different studies for the

previously discovered modification sites and for some modifications there are even conflicts on whether they appear in mRNA or not. Thus, the quality of training material could not be ensured. Third, the abundance of other modifications in mRNA are evenly lower than m^6A and pseudouridine, which means the prediction is based on an extremely unbalanced dataset and it could result in a large fraction of false positive predictions. If the modification is not preferred in any specific motifs, then the problem will be harder. Fewer abundance also means less training materials.

However, even if the technical problems could be solved, the major problem for other modifications remains. The currently available biological discoveries on other modifications are limited and it's not easy to find an application setting even if new methods are developed. In another word, NGS based methods seems to be plenty at this moment. If any of the other modifications could be found to have major functions in diseases or more biological pathways, then there will be more applications available for nanopore based methods. Also, in that case there will be more materials for cross validations and functional tests.

6.3 Coordinate of RNA modifications and other RNA events

One of the major advantages of nanopore sequencing is long read length. However, when we use nanopore sequencing for direct RNA modification detection, we seldom make the use of the advantage of long reads. Long reads enable the detection of large structure events or multiple events happening distant to each other on the same transcript, like alternative splicing and large structure variation of the genome. For nanopore sequencing data, it is possible to study whether the distributions of specific RNA modifications are correlated to those RNA events without the requirement of any extra pre-treatment on the samples. For example, m^6A at specific positions

are reported to be related to splicing regulation and such problems could be studied straightforwardly by nanopore direct RNA sequencing.

Also, currently multi-omics studies are prevalent, but RNA modifications are seldom considered as one of the omics. We have shown that RNA modifications play roles in biological processes like translation and there could be coordination. It is possible to coordinate the mRNA modifications with other genomic, transcriptomic and proteomic groups to better understand the process of gene expression and regulation.

6.4 Single read analysis

Whether the modifications affect each other on the same mRNA transcript remains a major biological question of the epitranscriptome field. Bulk RNA sequencing could only tell the overall modification fraction of each site, but the modification state in each molecule is unknown and is averaged out. Single read analysis could overcome these problems. We show the proof-of-concept results of our single read pseudouridine prediction model and apply it to analyze modification stoichiometry and multi-site linkage.

However, there are many unsolved problems in single read analysis. The prediction accuracy of single read pseudouridine model is fair, but not high enough to provide convincing biological discoveries. Also, although we observed exciting phenomena based on single read analysis, it is hard to use wet lab experiments to validate the observations. Third, the size of output is huge if the analysis is performed on the whole human transcriptome, and it is challenging to process and interpret the data. Currently, our compromised way is to show results within one specific transcript. It is hopeful that more approaches to interpret the single read data could be developed in the future.

The application of single read analysis is still promising. Personally, I do not recommend to use single read analysis to calculate stoichiometry for the whole transcriptome, as there are more practical ways to complete such tasks like using calibration curves for bulk reads. However, it will be useful to use single read analysis to investigate site-site linkage. The linkage of distant sites along a transcript may reflect secondary or higher dimension structure information, or regulation events happening spatially, which could not be revealed by one dimension information.

Reference

Addepalli, B., & Limbach, P. A. (2011). Mass spectrometry-based quantification of pseudouridine in RNA. *J Am Soc Mass Spectrom, 22*(8), 1363-1372. doi:10.1007/s13361-011-0137-5

Aganezov, S., Yan, S. M., Soto, D. C., Kirsche, M., Zarate, S., Avdeyev, P., . . . Xiao, C. (2022). A complete reference genome improves analysis of human genetic variation. *Science, 376*(6588), eabl3533.

Altemose, N., Logsdon, G. A., Bzikadze, A. V., Sidhwani, P., Langley, S. A., Caldas, G. V., . . . Shew, C. J. (2022). Complete genomic and epigenetic maps of human centromeres. *Science, 376*(6588), eabl4178.

Anderson, B. R., Muramatsu, H., Nallagatla, S. R., Bevilacqua, P. C., Sansing, L. H., Weissman, D., & Kariko, K. (2010). Incorporation of pseudouridine into mRNA enhances translation by diminishing PKR activation. *Nucleic acids research, 38*(17), 5884-5892.

Anderson, B. R., Muramatsu, H., Nallagatla, S. R., Bevilacqua, P. C., Sansing, L. H., Weissman, D., & Kariko, K. (2010). Incorporation of pseudouridine into mRNA enhances translation by diminishing PKR activation. *Nucleic Acids Res, 38*(17), 5884-5892. doi:10.1093/nar/gkq347

Arango, D., Sturgill, D., Alhusaini, N., Dillman, A. A., Sweet, T. J., Hanson, G., . . . Oberdoerffer, S. (2018). Acetylation of Cytidine in mRNA Promotes Translation Efficiency. *Cell, 175*(7), 1872-1886 e1824. doi:10.1016/j.cell.2018.10.030

Arango, D., Sturgill, D., Yang, R., Kanai, T., Bauer, P., Roy, J., . . . Oberdoerffer, S. (2022). Direct epitranscriptomic regulation of mammalian translation initiation through N4-acetylcytidine. *Molecular cell, 82*(15), 2797-2814. e2711.

Arnez, J. G., & Steitz, T. A. (1994). Crystal structure of unmodified tRNAGln complexed with glutaminyl-tRNA synthetase and ATP suggests a possible role for pseudo-uridines in stabilization of RNA structure. *Biochemistry, 33*(24), 7560-7567.

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., . . . Eppig, J. T. (2000). Gene ontology: tool for the unification of biology. *Nature genetics, 25*(1), 25-29.

Aw, J. G. A., Lim, S. W., Wang, J. X., Lambert, F. R., Tan, W. T., Shen, Y., . . . Ng, S. B. (2021). Determination of isoform-specific RNA structure with nanopore long reads. *Nature Biotechnology, 39*(3), 336-346.

Ayub, M., & Bayley, H. (2016). Engineered transmembrane pores. *Current Opinion in Chemical Biology, 34*, 117-126.

Bailey, T. L., Johnson, J., Grant, C. E., & Noble, W. S. (2015). The MEME suite. *Nucleic acids research, 43*(W1), W39-W49.

Bakin, A., & Ofengand, J. (1993). Four newly located pseudouridylate residues in Escherichia coli 23S ribosomal RNA are all at the peptidyltransferase center: analysis by the application of a new sequencing technique. *Biochemistry, 32*(37), 9754-9762.

Balacco, D. L., & Soller, M. (2018). The m6A writer: rise of a machine for growing tasks. *Biochemistry, 58*(5), 363-378.

Ballabio, A., & Gieselmann, V. (2009). Lysosomal disorders: from storage to cellular damage. *Biochimica et Biophysica Acta (BBA)-Molecular Cell Research, 1793*(4), 684-696.

Banerjee, A., Mikhailova, E., Cheley, S., Gu, L.-Q., Montoya, M., Nagaoka, Y., . . . Bayley, H. (2010). Molecular bases of cyclodextrin adapter interactions with engineered protein nanopores. *Proceedings of the National Academy of Sciences, 107*(18), 8165-8170.

Begik, O., Lucas, M. C., Pryszcz, L. P., Ramirez, J. M., Medina, R., Milenkovic, I., . . . Sas-Chen, A. (2021). Quantitative profiling of pseudouridylation dynamics in native RNAs with nanopore sequencing. *Nature Biotechnology*, 1-14.

Begik, O., Lucas, M. C., Pryszcz, L. P., Ramirez, J. M., Medina, R., Milenkovic, I., . . . Novoa, E. M. (2021). Quantitative profiling of pseudouridylation dynamics in native RNAs with nanopore sequencing. *Nat Biotechnol, 39*(10), 1278-1291. doi:10.1038/s41587-021-00915-6

Bi, Y., Jin, D., & Jia, C. (2020). EnsemPseU: identifying pseudouridine sites with an ensemble approach. *Ieee Access, 8*, 79376-79382.

Bizuayehu, T. T., Labun, K., Jakubec, M., Jefimov, K., Niazi, A. M., & Valen, E. (2022). Long-read single-molecule RNA structure sequencing using nanopore. *Nucleic acids research, 50*(20), e120-e120.

Boccaletto, P., Stefaniak, F., Ray, A., Cappannini, A., Mukherjee, S., Purta, E., . . . Bujnicki, J. M. (2022). MODOMICS: a database of RNA modification pathways. 2021 update. *Nucleic Acids Res, 50*(D1), D231-D235. doi:10.1093/nar/gkab1083

Bokar, J. A., Shambaugh, M. E., Polayes, D., Matera, A. G., & Rottman, F. M. (1997). Purification and cDNA cloning of the AdoMet-binding subunit of the human mRNA (N6-adenosine)-methyltransferase. *RNA, 3*(11), 1233-1247. Retrieved from https://www.ncbi.nlm.nih.gov/pubmed/9409616

Borchardt, E. K., Martinez, N. M., & Gilbert, W. V. (2020). Regulation and Function of RNA Pseudouridylation in Human Cells. *Annu Rev Genet, 54*, 309-336. doi:10.1146/annurev-genet-112618-043830

Borchardt, E. K., Martinez, N. M., & Gilbert, W. V. (2020). Regulation and function of RNA pseudouridylation in human cells. *Annual review of genetics, 54*, 309-336.

Brandmayr, C., Wagner, M., Brückl, T., Globisch, D., Pearson, D., Kneuttinger, A. C., . . . Thoma, I. (2012). Isotope-based analysis of modified tRNA nucleosides correlates modification density with translational efficiency. *Angewandte Chemie International Edition, 51*(44), 11162-11165.

Brown, C. G., & Clarke, J. (2016). Nanopore development at Oxford nanopore. *Nature Biotechnology, 34*(8), 810-811.

Brzezicha, B., Schmidt, M., Makałowska, I., Jarmołowski, A., Pieńkowska, J., & Szweykowska-Kulińska, Z. (2006). Identification of human tRNA: m5C methyltransferase catalysing intron-dependent m5C formation in the first position of the anticodon of the. *Nucleic acids research, 34*(20), 6034-6043.

Carlile, T. M., Rojas-Duran, M. F., & Gilbert, W. V. (2015). Pseudo-Seq: Genome-Wide Detection of Pseudouridine Modifications in RNA. *Methods Enzymol, 560*, 219-245. doi:10.1016/bs.mie.2015.03.011

Carlile, T. M., Rojas-Duran, M. F., Zinshteyn, B., Shin, H., Bartoli, K. M., & Gilbert, W. V. (2014). Pseudouridine profiling reveals regulated mRNA pseudouridylation in yeast and human cells. *Nature, 515*(7525), 143-146. doi:10.1038/nature13802

Charette, M., & Gray, M. W. (2000). Pseudouridine in RNA: what, where, how, and why. *IUBMB life, 49*(5), 341-352.

Chen, W., Tang, H., Ye, J., Lin, H., & Chou, K. C. (2016). iRNA-PseU: Identifying RNA pseudouridine sites. *Mol Ther Nucleic Acids, 5*(7), e332. doi:10.1038/mtna.2016.37

Cheng, S. H., Gregory, R. J., Marshall, J., Paul, S., Souza, D. W., White, G. A., . . . Smith, A. E. (1990). Defective intracellular transport and processing of CFTR is the molecular basis of most cystic fibrosis. *Cell, 63*(4), 827-834.

Choi, J., Ieong, K.-W., Demirci, H., Chen, J., Petrov, A., Prabhakar, A., . . . Soltis, S. M. (2016). N 6-methyladenosine in mRNA disrupts tRNA selection and translation-elongation dynamics. *Nature structural & molecular biology, 23*(2), 110-115.

Cohn, W. E. (1959). 5-Ribosyl uracil, a carbon-carbon ribofuranosyl nucleoside in ribonucleic acids. *Biochimica et biophysica acta, 32*, 569-571.

Cohn, W. E., & Volkin, E. (1951). Nucleoside-5′-phosphates from ribonucleic acid. *Nature, 167*(4247), 483-484.

Cui, X., Liang, Z., Shen, L., Zhang, Q., Bao, S., Geng, Y., . . . Lu, T. (2017). 5-Methylcytosine RNA methylation in Arabidopsis thaliana. *Molecular plant, 10*(11), 1387-1399.

Dai, Q., Ye, C., Irkliyenko, I., Wang, Y., Sun, H.-L., Gao, Y., . . . Goel, A. (2024). Ultrafast bisulfite sequencing detection of 5-methylcytosine in DNA and RNA. *Nature Biotechnology*, 1-12.

Dai, Q., Zhang, L.-S., Sun, H.-L., Pajdzik, K., Yang, L., Ye, C., . . . Zheng, Z. (2023). Quantitative sequencing using BID-seq uncovers abundant pseudouridines in mammalian mRNA at base resolution. *Nature Biotechnology, 41*(3), 344-354.

Davis, D. R. (1995). Stabilization of RNA stacking by pseudouridine. *Nucleic acids research, 23*(24), 5020-5026.

Davis, D. R., Veltri, C. A., & Nielsen, L. (1998). An RNA model system for investigation of pseudouridine stabilization of the codon-anticodon interaction in tRNALys, tRNAHis and tRNATyr. *Journal of Biomolecular Structure and Dynamics, 15*(6), 1121-1132.

Davis, F. F., & Allen, F. W. (1957). Ribonucleic acids from yeast which contain a fifth nucleotide. *J Biol Chem, 227*(2), 907-915.

Dean, J., & Monga'TensorFlow, R. (2015). Large-Scale Machine Learning on Heterogeneous Distributed Systems'. *TensorFlow. org*.

Decatur, W. A., & Fournier, M. J. (2002). rRNA modifications and ribosome function. *Trends in biochemical sciences, 27*(7), 344-351.

Delaunay, S., & Frye, M. (2019). RNA modifications regulating cell fate in cancer. *Nat Cell Biol, 21*(5), 552-559. doi:10.1038/s41556-019-0319-0

Desrosiers, R., Friderici, K., & Rottman, F. (1974). Identification of methylated nucleosides in messenger RNA from Novikoff hepatoma cells. *Proc Natl Acad Sci U S A, 71*(10), 3971-3975. doi:10.1073/pnas.71.10.3971

Ding, H., Bailey IV, A. D., Jain, M., Olsen, H., & Paten, B. (2020). Gaussian mixture model-based unsupervised nucleotide modification number detection using nanopore-sequencing readouts. *Bioinformatics, 36*(19), 4928-4934.

Dominissini, D., Moshitch-Moshkovitz, S., Schwartz, S., Salmon-Divon, M., Ungar, L., Osenberg, S., . . . Rechavi, G. (2012). Topology of the human and mouse m6A RNA methylomes revealed by m6A-seq. *Nature, 485*(7397), 201-206. doi:10.1038/nature11112

Dominissini, D., Nachtergaele, S., Moshitch-Moshkovitz, S., Peer, E., Kol, N., Ben-Haim, M. S., . . . Clark, W. C. (2016). The dynamic N 1-methyladenosine methylome in eukaryotic messenger RNA. *Nature, 530*(7591), 441-446.

Drexler, H. L., Choquet, K., & Churchman, L. S. (2020). Splicing kinetics and coordination revealed by direct nascent RNA sequencing through nanopores. *Molecular cell, 77*(5), 985-998. e988.

Dubin, D. T., & Taylor, R. H. (1975). The methylation state of poly A-containing-messenger RNA from cultured hamster cells. *Nucleic acids research, 2*(10), 1653-1668.

Dunn, D. (1961). The occurence of 1-methyladenine in ribonucleic acid. *Biochimica et biophysica acta, 46*(1), 198-200.

Durairaj, A., & Limbach, P. A. (2008). Mass spectrometry of the fifth nucleoside: a review of the identification of pseudouridine in nucleic acids. *Analytica chimica acta, 623*(2), 117-125.

El Yacoubi, B., Bailly, M., & de Crecy-Lagard, V. (2012). Biosynthesis and function of posttranscriptional modifications of transfer RNAs. *Annu Rev Genet, 46*, 69-95. doi:10.1146/annurev-genet-110711-155641

Engel, M., Eggert, C., Kaplick, P. M., Eder, M., Röh, S., Tietze, L., . . . Rex-Haffner, M. (2018). The role of m6A/m-RNA methylation in stress response regulation. *Neuron, 99*(2), 389-403. e389.

Euskirchen, P., Bielle, F., Labreche, K., Kloosterman, W. P., Rosenberg, S., Daniau, M., . . . Dehais, C. (2017). Same-day genomic and epigenomic diagnosis of brain tumors using real-time nanopore sequencing. *Acta neuropathologica, 134*, 691-703.

Everett, D. W. (1980). *Part i: reaction of pseudouridine with bisulfite. part ii: reaction of glyoxal with guanine derivatives: a spectrophotometric probe of molecular structure.* New York University,

Eyler, D. E., Franco, M. K., Batool, Z., Wu, M. Z., Dubuke, M. L., Dobosz-Bartoszek, M., . . . Koutmou, K. S. (2019). Pseudouridinylation of mRNA coding sequences alters translation. *Proc Natl Acad Sci U S A, 116*(46), 23068-23074. doi:10.1073/pnas.1821754116

Fleming, A. M., Mathewson, N. J., Howpay Manage, S. A., & Burrows, C. J. (2021). Nanopore dwell time analysis permits sequencing and conformational assignment of pseudouridine in SARS-CoV-2. *ACS Central Science*.

Fleming, A. M., Mathewson, N. J., Howpay Manage, S. A., & Burrows, C. J. (2021). Nanopore Dwell Time Analysis Permits Sequencing and Conformational Assignment of Pseudouridine in SARS-CoV-2. *ACS Cent Sci, 7*(10), 1707-1717. doi:10.1021/acscentsci.1c00788

Flusberg, B. A., Webster, D. R., Lee, J. H., Travers, K. J., Olivares, E. C., Clark, T. A., . . . Turner, S. W. (2010). Direct detection of DNA methylation during single-molecule, real-time sequencing. *Nature methods, 7*(6), 461-465.

Frankish, A., Diekhans, M., Jungreis, I., Lagarde, J., Loveland, J. E., Mudge, J. M., . . . Barnes, I. (2021). GENCODE 2021. *Nucleic acids research, 49*(D1), D916-D923.

Frankish, A., Diekhans, M., Jungreis, I., Lagarde, J., Loveland, J. E., Mudge, J. M., . . . Flicek, P. (2021). Gencode 2021. *Nucleic Acids Res, 49*(D1), D916-D923. doi:10.1093/nar/gkaa1087

Frye, M., Harada, B. T., Behm, M., & He, C. (2018). RNA modifications modulate gene expression during development. *Science, 361*(6409), 1346-1349. doi:10.1126/science.aau1646

Gao, Y., Liu, X., Wu, B., Wang, H., Xi, F., Kohnen, M. V., . . . Gu, L. (2021). Quantitative profiling of N(6)-methyladenosine at single-base resolution in stem-differentiating xylem of Populus trichocarpa using Nanopore direct RNA sequencing. *Genome Biol, 22*(1), 22. doi:10.1186/s13059-020-02241-7

Garalde, D. R., Snell, E. A., Jachimowicz, D., Sipos, B., Lloyd, J. H., Bruce, M., . . . Warland, A. (2018). Highly parallel direct RNA sequencing on an array of nanopores. *Nature methods, 15*(3), 201.

Garcia-Campos, M. A., Edelheit, S., Toth, U., Safra, M., Shachar, R., Viukov, S., . . . Brandis, A. (2019). Deciphering the “m6A code” via antibody-independent quantitative profiling. *Cell, 178*(3), 731-747. e716.

Garus, A., & Autexier, C. (2021). Dyskerin: an essential pseudouridine synthase with multifaceted roles in ribosome biogenesis, splicing, and telomere maintenance. *RNA, 27*(12), 1441-1458.

Ge, J., & Yu, Y.-T. (2013). RNA pseudouridylation: new insights into an old modification. *Trends in biochemical sciences, 38*(4), 210-218.

The Gene Ontology resource: enriching a GOld mine. (2021). *Nucleic Acids Research, 49*(D1), D325-D334.

Gershman, A., Sauria, M. E., Guitart, X., Vollger, M. R., Hook, P. W., Hoyt, S. J., . . . Koren, S. (2022). Epigenetic patterns in a complete human genome. *Science, 376*(6588), eabj5089.

Girardot, C., Scholtalbers, J., Sauer, S., Su, S.-Y., & Furlong, E. E. (2016). Je, a versatile suite to handle multiplexed NGS libraries with unique molecular identifiers. *BMC bioinformatics, 17*(1), 1-6.

Goll, M. G., Kirpekar, F., Maggert, K. A., Yoder, J. A., Hsieh, C.-L., Zhang, X., . . . Bestor, T. H. (2006). Methylation of tRNAAsp by the DNA methyltransferase homolog Dnmt2. *Science, 311*(5759), 395-398.

Greener, J. G., Kandathil, S. M., Moffat, L., & Jones, D. T. (2022). A guide to machine learning for biologists. *Nature reviews Molecular cell biology, 23*(1), 40-55.

Grünberger, F., Ferreira-Cerca, S., & Grohmann, D. (2022). Nanopore sequencing of RNA and cDNA molecules in Escherichia coli. *RNA, 28*(3), 400-417.

Hall, R. H. (1963). Method for isolation of 2′-O-methylribonucleosides and N1-methyladenosine from ribonucleic acid. *Biochimica et Biophysica Acta (BBA)-Specialized Section on Nucleic Acids and Related Subjects, 68*, 278-283.

Hamma, T., & Ferré-D'Amaré, A. R. (2006). Pseudouridine synthases. *Chemistry & biology, 13*(11), 1125-1135.

Harrington, K. M., Nazarenko, I. A., Dix, D. B., Thompson, R. C., & Uhlenbeck, O. C. (1993). In vitro analysis of translational rate and accuracy with an unmodified tRNA. *Biochemistry, 32*(30), 7617-7622.

Harris, C. R., Millman, K. J., van der Walt, S. J., Gommers, R., Virtanen, P., Cournapeau, D., . . . Oliphant, T. E. (2020). Array programming with NumPy. *Nature, 585*(7825), 357-362. doi:10.1038/s41586-020-2649-2

Hassan, D., Acevedo, D., Daulatabad, S. V., Mir, Q., & Janga, S. C. (2022). Penguin: A tool for predicting pseudouridine sites in direct RNA nanopore sequencing data. *Methods, 203*, 478-487. doi:10.1016/j.ymeth.2022.02.005

He, J., Fang, T., Zhang, Z., Huang, B., Zhu, X., & Xiong, Y. (2018). PseUI: Pseudouridine sites identification based on RNA sequence information. *BMC Bioinformatics, 19*(1), 306. doi:10.1186/s12859-018-2321-0

Hendra, C., Pratanwanich, P. N., Wan, Y. K., Goh, W. S. S., Thiery, A., & Goke, J. (2022). Detection of m6A from direct RNA sequencing using a multiple instance learning framework. *Nat Methods, 19*(12), 1590-1598. doi:10.1038/s41592-022-01666-1

Henley, R. Y., Carson, S., & Wanunu, M. (2016). Studies of RNA sequence and structure using nanopores. *Progress in molecular biology and translational science, 139*, 73-99.

Ho, L. L., Schiess, G. H., Miranda, P., Weber, G., & Astakhova, K. (2024). Pseudouridine and N 1-methylpseudouridine as potent nucleotide analogues for RNA therapy and vaccine development. *RSC Chemical Biology*.

Hou, Y., Zhang, W., McGilvray, P. T., Sobczyk, M., Wang, T., Weng, S. H. S., . . . Katanski, C. D. (2023). Engineered mischarged transfer RNAs for correcting pathogenic missense mutations. *Molecular Therapy*.

Hoyt, S. J., Storer, J. M., Hartley, G. A., Grady, P. G., Gershman, A., de Lima, L. G., . . . Rodriguez, M. (2022). From telomere to telomere: The transcriptional and epigenetic state of human repeat elements. *Science, 376*(6588), eabk3112.

Hsu, P. J., Zhu, Y., Ma, H., Guo, Y., Shi, X., Liu, Y., . . . Wang, J. (2017). Ythdc2 is an N6-methyladenosine binding protein that regulates mammalian spermatogenesis. *Cell research, 27*(9), 1115-1127.

Hu, L., Liu, S., Peng, Y., Ge, R., Su, R., Senevirathne, C., . . . He, C. (2022). m(6)A RNA modifications are measured at single-base resolution across the mammalian transcriptome. *Nat Biotechnol, 40*(8), 1210-1219. doi:10.1038/s41587-022-01243-z

Hu, Y.-X., Diao, L.-T., Hou, Y.-R., Lv, G., Tao, S., Xu, W.-Y., . . . Xiao, Z.-D. Pseudouridine synthase 1 promotes hepatocellular carcinoma through mRNA pseudouridylation to enhance the translation of oncogenic mRNAs. *Hepatology*, 10.1097.

Huang, S., Wylder, A. C., & Pan, T. (2024). Simultaneous nanopore profiling of mRNA m6A and pseudouridine reveals translation coordination. *Nature Biotechnology*, 1-5.

Huang, S., Zhang, W., Katanski, C. D., Dersh, D., Dai, Q., Lolans, K., . . . Pan, T. (2021). Interferon inducible pseudouridine modification in human mRNA by quantitative nanopore profiling. *Genome biology, 22*, 1-14.

Huang, T., Chen, W., Liu, J., Gu, N., & Zhang, R. (2019). Genome-wide identification of mRNA 5-methylcytosine in mammals. *Nature structural & molecular biology, 26*(5), 380-388.

Hunter, J. D. (2007). Matplotlib: A 2D graphics environment. *Computing in science & engineering, 9*(03), 90-95.

Hur, S., Stroud, R. M., & Finer-Moore, J. (2006). Substrate recognition by RNA 5-methyluridine methyltransferases and pseudouridine synthases: a structural perspective. *Journal of Biological Chemistry, 281*(51), 38969-38973.

Ingolia, N. T., Hussmann, J. A., & Weissman, J. S. (2019). Ribosome Profiling: Global Views of Translation. *Cold Spring Harb Perspect Biol, 11*(5). doi:10.1101/cshperspect.a032698

Ip, C. L., Loose, M., Tyson, J. R., de Cesare, M., Brown, B. L., Jain, M., . . . Urban, J. M. (2015). MinION Analysis and Reference Consortium: Phase 1 data release and analysis. *F1000Research, 4*.

Jackson, L. A., Anderson, E. J., Rouphael, N. G., Roberts, P. C., Makhene, M., Coler, R. N., . . . m, R. N. A. S. G. (2020). An mRNA Vaccine against SARS-CoV-2 - Preliminary Report. *N Engl J Med, 383*(20), 1920-1931. doi:10.1056/NEJMoa2022483

Jain, M., Abu-Shumays, R., Olsen, H. E., & Akeson, M. (2022). Advances in nanopore direct RNA sequencing. *Nature methods, 19*(10), 1160-1164.

Jain, M., Koren, S., Miga, K. H., Quick, J., Rand, A. C., Sasani, T. A., . . . Fiddes, I. T. (2018). Nanopore sequencing and assembly of a human genome with ultra-long reads. *Nature Biotechnology, 36*(4), 338-345.

Jenjaroenpun, P., Wongsurawat, T., Wadley, T. D., Wassenaar, T. M., Liu, J., Dai, Q., . . . Franco, A. T. (2021). Decoding the epitranscriptional landscape from native RNA sequences. *Nucleic Acids Research, 49*(2), e7-e7.

Jenjaroenpun, P., Wongsurawat, T., Wadley, T. D., Wassenaar, T. M., Liu, J., Dai, Q., . . . Nookaew, I. (2021). Decoding the epitranscriptional landscape from native RNA sequences. *Nucleic Acids Res, 49*(2), e7. doi:10.1093/nar/gkaa620

Jia, G., Fu, Y., Zhao, X., Dai, Q., Zheng, G., Yang, Y., . . . Yang, Y.-G. (2011). N 6-methyladenosine in nuclear RNA is a major substrate of the obesity-associated FTO. *Nature chemical biology, 7*(12), 885-887.

Jonkhout, N., Tran, J., Smith, M. A., Schonrock, N., Mattick, J. S., & Novoa, E. M. (2017). The RNA modification landscape in human disease. *RNA, 23*(12), 1754-1769. doi:10.1261/rna.063503.117

Karijolich, J., & Yu, Y.-T. (2011). Converting nonsense codons into sense codons by targeted pseudouridylation. *Nature, 474*(7351), 395-398.

Kariko, K., Muramatsu, H., Welsh, F. A., Ludwig, J., Kato, H., Akira, S., & Weissman, D. (2008). Incorporation of pseudouridine into mRNA yields superior nonimmunogenic vector with increased translational capacity and biological stability. *Mol Ther, 16*(11), 1833-1840. doi:10.1038/mt.2008.200

Katanski, C. D., Alshammary, H., Watkins, C. P., Huang, S., Gonzales-Reiche, A., Sordillo, E. M., . . . Simon, V. (2022). tRNA abundance, modification and fragmentation in nasopharyngeal swabs as biomarkers for COVID-19 severity. *Frontiers in Cell and Developmental Biology, 10*, 999351.

Khan, S. M., He, F., Wang, D., Chen, Y., & Xu, D. (2020). MU-PseUDeep: A deep learning method for prediction of pseudouridine sites. *Comput Struct Biotechnol J, 18*, 1877-1883. doi:10.1016/j.csbj.2020.07.010

Khoddami, V., Yerra, A., Mosbruger, T. L., Fleming, A. M., Burrows, C. J., & Cairns, B. R. (2019). Transcriptome-wide profiling of multiple RNA modifications simultaneously at single-base resolution. *Proceedings of the National Academy of Sciences, 116*(14), 6784-6789.

Kim, D., Lee, J.-Y., Yang, J.-S., Kim, J. W., Kim, V. N., & Chang, H. (2020). The architecture of SARS-CoV-2 transcriptome. *Cell, 181*(4), 914-921. e910.

Koh, C. W., Goh, Y. T., & Goh, W. S. (2019). Atlas of quantitative single-base-resolution N 6-methyl-adenine methylomes. *Nature Communications, 10*(1), 5636.

Körtel, N., Rücklé, C., Zhou, Y., Busch, A., Hoch-Kraft, P., Sutandy, F. R., . . . Ostareck, D. (2021). Deep and accurate detection of m6A RNA modifications using miCLIP2 and m6Aboost machine learning. *Nucleic acids research, 49*(16), e92-e92.

Kumbhar, B. V., Kamble, A. D., & Sonawane, K. D. (2013). Conformational preferences of modified nucleoside N (4)-acetylcytidine, ac 4 C Occur at “Wobble” 34th position in the anticodon loop of tRNA. *Cell Biochemistry and Biophysics, 66*, 797-816.

Langmead, B., & Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. *Nature methods, 9*(4), 357-359.

Laszlo, A. H., Derrington, I. M., Brinkerhoff, H., Langford, K. W., Nova, I. C., Samson, J. M., . . . Gundlach, J. H. (2013). Detection and mapping of 5-methylcytosine and 5-hydroxymethylcytosine with nanopore MspA. *Proceedings of the National Academy of Sciences, 110*(47), 18904-18909.

Lee, A. J., & Ashkar, A. A. (2018). The dual nature of type I and type II interferons. *Frontiers in immunology, 9*, 2061.

Leger, A., Amaral, P. P., Pandolfini, L., Capitanchik, C., Capraro, F., Miano, V., . . . Kouzarides, T. (2021). RNA modifications detection by comparative Nanopore direct RNA sequencing. *Nat Commun, 12*(1), 7198. doi:10.1038/s41467-021-27393-3

Li, F., Guo, X., Jin, P., Chen, J., Xiang, D., Song, J., & Coin, L. J. M. (2021). Porpoise: a new approach for accurate prediction of RNA pseudouridine sites. *Brief Bioinform, 22*(6). doi:10.1093/bib/bbab245

Li, H. (2018). Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics, 34*(18), 3094-3100.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., . . . Durbin, R. (2009). The sequence alignment/map format and SAMtools. *Bioinformatics, 25*(16), 2078-2079.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., . . . Genome Project Data Processing, S. (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics, 25*(16), 2078-2079. doi:10.1093/bioinformatics/btp352

Li, X., Ma, S., & Yi, C. (2016). Pseudouridine: the fifth RNA nucleotide with renewed interests. *Current Opinion in Chemical Biology, 33*, 108-116.

Li, X., Xiong, X., Wang, K., Wang, L., Shu, X., Ma, S., & Yi, C. (2016). Transcriptome-wide mapping reveals reversible and dynamic N 1-methyladenosine methylome. *Nature chemical biology, 12*(5), 311-316.

Li, X., Zhu, P., Ma, S., Song, J., Bai, J., Sun, F., & Yi, C. (2015). Chemical pulldown reveals dynamic pseudouridylation of the mammalian transcriptome. *Nat Chem Biol, 11*(8), 592-597. doi:10.1038/nchembio.1836

Li, Y. H., Zhang, G., & Cui, Q. (2015). PPUS: a web server to predict PUS-specific pseudouridine sites. *Bioinformatics, 31*(20), 3362-3364. doi:10.1093/bioinformatics/btv366

Liao, Y., Liu, Z., Zhang, Y., Lu, P., Wen, L., & Tang, F. (2023). High-throughput and high-sensitivity full-length single-cell RNA-seq analysis on third-generation sequencing platform. *Cell Discovery, 9*(1), 5.

Linder, B., Grozhik, A. V., Olarerin-George, A. O., Meydan, C., Mason, C. E., & Jaffrey, S. R. (2015). Single-nucleotide-resolution mapping of m6A and m6Am throughout the transcriptome. *Nat Methods, 12*(8), 767-772. doi:10.1038/nmeth.3453

Liu, C., Sun, H., Yi, Y., Shen, W., Li, K., Xiao, Y., . . . Wang, J. (2023). Absolute quantification of single-base m(6)A methylation in the mammalian transcriptome using GLORI. *Nat Biotechnol, 41*(3), 355-366. doi:10.1038/s41587-022-01487-9

Liu, H., Begik, O., Lucas, M. C., Ramirez, J. M., Mason, C. E., Wiener, D., . . . Novoa, E. M. (2019). Accurate detection of m 6 A RNA modifications in native RNA sequences. *Nature communications, 10*(1), 1-9.

Liu, H., Begik, O., Lucas, M. C., Ramirez, J. M., Mason, C. E., Wiener, D., . . . Novoa, E. M. (2019). Accurate detection of m(6)A RNA modifications in native RNA sequences. *Nat Commun, 10*(1), 4079. doi:10.1038/s41467-019-11713-9

Liu, H., Begik, O., & Novoa, E. M. (2021). EpiNano: Detection of m(6)A RNA Modifications Using Oxford Nanopore Direct RNA Sequencing. *Methods Mol Biol, 2298*, 31-52. doi:10.1007/978-1-0716-1374-0_3

Liu, J., Yue, Y., Han, D., Wang, X., Fu, Y., Zhang, L., . . . He, C. (2014). A METTL3-METTL14 complex mediates mammalian nuclear RNA N6-adenosine methylation. *Nat Chem Biol, 10*(2), 93-95. doi:10.1038/nchembio.1432

Liu, K., Chen, W., & Lin, H. (2020). XG-PseU: an eXtreme Gradient Boosting based method for identifying pseudouridine sites. *Mol Genet Genomics, 295*(1), 13-21. doi:10.1007/s00438-019-01600-9

Liu, N., Dai, Q., Zheng, G., He, C., Parisien, M., & Pan, T. (2015). N 6-methyladenosine-dependent RNA structural switches regulate RNA–protein interactions. *Nature, 518*(7540), 560-564.

Liu, N., Zhou, K. I., Parisien, M., Dai, Q., Diatchenko, L., & Pan, T. (2017). N 6-methyladenosine alters RNA structure to regulate binding of a low-complexity protein. *Nucleic acids research, 45*(10), 6051-6063.

Liu, Q., Fang, L., Yu, G., Wang, D., Xiao, C.-L., & Wang, K. (2019). Detection of DNA base modifications by deep recurrent neural network on Oxford Nanopore sequencing data. *Nature Communications, 10*(1), 2449.

Liu, R., Ou, L., Sheng, B., Hao, P., Li, P., Yang, X., . . . Feng, D. D. (2022). Mixed-Weight Neural Bagging for Detecting m(6)A Modifications in SARS-CoV-2 RNA Sequencing. *IEEE Trans Biomed Eng, 69*(8), 2557-2568. doi:10.1109/TBME.2022.3150420

Lorenz, D. A., Sathe, S., Einstein, J. M., & Yeo, G. W. (2020). Direct RNA sequencing enables m6A detection in endogenous transcript isoforms at base-specific resolution. *RNA, 26*(1), 19-28.

Lorenz, D. A., Sathe, S., Einstein, J. M., & Yeo, G. W. (2020). Direct RNA sequencing enables m(6)A detection in endogenous transcript isoforms at base-specific resolution. *RNA, 26*(1), 19-28. doi:10.1261/rna.072785.119

Louloupi, A., Ntini, E., Conrad, T., & Ørom, U. A. V. (2018). Transient N-6-methyladenosine transcriptome sequencing reveals a regulatory role of m6A in splicing efficiency. *Cell reports, 23*(12), 3429-3437.

Lovejoy, A. F., Riordan, D. P., & Brown, P. O. (2014). Transcriptome-wide mapping of pseudouridines: pseudouridine synthases modify specific mRNAs in S. cerevisiae. *PLoS One, 9*(10), e110799. doi:10.1371/journal.pone.0110799

Lucas, M. C., Pryszcz, L. P., Medina, R., Milenkovic, I., Camacho, N., Marchand, V., . . . Novoa, E. M. (2024). Quantitative analysis of tRNA abundance and modifications by nanopore RNA sequencing. *Nature Biotechnology, 42*(1), 72-86.

Luo, N., Huang, Q., Dong, L., Liu, W., Song, J., Sun, H., . . . Yi, C. (2024). Near-cognate tRNAs increase the efficiency and precision of pseudouridine-mediated readthrough of premature termination codons. *Nature Biotechnology*, 1-10.

Lv, Z., Zhang, J., Ding, H., & Zou, Q. (2020). RF-PseU: A Random Forest Predictor for RNA Pseudouridine Sites. *Front Bioeng Biotechnol, 8*, 134. doi:10.3389/fbioe.2020.00134

Mahesh, B. (2020). Machine learning algorithms-a review. *International Journal of Science and Research (IJSR).[Internet], 9*(1), 381-386.

Martinez, N. M., Su, A., Burns, M. C., Nussbacher, J. K., Schaening, C., Sathe, S., . . . Gilbert, W. V. (2022). Pseudouridine synthases modify human pre-mRNA co-transcriptionally and affect pre-mRNA processing. *Mol Cell, 82*(3), 645-659 e649. doi:10.1016/j.molcel.2021.12.023

McKinney, W. (2010). *Data structures for statistical computing in python.* Paper presented at the Proceedings of the 9th Python in Science Conference.

Meyer, K. D. (2019). DART-seq: an antibody-free method for global m(6)A detection. *Nat Methods, 16*(12), 1275-1280. doi:10.1038/s41592-019-0570-0

Meyer, K. D., Patil, D. P., Zhou, J., Zinoviev, A., Skabkin, M. A., Elemento, O., . . . Jaffrey, S. R. (2015). 5′ UTR m6A promotes cap-independent translation. *Cell, 163*(4), 999-1010.

Meyer, K. D., Saletore, Y., Zumbo, P., Elemento, O., Mason, C. E., & Jaffrey, S. R. (2012). Comprehensive analysis of mRNA methylation reveals enrichment in 3' UTRs and near stop codons. *Cell, 149*(7), 1635-1646. doi:10.1016/j.cell.2012.05.003

Molinie, B., Wang, J., Lim, K. S., Hillebrand, R., Lu, Z.-x., Van Wittenberghe, N., . . . Dedon, P. (2016). m6A-LAIC-seq reveals the census and complexity of the m6A epitranscriptome. *Nature methods, 13*(8), 692-698.

Morais, P., Adachi, H., & Yu, Y.-T. (2021). The critical contribution of pseudouridine to mRNA COVID-19 vaccines. *Frontiers in Cell and Developmental Biology, 9*, 3187.

Morton, D., Mortezaei, S., Yemenicioglu, S., Isaacman, M. J., Nova, I. C., Gundlach, J. H., & Theogarajan, L. (2015). Tailored polymeric membranes for Mycobacterium smegmatis porin A (MspA) based biosensors. *Journal of materials chemistry B, 3*(25), 5080-5086.

Mulroney, T. E., Pöyry, T., Yam-Puc, J. C., Rust, M., Harvey, R. F., Kalmar, L., . . . Stoneley, M. (2024). N 1-methylpseudouridylation of mRNA causes+ 1 ribosomal frameshifting. *Nature, 625*(7993), 189-194.

Newby, M. I., & Greenbaum, N. L. (2001). A conserved pseudouridine modification in eukaryotic U2 snRNA induces a change in branch-site architecture. *RNA, 7*(6), 833-845.

Newby, M. I., & Greenbaum, N. L. (2002). Sculpting of the spliceosomal branch site recognition motif by a conserved pseudouridine. *Nature structural biology, 9*(12), 958-965.

Nguyen, T. A., Heng, J. W. J., Kaewsapsak, P., Kok, E. P. L., Stanojević, D., Liu, H., . . . Lin, M. (2022). Direct identification of A-to-I editing sites with nanopore native RNA sequencing. *Nature methods, 19*(7), 833-844.

Ni, P., Huang, N., Zhang, Z., Wang, D.-P., Liang, F., Miao, Y., . . . Wang, J. (2019). DeepSignal: detecting DNA methylation state from Nanopore sequencing reads using deep-learning. *Bioinformatics, 35*(22), 4586-4595.

Nurk, S., Koren, S., Rhie, A., Rautiainen, M., Bzikadze, A. V., Mikheenko, A., . . . Gershman, A. (2022). The complete sequence of a human genome. *Science, 376*(6588), 44-53.

Pan, T. (2018). Modifications and functional genomics of human transfer RNA. *Cell Res, 28*(4), 395-404. doi:10.1038/s41422-018-0013-y

Parker, M. T., Knop, K., Sherwood, A. V., Schurch, N. J., Mackinnon, K., Gould, P. D., . . . Simpson, G. G. (2020). Nanopore direct RNA sequencing maps the complexity of Arabidopsis mRNA processing and m(6)A modification. *Elife, 9*. doi:10.7554/eLife.49658

Patil, A., Dyavaiah, M., Joseph, F., Rooney, J. P., Chan, C. T., Dedon, P. C., & Begley, T. J. (2012). Increased tRNA modification and gene-specific codon usage regulate cell cycle progression during the DNA damage response. *Cell cycle, 11*(19), 3656-3665.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., . . . Dubourg, V. (2011). Scikit-learn: Machine learning in Python. *the Journal of machine Learning research, 12*, 2825-2830.

Piechotta, M., Naarmann-de Vries, I. S., Wang, Q., Altmuller, J., & Dieterich, C. (2022). RNA modification mapping with JACUSA2. *Genome Biol, 23*(1), 115. doi:10.1186/s13059-022-02676-0

Ping, X.-L., Sun, B.-F., Wang, L., Xiao, W., Yang, X., Wang, W.-J., . . . Chen, Y.-S. (2014). Mammalian WTAP is a regulatory subunit of the RNA N6-methyladenosine methyltransferase. *Cell research, 24*(2), 177-189.

Pratanwanich, P. N., Yao, F., Chen, Y., Koh, C. W. Q., Wan, Y. K., Hendra, C., . . . Goke, J. (2021). Identification of differential RNA modifications from nanopore direct RNA sequencing with xPore. *Nat Biotechnol, 39*(11), 1394-1402. doi:10.1038/s41587-021-00949-w

Price, A. M., Hayer, K. E., McIntyre, A. B. R., Gokhale, N. S., Abebe, J. S., Della Fera, A. N., . . . Weitzman, M. D. (2020). Direct RNA sequencing reveals m(6)A modifications on adenovirus RNA are necessary for efficient splicing. *Nat Commun, 11*(1), 6016. doi:10.1038/s41467-020-19787-6

Qin, H., Ou, L., Gao, J., Chen, L., Wang, J. W., Hao, P., & Li, X. (2022). DENA: training an authentic neural network model using Nanopore sequencing data of Arabidopsis transcripts for detection and quantification of N(6)-methyladenosine on RNA. *Genome Biol, 23*(1), 25. doi:10.1186/s13059-021-02598-3

Quinlan, A. R., & Hall, I. M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics, 26*(6), 841-842.

Ramasamy, S., Mishra, S., Sharma, S., Parimalam, S. S., Vaijayanthi, T., Fujita, Y., . . . Pandian, G. N. (2022). An informatics approach to distinguish RNA modifications in nanopore direct RNA sequencing. *Genomics, 114*(3), 110372. doi:10.1016/j.ygeno.2022.110372

Ramasamy, S., Sahayasheela, V. J., Sharma, S., Yu, Z., Hidaka, T., Cai, L., . . . Pandian, G. N. (2022). Chemical Probe-Based Nanopore Sequencing to Selectively Assess the RNA Modifications. *ACS Chem Biol, 17*(10), 2704-2709. doi:10.1021/acschembio.2c00221

Rand, A. C., Jain, M., Eizenga, J. M., Musselman-Brown, A., Olsen, H. E., Akeson, M., & Paten, B. (2017). Mapping DNA methylation with high-throughput nanopore sequencing. *Nature methods, 14*(4), 411-413.

Robinson, J. T., Thorvaldsdóttir, H., Winckler, W., Guttman, M., Lander, E. S., Getz, G., & Mesirov, J. P. (2011). Integrative genomics viewer. *Nature biotechnology, 29*(1), 24-26.

Roundtree, I. A., Evans, M. E., Pan, T., & He, C. (2017). Dynamic RNA Modifications in Gene Expression Regulation. *Cell, 169*(7), 1187-1200. doi:10.1016/j.cell.2017.05.045

Roundtree, I. A., Luo, G.-Z., Zhang, Z., Wang, X., Zhou, T., Cui, Y., . . . Xie, P. (2017). YTHDC1 mediates nuclear export of N6-methyladenosine methylated mRNAs. *Elife, 6*, e31311.

Rousseau-Gueutin, M., Belser, C., Da Silva, C., Richard, G., Istace, B., Cruaud, C., . . . Delourme, R. (2020). Long-read assembly of the Brassica napus reference genome Darmor-bzh. *GigaScience, 9*(12), giaa137.

Safra, M., Nir, R., Farouq, D., Slutskin, I. V., & Schwartz, S. (2017). TRUB1 is the predominant pseudouridine synthase acting on mammalian mRNA via a predictable and conserved code. *Genome research, 27*(3), 393-406.

Safra, M., Nir, R., Farouq, D., Vainberg Slutskin, I., & Schwartz, S. (2017). TRUB1 is the predominant pseudouridine synthase acting on mammalian mRNA via a predictable and conserved code. *Genome Res, 27*(3), 393-406. doi:10.1101/gr.207613.116

Sakakibara, Y., & Chow, C. S. (2012). Role of pseudouridine in structural rearrangements of helix 69 during bacterial ribosome assembly. *ACS chemical biology, 7*(5), 871-878.

Sas-Chen, A., Thomas, J. M., Matzov, D., Taoka, M., Nance, K. D., Nir, R., . . . Burkhart, B. W. (2020). Dynamic RNA acetylation revealed by quantitative cross-evolutionary mapping. *Nature, 583*(7817), 638-643.

Schaefer, M., Pollex, T., Hanna, K., Tuorto, F., Meusburger, M., Helm, M., & Lyko, F. (2010). RNA methylation by Dnmt2 protects transfer RNAs against stress-induced cleavage. *Genes & development, 24*(15), 1590-1595.

Schreiber, J., Wescoe, Z. L., Abu-Shumays, R., Vivian, J. T., Baatar, B., Karplus, K., & Akeson, M. (2013). Error rates for nanopore discrimination among cytosine, methylcytosine, and hydroxymethylcytosine along individual DNA strands. *Proceedings of the National Academy of Sciences, 110*(47), 18910-18915.

Schwartz, S., Bernstein, D. A., Mumbach, M. R., Jovanovic, M., Herbst, R. H., Leon-Ricardo, B. X., . . . Regev, A. (2014). Transcriptome-wide mapping reveals widespread dynamic-regulated pseudouridylation of ncRNA and mRNA. *Cell, 159*(1), 148-162. doi:10.1016/j.cell.2014.08.028

Sedlazeck, F. J., Rescheneder, P., Smolka, M., Fang, H., Nattestad, M., Von Haeseler, A., & Schatz, M. C. (2018). Accurate detection of complex structural variations using single-molecule sequencing. *Nature methods, 15*(6), 461-468.

Sereika, M., Kirkegaard, R. H., Karst, S. M., Michaelsen, T. Y., Sørensen, E. A., Wollenberg, R. D., & Albertsen, M. (2022). Oxford Nanopore R10. 4 long-read sequencing enables the generation of near-finished bacterial genomes from pure cultures and metagenomes without short-read or reference polishing. *Nature methods, 19*(7), 823-826.

Shi, H., Wang, X., Lu, Z., Zhao, B. S., Ma, H., Hsu, P. J., . . . He, C. (2017). YTHDF3 facilitates translation and decay of N6-methyladenosine-modified RNA. *Cell research, 27*(3), 315-328.

Shu, X., Cao, J., Cheng, M., Xiang, S., Gao, M., Li, T., . . . Lu, Z. (2020). A metabolic labeling method detects m6A transcriptome-wide at single base resolution. *Nature chemical biology, 16*(8), 887-895.

Simpson, J. T., Workman, R. E., Zuzarte, P., David, M., Dursi, L., & Timp, W. (2017). Detecting DNA cytosine methylation using nanopore sequencing. *Nature methods, 14*(4), 407-410.

Singhal, R. P. (1974). Chemical probe of structure and function of transfer ribonucleic acids. *Biochemistry, 13*(14), 2924-2932.

Śledź, P., & Jinek, M. (2016). Structural insights into the molecular mechanism of the m6A writer complex. *Elife, 5*, e18434.

Song, B., Chen, K., Tang, Y., Ma, J., Meng, J., & Wei, Z. (2020). PSI-MOUSE: Predicting Mouse Pseudouridine Sites From Sequence and Genome-Derived Features. *Evol Bioinform Online, 16*, 1176934320925752. doi:10.1177/1176934320925752

Song, B., Tang, Y., Wei, Z., Liu, G., Su, J., Meng, J., & Chen, K. (2020). PIANO: A Web Server for Pseudouridine-Site (Psi) Identification and Functional Annotation. *Front Genet, 11*, 88. doi:10.3389/fgene.2020.00088

Spenkuch, F., Motorin, Y., & Helm, M. (2014). Pseudouridine: still mysterious, but never a fake (uridine)! *RNA biology, 11*(12), 1540-1554.

Squires, J. E., Patel, H. R., Nousch, M., Sibbritt, T., Humphreys, D. T., Parker, B. J., . . . Preiss, T. (2012). Widespread occurrence of 5-methylcytosine in human coding and non-coding RNA. *Nucleic acids research, 40*(11), 5023-5033.

Stephenson, W., Razaghi, R., Busan, S., Weeks, K. M., Timp, W., & Smibert, P. (2022). Direct detection of RNA modifications and structure using single-molecule nanopore sequencing. *Cell genomics, 2*(2).

Stern, L., & LH, S. (1978). The role of the minor base N4-acetylcytidine in the function of the Escherichia coli noninitiator methionine transfer RNA.

Stoiber, M., Quick, J., Egan, R., Eun Lee, J., Celniker, S., Neely, R. K., . . . Brown, J. (2016). De novo identification of DNA modifications enabled by genome-guided nanopore signal processing. *BioRxiv*, 094672.

Sun, H., Li, K., Liu, C., & Yi, C. (2023). Regulation and functions of non-m(6)A mRNA modifications. *Nat Rev Mol Cell Biol, 24*(10), 714-731. doi:10.1038/s41580-023-00622-x

Suzuki, M. M., & Bird, A. (2008). DNA methylation landscapes: provocative insights from epigenomics. *Nature reviews genetics, 9*(6), 465-476.

Svitkin, Y. V., Cheng, Y. M., Chakraborty, T., Presnyak, V., John, M., & Sonenberg, N. (2017). N1-methyl-pseudouridine in mRNA enhances translation through eIF2α-dependent and independent mechanisms by increasing ribosome density. *Nucleic acids research, 45*(10), 6023-6036.

Tahir, M., Tayara, H., & Chong, K. T. (2019). iPseU-CNN: Identifying RNA Pseudouridine Sites Using Convolutional Neural Networks. *Mol Ther Nucleic Acids, 16*, 463-470. doi:10.1016/j.omtn.2019.03.010

Taoka, M., Nobe, Y., Yamaki, Y., Sato, K., Ishikawa, H., Izumikawa, K., . . . Takahashi, N. (2018). Landscape of the complete RNA chemical modifications in the human 80S ribosome. *Nucleic acids research, 46*(18), 9289-9298.

Taoka, M., Nobe, Y., Yamaki, Y., Sato, K., Ishikawa, H., Izumikawa, K., . . . Isobe, T. (2018). Landscape of the complete RNA chemical modifications in the human 80S ribosome. *Nucleic Acids Res, 46*(18), 9289-9298. doi:10.1093/nar/gky811

Tarca, A. L., Carey, V. J., Chen, X.-w., Romero, R., & Drăghici, S. (2007). Machine learning and its applications to biology. *PLoS computational biology, 3*(6), e116.

Taucher, M., Ganisl, B., & Breuker, K. (2011). Identification, localization, and relative quantitation of pseudouridine in RNA by tandem mass spectrometry of hydrolysis products. *International Journal of Mass Spectrometry, 304*(2-3), 91-97.

Tavakoli, S., Nabizadeh, M., Makhamreh, A., Gamper, H., McCormick, C. A., Rezapour, N. K., . . . Rouhanifard, S. H. (2023). Semi-quantitative detection of pseudouridine modifications and type I/II hypermodifications in human mRNAs using direct long-read sequencing. *Nat Commun, 14*(1), 334. doi:10.1038/s41467-023-35858-w

Thalalla Gamage, S., Sas-Chen, A., Schwartz, S., & Meier, J. L. (2021). Quantitative nucleotide resolution profiling of RNA cytidine acetylation by ac4C-seq. *Nature Protocols, 16*(4), 2286-2307.

Torres, A. G., Batlle, E., & Ribas de Pouplana, L. (2014). Role of tRNA modifications in human diseases. *Trends Mol Med, 20*(6), 306-314. doi:10.1016/j.molmed.2014.01.008

Tuorto, F., Liebers, R., Musch, T., Schaefer, M., Hofmann, S., Kellner, S., . . . Lyko, F. (2012). RNA cytosine methylation by Dnmt2 and NSun2 promotes tRNA stability and protein synthesis. *Nature structural & molecular biology, 19*(9), 900-905.

Viehweger, A., Krautwurst, S., Lamkiewicz, K., Madhugiri, R., Ziebuhr, J., Hölzer, M., & Marz, M. (2019). Direct RNA nanopore sequencing of full-length coronavirus genomes provides novel insights into structural variants and enables modification analysis. *Genome research, 29*(9), 1545-1554.

Vollger, M. R., Guitart, X., Dishuck, P. C., Mercuri, L., Harvey, W. T., Gershman, A., . . . Lewis, A. P. (2022). Segmental duplications and their variation in a complete human genome. *Science, 376*(6588), eabj6965.

Wang, K., Zhang, S., Zhou, X., Yang, X., Li, X., Wang, Y., . . . Zhang, P. (2024). Unambiguous discrimination of all 20 proteinogenic amino acids and their modifications by nanopore. *Nature methods, 21*(1), 92-101.

Wang, P., Doxtader, K. A., & Nam, Y. (2016). Structural basis for cooperative function of Mettl3 and Mettl14 methyltransferases. *Molecular cell, 63*(2), 306-317.

Wang, X., Feng, J., Xue, Y., Guan, Z., Zhang, D., Liu, Z., . . . Tang, C. (2016). Structural basis of N 6-adenosine methylation by the METTL3–METTL14 complex. *Nature, 534*(7608), 575-578.

Wang, X., Lu, Z., Gomez, A., Hon, G. C., Yue, Y., Han, D., . . . He, C. (2014). N6-methyladenosine-dependent regulation of messenger RNA stability. *Nature, 505*(7481), 117-120. doi:10.1038/nature12730

Wang, X., Zhao, B. S., Roundtree, I. A., Lu, Z., Han, D., Ma, H., . . . He, C. (2015). N6-methyladenosine modulates messenger RNA translation efficiency. *Cell, 161*(6), 1388-1399.

Wang, Y., Li, Y., Toth, J. I., Petroski, M. D., Zhang, Z., & Zhao, J. C. (2014). N6-methyladenosine modification destabilizes developmental regulators in embryonic stem cells. *Nat Cell Biol, 16*(2), 191-198. doi:10.1038/ncb2902

Wang, Y., Xiao, Y., Dong, S., Yu, Q., & Jia, G. (2020). Antibody-free enzyme-assisted chemical approach for detection of N 6-methyladenosine. *Nature chemical biology, 16*(8), 896-903.

WATKINS, N. J., GOTTSCHALK, A., NEUBAUER, G., KASTNER, B., FABRIZIO, P., MANN, M., & LUeHRMANN, R. (1998). Cbf5p, a potential pseudouridine synthase, and Nhp2p, a putative RNA-binding protein, are present together with Gar1p in all H BOX/ACA-motif snoRNPs and constitute a common bipartite structure. *RNA, 4*(12), 1549-1568.

Wei, C., Gershowitz, A., & Moss, B. (1975). N6, O2'-dimethyladenosine a novel methylated ribonucleoside next to the 5' terminal of animal cell and virus mRNAs. *Nature, 257*(5523), 251-253. doi:10.1038/257251a0

Wick, R., Judd, L., & Holt, K. (2019). Performance of neural network basecalling tools for Oxford Nanopore sequencing. BioRxiv, 543439. In.

Wongsurawat, T., Jenjaroenpun, P., Taylor, M. K., Lee, J., Tolardo, A. L., Parvathareddy, J., . . . Athipanyasilp, N. (2019). Rapid sequencing of multiple RNA viruses in their native form. *Frontiers in microbiology, 10*, 260.

Workman, R. E., Tang, A. D., Tang, P. S., Jain, M., Tyson, J. R., Razaghi, R., . . . Quick, J. (2019). Nanopore native RNA sequencing of a human poly (A) transcriptome. *Nature methods, 16*(12), 1297-1305.

Workman, R. E., Tang, A. D., Tang, P. S., Jain, M., Tyson, J. R., Razaghi, R., . . . Timp, W. (2019). Nanopore native RNA sequencing of a human poly(A) transcriptome. *Nat Methods, 16*(12), 1297-1305. doi:10.1038/s41592-019-0617-2

Wu, G., Adachi, H., Ge, J., Stephenson, D., Query, C. C., & Yu, Y. T. (2016). Pseudouridines in U2 snRNA stimulate the ATPase activity of Prp5 during spliceosome assembly. *EMBO J, 35*(6), 654-667. doi:10.15252/embj.201593113

Xiang, Y., Laurent, B., Hsu, C.-H., Nachtergaele, S., Lu, Z., Sheng, W., . . . Wang, S. (2017). RNA m6A methylation regulates the ultraviolet-induced DNA damage response. *Nature, 543*(7646), 573-576.

Xiao, W., Adhikari, S., Dahal, U., Chen, Y. S., Hao, Y. J., Sun, B. F., . . . Yang, Y. G. (2016). Nuclear m(6)A Reader YTHDC1 Regulates mRNA Splicing. *Mol Cell, 61*(4), 507-519. doi:10.1016/j.molcel.2016.01.012

Xiao, Y.-L., Liu, S., Ge, R., Wu, Y., He, C., Chen, M., & Tang, W. (2023). Transcriptome-wide profiling and quantification of N 6-methyladenosine by enzyme-assisted adenosine deamination. *Nature Biotechnology*, 1-11.

Xu, K., Yang, Y., Feng, G.-H., Sun, B.-F., Chen, J.-Q., Li, Y.-F., . . . Jiang, L.-Y. (2017). Mettl3-mediated m6A regulates spermatogonial differentiation and meiosis initiation. *Cell research, 27*(9), 1100-1114.

Yan, S., Lu, Z., Yang, W., Xu, J., Wang, Y., Xiong, W., . . . Wei, Q. (2023). Antibody-Free Fluorine-Assisted Metabolic Sequencing of RNA N 4-Acetylcytidine. *Journal of the American Chemical Society, 145*(40), 22232-22242.

Yankova, E., Blackaby, W., Albertella, M., Rak, J., De Braekeleer, E., Tsagkogeorga, G., . . . Hendrick, A. G. (2021). Small-molecule inhibition of METTL3 as a strategy against myeloid leukaemia. *Nature, 593*(7860), 597-601.

Yu, C.-T., & Allen, F. W. (1959). Studies of an isomer of uridine isolated from ribonucleic acids. *Biochimica et biophysica acta, 32*, 393-406.

Yu, F., Qi, H., Gao, L., Luo, S., Njeri Damaris, R., Ke, Y., . . . Yang, P. (2023). Identifying RNA Modifications by Direct RNA Sequencing Reveals Complexity of Epitranscriptomic Dynamics in Rice. *Genomics Proteomics Bioinformatics*. doi:10.1016/j.gpb.2023.02.002

Zhang, L.-S., Ju, C.-W., Jiang, B., & He, C. (2023). Base-resolution quantitative DAMM-seq for mapping RNA methylations in tRNA and mitochondrial polycistronic RNA. *Methods in Enzymology, 692*, 39-54.

Zhang, L.-S., Liu, C., Ma, H., Dai, Q., Sun, H.-L., Luo, G., . . . Dong, X. (2019). Transcriptome-wide mapping of internal N7-methylguanosine methylome in mammalian mRNA. *Molecular cell, 74*(6), 1304-1316. e1308.

Zhang, L.-S., Ye, C., Ju, C.-W., Gao, B., Feng, X., Sun, H.-L., . . . He, C. (2023). BID-seq for transcriptome-wide quantitative sequencing of mRNA pseudouridine at base resolution. *Nature Protocols*, 1-22.

Zhang, M., Jiang, Z., Ma, Y., Liu, W., Zhuang, Y., Lu, B., . . . Yi, C. (2023). Quantitative profiling of pseudouridylation landscape in the human transcriptome. *Nature chemical biology*, 1-11.

Zhang, S., Cao, Z., Fan, P., Sun, W., Xiao, Y., Zhang, P., . . . Huang, S. (2023). Discrimination of Disaccharide Isomers of Different Glycosidic Linkages Using a Modified MspA Nanopore. *Angew Chem Int Ed Engl*, e202316766. doi:10.1002/anie.202316766

Zhang, W., Eckwahl, M. J., Zhou, K. I., & Pan, T. (2019). Sensitive and quantitative probing of pseudouridine modification in mRNA and long noncoding RNA. *Rna, 25*(9), 1218-1225.

Zhang, Y., Huang, D., Wei, Z., & Chen, K. (2022). Primary sequence-assisted prediction of m(6)A RNA methylation sites from Oxford nanopore direct RNA sequencing data. *Methods, 203*, 62-69. doi:10.1016/j.ymeth.2022.04.003

Zhang, Y., Yi, Y., Li, Z., Zhou, K., Liu, L., & Wu, H.-C. (2024). Peptide sequencing based on host–guest interaction-assisted nanopore sensing. *Nature methods, 21*(1), 102-109.

Zhang, Z., Chen, L. Q., Zhao, Y. L., Yang, C. G., Roundtree, I. A., Zhang, Z., . . . Luo, G. Z. (2019). Single-base mapping of m(6)A by an antibody-independent method. *Sci Adv, 5*(7), eaax0250. doi:10.1126/sciadv.aax0250

Zhao, B. S., Roundtree, I. A., & He, C. (2017). Post-transcriptional gene regulation by mRNA modifications. *Nat Rev Mol Cell Biol, 18*(1), 31-42. doi:10.1038/nrm.2016.132

Zhao, L. Y., Song, J., Liu, Y., Song, C. X., & Yi, C. (2020). Mapping the epigenetic modifications of DNA and RNA. *Protein Cell, 11*(11), 792-808. doi:10.1007/s13238-020-00733-7

Zheng, G., Dahl, J. A., Niu, Y., Fedorcsak, P., Huang, C.-M., Li, C. J., . . . Song, S.-H. (2013). ALKBH5 is a mammalian RNA demethylase that impacts RNA metabolism and mouse fertility. *Molecular cell, 49*(1), 18-29.

Zhong, Z. D., Xie, Y. Y., Chen, H. X., Lan, Y. L., Liu, X. H., Ji, J. Y., . . . Luo, G. Z. (2023). Systematic comparison of tools used for m(6)A mapping from nanopore direct RNA sequencing. *Nat Commun, 14*(1), 1906. doi:10.1038/s41467-023-37596-5

Zhou, H., Rauch, S., Dai, Q., Cui, X., Zhang, Z., Nachtergaele, S., . . . Dickinson, B. C. (2019). Evolution of a reverse transcriptase to map N 1-methyladenosine in human messenger RNA. *Nature methods, 16*(12), 1281-1288.

Zhou, K. I., Clark, W. C., Pan, D. W., Eckwahl, M. J., Dai, Q., & Pan, T. (2018). Pseudouridines have context-dependent mutation and stop rates in high-throughput sequencing. *RNA biology, 15*(7), 892-900.

Zhou, K. I., & Pan, T. (2018). An additional class of m6A readers. *Nature cell biology, 20*(3), 230-232.

Figure S4.1 Full scans of Western blot gels
(a) Full scan for siCTRL and siTRUB1, also see Fig. 4.1b. (b) Full scan for siCTRL and siMETTL3, also see Fig. 4.1b.