Photo-Oxidative Sequencing Approach for Methylated Guanosine in RNA

Kimberley Alexandra Chung Kim Chung

A DISSERTATION SUBMITTED TO THE FACULTY OF GRADUATE STUDIES
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

GRADUATE PROGRAM IN CHEMISTRY
YORK UNIVERSITY
TORONTO, ONTARIO

November 2024

Abstract

In recent years, RNA-modifying enzymes have gained significant attention due to their impact on critical RNA-based processes, and consequently human pathology. However, identifying sites of modifications throughout the cellular transcriptome remains challenging largely due to the lack of accurate and sensitive detection technologies. Indeed, N^2-methylation at guanosine such as N^2-methylated guanosine (m^2G) and N^2,N^2-dimethylated guanosine ($m^2{}_2G$) are prevalent modifications found in RNA but their functional studies are limited due to the absence of sequencing methods. While $m^2{}_2G$ creates termination during reverse transcription and can often be detected during sequencing due to its high error signature, m^2G does not affect Watson−Crick−Franklin base pairing, and therefore cannot be distinguished from unmodified guanosine. Although computational predictors exist for m^2G, there are no direct sequencing method available. In this thesis, the first chemistry-based sequencing method for N^2-methylated guanosine at the single-nucleotide level is described. This method takes advantage of the chemoselectivity of photoredox chemistry to enable selective photo-oxidation of m^2G and $m^2{}_2G$ in the presence of guanosine. Under blue light, using riboflavin as a photocatalyst and selectfluor as an oxidant, N^2-methylated guanosines convert into their respective 2,5-diamino-4H-imidazol-4-one (Iz) and 8-oxoguanine (OG) products resulting in different mutation signature after sequencing. The limitation and scope of this method were further evaluated and extended to explore other methylated guanosine types such as m^1G and generate the first database of high confidence m^2G sites in multiple RNA types in human lung cancer cell line.

Acknowledgment

I want to thank my supervisor, Professor Ryan Hili, for giving me the opportunity to pursue my PhD in his lab. Thank you for your guidance and feedback throughout these years.

I would also like to thank my committee members, Professor Derek Wilson and Professor Philip Johnson, for their valuable feedback throughout my graduate studies.

I am grateful to Molly Hu and Yasaman Mahdavi-Amiri for training me during my undergraduate and graduate studies, which helped shape me as a scientist.

To my lab mates, Morgan, Molly, Matina, Natalie, Ian, Emily, Luca, and Nicole, thank you for your friendship and support. It has been fun working with you.

Finally, I would like to thank my family and friends for always believing in me and encouraging me.

Table of Contents

List of Tables

List of Figures

Chapter 1: Introduction to RNA Modifications

1.1. Nucleic acids

Nucleic acids such as deoxyribonucleic acid (DNA) and ribonucleic acid (RNA) are large biomolecules that play crucial roles in living organisms. First introduced in the 1950s, the central dogma proposed that RNA carries the genetic information coded in DNA to make proteins. In reality, the biological functions of RNA expand far beyond serving as a template for protein expression; these molecules are involved in a wide range of processes including as ribozymes in protein synthesis and RNA splicing, or as riboswitches in gene regulation. New functions and forms of RNA continue to be discovered and contribute to the diverse nature of the RNA molecule. [1]

1.1.1. Basic structure of RNA

RNAs are polymers of nucleotides composed of a phosphate group, a 5-carbon ribose sugar, and a nitrogenous base. There are four canonical nucleobases categorized into two types of heterocyclic rings: purines (adenine (A), guanine (G)) and pyrimidines (cytosine (C), uracil (U)). Phosphodiester bonds link the sugars and phosphates in the chain. The 3'-hydroxyl group of one sugar connects a phosphate group to the 5'-hydroxyl group of the next sugar in the polymer. An *N*-glycosidic bond links the nucleobase ring nitrogen (*N*1 for pyrimidines and *N*9 for purines) to the 1' carbon of the sugar ring (Figure 1-1).[1] Several classes of RNA exist, of which the most relevant ones are discussed below.

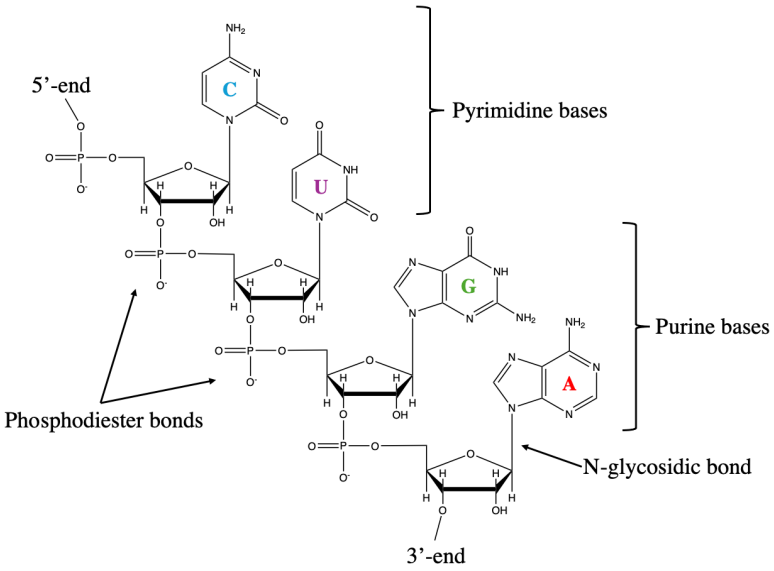


Figure 1-1: RNA structure showing 4 possible bases, ribose sugar, and phosphate backbone.

1.1.2. Ribosomal RNA (rRNA)

Ribosomal ribonucleic acid (rRNA) is a non-coding RNA type that forms the catalytic component of the ribosome essential for protein synthesis in all cells. In the nucleolus, pre-rRNA is synthesized by RNA polymerase I from ribosomal DNA (rDNA). Pre-rRNA undergoes processing to become a mature rRNA, which is used as a building block for ribosomes. The eukaryotic ribosome consists of two subunits with four rRNA species and more than 70 proteins. The large subunit (LSU) is composed of the 28S, the 5.8S, and the 5S rRNAs whereas the small subunit (SSU) only has the 18S rRNA. Both subunits are assembled to form the 80S unit and exported to the cytoplasm where they act as a

physical structure to push transfer RNA (tRNA) and messenger RNA (mRNA) through the ribosome for translation (Figure 1-2).[2]

1.1.3. Messenger RNA (mRNA)

Messenger ribonucleic acid (mRNA) is a single-stranded RNA molecule containing a sequence corresponding to a gene and read by the ribosome during protein synthesis. Pre-mRNA is transcribed from DNA by RNA polymerase II and undergoes a series of processing simultaneously in the nucleus. Pre-mRNA goes through 5' capping and polyadenylation at the 3' end that protects the transcript and helps it get exported to the cytoplasm. The introns on the pre-mRNA get spliced out by the spliceosome, leaving the exons that encode for a protein. The mature mRNA containing genetic information in the form of a nucleotide sequence is then exported to the cytoplasm for protein translation (Figure 1-2).[3]

1.1.4. Transfer RNA (tRNA)

Transfer ribonucleic acid (tRNA) is a small non-coding RNA that decodes mRNA sequence to synthesize proteins during translation. Pre-tRNA is transcribed from tRNA genes by RNA polymerase III in the nucleus that undergoes a series of tRNA maturation processes such as the removal of 5'-leaders and 3'-trailers sequences, nucleotide modifications, tRNA splicing, and aminoacylation to become functional. Each mature

tRNA contains a three-nucleotide anticodon that encodes the amino acid it carries and is complementary to a three-nucleotide codon on mRNA carrying a specific nucleotide sequence for a protein. During protein translation, tRNA is carried into the ribosome by elongation factors. If the tRNA anticodon is complementary to the mRNA codon, a peptide bond is formed between the amino acid on the existing tRNA and that of the incoming tRNA. The process continues to grow a polypeptide chain until a stop codon is read, which terminates translation thereby releasing a protein (Figure 1-2). Moreover, the nucleotides found on tRNA are heavily modified, which is important to its structure, as well as mRNA translation, and will be discussed in more detail later in this chapter.[4]

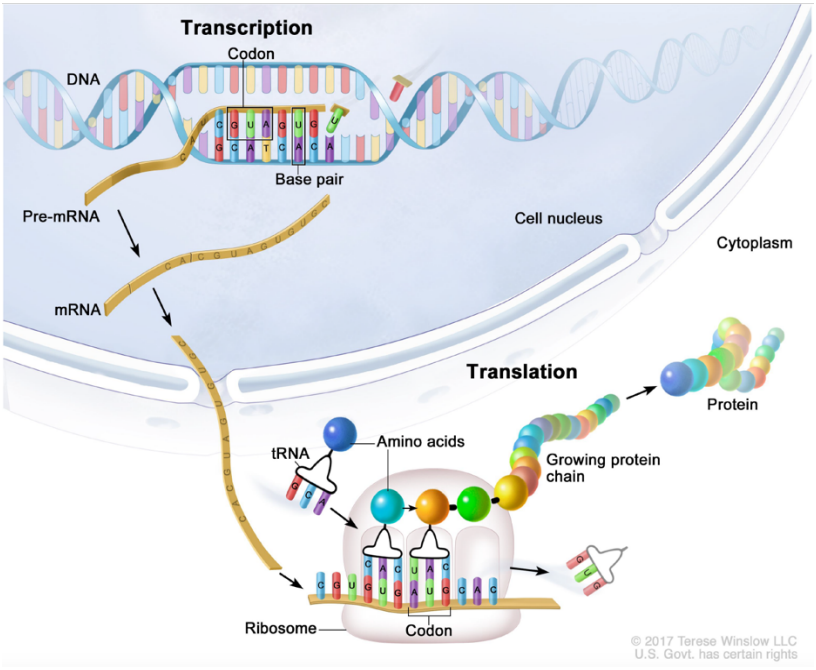


Figure 1-2: Transcription and translation processes involving several RNA species. Courtesy: National Human Genome Research Institute (https://www.cancer.gov/publications/dictionaries/cancer-terms/def/transcription).

1.1.5. MicroRNA (miRNA)

MicroRNA (miRNA) is a small non-coding RNA molecule ranging between 21-23 nucleotides involved in RNA silencing and regulation of gene expression. Most miRNAs are synthesized from DNA by RNA polymerase II/III to form a hairpin loop primary miRNA (pri-miRNA) capped at the 5' end, polyadenylated, and spliced post- or co-transcriptionally. The primary miRNA undergoes processing into a precursor miRNA (pre-miRNA) in the nucleus and is exported to the cytoplasm where it is cleaved into a mature miRNA. miRNA base pairs with mRNA to silence or regulate gene expression by cleaving the mRNA strand, shortening the poly-A tail, or reducing mRNA translation.[5]

1.1.6. Long non-coding RNAs (lncRNA)

Long non-coding RNA (lncRNA) consists of more than 200 nucleotides that regulate gene expression by interacting mostly with mRNA, DNA, protein, and miRNA. Most lncRNA are transcribed from intergenic region, intronic region, sense, and antisense transcripts by RNA polymerases. Similar to mRNA, lncRNA undergoes 5' capping, splicing, and polyadenylation; however, it does not translate into proteins. The biogenesis of lncRNA depends on their localization and function. LncRNA can modify chromatin structure and function, gene transcription, and affect RNA splicing, stability, and translation.[6]

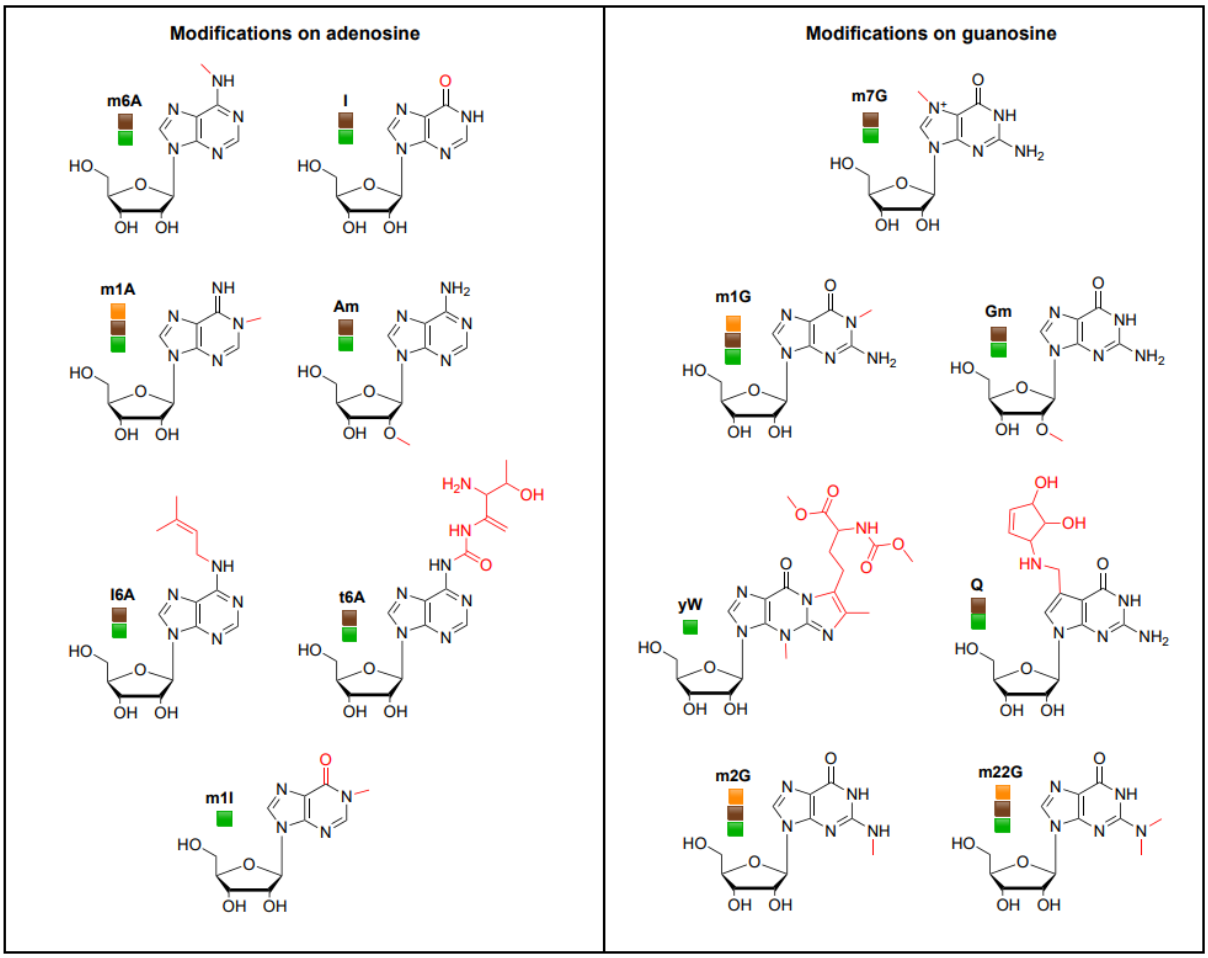
1.2. Epitranscriptomic modifications

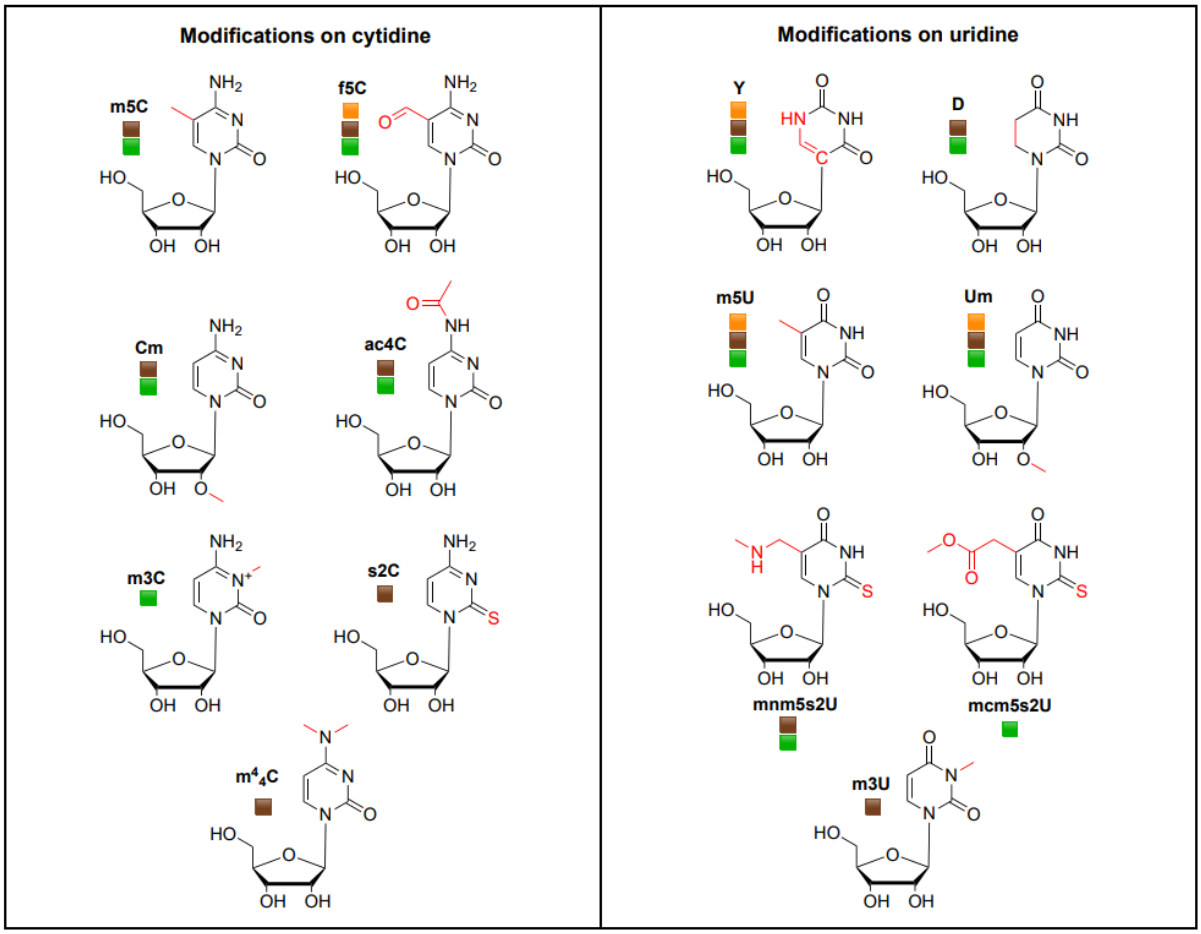
As discussed above, various types of RNA such as coding (mRNA) and non-coding (rRNA, tRNA, miRNA, lncRNA, etc.) RNA exists in the human transcriptome. These biological macromolecules can regulate their functions efficiently through chemical modifications. In fact, over 170 modifications have been identified on RNA that arise from a range of reactions such as methylation, deamination, thiolation, etc. All four RNA bases (adenine, cytosine, guanine, uracil) as well as the ribose sugar can be modified mostly for post-transcriptional regulation of gene expression.[7]

1.2.1. Brief history of RNA modifications

Pseudouridine (Ψ) was the first RNA modification discovered in 1957 by Davis and Allen who performed 2D chromatography on ribonucleic acids in *Saccharomyces cerevisiae*.[8] Years later, the oligonucleotide composition of yeast $tRNA^{ala}$ was elucidated, confirming 10 more modifications.[9] By 1995, 93 modified nucleosides were identified mostly in rRNA and tRNA; however, their function and prevalence were also unknown due to limitation in the technology available at the time. In the past few decades, several tools such as LC-MS/MS and next-generation sequencing (NGS) have characterized hundreds of modifications across all RNA types and domains of life. A curated list of RNA modifications, their location, modifying enzymes, and other important information are now available in the MODOMICS database[10]. Some examples of RNA modifications are illustrated in Table 1-1.[10,11]

Table 1-1: Examples of RNA modifications taken from MODOMICS database[10].





Domains: Archaea, Bacteria, Eukaryotes

1.2.2. Role of RNA modifications

Almost all RNA species contain many modified nucleotides essential for proper functioning such as folding, transcription, splicing, translation, RNA transport, and immune responses. While some of the modifications are always present at a specific position on RNA, other modifications are dynamic and can occur as a response to stimuli, such as heat shock or different toxins.[11] The significance of epitranscriptomic modification

research is tightly associated with their ability to promote or inhibit the formation of cancer by regulating cell proliferation, differentiation, invasion, migration, stemness, metabolism, and drug resistance.[7] Table 1-2 shows a list of modifications along with their functions, associated diseases, and methods to identify them.[11,12]

Table 1-2: Available information on selected RNA modifications.[11,12]

Base	Modification	Code	RNA classes	Function	Diseases	Method
A	N6-methyladenosine	m6A	rRNA, tRNA, mRNA, miRNA, lncRNA, circRNA, snRNA	mRNA: regulates splicing events, induces mRNA instability, increases translation efficiency; lncRNA: plays a role in lncRNA-mediated transcriptional repression	Osteoporosis, HCC, obesity	DRS, HPLC, MeRIP-seq, MiCLIP, DART-seq, Nitrite-seq
	Inosine	I	tRNA, mRNA, miRNA	tRNA: required for wobble codon recognition; mRNA: affects stability and localization, can change protein sequence	HCC; gastric, colorectal, esophageal, glioblastoma, and lung cancers	LC-MS/MS, RT-PCR, DRS, de novo RNA-seq
	N1-methyladenosine	m1A	rRNA, tRNA, mRNA, miRNA, lncRNA	rRNA: ensures proper folding; tRNA: ensures stability and proper folding	HCC; cervical, pancreatic, breast, and ovarian cancers	ARM-seq, m1A-seq, m1A-quant-seq, m1A-ID-seq, m1A-seq-SS, m1A-seq-TGIRT, m1A-MAP, m1A-IP-seq
	N6-isopentenyladenosine	I6A	tRNA	Increases translation fidelity and efficiency of cognate codons	Lung cancer	PHA6 assay, HPLC, Sanger
	N6-threonylcarbamoyladenosine	t6A	tRNA	Facilitates codon–anticodon pairing and prevents frameshift during protein synthesis	MERFF, neurodegeneration, diabetes	LC-MS/MS
	N1-methylinosine	m1I	tRNA	Unknown	Myelomonocytic leukemia, large cell lung carcinoma, breast cancer	LC-MS/MS
G	N7-methylguanosine	m7G	miRNA, rRNA, mRNA, tRNA	tRNA: increases stability; mRNA: regulates translation	HCC, PAD, lung cancer	m7G-MeRIP-seq, m7G-seq, DRS, m7G-miCLIP-seq, m7G-MaP-seq
	N1-methylguanosine	m1G	tRNA, mRNA	tRNA: prevents frameshifting; mRNA: reduces translation fidelity in a position- and codon-dependent manner	Several types of cancer	PhOxi-seq, MR-FISH, LC-MS/MS, DRS
	Wybutosine	yW	tRNA	Stabilizes codon–anticodon interactions	Breast cancer, leukemia	LC-MS/MS
	Queuosine	Q	tRNA	Impacts coding potential of tRNA, protects tRNA from ribonuclease cleavage	T-cell lymphoma, colon cancer	UHPLC-MS/MS, LC-MC
	N2-methylguanosine	m2G	tRNA, mRNA, snRNA	tRNA: plays a role in tRNA folding and prevents tRNA from adopting wrong conformation. Involved in cell proliferation, protein translation, pre-mRNA splicing	Prostate cancer	PhOxi-seq, LC-MS/MS
	N2, N2-dimethylguanosine	m22G	tRNA, mRNA	tRNA: plays a role in tRNA folding and prevents tRNA from adopting wrong conformation; mRNA: reduces translation fidelity in a position- and codon-dependent manner	Intellectual disability	PhOxi-seq, LC-MS/MS
C	N5-methylcytidine	m5C	rRNA, tRNA, mRNA, eRNA, miRNA, lncRNA, viral RNA, vault RNA, snRNA, snoRNA	rRNA: ensures proper folding and translational fidelity; tRNA: ensures proper folding and stability, codon–anticodon interactions, and reading frame maintenance; mRNA: affects stability, translation rate; lncRNA: affects stability; vault RNA: required for biogenesis	ARID, DS, lactic acidosis, breast cancer, hypotonia /floppy baby syndrome, metabolism	m5C-RIP-seq, Aza-IP-seq, miCLIP-seq, TAWO-seq, RNA-BisSeq, RBS-seq, DRS, MeRIP-seq
	5-formylcytidine	f5C	tRNA	Required for decoding of AUA codon in mitochondria	NA	HPLC-MS
	N4-acetylcytidine	ac4C	rRNA, tRNA, mRNA	tRNA: increases translation efficiency and accuracy; mRNA: increases stability	Laminopathy, Hutchinson-Gilford progeria syndrome, Malignant disease	acRIP-seq, ac4C-seq
	N3-methylcytidine	m3C	tRNA	Important for structural tuning, ribosome binding, and translation fidelity	Several types of cancer	AlkAniline-seq, HAC-seq, ARM-seq, DM-tRNA-seq, mim-tRNA-seq, hydro-tRNA-seq, LC-MS/MS, NAIL-MS
	2-thiocytidine	s2C	tRNA	Restricts recognition of certain wobble codons	NA	HPLC-MS
U	Pseudouridine	Y	rRNA, tRNA, mRNA, snRNA, MT-tRNA, scaRNA, snoRNA, miRNA, lincRNA	rRNA: plays a role in ribosome assembly and translational fidelity; tRNA: increases stability	Prostate, breast, and lung cancers; HCC	Pseudo-Seq, DRS, PSI-seq, CeU-Seq, Y-Seq, RBS-seq
	Dihydrouridine	D	tRNA, mRNA, snoRNA	tRNA: destabilizes the structure; mRNA: affects splicing and translation	Lung cancer	D-seq, Rho-seq, LC-MS
	5-methyluridine	m5U	rRNA, tRNA	Ensures fidelity of translation, folding, and stability of tRNA	Breast cancer, systemic lupus erythematosus	FICC-Seq, iCLIP, miCLIP-seq
	5-methylaminomethyl2-thiouridine	mnm5s2U	tRNA	Required for decoding of the lysine codons AAA and AAG by $tRNA^{Lys3}_{UUU}$	NA	LC-MS/MS
	5-methoxycarbonylmethyl-2-thiouridine	mcm5s2U	tRNA	Improves decoding efficiency of tRNA	Several types of cancer	HPLC-MS
	N3-methyluridine	m3U	rRNA, tRNA	Unknown	Ovarian cancer	MR-FISH, RP-HPLC
All	2'-O-methylation	Nm	rRNA, tRNA, mRNA, snRNA, snoRNA	Nm increases the thermodynamic stability of RNA:RNA base pairs and stabilizes A-form RNA duplexes	Asthma, Alzheimer's disease	RiboMethSeq, 2D-TLC, Nm-seq, 2'OMe-seq, RTL-P, DRS, RibOxi-seq

NA—information not found. **RNA classes:** mRNA—messenger RNA; rRNA—ribosomal RNA; tRNA—transfer RNA; miRNA—microRNA; lncRNA—long noncoding RNA; snRNA—small nuclear RNA; snoRNA—small nucleolar RNA; eRNA—enhancer RNA; Mt-tRNA—mitochondrial tRNA; scaRNA—small Cajal body-specific RNA; lincRNA—long intergenic noncoding RNA. **Diseases:** HCC—hepatocellular carcinoma; MERFF—myoclonic epilepsy with ragged red fibers; PAD—peripheral arterial disease; ARID—autosomal recessive forms of intellectual disabilities; DS—Dubowitz syndrome. **Method:** LC-MS/MS—liquid chromatography–mass spectrometry; RP-HPLC—reversed-phase liquid chromatography–tandem mass spectrometry; MeRIP-seq—methylated RNA immunoprecipitation sequencing; miCLIP—methylated individual-nucleotide-resolution crosslinking and immunoprecipitation; PA-m6A-seq—photo-crosslinking-assisted m6A sequencing; m6A-CLIP—m6A individual-nucleotide-resolution crosslinking and immunoprecipitation; DRS—direct RNA-seq; PhOxi-seq—Photo-oxidative sequencing.

1.3. RNA modifying enzymes

Many RNA modifications can be reversibly installed by RNA-binding proteins (RBPs) classified as writers (enzymes that deposit modification) and erasers (enzymes that remove modification). Reader proteins (RBPs that bind to modification) can detect the modification on RNA and execute their biological function. The dysregulation of RNA modifications and RNA-modifying enzymes are involved in diseases such as cancer, neurological diseases, and other disorders.[12]

1.3.1. Writers

Methylation is one of the most common types of RNA modification and is catalyzed by methyltransferases (MTs), which can be grouped into several subclasses. Class I is the most common class of methyltransferases containing a Rossmann fold. MTs use this

tertiary structure for binding the co-substrate S-adenosyl-methionine (SAM), which acts as an electrophilic methyl donor. Class II methyltransferases consist of a SET (suppressor of variegation, enhancer of zeste and trithorax) domain and include histone methyltransferases whereas class III methyltransferases are membrane-associated enzymes.[13]

In 1994, METTL3 was the first methyltransferase-like gene family member to be identified, which transfers a methyl group from SAM to the N^6 position of the adenosine base forming m^6A together with other writers and co-factors such as METTL14, Wilms Tumor 1 Associated Protein (WTAP), KIAA1429, RNA Binding Motif Protein 15 (RBM15), and zinc finger CCCH domain-containing protein 13 (ZC3H13).[14] METTL3 is the catalytic subunit that plays a role in cancer as an oncogene or tumor suppressor by installing m^6A on critical transcripts. METTL3 is upregulated in liver cancer, more specifically, hepatocellular carcinoma (HCC) resulting in poor patient survival. METTL3 deposits m^6A on a tumor suppressor gene called suppressor of cytokine signaling 2 (SOCS2) promoting its mRNA degradation resulting in metastasis of HCC. Conversely, METTL3 knockdown in glioblastoma stem cells (GSC) leads to tumor progression suggesting that m^6A on mRNA is involved in regulating self-renewal and tumorigenesis in GSC.[15]

Many other RNA methyltransferases exist such as NSUN, MePCE, BCDIN3D, ALKBH8, RNMT, and TRMT. Moreover, different writers recognize different sites and RNA structures to install modifications. For example, METTL3/14 complex methylate adenosine within the motif RRACH (R stands for G, A or U; H stands for U, A or C) in mRNA, while METTL16 prefers the UAC(m6A)GAGAA sequence in the stem-loop of U6

small nuclear RNA. Other non-methyltransferase writers include DKC1, NAT10, and TRIT1, which are responsible for installing pseudouridine (Ψ), N^4-acetylcytosine (ac^4C), and N^6-isopentenyladenosine (I^6A), respectively.[12]

1.3.2. Erasers

In 2011, the first m^6A demethylase, fat mass and obesity-associated protein (FTO) was found to efficiently demethylate m^6A in both DNA and RNA. FTO belongs to the ALKB family of Fe(II)/α-ketoglutarate–dependent dioxygenases and binds preferentially to introns of pre-mRNA. FTO can demethylate other methylated RNA substrates such as m^6Am affecting mRNA splicing, stability, decay, and translation.[16] Found mostly in the nucleus and cytoplasm, FTO regulates fatty acid mobilization in fat cells and consequently controls body weight. Indeed, a deficiency in FTO leads to an increase in m^6A level in Angptl4 (angiopoietin-like protein 4) transcripts, which, in turn, lowers Angptl4 protein expression level. Angptl4 is known to regulate glucose homeostasis, lipid metabolism, and insulin sensitivity. As a result, low expression of Angptl4 leads to an increase in fatty acid deposition in adipose tissues and eventually obesity.[17]

AlkB homolog 5 (ALKBH5) is the second type of m^6A demethylase, which displays comparable demethylation activity to FTO. However, ALKBH5 prefers demethylating m^6A in a consensus sequence rather than other methylated RNA substrates. ALKBH5 is found in the nucleus and its depletion results in reduced mRNA levels in the cell indicating its role in regulating mRNA nuclear export. ALKBH5 is expressed in most tissues and is

dysregulated in several types of human cancers and non-cancers. For example, ALKBH5 is elevated in non-small cell lung cancer (NSCLC) and the knockdown of ALKBH5 increased m^6A level in FOXM1 (Forkhead Box M1), which controls cell proliferation. This, in turn, stops the proliferation and invasion of lung adenocarcinoma cells suggesting that ALKBH5 was upregulated in NSCLC cells.

ALKBH5 is also associated with non-cancers such as reproductive system diseases. A deficiency in ALKBH5 leads to an increase in m^6A level in mRNA, which is connected to compromised spermatogenesis in male mice thus affecting RNA metabolism. Moreover, depletion of ALKBH5 leads to infertility in male mice due to abnormal splicing and shorter transcript production.[18] There are other erasers such as ALKBH3 and H29K that remove N^1-methyladenosine (m^1A) and N^7-methylguanosine (m^7G), respectively. Erasers of several modifications are either unknown or do not exist due to an irreversible modification process.[19]

1.3.3. Readers

Epitranscriptome modifications can be recognized by proteins called readers producing a cellular response. The most well-studied family of readers, YT521-B homology (YTH) domain proteins, reads m^6A by recognizing RRACH sequences and fulfills different functions depending on its location. If the mRNA is found in the nucleus, YTH N^6-Methyladenosine RNA Binding Protein C1 (YTHDC1) binds m^6A and helps exon selection

during splicing. As for mRNA in the cytoplasm, YTH N^6-Methyladenosine RNA Binding Protein F1 (YTHDF1) binds m^6A to increase translation efficiency.

Another reader, YTH N^6-Methyladenosine RNA Binding Protein F2 (YTHDF2) binds to m^6A and recruits Carbon Catabolite Repression—Negative On TATA-less (CCR4-NOT) to deadenylate and further degrade m^6A mRNA. Moreover, m^6A can also be read by the Insulin-like growth factor 2 mRNA-binding proteins (IGF2BP) family of proteins increasing mRNA stability. YTH domain proteins such as YTHDF1–3 and YTHDC1 can also read N^1-methyladenosine (m^1A). N^5-methylcytidine (m^5C) can be read by Aly/REF Export Factor (ALYREF) to promote nuclear export and Y-box-binding protein 1 (YBX1) in the cytoplasm, which recruits ELAV like RNA binding protein 1 (ELAVL1) to improve mRNA stability.[12,19]

Due to their governance of root cellular processes, dysregulation of epitranscriptome readers is associated with diseases. Recent studies have found that YTHDF2 is upregulated in triple-negative breast cancer (TNBC). YTHDF2 negatively regulates cell death by promoting the degradation of Serine Protease 23 (Prss23) mRNA, a gene involved in translation. Knocking-down YTHDF2 increases Prss23 expression level and protein translation triggering cancer cell death. In chondrocytes, YTHDF1 also negatively regulates cell death by stabilizing mRNA of anti-apoptotic B-cell lymphoma 2 (Bcl-2).[20] Readers of several other modifications remain unknown.

1.4. Tools to map nucleotide modifications

Throughout the past few decades, multiple techniques to detect, quantify, and map the precise location of RNA modifications have been developed as a means to further understand their biological role across the transcriptome. Each modification has unique properties that have been explored for their detection but also contributed to the limitations of their respective methods. The existing detection approach can be divided into several groups such as antibody (Ab)-based techniques, natural or induced reverse-transcriptase (RT) signatures, and chemistry-based methods. These techniques can be used in parallel to validate one another.

1.4.1. Antibody-based methods

The first m^6A-Ab-based enrichment method was developed in the 1980s[21] but only started being revolutionary when it was coupled to deep sequencing. The mapping of m^6A in human transcriptome by methylated RNA immunoprecipitation (meRIP)-seq immunocapturing and sequencing remains one of the most frequently used technologies in the field (Figure 1-3). The basic principle of this method relies on immunocapturing RNA fragments containing the modification of interest followed by elution of modification-enriched RNA by either free modified nucleoside competition or digestion with ionic detergent and proteinase K. The eluted enriched RNA is then subjected to high-throughput sequencing for analysis.[22] Initially, the resolution of meRIP-seq was limited to

100-200 nt where the exact location is unknown. For peak calling, bioinformatic pipelines are used to look for read accumulation in immunoprecipitated RNA regions.

The resolution of meRIP-seq was further improved by combining it with Ab cross-linking, which allowed for single-nucleotide resolution due to the analysis of reverse transcriptase (RT)-stops and/or misincorporations. This method is called methylation individual-nucleotide resolution UV cross-linking and immunoprecipitation (miCLIP) (Figure 1-3) adapted from cross-linking immunoprecipitation (CLIP). miCLIP uses an antibody for m^5C writer protein to identify m^5C modification in non-coding RNA.[23] Antibody-related techniques have also been applied to detect modifications such as ac^4C, m^1A, and m^7G. However, these antibody-based methods have inherent disadvantages including unpredictable specificity of the antibody due to batch-to-batch variability and non-specific binding of RNA to antibody.

In addition, these antibody-based techniques are not quantitative since the signals obtained from peak enrichment do not correspond linearly to the modification level. A quantification method, m^6A-level and isoform-characterization sequencing (m^6A-LAIC-seq), that spikes in mRNA standard allowing for a more precise approximation of m^6A stoichiometry was later introduced. Combining antibody-based methods with enzyme/chemical treatments to induce reverse-transcriptase (RT)-stops and/or misincorporation can improve detection resolution or lead to a more robust approach.[24,25]

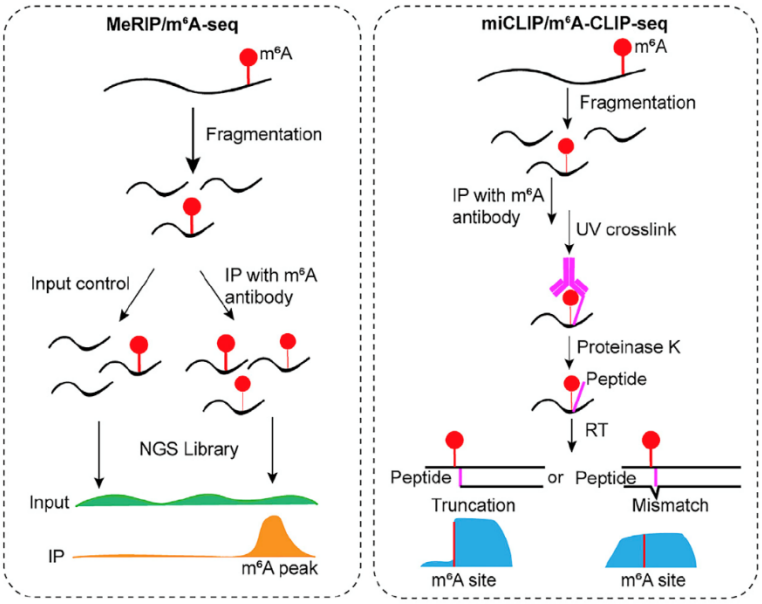


Figure 1-3: Schematic diagram of antibody-based sequencing technologies: MeRIP-seq/m6A-seq and m6A-CLIP-seq/miCLIP. [26]

1.4.2. Natural reverse-transcriptase (RT) signature-based methods

Some modified nucleotides are not recognized by reverse transcriptases due to steric hindrance and thus can no longer base pair with canonical bases. As a result, some modifications cause reverse transcriptases to misincorporate a non-Watson-Crick compatible deoxynucleotide triphosphate (dNTP), others can cause reverse transcriptases to stop during complementary DNA (cDNA) synthesis. These create an RT signature depending on the type of modification, reaction conditions, and reverse transcriptase, which can be useful for identifying modification sites.[24]

A computational technique, high-throughput annotation of modified ribonucleotides (HAMR), was developed to detect sequence patterns of modifications after reverse transcription at the single nucleotide level. This method can both locate modifications transcriptome-wide and distinguish classes of modifications. However, if the mutation signatures of different modifications are too similar, they must be merged to be resolved. For instance, two sets of adenosine modifications were grouped into combined classes $m^1A|m^1I|ms^2i^6A$ and $i^6A|t^6A$, which could be distinguished with a 98 % accuracy. Conversely, guanosine modification types m^1G and $m^2G|m^2{}_2G$ could only be predicted with a 78 % accuracy. Lower accuracy is observed since both m^1G and $m^2{}_2G$ strongly favor the incorporation of A with varying amounts of G and T. Despite the limitations of this software, HAMR identified potential m^3C, m^1A, m^1I, $m^2{}_2G$, and m^1G sites as well as validated three m^3C sites in human tRNAs.[27]

Some modifications can still form Watson-Crick base-pairing with canonical bases during reverse transcription remaining RT-silent. However, certain conditions were found to make the RT signature visible. In the case of 2'-O-methylations, reduced dNTP concentrations during primer extension create RT-stops used to map those modifications in rRNAs. As for m^6A, substituting O4 of thymidine with selenium (4Se-dTTP) induces RT-stops resulting in a change in signature. Combined with FTO demethylation, another m^6A detection strategy with single nucleotide resolution was developed.[28] Although RT-based detection methods can be useful for certain modifications, introducing an RT signature can be sequence dependent and is not always reliable due to reverse transcriptase bias.[24,29]

1.4.3. Engineered reverse-transcriptase (RT) signature-based methods

Instead of exploiting reaction conditions to produce RT signatures or relying on modifications to induce RT-stops, RT enzymes can be engineered to display sensitivity to specific modifications. For example, KlenTaq DNA polymerase was generated to show preference for 2'-O-methylations in RNA at normal dNTP concentrations but has yet to be used with high-throughput RNA sequencing.[30] Additionally, another KlenTaq DNA polymerase mutant was developed to increase misincorporation at m^6A sites, which is also awaiting transcriptome-wide analysis.[31]

Known demethylase enzymes such as the AlkB family, can be used to remove methyl groups. Indeed, AlkB-facilitated RNA methylation sequencing (ARM-seq), uses *Escherichia coli* AlkB to demethylate m^1A, m^3C, and m^1G in tRNAs, thereby removing RT-blocking modifications.[32] Wild-type AlkB was reported to efficiently demethylate m^1A and m^3C but works poorly on m^1G. To overcome this, another sequencing method, DM-tRNA-Seq added an AlkB D135S mutant of improved activity to efficiently convert m^1G to G.[33] N^2,N^2-dimethylguanosine ($m^2{}_2G$) was found to create stops upon demethylase treatment, therefore AlkB D135S/L118V mutant was generated to partially demethylate $m^2{}_2G$ into m^2G, which can be read as unmodified G by reverse transcriptases for sequencing purposes.[34]

In addition to eraser proteins, readers can also be explored to target modified nucleotides. In fact, deamination adjacent to RNA modification targets (DART-seq) fuses cytidine

deaminase APOBEC1 to m^6A-binding YTH domain resulting in C-to-U deamination at sites adjacent to m^6A residues, which appear as C-to-T mismatches in sequencing data.[35] Some RNA exonucleases can bind modified nucleotides instead of cleaving them under certain conditions such as *Escherichia coli* Endonuclease V (eEndoV) binds inosine in the presence of Ca^{2+}. Endonuclease V immunoprecipitation enrichment sequencing (EndoVIPER-seq) was developed to enrich A-to-I editing transcripts for sequencing.[36] Lastly, in RNA-endoribonuclease-facilitated sequencing (m^6A-REF-seq)/MAZTER-seq, the cleaving properties of *E. coli* RNA endoribonuclease MazF target only the unmethylated 5'-ACA-3' motif allowing for m^6A detection.[37] Enzymes provide an additional outlet for developing novel sequencing methods to identify modifications.

1.4.4. Chemistry-based methods

An alternative to antibody-based or enzyme-based detection methods is taking advantage of the physical properties of certain modifications. The chosen reagent selectively reacts with either the modified or unmodified nucleotide. Hence, chemical treatment prior to RT reaction can modulate the incorporation of dNTP or RT-stops during cDNA synthesis. One of the most common applications of this strategy is bisulfite sequencing (BS-seq), which involves Michael addition to deaminate cytosine to uracil leaving m^5C unreacted (Figure 1-4). This method is suitable for quantification and single-nucleotide mapping of m^5C in tRNA and rRNA.[38]

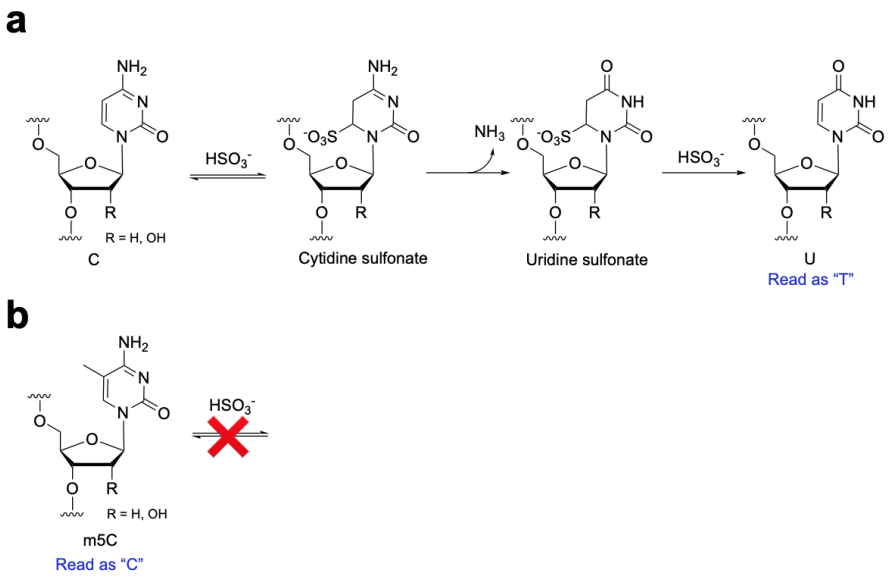


Figure 1-4: Schematic diagram of bisulfite reaction chemoselectively deaminates C into U, leaving m^5C unreacted. a) Unmodified C is read as T and b) m^5C is read as C after bisulfite treatment.

Inspired by the concept of bisulfite sequencing, our lab developed an m^6A sequencing method, nitrite-sequencing, to chemoselectively deaminate adenine leaving m^6A in the nitrosamine form. Under mild aqueous acidic conditions, nitrite forms a nitrosonium ion. The aromatic amine of nucleotides attacks the highly reactive nitrosonium ion to form nitrosamine. The primary aromatic amine of adenine can dehydrate into the diazonium ion whereas the secondary aromatic amine like m^6A lacks the N-H needed for dehydration. The diazonium ion of adenine subsequently hydrolyzes into inosine (Figure 1-5). As a result, the nitrite reaction is selective for the primary exocyclic amine of unmodified bases over the secondary exocyclic amine of methylated bases. Polymerases can read I as G resulting in an A → G transition and the nitrosamine m^6A is read as A,

which can be distinguished during high-throughput DNA sequencing.[39] Shortly after, another lab applied nitrite sequencing to the mammalian transcriptome and developed GLORI-seq (glyoxal and nitrite-mediated deamination of unmethylated adenosines).[40]

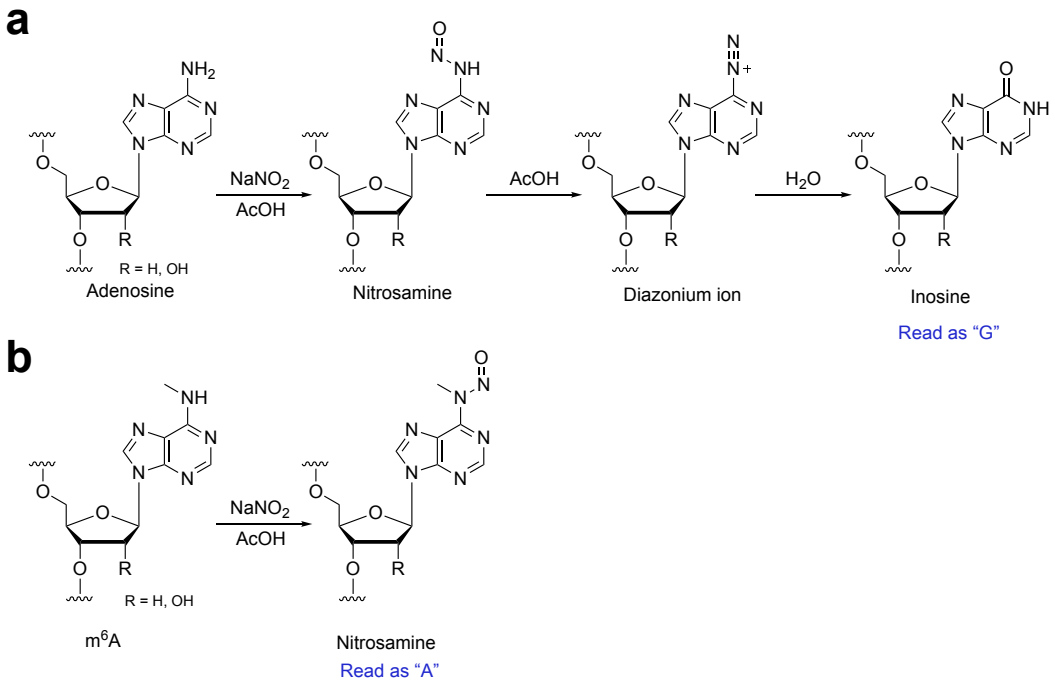


Figure 1-5: Nitrite reaction selectively deaminates adenosine into inosine leaving m^6A in its nitrosamine form. a) Unmodified A is read as G and b) m^6A is read as A after nitrite treatment.

There are other sequencing methods that exploit the chemical reactivity of modified or unmodified nucleobases to explore methylomes. Thiol(SH)-linked alkylation for the metabolic sequencing of RNA (SLAM-seq) detects 4-thiouridine (s^4U) by alkylation with iodoacetamide thereby converting s^4U into C.[41] This method was used to identify s^4U in tRNAs and mostly in metabolic labeling experiments. Furthermore, polymerases and

reverse transcriptases can read inosine as guanosine and do not cause RT-stops. Therefore, inosine chemical erasing sequencing (ICE-seq) was developed to identify inosine by alkylation with acrylonitrile thereby causing RT-arrest. This technique is reliable for identifying inosines transcriptome-wide.[42] Lastly, tRNA reduction and cleavage sequencing (TRAC-seq) allows the reduction of m^7G with $NaBH_4$ making it prone to aniline-induced cleavage and is identified as RT-stops. This approach provides single-nucleotide profiling of m^7G modification sites in tRNA.[43] Each of these chemical reaction-based methods has its challenges and limitations that may or may not be resolved by coupling them with other experimental methods.

1.5. Importance of studying modifications

As discussed earlier, epitranscriptome modifications are essential for the proper functioning of all living organisms. Both RNA modifications and RNA modifying enzymes can become dysregulated causing a plethora of diseases ranging from disorders to cancers. These diseases occur mainly through the downregulation of suppressor gene or the activation of genes that contribute to those diseases.[44] Cancer being one of the deadliest types of diseases, also encounters challenges such as drug resistance, limited therapeutic success, and toxic side effects from treatments.

Many anti-cancer strategies are continuously emerging, especially gene therapy that can regulate the expression of specific genes to treat those life-threatening diseases. Gene expression consists of several steps that can be regulated such as transcription, RNA

splicing, translation, and post-translational modification. Most of them involve the help of RNA, which is a crucial component of the central dogma. Detecting and identifying the exact location of RNA modifications lays the foundation for understanding their 'writing-erasing-reading' mechanism in diseases thus highlighting their potential in the diagnosis, treatment, and prevention of those diseases.[45]

The structures of many RNA modifying enzymes have been solved leaving a potential window for drug design. Indeed, small molecule inhibitors or activators have been developed for mostly m^6A-related enzymes such as METTL3, FTO, and ALKBH5 that show promising properties for developing drugs in the future.[46] Other anti-tumor strategies use RNA as therapeutic targets such as siRNA-mediated RNA interference (RNAi), RNA-targeted small molecules, and covalent chemical modification of RNAs.[45]

1.6. Aim of PhD thesis

The limited availability of antibodies and challenges of engineering enzymes for reverse transcriptase signature-based detection methods still impede the functional studies of RNA modifications. Due to the limited availability of sequencing methods for methylated guanines, this thesis aims to take advantage of the chemical properties of guanine and its modifications to develop a novel chemistry-based sequencing method at the single-nucleotide level.

Chapter 2: PhOxi-Seq: Single-Nucleotide Resolution Sequencing of N^2-Methylation at Guanosine in RNA by Photoredox Catalysis[1]

[1] Reprinted with permission from *J. Am. Chem. Soc.* 2022, 144, 13, 5723–5727. Copyright 2022 American Chemical Society.

My contribution to the article was: (i) planning all the experiments, (ii) performing all experiments, (iii) interpreting the results, and (iv) preparing some of the figures. My lab mentor, Yasaman Mahdavi-Amiri, was working on a different project involving the blue light reaction and the insights of the optimization steps were used to develop PhOxi-seq. Christopher Korfmann, an undergraduate student in our lab, wrote the script to analyze the sequencing data.

2.1. Abstract

Chemical modifications regulate the fate and function of cellular RNAs. Newly developed sequencing methods have allowed a deeper understanding of the biological role of RNA modifications; however, the vast majority of post-transcriptional modifications lack a well-defined sequencing method. Here, I report a photo-oxidative sequencing (PhOxi-seq) approach for guanosine N^2-methylation, a common methylation mark seen in N^2-methylguanosine (m^2G) and N^2,N^2-dimethylguanosine ($m^2{}_2G$). Using visible light-mediated organic photoredox catalysis, m^2G and $m^2{}_2G$ are chemoselectively oxidized in the presence of canonical RNA nucleosides, which results in a strong mutation signature observed during sequencing. PhOxi-seq was demonstrated on various tRNAs and rRNA to reveal N^2-methylation with excellent response and markedly improved read-through at $m^2{}_2G$ sites.

2.2. Introduction

Post-transcriptional modifications can have profound effects on the function and processing of RNA.[47] Not surprisingly, their dysregulation has been implicated in numerous human diseases,[48] especially various types of cancer.[49] With over 160 known RNA modifications identified across all life,[50] deciphering their biological roles remains a significant unmet challenge. Sequencing technologies that can resolve RNA modifications at single-nucleotide resolution have been instrumental to functional studies.[51] Indeed, methods that have mapped the locations of N^6-methyladenosine (m^6A), 5-methylcytidine (m^5C), inosine (I), pseudouridine (Ψ), N^1-methyladenosine (m^1A), and 5-hydroxylmethylcytidine (hm^5C) have provided deeper insights into their impact on RNA function and metabolism[52] and found application to disease diagnostics, prognostics, and epigenetic therapy.[53] Nevertheless, most post-transcriptional modifications lack a defined sequencing approach, which has hampered their functional elucidation.

N^2-Methylguanosine (m^2G) and N^2,N^2-dimethylguanosine ($m^2{}_2G$) are prevalent modifications in RNA (Figure 2-1). While mostly believed to be involved in structural tuning of RNAs,[54] recent reports have begun to shed light on the biological significance of N^2-methylation of guanosine. Studies have shown that upregulated m^2G in sperm tRNA-derived small RNAs (tsRNAs) is an epigenetic factor that mediates intergenerational inheritance of metabolic diseases.[55,56] Furthermore, TRMT11, an m^2G writer, has been implicated in poor prognosis in prostate cancer, while mutations in TRMT1, an $m^2{}_2G$ writer, have been tied to aberrant human brain development[57] and autosomal-recessive intellectual disabilities.[58] Developing robust functional studies on N^2-methylation of

guanosine has been limited by the absence of sequencing methods. While $m^2{}_2G$ blocks Watson−Crick−Franklin base-pairing and can often be detected during sequencing due to its high error signature,[59] only computational predictors exist for m^2G,[60−62] with no direct sequencing method available. The development of a robust single-nucleotide resolution sequencing method for N^2-methylation of guanosine in RNAs is expected to enable a better understanding of these modifications in human health and disease.

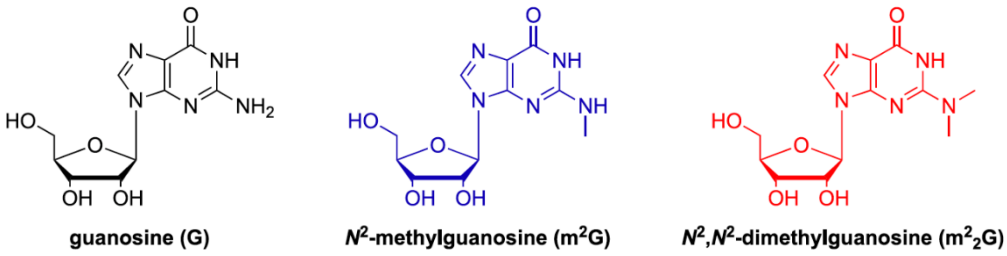


Figure 2-1: Guanosine and its N^2-methylated derivatives.

Guanosine is the most readily oxidizable nucleoside; its photo-oxidation in nucleic acids is known to occur by single electron oxidation to generate the guanine radical cation (Figure 2-2a).[63] This species can degrade via two oxidative pathways, and the outcome is governed largely by the nucleic acid environment.[64] In single-stranded or double-stranded forms, guanine is oxidized to 2,5-diamino-4H-imidazol-4-one (Iz), while in G-quadruplex form, guanine is preferentially oxidized to 8-oxoguanine (OG). Both Iz and OG are mutagenic forms of guanine. OG is known to base-pair with adenine, resulting in a G → T transversion, while Iz is known to base-pair with guanine, resulting in a G → C transversion (Figure 2-2a).[65] With these oxidation outcomes in mind and inspired by

examples of visible light photoredox chemistry on RNA[66,67] and DNA,[68] I sought to explore the prospect of visible light photoredox catalysis as a means to detect N^2-methylation at guanosine sites by oxidative mutagenesis (Figure 2-2b).

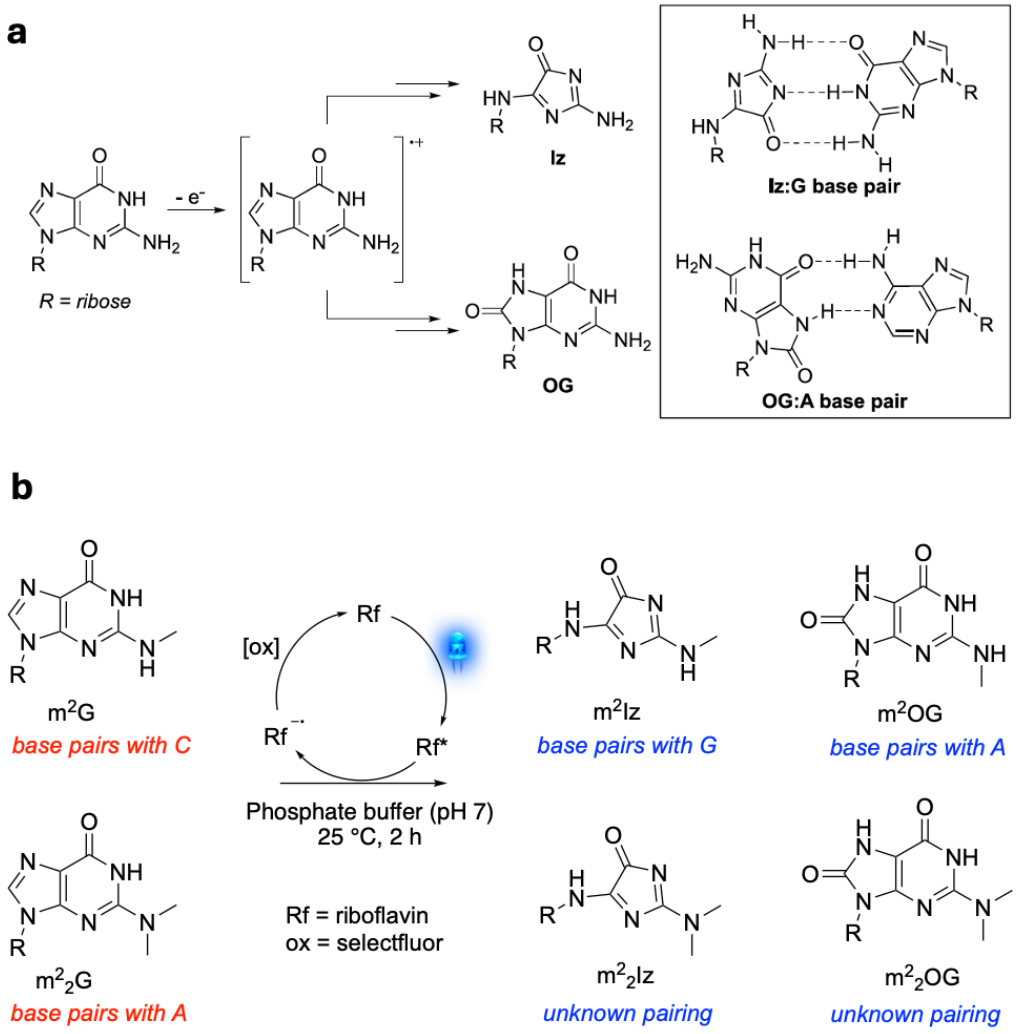


Figure 2-2: a) Oxidation pathway of guanosine into mutagenic derivatives along with their base-pairing partners (inset). b) Potential products formed during the photo-oxidation of N^2-methylated guanosine derivatives.

2.3. Results and discussion

2.3.1. Photocatalysis reaction with N^2-methylated guanosines

I first determined the oxidation potential of N^2-methylated guanosines compared with unmodified guanosine using cyclic voltammetry (Figure 2-3). The results showed that the oxidation potential of both $m^2{}_2G$ and m^2G were approximately 0.1−0.2 V below that of guanosine, providing a potential window for chemoselective oxidation. Using $m^2{}_2G$ as a model, I surveyed various photoredox conditions that were previously shown to be RNA compatible and found that riboflavin under blue light as a photocatalyst and Selectfluor as an oxidant were optimal.

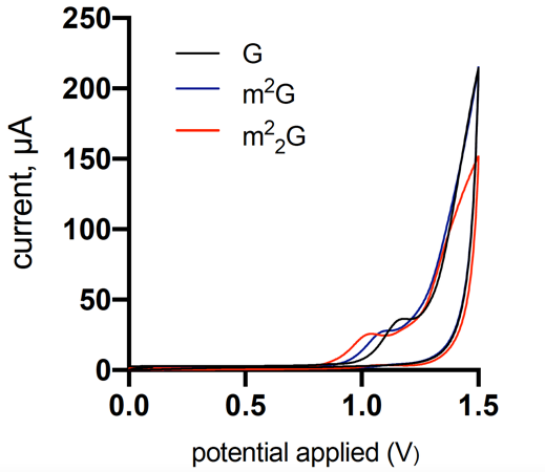
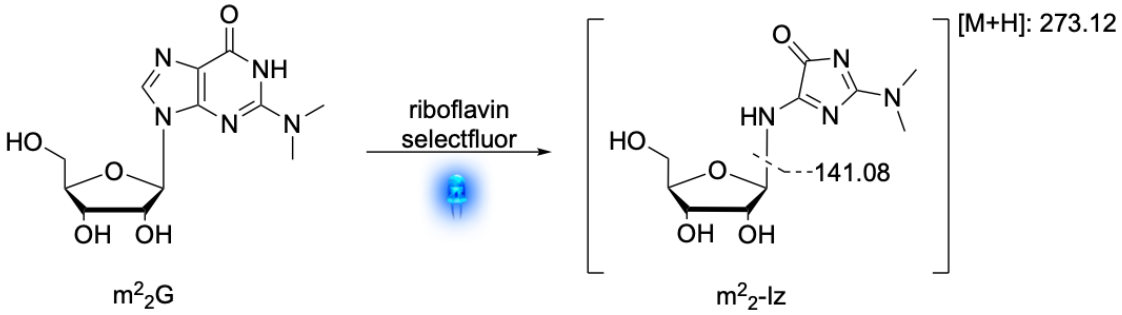
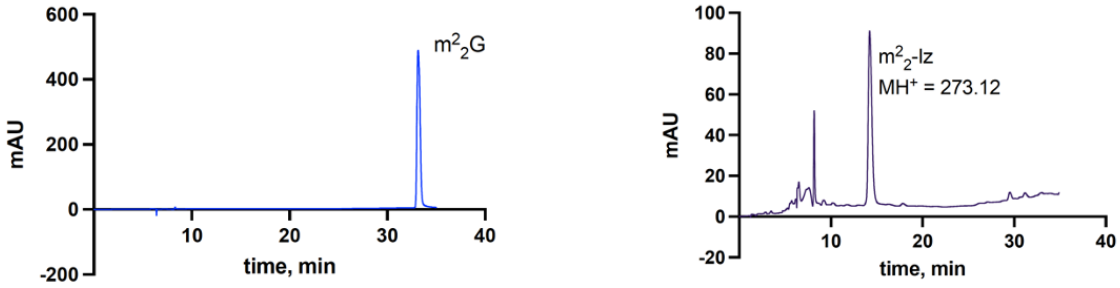


Figure 2-3: Cyclic voltammograms of guanosine (black line), m^2G (blue line), and $m^2{}_2G$ (red line) in 1 M KCl. Scan rate, 10 mV/s versus Ag/AgCl.

Under these conditions, full conversion of m^2_2G occurred within 2 h, resulting in the formation of the corresponding m^2_2-Iz, which was confirmed by mass spectrometry (Figure 2-4a-b). Formation of m^2_2-OG or N^2-demethylated by-products were not observed. On the other hand, the oxidation product of m^2G is unstable, and isolation of the pure product proved challenging by HPLC. Therefore, crude LC-MS analysis of photoredox-treated m^2G was performed showing conversion into m^2-Iz (Figure 2-5).



a



b

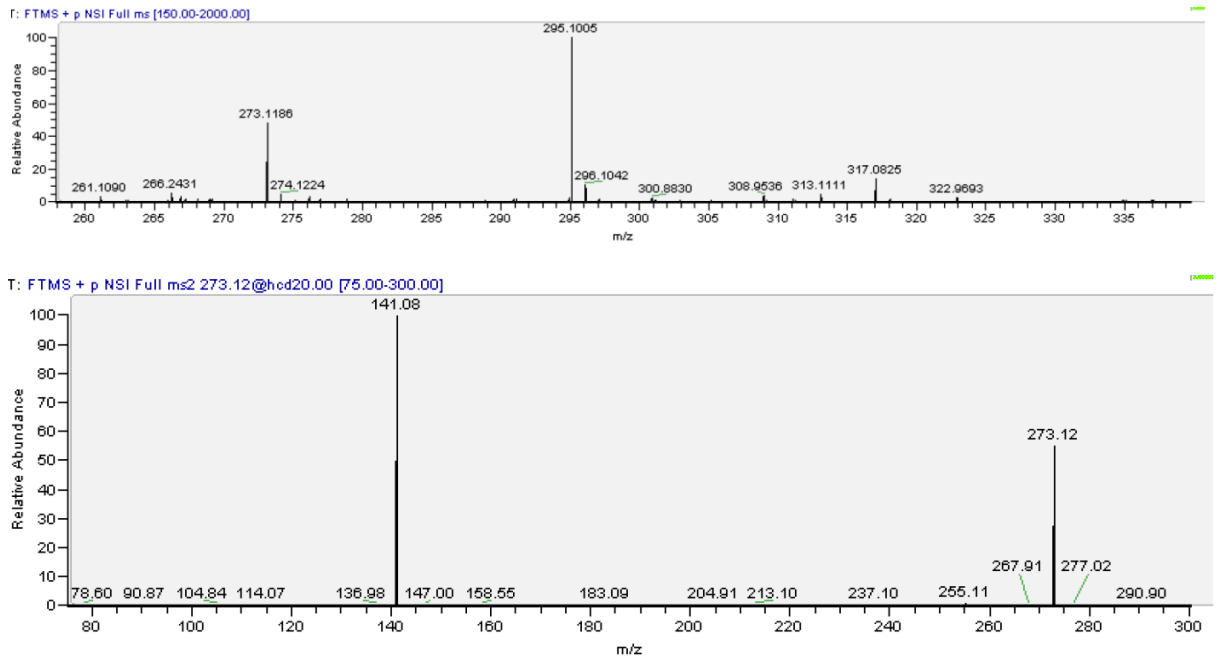


Figure 2-4: a) HPLC analysis showing conversion of m^2_2G into m^2_2-Iz within 2 hrs; note that peaks less than 10 minutes are by-products of the photoredox reagents and catalyst. b) Analysis by both mass spectrometry and MS/MS confirms m^2_2-Iz product.

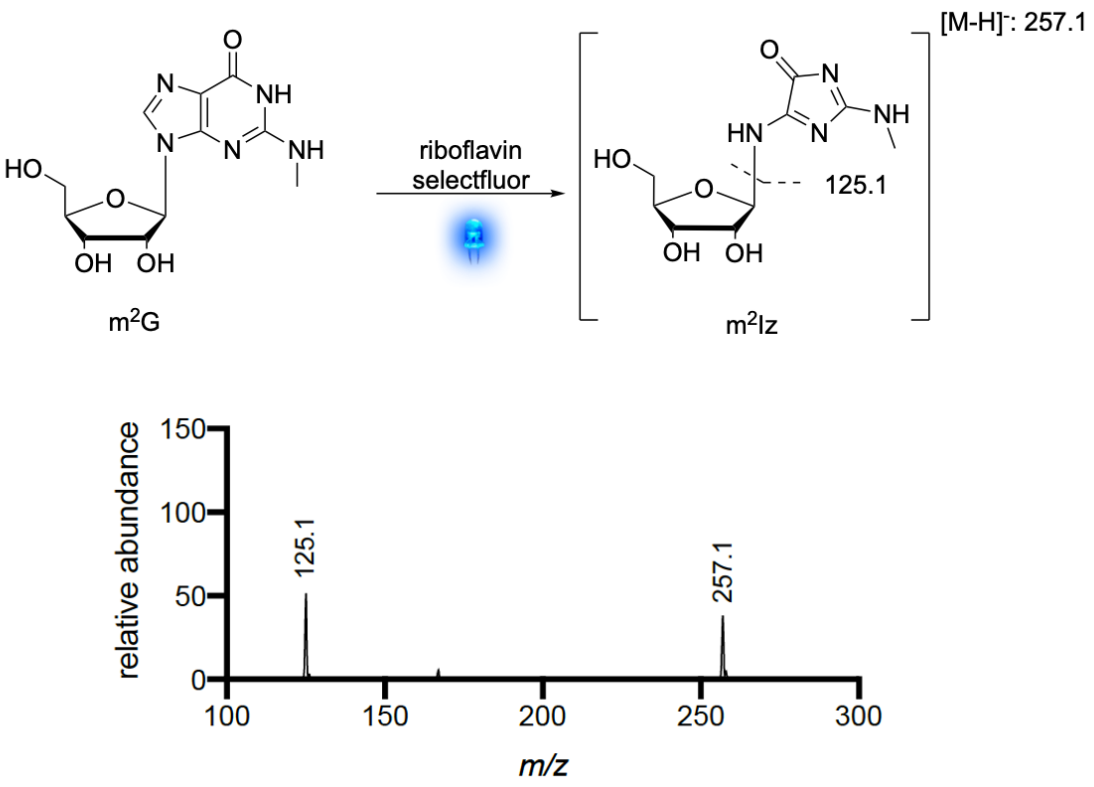


Figure 2-5: Crude LC-MS analysis in negative-ESI mode of photooxidation of m^2G into m^2Iz. Conditions: 0.5 mL/min 5 mM ammonium acetate with 5% MeOH, InfinityLab Poroshell 120 EC-C18 (4.6 × 100 mm, 4 µm) on an Agilent single quadrupole LC/MSD, frag = 200.

2.3.2. Photocatalysis reaction with unmodified RNA

Riboflavin exhibits an excited state reduction potential of E^{T1}= 1.37 V vs SCE, which could oxidize guanosine in RNA at low levels, potentially causing issues in some sequencing applications. Indeed, guanosine oxidation with riboflavin with higher energy (UVA) photoexcitation has been observed.[69] To test if this undesired process was occurring in our system, I subjected an RNA oligomer to the photoredox conditions and analyzed the

nucleoside composition by HPLC after nuclease digestion; no guanosine oxidation products were identified (Figure 2-6a). To confirm that guanosine was not undergoing undesired oxidation in more complex RNAs, I performed sequencing to analyze the changes in error at unmodified bases in yeast $tRNA^{Glu}_{UUC}$ after photoredox reaction (Figure 2-6b). No marked change in error was observed during sequencing analysis in any canonical nucleobase group, supporting the conclusion that conversion of guanosine into OG or other mutagenic forms was not observed under the tested conditions.

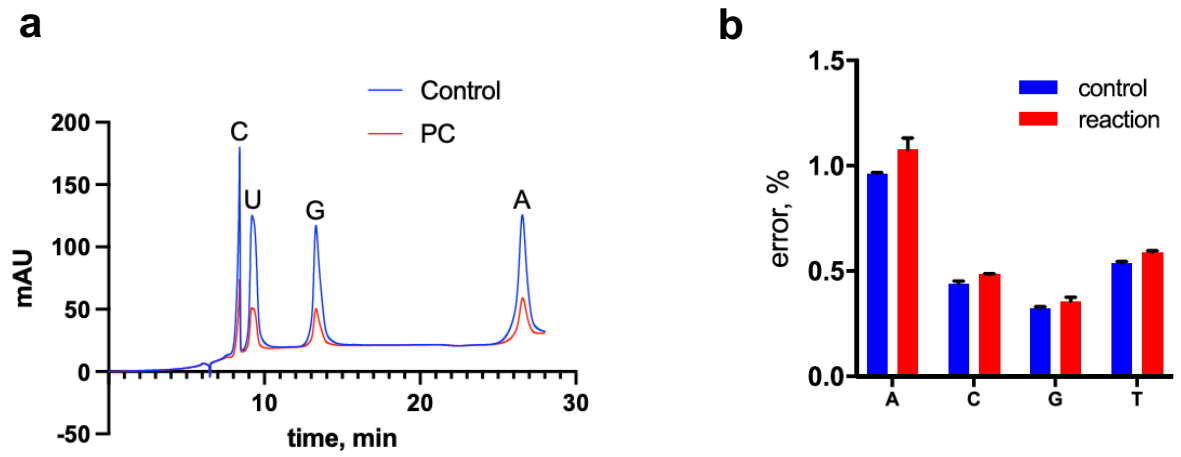


Figure 2-6: a) HPLC analysis of digested unmethylated RNA before (control) and after (PC) photocatalysis reaction showing no guanosine oxidation products. b) Comparative error analysis of unmodified nucleobases of $tRNA^{Glu}_{UUC}$ when subjected to photo-oxidation. Control represents direct sequencing without photoredox catalysis.

2.3.3. Stability test of RNA under optimized photocatalysis reaction condition

I also examined the stability of RNA under the optimized photoredox conditions and found that a phosphate buffer was essential to minimize RNA degradation, with unbuffered aqueous media resulting in rapid decomposition of the RNA (Figure 2-7, Table 2-1). Analysis by RT-qPCR demonstrated only 4 % loss of RNA compared to the no reaction control (Figure 2-8).

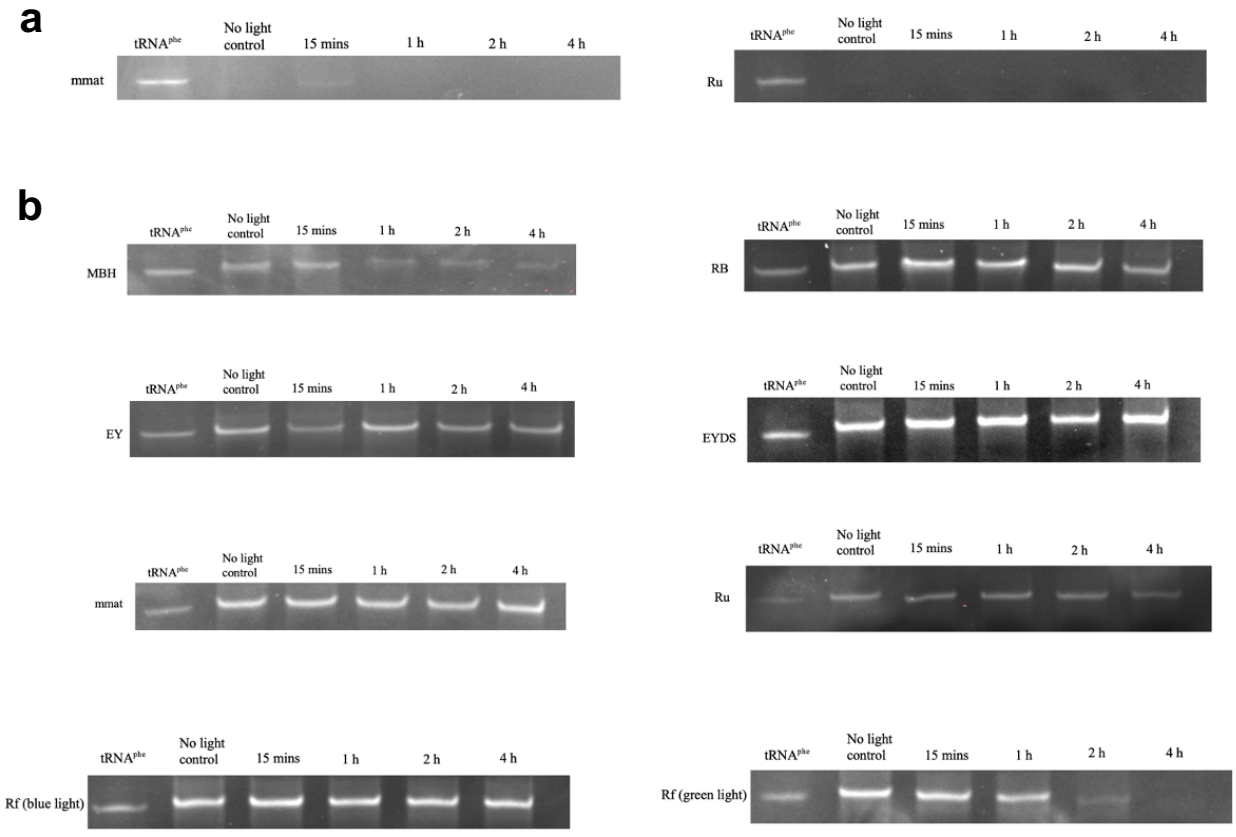


Figure 2-7: Stability of $tRNA^{phe}_{GAA}$ was tested under photoredox conditions in a) unbuffered aqueous media (nuclease-free water only) and b) phosphate buffer.

Photocatalysis reaction was performed with $tRNA^{Phe}_{GAA}$ using various photocatalysts such as riboflavin (Rf) (Alfa Aesar, A11764) under blue and green light, 9-mesityl-10-methylacridinium tetrafluoroborate (mmat) (Sigma-Aldrich, 794171) under blue light, Tris(2,2'bipyridyl)dichlororuthenium(II) hexahydrate (Ru) (Sigma-Aldrich, 544981) under blue light, methylene blue hydrate (MBH) (Sigma-Aldrich, 66720) under red light, Rose Bengal (RB) (SigmaAldrich, 330000) under green light, Eosin Y (EY) (Sigma-Aldrich, E4009) under green light, and Eosin Y disodium salt (EYDS) (Sigma-Aldrich, E6003) under green light.

Table 2-1: Optimisation of photooxidation of m^2_2G.

Photocatalyst	**Light source**	**Conditions**	**Conversion**	**Product**	**RNA degradation[a]**
Riboflavin	*Blue LED*	*Optimized[b]*	*100%*	*m^2_2-Iz*	*No*
None	Blue LED	Optimized	0%	N/A	No
Riboflavin	Dark	Optimized	0%	N/A	No
Riboflavin	Blue LED	No selectfluor	30%	*m^2_2-Iz*	No
Riboflavin	Blue LED	Water	100%	*multiple*	Yes
Riboflavin	Green LED	Optimized	100%	*m^2_2-Iz*	Yes
mmat	Blue LED	Water	95%	multiple	Yes

mmat	Blue LED	Optimized	0%	N/A	No
Ru	Blue LED	Optimized	0%	N/A	Yes
MBH	Red LED	Optimized	100%	multiple	Yes
RB	Green LED	Optimized	90%	multiple	No
EY	Green LED	Optimized	100%	multiple	No
EYDS	Green LED	Optimized	100%	multiple	No

[a]determined by time course experiment using polyacrylamide electrophoresis. Performed on 10 pmol yeast $tRNA^{Phe}_{GAA}$, 13.2 nmol of Selectfluor, 0.6 nmol of photocatalyst in 30 μL of 0.1 M phosphate buffer pH 7. See Figure 2-7 for details. [b]Optimized conditions: 6 nmol m^2_2G, 6 nmol photocatalyst, 13.2 nmol of riboflavin, in 0.1 M phosphate buffer pH 7. Abbreviations: 9-mesityl-10-methylacridinium tetrafluoroborate (**mmat**); Tris(2,2'-bipyridyl)dichlororuthenium(II) hexahydrate (**Ru**); methylene blue hydrate (**MBH**); Rose Bengal (**RB**); Eosin Y (**EY**) ; Eosin Y disodium salt (**EYDS**).

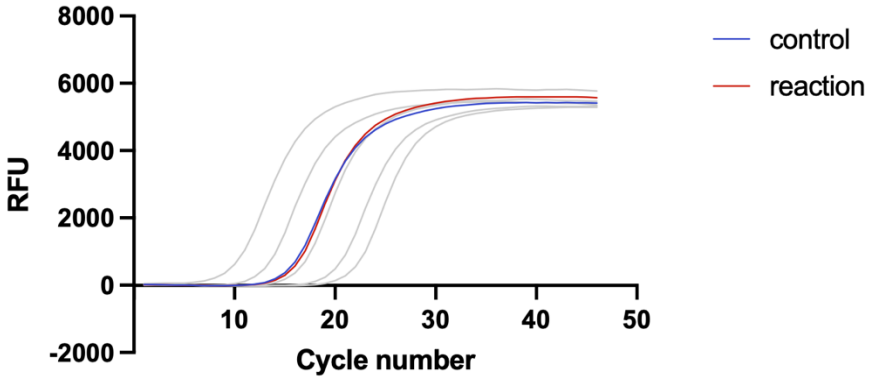


Figure 2-8: RT-qPCR graph showing 4% loss of RNA after reaction

2.3.4. Evaluation of photo-oxidative sequencing of N^2-methylated guanosine sites in RNA

Having optimized the photoredox conditions and demonstrated its compatibility with RNA, I sought to examine its chemoselectivity in photo-oxidative sequencing (PhOxi-seq) of m^2G and $m^2{}_2G$ in various RNA contexts. I first tested PhOxi-seq on yeast $tRNA^{Lys}_{UUU}$, which contains an m^2G and an $m^2{}_2G$ at positions 10 and 26, respectively. tRNA was prepared using the YAMAT-seq[70] workflow to remove the charged amino acids and install sequencing adapters to the no reaction control and photoredox-treated samples. Following high-throughput sequencing, sequences were trimmed for length and quality and aligned to the reference sequence using Bowtie 1[71] to enable induced single-nucleotide polymorphism (SNP) calling.[72] The error at each expected guanosine position was plotted for the control and treated samples (Figure 2-9a). Consistent with our hypothesis, m^2G_{10} was readily detected with a marked increase in error to 88%. However,

surprisingly, $m^2{}_2G_{26}$ experienced a precipitous drop in error from 85% to 16% (Figure 2-9b).

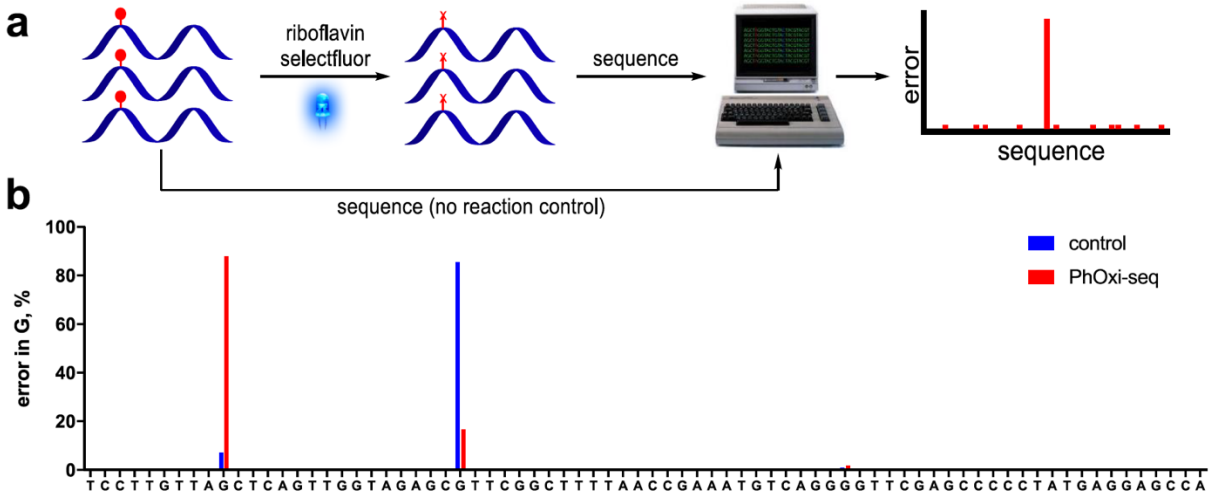


Figure 2-9: a) Workflow for PhOxi-seq for detection of m^2G and $m^2{}_2G$ in RNA. b) Observed error at each guanosine position in yeast $tRNA^{Lys}{}_{UUU}$ before (blue) and after (red) photo-oxidation.

Comparing the error distribution at m^2G_{10} before and after photo-oxidation revealed mostly G → T and G → C transitions, which is consistent with m^2-Iz and m^2-OG formation (Figure 2-10a). Error analysis at $m^2{}_2G_{26}$ showed conversion from a mostly adenosine incorporation typically observed during read-through of bulky bases[62] to a more balanced error distribution (Figure 2-10b), with no clear base-pairing preference.

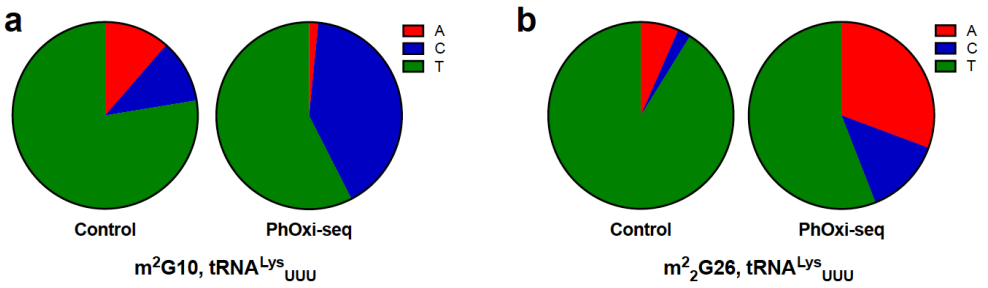


Figure 2-10: Error distribution observed before and after photooxidation reaction for a) m^2G at position 10 in $tRNA^{Lys}_{UUU}$ and b) m^2_2G at position 26 in $tRNA^{Lys}_{UUU}$.

Encouraged by these results, I examined other readily sequenced yeast tRNAs. $tRNA^{Val}_{UAC}$ and $tRNA^{Asn}_{GUU}$, both of which contain m^2G_{10} and $m^2_2G_{26}$ modifications, showed a large increase in error at m^2G sites, with a concomitant decrease in error at m^2_2G sites (Figure 2-11a-b). $tRNA^{Val}_{CAC}$, which contains two instances of m^2G at positions 10 and 26, was effectively sequenced by PhOxi-seq (Figure 2-11c), and $tRNA^{Arg}_{ACG}$, which contains two consecutive m^2G sites at positions 9 and 10 and an m^2_2G located at position 26, was also found to be within the scope of this method (Figure 2-11d). While there does appear to be a lower response for m^2G_{10} in $tRNA^{Arg}_{ACG}$, PhOxi-seq resulted in a 40-fold increase in error at this position from 0.3% to 13% error.

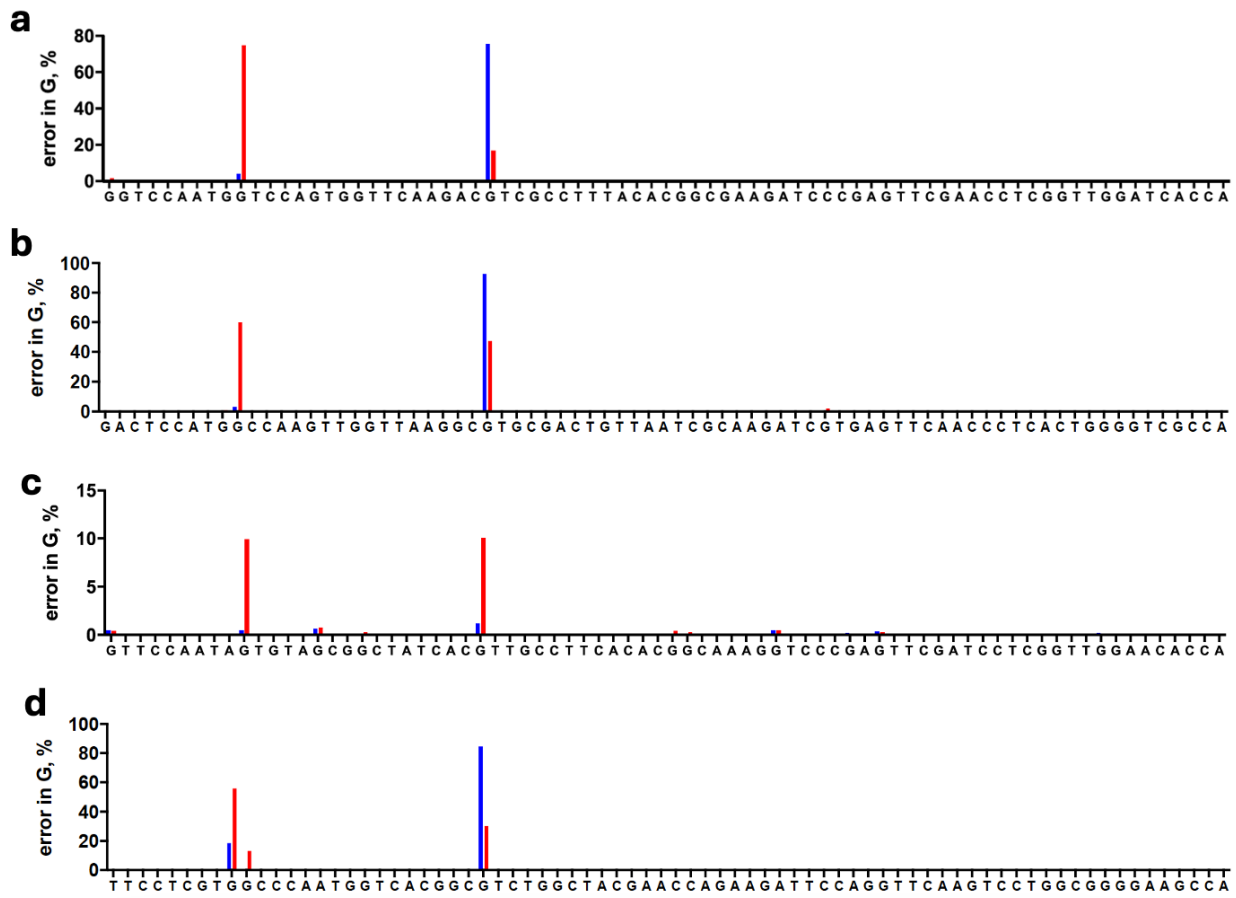


Figure 2-11: Evaluation of PhOxi-seq on various RNAs. Plots of guanosine error at each position mapped for a) $tRNA^{Val}_{UAC}$, b) $tRNA^{Asn}_{GUU}$, c) $tRNA^{Val}_{CAC}$, d) $tRNA^{Arg}_{ACG}$. Blue bars denote control sequencing run; red bars denote PhOxi-seq run.

Other known m^2G and m^2_2G modifications on yeast tRNAs were also found to be effectively sequenced (Figure 2-12a-b). I also used PhOxi-seq to detect the m^2G modification at position 1516 of *E. coli* 16S rRNA with good response and discrimination over flanking guanosines (Figure 2-12c). While *E. coli* rRNA was denatured before photo-oxidation, the observed error was generally lower than those observed in tRNAs, which may be due to the reaction being performed on the whole ribosome. It is likely that larger

error responses would be observed for smaller RNA fragments. PhOxi-seq analysis on other available tRNA modifications showed no major changes in error, suggesting that this method is compatible with a wide scope of functional groups present on RNA (Figure 2-12d).

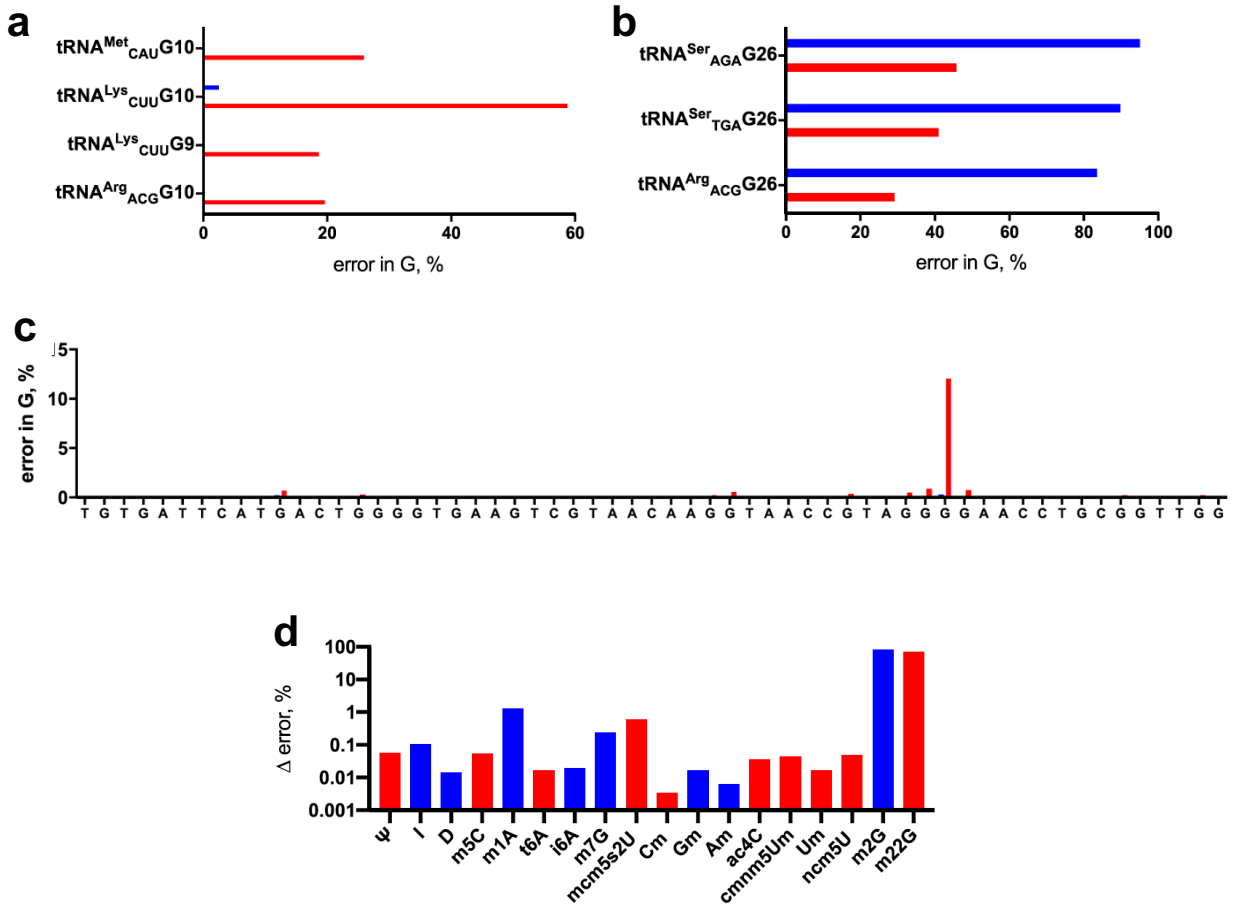


Figure 2-12: Guanosine error at known a) m^2G and b) $m^2{}_2G$ in yeast tRNA; Blue bars denote control sequencing run; red bars denote PhOxi-seq run. c) Plot of guanosine error for *E. coli* 16S rRNA between bases 1471 and 1571; Blue bars denote control sequencing run; red bars denote PhOxi-seq run. d) Difference in error between control and PhOxi-seq at various modified bases in tRNA; blue bars denote increase in error; red bars denote decrease in error.

However, it is important to note that PhOxi-seq alone cannot effectively distinguish between m^2G and m^1G as an increase in error at m^1G sites was observed in all tested cases (Figure 2-13a). While error distributions differ from those seen with m^2G (Figure 2-13b), and native error rates at m^1G sites are significantly higher than those of m^2G, which could enable discrimination between m^1G and m^2G sites, the use of selective m^1G demethylases[73] are likely to be an integral component of PhOxi-seq for N^2-methylation. This is further explored in Chapter 3 of this thesis.

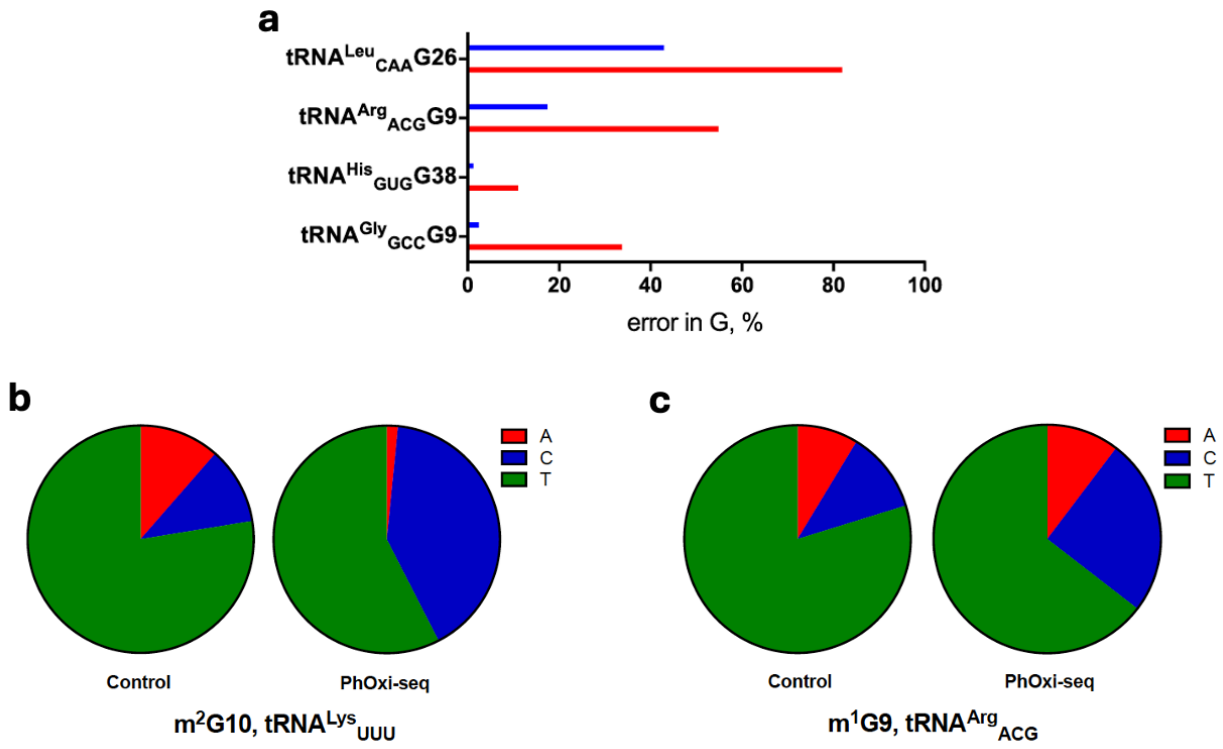


Figure 2-13: a) Guanosine error at known m^1G sites in yeast tRNA. Error distribution observed before and after photooxidation reaction for b) m^2G at position 10 in $tRNA^{Lys}_{UUU}$ and c) m^1G at position 9 in $tRNA^{Arg}_{ACG}$.

2.3.5. Photocatalysis reaction improves read-through at m^2_2G sites

During PhOxi-seq analysis, I found that amplification of tRNAs occurred more efficiently after photoredox reaction. It is known that m^2_2G is a terminator site during reverse transcription and can often be challenging to read through.[28] Indeed, enzymatic methods have recently been developed that enable read-through of m^2_2G sites during tRNA sequencing.[74] I hypothesized that the conversion of m^2_2G into m^2_2-Iz during PhOxi-seq enables more efficient read-through, perhaps due to the reduced size or via alternative base-pairing not available to m^2_2G. To test this on a model system, I used purified yeast $tRNA^{Phe}_{GAA}$ and monitored read-through of m^2_2G by primer extension with and without photo-oxidation. PAGE analysis revealed that superscript III reverse transcriptase read through m^2_2G six times more efficiently after photo-oxidation (Figure 2-14a), which was confirmed using RT-qPCR (Figure 2-14b). Read-through of m^2G sites in $tRNA^{Phe}_{GAA}$ was unaffected after photoredox treatment (Figure 2-14c). Such photoredox chemistry could find applications to sequencing of m^2_2G-containing transcripts as an alternative to engineered polymerases.

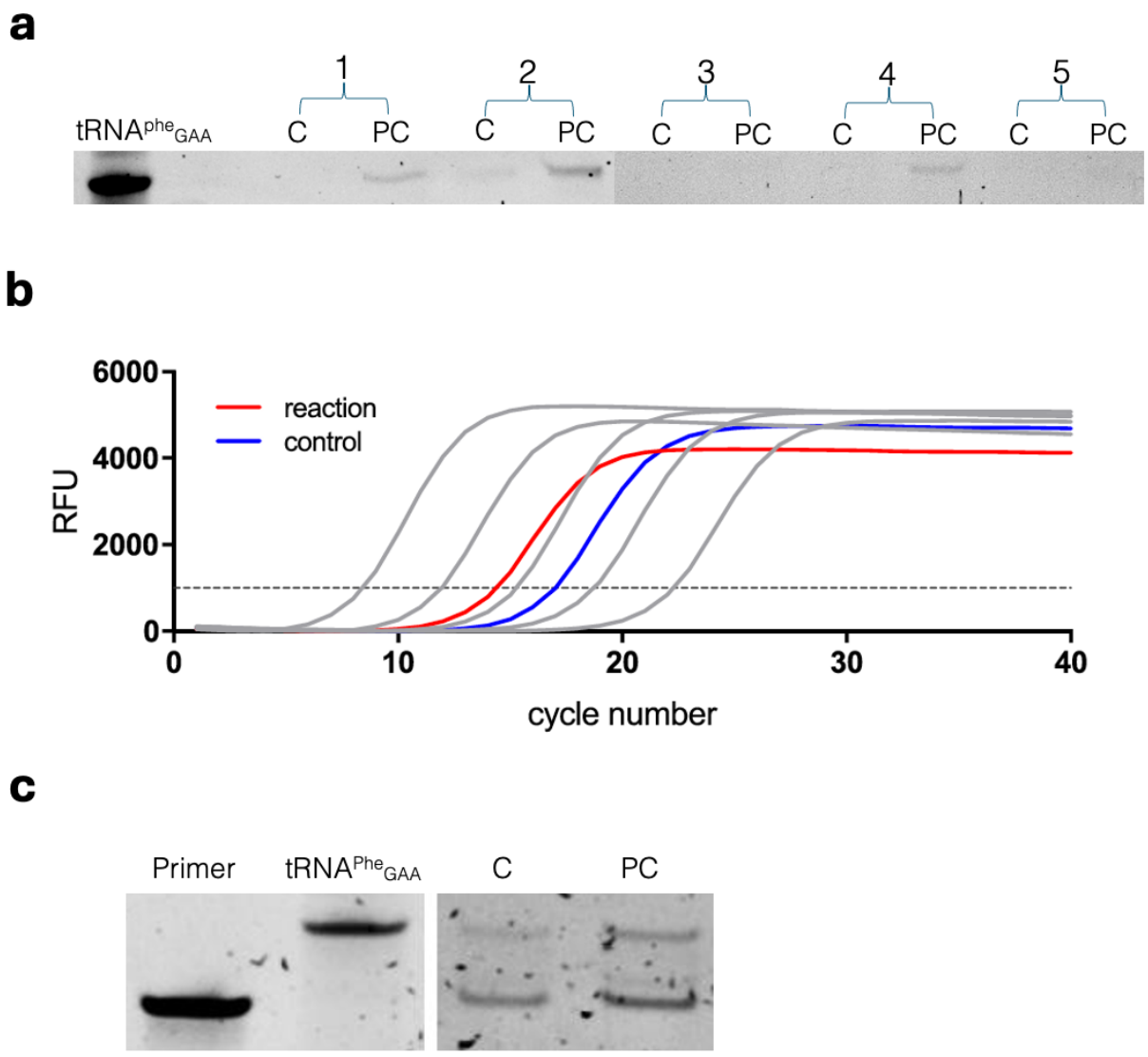


Figure 2-14: a) Reverse transcription of tRNA$^{Phe}_{GAA}$ with (PC) and without (C) photocatalysis reaction was performed using different reverse transcriptase classes such as: 1) GoScript™ Reverse Transcriptase, 2) SuperScript™ III Reverse Transcriptase, 3) Reverse Transcriptase AMV, 4) HIV Reverse Transcriptase, and 5) TGIRT™-III Reverse Transcriptase. b) RT-qPCR analysis of read-through efficiency of $m^2{}_2G$ at position 26 in tRNA$^{Phe}_{GAA}$. c) Reverse transcription was performed with (PC) and without (C) photocatalysis reaction using SuperScript™ III RT, yeast tRNA$^{Phe}_{GAA}$, and tRNA-phe_63 primer.

2.4. Conclusion

In summary, PhOxi-seq provides a robust approach to the single-nucleotide resolution sequencing of N^2-methylation of guanosine by harnessing the chemoselectivity of photoredox chemistry to enable selective photo-oxidation of m^2G and $m^2{}_2G$ in the presence of other RNA bases. I have shown that error markedly increases in the case of m^2G by the hypothesized formation of m^2-Iz and m^2-OG that pair with G and A, respectively. Error decreased in the case of $m^2{}_2G$ with concomitant 6-fold enhanced read-through by the formation of $m^2{}_2$-Iz. The method was shown to be compatible within a variety of sequence contexts in both tRNA and rRNA. I anticipate that PhOxi-seq will find broad use as a straightforward sequencing approach to enable functional studies of N^2-methylation of guanosine in RNA. As the only chemistry-based sequencing approach available for N^2-methylation of guanosine, PhOxi-seq could be used as a diagnostic detection tool for m^2G and $m^2{}_2G$ sites involved in human disease and aid in selecting the appropriate treatments in the future.

2.5. Experimental details

2.5.1. General information

Unless otherwise noted, water was purified with the MilliQ Direct Q3. DNA and RNA oligonucleotides were purchased from Integrated DNA Technologies, with HPLC purification. Monarch RNA cleanup kit and E.Z.N.A. Cycle Pure Kit were purchased from

New England BioLabs (T2030L) and Omega Bio-tek (D6492), respectively. Different reverse transcriptases such as GoScript™ Reverse Transcriptase, SuperScript™ III Reverse Transcriptase, Reverse Transcriptase AMV, HIV Reverse Transcriptase, and TGIRT™-III Reverse Transcriptase were purchased from Promega (9PIA500), Invitrogen/Thermo Fisher Scientific (18080093), Sigma-Aldrich (10109118001), EMD Millipore Corp. (382129), and InGex, LLC (TGIRT), respectively. Nucleoside analysis was performed by reverse-phase high-performance liquid chromatography (HPLC, Agilent 1260 Infinity II) using a C18 stationary phase (Phenomenex, Luna® 5 µm C18(2) 100 Å, 250 x 4.6 mm) and an acetonitrile/100 mM triethylammonium acetate gradient. High-throughput DNA sequencing samples were quantified using a Qubit 4 Fluorometer, prepared on an IonChef instrument and sequenced on an Ion Torrent GeneStudio S5 Plus using Ion 540 Chips.

2.5.2. DNA Sequences

Note: 5phos/ = 5'-phosphate, /3ddC = 3'-dideoxycytidine

The sequences below are written from 5' → 3':

T7 TAA TAC GAC TCA CTA TAG GG

T7_term GCT AGT TAT TGC TCA GCG G

tRNA-phe_45 TGG TGC GAA TTC TGT GGA TCG AAC ACA GGA CCT CCA GAT CTT CAG

tRNA-phe_31_rev GCG GAT TTA GCT CAG TTG GGA GAG CGC CAG A

tRNA_phe_cDNA-qPCR TGG TGC GAA TTC TGT GGA TCG AAC ACA GGA CGA CCG AAT CTT CAG TCT GGC GCT CTC CCA ACT GAG CTA AAT CCG C

Y-5'-AD-N GTT CAG AGT TCT ACA GTC CGA CGA TCA CTG GAT ACT GrGrN

Y-3'-AD 5Phos/GTA TCC AGT TGG AAT TCT CGG GTG CCA AGG/3ddC

2.5.3. RNA Sequences

The sequence below is written from 5' → 3':

temp_3_unmeth rGrCrU rArGrU rUrArU rUrGrC rUrCrA rGrCrG rGrUrA rUrArC rGrUrC rUrGrC rArArU rArCrA rGrCrG rArCrC rCrUrA rUrArG rUrGrA rGrUrC rGrUrA rUrUrA

2.5.4. IonCode adapters

The sequences below are written from 5' → 3':

Y-AD-P1 CCA CTA CGC CTC CGC TTT CCT CTC TAT GGG CAG TCG GTG ATG CCT TGG CAC CCG AGA

tRNA_arg3_P1 CCA CTA CGC CGC TTT CCT CTC TAT GGG CAG TCG GTG ATT GGC TTC CCC GCC AGG ACT TGA A

tRNA_asn1_P1 CCA CTA CGC CTC TTT CCT CTC TAT GGG CAG TCG GTG ATT GGC GAC CCC AGT GAG GGT TGA A

tRNA_lys1_P1 CCA CTA CGC CTC CGC TTT CCT CTC TAT GGG CAG TCG GTG ATT GGC TCC TCA TAG GGG GCT CGA A

tRNA_val1_P1 CCA CTA CGC CTC CGC TTT CCT CTC TAT GGG CAG TCG GTG ATT GGT GAT CCA ACC GAG GTT CGA A

Y-AD_IC_0101 CCA TCT CAT CCC TGC GTG TCT CCG ACT CAG CTA AGG TAA CGG TGA TGT TCA GAG TTC TAC AGT CCG ACG ATC

Y-AD_IC_0102 CCA TCT CAT CCC TGC GTG TCT CCG ACT CAG TAA GGA GAA CGG TGA TGT TCA GAG TTC TAC AGT CCG ACG ATC

Y-AD_IC_0103 CCA TCT CAT CCC TGC GTG TCT CCG ACT CAG AAG AGG ATT CGG TGA TGT TCA GAG TTC TAC AGT CCG ACG ATC

Y-AD_IC_0104 CCA TCT CAT CCC TGC GTG TCT CCG ACT CAG TAC CAA GAT CGG TGA TGT TCA GAG TTC TAC AGT CCG ACG ATC

Y-AD_IC_0105 CCA TCT CAT CCC TGC GTG TCT CCG ACT CAG CAG AAG GAA CGG TGA TGT TCA GAG TTC TAC AGT CCG ACG ATC

Y-AD_IC_0106 CCA TCT CAT CCC TGC GTG TCT CCG ACT CAG CTG CAA GTT CGG TGA TGT TCA GAG TTC TAC AGT CCG ACG ATC

Y-AD_IC_0107 CCA TCT CAT CCC TGC GTG TCT CCG ACT CAG TTC GTG ATT CGG TGA TGT TCA GAG TTC TAC AGT CCG ACG ATC

Y-AD_IC_0108 CCA TCT CAT CCC TGC GTG TCT CCG ACT CAG TTC CGA TAA CGG TGA TGT TCA GAG TTC TAC AGT CCG ACG ATC

16S-_1521_P1 CCA CTA CGC CTC CGC TTT CCT CTC TAT GGG CAG TCG GTG ATA AGG AGG TGA TCC AAC CGC A

16S+_1471_IC_0105 CCA TCT CAT CCC TGC GTG TCT CCG ACT CAG CAG AAG GAA CGG TGA TTG TGA TTC ATG ACT GGG GTG

16S+_1471_IC_0106 CCA TCT CAT CCC TGC GTG TCT CCG ACT CAG CTG CAA GTT CGG TGA TTG TGA TTC ATG ACT GGG GTG

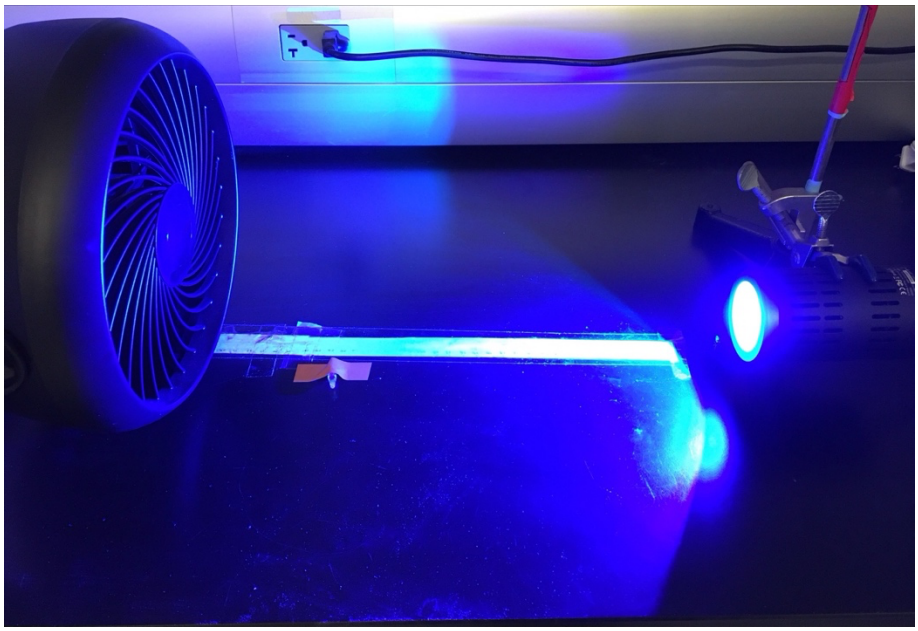
2.5.5. Photocatalysis reaction with nucleosides

The photocatalysis reaction was performed in phosphate buffer (100 mM, pH 7). In a PCR tube, 6 µL of 1 mM modified or unmodified nucleoside, 13.2 µL of 1 mM selectfluor (Sigma-Aldrich, 439479), and 12 µL of 0.5 mM riboflavin (Alfa Aesar, A11764) were added. The reaction was placed at 65 cm from blue light for 2 h, followed by quenching with 3 µL of saturated $NaHCO_3$ and HPLC analysis.

2.5.6. Photocatalysis reaction with RNA

The photocatalysis reaction was performed in a PCR tube by adding 5 µL of 2 µM RNA [in nuclease-free water (Ambion, AM9937)], 13.2 µL of 1 mM selectfluor (in phosphate buffer 100 mM, pH 7), 1.2 µL of 0.5 mM riboflavin (in phosphate buffer 100 mM, pH 7), and phosphate buffer (100 mM, pH 7) up to 30 µL. Reaction was left at 65 cm from blue

light for 2 h, followed by quenching with 3 µL of saturated $NaHCO_3$ and purification with Monarch RNA cleanup kit.



Reaction set up: Sample was secured 65 cm from Kessil Tuna Blue A160WE lamp set to the lowest intensity. A fan was situated directly behind the reaction vessel to dissipate heat. Note that the photo shows the reaction vessel at less than 65 cm for clarity.

2.5.7. HPLC method

Flow rate: 0.5 mL/min

Detection wavelength: 260 nm

Mobile phase A: 3% acetonitrile in 0.1 M triethylammonium acetate (92:5:3 deionized water: 2 M TEAA: acetonitrile)

Mobile phase B: 90% acetonitrile (9:1 acetonitrile: water)

Elapsed time (min)	%B
0	0
15	0
35	10
65	100
70	100
72	0
85	0

Column: Phenomenex, Luna® 5 µm C18(2) 100 Å, 250 x 4.6 mm

2.5.8. RT-qPCR with $tRNA^{Phe}_{GAA}$

Photocatalysis reaction was scaled up by 3-fold and performed with $tRNA^{Phe}_{GAA}$ (Sigma-Aldrich, R4018).

In a PCR tube, $tRNA^{Phe}_{GAA}$ was reverse transcribed using SuperScript™ III Reverse Transcriptase (Invitrogen, Thermo Fisher Scientific, 18080093) protocol.

Complementary DNA was diluted to 1 nM for qPCR. Serial dilution of the cDNA of $tRNA^{Phe}_{GAA}$ (tRNA_phe_cDNA-qPCR) were also used for qPCR standard curve.

qPCR was performed with $tRNA^{Phe}_{GAA}$ and standards using the Luna Universal qPCR Master Mix protocol (NEB, M3003) in a thermal cycler (Biorad, 1855201). Primers used were tRNA-phe_45 and tRNA-phe_31_rev. The qPCR cycles were as follows:

Cycle step	**Temperature, °C**	**Time, s**	**Cycles**
Initial denaturation	95	60	1
Denaturation	95	15	40
Extension	60	30 + plate read	
Melt curve	60 – 95	5	1

2.5.9. PhOxi-seq with Baker's Yeast total tRNA

In a PCR tube, 1 µL of 10 µM Baker's Yeast total RNA (Sigma-Aldrich, 10109509001) was deacylated using 25 µL of 20 mM Tris-HCl (pH 9) in a thermal cycler (Biorad, T100) at 37 °C for 40 min, followed by purification with Monarch RNA cleanup kit. Photocatalysis reaction was then performed with total tRNA.

Ligation of adapters was performed by incubating a PCR tube containing 4 µL of 10 µM Y-5'-AD-N, 4 µL of 10 µM Y-3'-AD, and 1 µL of 10 µM deacylated Yeast total tRNA at 90 °C for 2 min in a thermal cycler (Biorad, T100). To the same PCR tube, 1 µL of 10× annealing buffer [50 mM Tris-HCl (pH 8), 5 mM EDTA, 100 mM $MgCl_2$] was then added,

followed by incubation at 37 °C for 15 min. A 10 µL of 1× reaction buffer containing 10 units of T4 RNA ligase 2 (NEB, #M0239), was then added and incubated at 37 °C for 1 hr and 4 °C overnight. Ligated RNA was purified with Monarch RNA cleanup kit.

The purified ligated tRNA was prepared for sequencing by RT-PCR using IonCode adapters and SuperScript III One-Step RT-PCR System with platinum Taq DNA Polymerase (Invitrogen, Thermo Fisher Scientific, 12574-018) using the following RT-PCR cycles:

1. 55 °C, 60 min
2. 94 °C, 2 min
3. 94 °C, 15 s
4. 50 °C, 30 s
5. 68 °C, 60 s
6. GOTO step 3 (variable)
7. 68 °C, 5 min
8. 4 °C, ∞

The reverse transcribed DNA was purified using E.Z.N.A. Cycle Pure Kit, followed by purification using 10% native polyacrylamide gel. After staining the gel for 15 minutes with SYBR safe DNA gel stain (Invitrogen, 33100), the gel was visualized on BluPAD Dual LED Blue/White Light Transilluminator (bio-helix, BP001CU), and the desired DNA amplicon was excised from the gel. The excised band was crushed into a slurry, 100 µL of 0.3 M NaCl was added to the slurry and incubated overnight at 37 °C followed by

purification with E.Z.N.A. Cycle Pure Kit (Omega Biotek, D6492). The concentration of the DNA was then measured using a Qubit 4.0 Fluorometer (Thermo Fisher Scientific) using the dsDNA HS Assay Kit (Invitrogen, Q32851) and then diluted to 50 pM. The prepped and pooled DNA libraries were loaded onto an Ion Chef with Ion 540 Chips (Thermo Fisher Scientific, A27765). The prepared chips were then sequenced on an Ion GeneStudio™ S5 Plus DNA sequencing system (Thermo Fisher Scientific).

2.5.10. PhOxi-seq with *E. Coli* Ribosome

E. Coli Ribosome (NEB, P0763S) was purified with Monarch cleanup kit and denatured by incubating at 70 °C for 5 minutes and snap-cooling on ice.

Photocatalysis reaction was then performed with *E. Coli* Ribosome.

The purified *E. Coli* Ribosome was prepared for sequencing by RT-PCR using IonCode adapters and SuperScript III One-Step RT-PCR System with platinum Taq DNA Polymerase (Invitrogen, Thermo Fisher Scientific, 12574-018) using the following RT-PCR cycles:

1. 55 °C, 30 min
2. 94 °C, 2 min
3. 94 °C, 15 s
4. 50 °C, 30 s
5. 68 °C, 60 s
6. GOTO step 3 (variable)

7. 68 °C, 5 min
8. 4 °C, ∞

The reverse transcribed DNA was purified using E.Z.N.A. Cycle Pure Kit, followed by purification using 10% native polyacrylamide gel. After staining the gel for 15 minutes with SYBR safe DNA gel stain (Invitrogen, 33100), the gel was visualized on BluPAD Dual LED Blue/White Light Transilluminator (bio-helix, BP001CU), and the desired DNA amplicon was excised from the gel. The excised band was crushed into a slurry, 100 µL of 0.3 M NaCl was added to the slurry and incubated overnight at 37 °C followed by purification with E.Z.N.A. Cycle Pure Kit (Omega Biotek, D6492). The concentration of the DNA was then measured using a Qubit 4.0 Fluorometer (Thermo Fisher Scientific) using the dsDNA HS Assay Kit (Invitrogen, Q32851) and then diluted to 50 pM. The prepped and pooled DNA libraries were loaded onto an Ion Chef with Ion 540 Chips (Thermo Fisher Scientific, A27765). The prepared chips were then sequenced on an Ion GeneStudio™ S5 Plus DNA sequencing system (Thermo Fisher Scientific).

2.5.11. Photocatalysis reaction with unmethylated RNA followed by digestion

Photocatalysis reaction was performed with unmethylated RNA (temp_3_unmeth) followed by digestion using Nucleoside Digestion Mix kit (NEB, M0649) and HPLC analysis.

2.5.12. RT-qPCR analysis of RNA degradation by PhOxi-seq

Photocatalysis reaction was performed with unmethylated RNA (temp_3_unmeth).

RT-qPCR was performed according to Luna Universal One-Step RT-qPCR kit protocol (NEB, E3005) in a thermal cycler (Biorad, 1855201) using T7 and T7_term primers. RT-qPCR cycles were as follows:

Cycle Step	**Temperature, °C**	**Time**	**Cycles**
Reverse transcription	55	10 min	1
Initial denaturation	95	1 min	1
Denaturation	95	10 s	45
Extension	60	30 s + plate read	
Melt Curve	60-95	various	1

Chapter 3: Identification of N^1-Methylated Guanosine (m^1G) in RNA with Various Sequence Contexts by PhOxi-Seq

3.1. Abstract

PhOxi-seq was initially developed to confirm known sites of m^2G and $m^2{}_2G$ within abundant classes of RNA, namely purified rRNA and purified tRNA. However, after photooxidation, the same mutations occur at both m^2G and m^1G sites making it challenging to differentiate between these modifications. To determine if m^1G could be resolved from m^2G sites, I further explored the selectivity and limitations of PhOxi-seq. To this end, I demonstrated that PhOxi-seq is highly dependent on sequence context and is not quantitative below 50% m^1G stoichiometry.

3.2. Introduction

N^1-Methylguanosine (m^1G) is a modification found 3' of the anticodon at position 37 of tRNA in all organisms. This modification is known to prevent frameshifting, which is detrimental to protein synthesis. In fact, studies have revealed that the lack of m^1G in tRNA of either Bacteria or Eukarya affects growth and is thus essential for survival of organisms.[75] In human tRNA, TRMT5 and TRMT10 are known methyltransferases that install m^1G at positions 37 and 9, respectively. TRMT5 mutations are found in patients with lactic acidosis and multiple mitochondrial respiratory-chain-complex deficiencies in skeletal muscle whereas TRMT10 dysfunction is associated with patients suffering from diabetes.[76-78] Additionally, high levels of m^1G have been found in the urine of tumor patients and used as biomarkers of cancer.[79] Functional studies about m^1G are limited due to the lack of sequencing methods available. Fortuitously, when PhOxi-seq was

initially developed, known m^1G sites in yeast tRNA were mutated, providing a potential tool to study m^1G.[80]

The sequencing of tRNAs provide a good testing ground to explore the selectivity of sequencing methods due to the rich diversity of modifications present. Indeed, our initial assessment of PhOxi-seq showed that mutagenic photooxidation occurs at m^1G, m^2G, and m^2_2G. In the case of m^2_2G, which is read through with high error due to the modification blocking Watson-Crick-Franklin hydrogen bonding, photooxidation was shown to markedly decrease the error, which can be easily distinguished from m^2G.

However, m^1G and m^2G both showed an increase in error in yeast total tRNA with mostly G → T and G → C transitions making it challenging to discriminate between both modifications. When various tRNAs with known m^1G sites were examined, variation in the native read-through error were observed, pointing to either read-through error being sequence dependent or perhaps these sites had varying m^1G stoichiometries (Figure 3-1a-b). As a result, I sought to determine if this observation was general in other RNA models and to establish if m^1G could be easily distinguished from m^2G signatures.

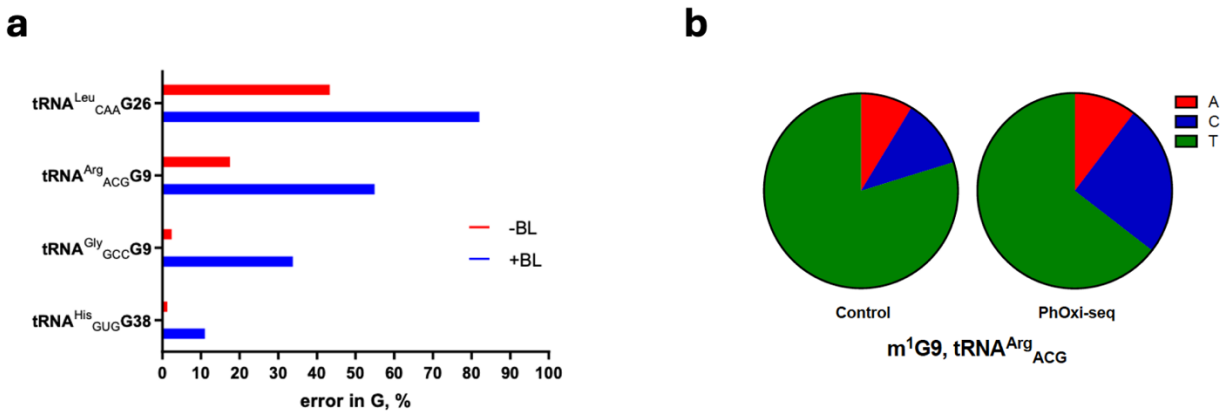
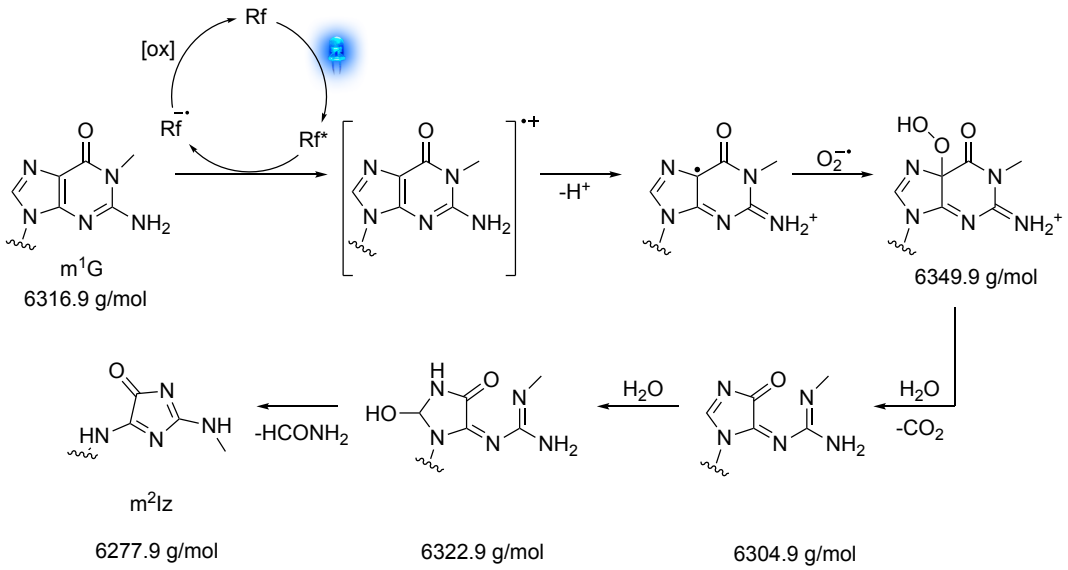


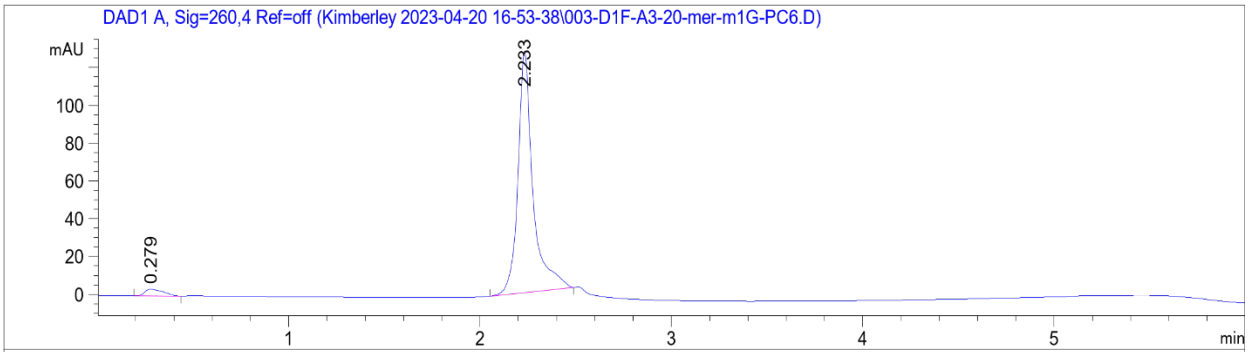
Figure 3-1: a) Detection of known m^1G sites in yeast tRNA. Red bars denote sequencing error without blue light-mediated photooxidation (-BL), while blue bars denote the sequencing error with blue light-mediate photooxidation (+BL). b) Error distribution observed before and after photooxidation reaction for m^1G at position 9 in $tRNA^{Arg}_{ACG}$.

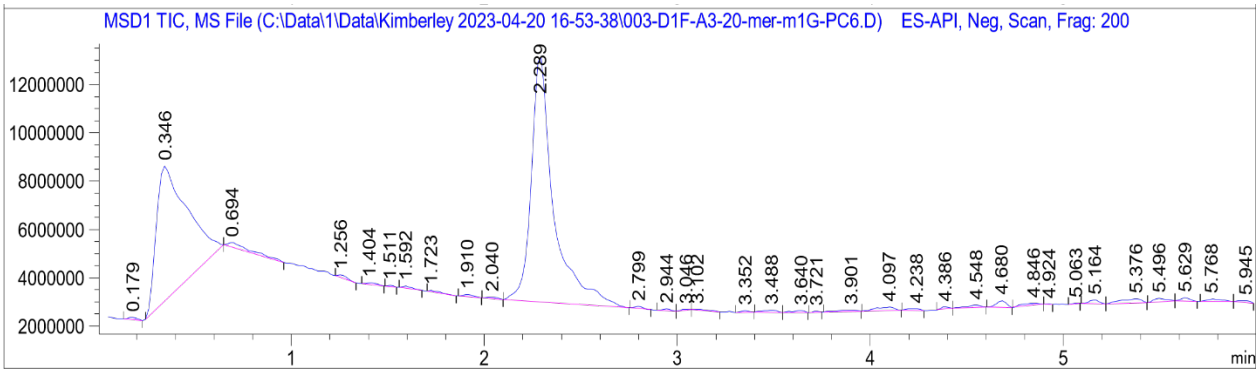
3.3. Results and discussion

3.3.1. PhOxi-seq with m^1G containing RNA

Due to its similar sequencing data outcome and error profiles as m^2G, I first hypothesized that both m^1G and m^2G oxidize into m^2-Iz. Knowing that m^2-Iz product is potentially unstable due to its challenging detection by HPLC, to overcome this issue, I used a 20-mer oligonucleotide containing an m^1G site as a model system: 5'-AACUACUAC(m^1G)UAACUACUAA-3'. LC-MS analysis following photoredox treatment confirms our proposed reaction pathway indicating that the sequencing data obtained was due to both m^1G and m^2G resulting in the same mutagenic products (Figure 3-2).


[ox]
Rf
Rf•−
Rf*
m1G
6316.9 g/mol
-H+
O2−•
6349.9 g/mol
H2O
-CO2
6304.9 g/mol
H2O
6322.9 g/mol
-HCONH2
m2Iz
6277.9 g/mol


DAD1 A, Sig=260,4 Ref=off (Kimberley 2023-04-20 16-53-38\003-D1F-A3-20-mer-m1G-PC6.D)
mAU
2.233
0.279
min


MSD1 TIC, MS File (C:\Data\1\Data\Kimberley 2023-04-20 16-53-38\003-D1F-A3-20-mer-m1G-PC6.D) ES-API, Neg, Scan, Frag: 200
2.289
0.346
0.694
0.179
1.256
1.404
1.511
1.592
1.723
1.910
2.040
2.799
2.944
3.046
3.102
3.352
3.488
3.640
3.721
3.901
4.097
4.238
4.386
4.548
4.680
4.846
4.924
5.063
5.164
5.376
5.496
5.629
5.768
5.945
min

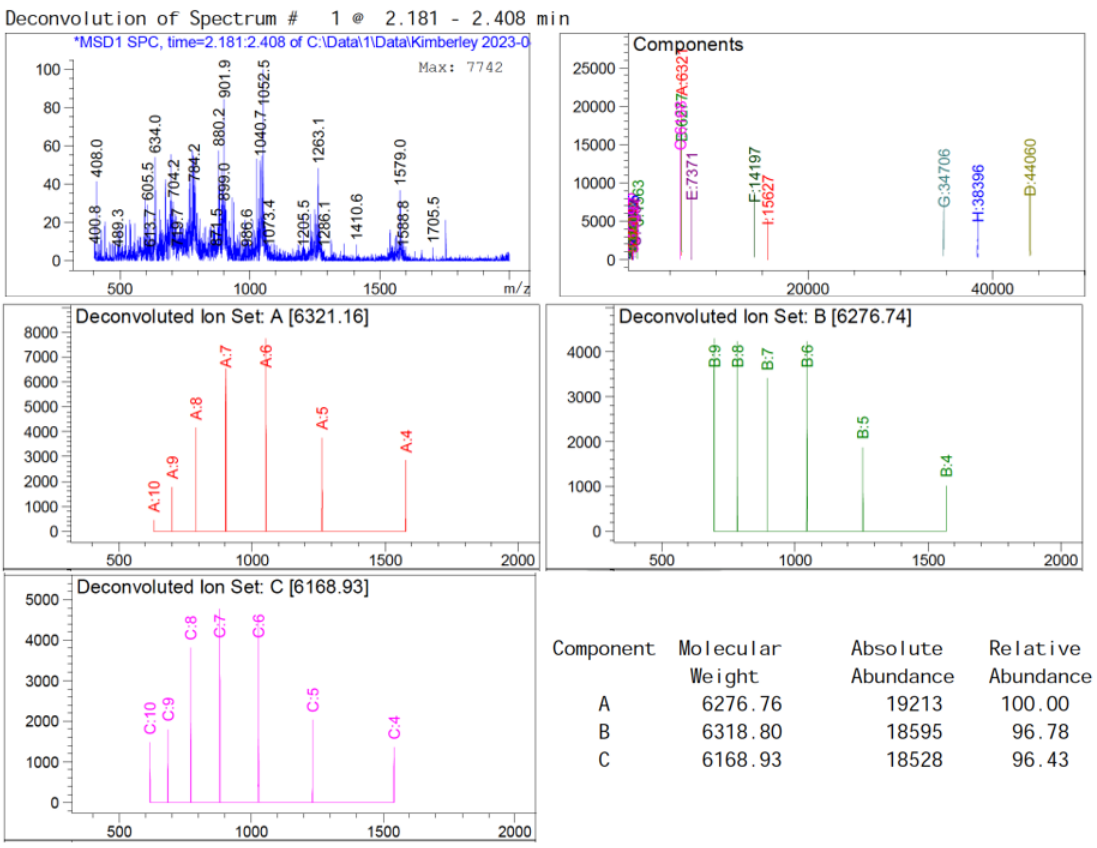


Figure 3-2: LC-MS analysis of photoredox-treated m^1G containing 20-mer oligonucleotide showing m^2-Iz product.

3.3.2. PhOxi-seq with m^1G containing RNA in stem and loop

To further investigate if read-through error at m^1G sites were sequence dependent, I conducted sequence analysis on model systems with controlled m^1G stoichiometries in different sequence context. RNA sequences with m^1G positioned in either a predicted stem or loop structure were examined with PhOxi-seq to evaluate read-through and photooxidation preferences. When m^1G was located in a predicted stem (Figure 3-3a), low read-through error was observed (2%), with an expected high error (61%) after

photooxidation. However, when m^1G was positioned in a predicted loop (Figure 3-3b), high read-through error was observed (35%) with a further increase in error upon photooxidation (58%). This further confirmed that PhOxi-seq analysis of m^1G sites will be highly dependent on sequence context and pose potential issues with downstream analysis.

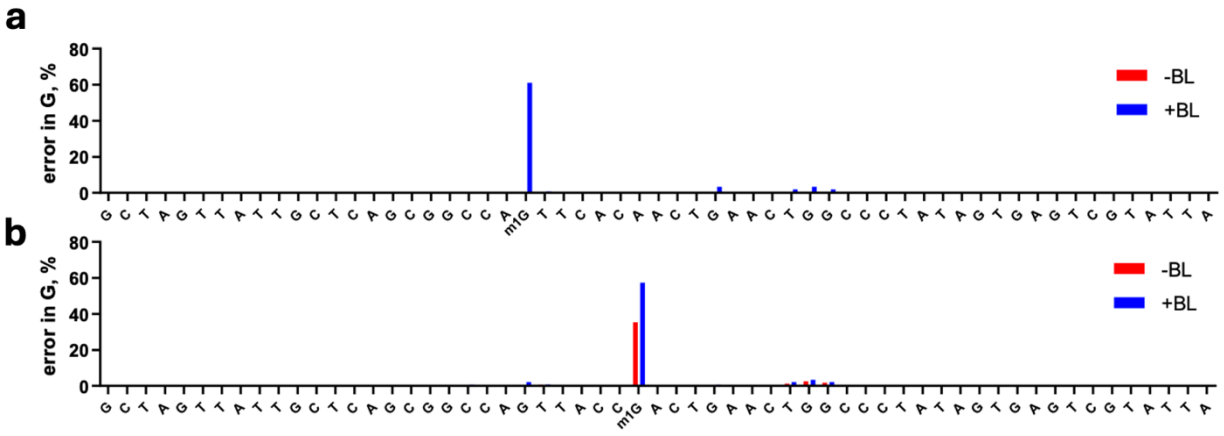


Figure 3-3: a) PhOxi-seq of an RNA containing a single m^1G in a predicted[81] stem region. b) PhOxi-seq of an RNA containing a single m^1G in a predicted loop region.

3.3.3. PhOxi-seq with sequences containing m^1G flanked by random degenerate bases

Indeed, a library of m^1G sequences comprising NN(m^1G)NN demonstrated high variability of read-through error for m^1G, with errors up to 90% (Figure 3-4a); however, in almost all sequence contexts, PhOxi-seq did provide a marked increase in error (Figure 3-4b). These data suggest that m^1G may be readily distinguished from m^2G in cases where read-through error at m^1G sites prior to photooxidation is high. Sites that exhibit this

signature can be removed from putative m^2G site lists during sequencing analysis; however, false positives are likely to occur due to the high variability of m^1G read-through error.

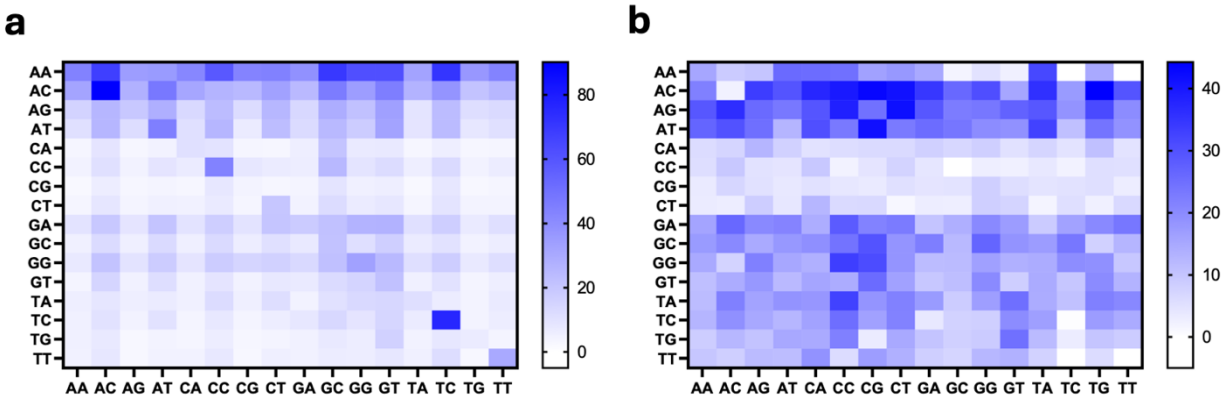


Figure 3-4: a) Heat map of read-through error of an m^1G-containing RNA flanked by two degenerate bases on both sides (NN(m^1G)NN). The x-axis reflects the dinucleotide 5' to the m^1G, while the y-axis reflects the dinucleotide 3' to the m^1G. b) Similar to the heat map in a), but showing the difference in error between the no blue light photooxidation control and the photo-oxidized sample.

3.3.4. PhOxi-seq with sequences containing m^1G mixed in different fractions ranging from 0 – 100 %

Lastly, I also explored the effect of m^1G stoichiometry during PhOxi-seq analysis. To do this, I evaluated error profiles for a model template with a single m^1G at stoichiometries ranging from 0% to 100%. I observed that increases in error due to photooxidation became significant above 50% stoichiometry, increasing markedly as stoichiometries reached 100% (Figure 3-5a). One explanation for this is read-through bias, whereby the

unmodified guanosine is reverse transcribed more efficiently than m^1G (Figure 3-5b); however, approaches to circumvent this issue were not successful.

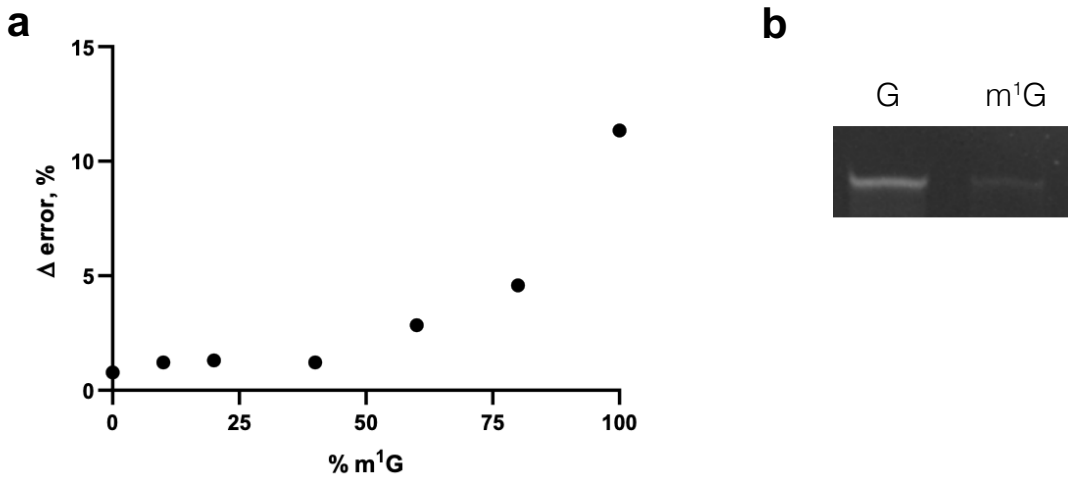


Figure 3-5: a) PhOxi-seq analysis on m^1G-containing RNA with varying stoichiometries of m^1G at a single position, ranging from 0-100%. The difference in error is between the no photooxidation control and the photooxidized sample. b) Reverse transcription of oligonucleotide with (m^1G) and without m^1G (G).

3.4. Conclusion

To conclude, I explored the possibility of using PhOxi-seq to distinguish between m^2G and m^1G. I found that while m^1G sites in RNA are sensitive to photooxidation and are cleanly detected using PhOxi-seq, the variability of read-through error in the no-photooxidation control was considered a potential confounding issue. To overcome the limitations of PhOxi-seq, an m^1G antibody can be used prior to PhOxi treatment to enrich m^1G sequences. However, if both m^2G and m^1G are present on the same transcript, false

positives might occur. On the other hand, using an m^1G demethylase to fully convert m^1G into G will be explored in the near future. While I am currently investigating more robust approaches to definitively resolve m^1G from m^2G in PhOxi-seq data, due to the above uncertainties, I sought a different approach that would yield higher confidence sites for m^2G within a human cancer cell line discussed in more detail in chapter 4.

3.5. Experimental details

3.5.1. General information

Unless otherwise noted, water is purified with the MilliQ Direct Q3. DNA and RNA oligonucleotides are purchased from Integrated DNA Technologies and Dharmacon. Monarch RNA cleanup kit and E.Z.N.A. Cycle Pure Kit are purchased from NEW ENGLAND BioLabs (T2030L) and Omega Bio-tek (D6492), respectively. High-throughput DNA sequencing samples are quantified using a Qubit 4 Fluorometer, prepared on an IonChef instrument and sequenced on an Ion Torrent GeneStudio S5 Plus using Ion 530 Chips.

3.5.2. RNA sequences

The sequences below are written from 5’ → 3’:

G_nestin_RNA AAU GCC GCU AGU CUC UGA GGG AGA CCG AGG GAG CCC CUU UCA GGA GGA GGA GGG GAG UGC UCU GAA GAC CUC UUG GGC AGG G

m1G_nestin_RNA AAU GCC GCU AGU CUC UGA GGG AGA CCG AGG GAG CCC CUU UCA GGA Gm1GA GGAGGG GAG UGC UCU GAA GAC CUC UUG GGC AGG G

m1G_stem_RNA GCU AGU UAU UGC UCA GCG GCC Am1GU UCA CAA CUG AAC UGG CCC UAU AGU GAG UCG UAU UA

m1G_loop_RNA GCU AGU UAU UGC UCA GCG GCC AGU UAC Cm1GA CUG AAC UGG CCC UAU AGU GAG UCG UAU UA

NN(m1G)NN_RNA GCU AGU UAU UGC UCA GCG GAU GCU ACrN rNm1GrN rNAU AGA CUG CCC UAU AGU GAG UCG UAU UA

3.5.3. IonCode adapters

The sequences below are written from 5' → 3':

nestin_IC_0118 CCA TCT CAT CCC TGC GTG TCT CCG ACT CAG AGG CAA TTG CGG TGA TAA TGC CGC TAG TCT CTG AGG

nestin_IC_0119 CCA TCT CAT CCC TGC GTG TCT CCG ACT CAG TTA GTC GGA CGG TGA TAA TGC CGC TAG TCT CTG AGG

nestin_IC_0120 CCA TCT CAT CCC TGC GTG TCT CCG ACT CAG CAG ATC CAT CGG TGA TAA TGC CGC TAG TCT CTG AGG

nestin_IC_0121 CCA TCT CAT CCC TGC GTG TCT CCG ACT CAG TCG CAA TTA CGG TGA TAA TGC CGC TAG TCT CTG AGG

nestin_IC_0122 CCA TCT CAT CCC TGC GTG TCT CCG ACT CAG TTC GAG ACG CGG TGA TAA TGC CGC TAG TCT CTG AGG

nestin_IC_0123 CCA TCT CAT CCC TGC GTG TCT CCG ACT CAG TGC CAC GAA CGG TGA TAA TGC CGC TAG TCT CTG AGG

nestin_IC_0124 CCA TCT CAT CCC TGC GTG TCT CCG ACT CAG AAC CTC ATT CGG TGA TAA TGC CGC TAG TCT CTG AGG

nestin_IC_0125 CCA TCT CAT CCC TGC GTG TCT CCG ACT CAG CCT GAG ATA CGG TGA TAA TGC CGC TAG TCT CTG AGG

nestin_IC_0126 CCA TCT CAT CCC TGC GTG TCT CCG ACT CAG TTA CAA CCT CGG TGA TAA TGC CGC TAG TCT CTG AGG

nestin_IC_0127 CCA TCT CAT CCC TGC GTG TCT CCG ACT CAG AAC CAT CCG CGG TGA TAA TGC CGC TAG TCT CTG AGG

nestin_IC_0128 CCA TCT CAT CCC TGC GTG TCT CCG ACT CAG ATC CGG AAT CGG TGA TAA TGC CGC TAG TCT CTG AGG

nestin_IC_0129 CCA TCT CAT CCC TGC GTG TCT CCG ACT CAG CGA GGT TAT CGG TGA TAA TGC CGC TAG TCT CTG AGG

nestin_IC_0130 CCA TCT CAT CCC TGC GTG TCT CCG ACT CAG TCC AAG CTG CGG TGA TAA TGC CGC TAG TCT CTG AGG

nestin_IC_0131 CCA TCT CAT CCC TGC GTG TCT CCG ACT CAG TCT TAC ACA CGG TGA TAA TGC CGC TAG TCT CTG AGG

p1_nestin CCA CTA CGC CTC CGC TTT CCT CTC TAT GGG CAG TCG GTG ATC CCT GCC CAA GAG GTC TTC

T7_P1 CCA CTA CGC CTC CGC TTT CCT CTC TAT GGG CAG TCG GTG ATT AAT ACG ACT CAC TAT AGG G

T7_term_IC_0101 CCA TCT CAT CCC TGC GTG TCT CCG ACT CAG CTA AGG TAA C GGT GAT GCT AGT TAT TGC TCA GCG G

T7_term_IC_0102 CCA TCT CAT CCC TGC GTG TCT CCG ACT CAG TAA GGA GAA C GGT GAT GCT AGT TAT TGC TCA GCG G

T7_term_IC_0103 CCA TCT CAT CCC TGC GTG TCT CCG ACT CAG AAG AGG ATT C GGT GAT GCT AGT TAT TGC TCA GCG G

T7_term_IC_0104 CCA TCT CAT CCC TGC GTG TCT CCG ACT CAG TAC CAA GAT C GGT GAT GCT AGT TAT TGC TCA GCG G

T7_term_IC_0105 CCA TCT CAT CCC TGC GTG TCT CCG ACT CAG CAG AAG GAA C GGT GAT GCT AGT TAT TGC TCA GCG G

T7_term_IC_0106 CCA TCT CAT CCC TGC GTG TCT CCG ACT CAG CTG CAA GTT C GGT GAT GCT AGT TAT TGC TCA GCG G

3.5.4. LC-MS method for m^1G containing oligonucleotide

The photocatalysis reaction is performed in a PCR tube by adding 6 µL of 50 µM of RNA, 13.2 µL of 1 mM selectfluor (Sigma-Aldrich, 439479) (in phosphate buffer 100 mM, pH 7), and 24 µL of 0.5 mM riboflavin (Alfa Aesar, A11764) (in Milli-Q). The reaction is placed at 65 cm from blue light (Kessil Tune Blue A160WE lamp set to lowest intensity) for 2 h. The reaction is cleaned up with Monarch kit followed by LC-MS analysis. LCMS analyses were performed using Agilent Infinity Lab LC/MSD system. Negative ionization mode.

Flow rate: 0.5 mL/min

Detection wavelength: 260 nm

Mobile phase A: 10 µM EDTA, 0.038% TEA, 0.75% HFIP, in MilliQ water

Mobile phase B: 10 µM EDTA, 0.038% TEA, 0.75% HFIP, in 90:10 Methanol:MilliQ water

Elapsed time (min)	% B
0	10
4	90
5	10
6	10

Column: HALO 400 A, ES-C18, 3.4 µM, 2.1 x 30 mm

3.5.5. Photocatalysis reaction with RNA

The photocatalysis reaction was performed in a PCR tube by adding 5 µL of 2 µM RNA [in nuclease-free water (Ambion, AM9937)], 13.2 µL of 1 mM selectfluor (in 100 mM phosphate buffer, pH 7), 1.2 µL of 0.5 mM riboflavin (in 100 mM phosphate buffer, pH 7), and phosphate buffer (100 mM, pH 7) up to 30 µL. Reaction was left at 65 cm from blue light (Kessil Tune Blue A160WE lamp set to lowest intensity) for 2 h, followed by purification with Monarch RNA cleanup kit.

3.5.6. PhOxi-seq with RNA containing m^1G in various sequence contexts

Photocatalysis reaction was performed with RNA as mentioned above.

The purified RNA was prepared for sequencing by RT-PCR using IonCode adapters and SuperScript III One-Step RT-PCR System with platinum Taq DNA Polymerase (Invitrogen, Thermo Fisher Scientific, 12574-018) using the following RT-PCR cycles:

1. 55 °C, 30 min
2. 94 °C, 2 min
3. 94 °C, 15 s
4. 50 °C, 30 s
5. 68 °C, 60 s
6. GOTO step 3 (variable)
7. 68 °C, 5 min
8. 4 °C, ∞

The reverse transcribed DNA was purified using E.Z.N.A. Cycle Pure Kit, followed by purification using 10% native polyacrylamide gel. After staining the gel for 15 minutes with SYBR safe DNA gel stain (Invitrogen, 33100), the gel was visualized on BluPAD Dual LED Blue/White Light Transilluminator (bio-helix, BP001CU), and the desired DNA amplicon was excised from the gel. The excised band was crushed into a slurry, 100 µL of 0.3 M NaCl was added to the slurry and incubated overnight at 37 °C followed by purification with E.Z.N.A. Cycle Pure Kit (Omega Bio-tek, D6492). The concentration of the DNA was then measured using a Qubit 4.0 Fluorometer (Thermo Fisher Scientific) using the dsDNA HS Assay Kit (Invitrogen, Q32851) and then diluted to 50 pM. The prepped and pooled DNA libraries were loaded onto an Ion Chef with Ion 530 Chips

(Thermo Fisher Scientific, A27765). The prepared chips were then sequenced on an Ion GeneStudio™ S5 Plus DNA sequencing system (Thermo Fisher Scientific).

3.5.7. Read-through test with RNA containing m^1G

Reverse transcription using SuperScript III RT (Invitrogen, Thermo Fisher Scientific, 18080093) was performed. In a PCR tube, 2.5 µL of 100 µM (250 pmol) Nestin_primer_F primer, 2.5 µL of 10 µM (25 pmol) G_Nestin_RNA/m^1G_Nestin RNA (with or without PC), 1 µL of 10 mM (10 nmol) dNTP, and Nuclease-free water up to 13 µL were added. The reaction was incubated at 65 °C for 5 mins, then placed on ice followed by the addition of 4 µL 5× First-Strand buffer, 1 µL of 0.1 M (100 nmol) DTT, 1 µL (200 units) SuperScript III RT (200 U/µL), and 1 µL Nuclease-free water. The reaction was incubated at 50 °C for 1 hr, then 70 °C for 15 mins.

RNA strand was hydrolyzed by adding 1 µL (2 units) RNAse H (M0297) and incubating at 37 °C for 20 min. Reaction was cleaned up with E.Z.N.A DNA kit and loaded on a 10% denaturing gel at 150 V for 30 min.

Chapter 4: Identification of Enzyme-Dependent m^2G in Multiple RNA Types by PhOxi-Seq

The presented material was submitted to *ACS chemical biology* (currently under review): Klimontova, M.; **Chung Kim Chung, K.**; Zhang, H.; Kouzarides, T.; Bannister, A. J.; Hili, R. PhOxi-seq detects enzyme-dependent m^2G in multiple RNA types. My contributions to this manuscript were performing the PhOxi reaction and tRNA sequencing. Marie Klimontova (Kouzarides lab) performed the cell culture, THUMPD3 knockdown experiments, and transcriptome sequencing. Han Zhang (Kouzarides lab) generated the bioinformatic pipeline and analyzed the sequencing data.

4.1. Abstract

In chapter 2, I introduced PhOxi-seq as a chemistry-based sequencing method that identifies known sites of m^2G within multiple RNA types, namely purified rRNA and purified tRNA. Here, in collaboration with the Kouzarides lab at The University of Cambridge (UK), we describe an optimized PhOxi-seq workflow, coupled to a novel bioinformatic pipeline, that is capable of detecting enzyme-dependent m^2G sites throughout the transcriptome, including low abundant mRNAs. In this way, we generated the first database of high confidence sites of THUMPD3-dependent m^2G in multiple RNA classes within a human cancer cell line (A549) and further identified non-THUMPD3 controlled sites throughout the transcriptome.

4.2. Introduction

N^2-Methylguanosine (m^2G) has been identified in various RNA molecules across all three domains of life: Archaea, Bacteria, and Eukarya. While most often associated with tRNA

and rRNA, m^2G was recently detected by mass spectrometry analysis of mRNA isolated from *S. cerevisiae*, suggesting that it may have regulatory function at the transcriptome level.[82] Recently, three m^2G-catalysing RNA methyltransferases – THUMPD2, THUMPD3, and TRMT11 – have been characterized.[83-85] The dysregulation of these methyltransferases has been associated with various human diseases. THUMPD2 is known to install a critical m^2G modification in U6 snRNA, the catalytic centre of the spliceosome, necessary for efficient splicing activity; its dysregulation is linked to age-related macular degeneration and retinal function.[85]

Upregulated THUMPD3 has been shown to promote lung cancer cell proliferation and migration by modulating the extracellular matrix through mechanisms such as alternative splicing.[86] Lastly, TRMT11 fusions have been associated with aggressive and recurring prostate cancers.[87] Further functional studies are required to determine the role of m^2G in human health; however, limited tools are available to experimentally study m^2G. Early transcriptome maps of other post-transcriptional modifications, such as m^6A, relied on established antibodies, well-characterised demethylases, and a relatively high natural abundance of the modification.[88] On the contrary, m^2G lacks these essential tools making its study particularly challenging.

I previously described photooxidative sequencing (PhOxi-seq), which enables the single-nucleotide resolution sequencing of m^2G by harnessing the chemoselectivity of photoredox chemistry to enable selective photo-oxidation of m^2G in the presence of guanosine.[80] The method relies upon the lower oxidation potential of methylated guanine over that of canonical guanine. In the presence of blue light and riboflavin as a

photocatalyst, m^2G is mutated into m^2-Iz and m^2-OG, which base pair with G and A, respectively (Figure 4-1).

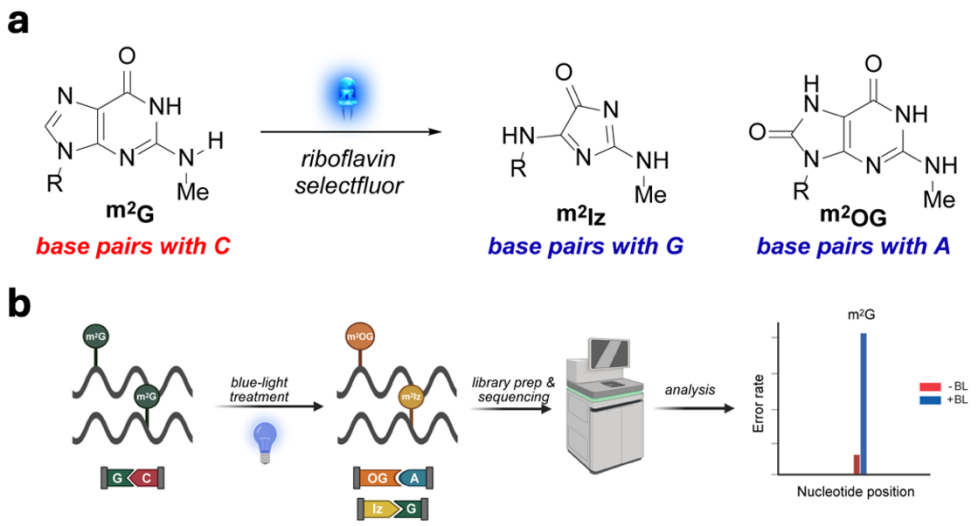


Figure 4-1: a) Photooxidation reaction of N^2-Methylguanosine into N^2-imidazolone (m^2Iz) and N^2-oxoguanine (m^2OG), which is the chemical basis for photooxidative sequencing (PhOxi-seq). b) PhOxi-seq workflow to detect m^2G in RNA.

PhOxi-seq was shown to readily identify m^2G sites in model synthetic RNAs, as well as known sites in highly abundant RNAs, such as purified yeast tRNAs and purified *E. coli* rRNA. One shortcoming of PhOxi-seq is that preliminary evidence suggested that m^1G may form the same mutagenic oxidation products as m^2G, and thus convolute transcriptome-wide mapping of m^2G.[80] Given the limited understanding of the biological role of m^2G and the potential of m^2G methyltransferases as therapeutic targets, in collaboration with the Kouzarides lab at the University of Cambridge (UK), we sought to adapt PhOxi-seq to identify potential m^2G sites within the human transcriptome to

facilitate their functional studies. To this end, we explored the effect of knocking down an m^2G methyltransferase, THUMPD3, to assist in m^2G site confirmation from PhOxi-seq data in human lung cancer cells (A549) (Figure 4-2). THUMPD3 knockdown was performed by the Kouzarides lab and PhOxi treatment was performed by me in this collaboration.

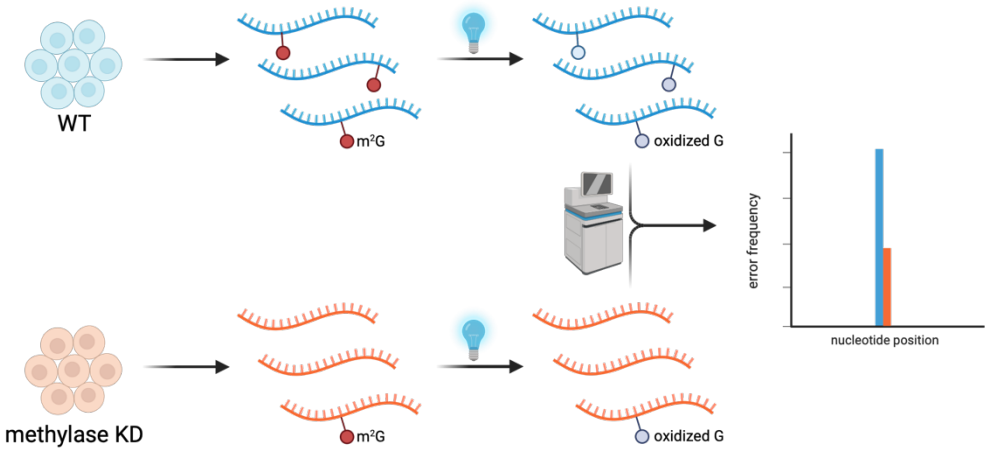


Figure 4-2: Workflow of PhOxi-seq with wild-type (WT) and THUMPD3 knockdown (methylase KD) of human lung cancer cells (A549).

4.3. Results and discussion

4.3.1. PhOxi-seq can detect enzyme-dependent reduction in m^2G levels in tRNA

We first considered whether PhOxi-seq could be used to detect specific enzyme-dependent sites of m^2G within tRNAs isolated from human lung cancer cell line A549. I previously used PhOxi-seq to identify m^2G within tRNAs isolated from yeast.[80] Other works using mass spectrometry-coupled methods later showed that human THUMPD3 installs m^2G at position 6 within specific tRNAs, including isodecoders *tRNA-Gly-GCC-2* and *tRNA-Gly-CCC-2*. Importantly, m^2G_6 within these isodecoders was shown to be significantly decreased by loss of THUMPD3.[83] We first asked whether PhOxi-seq could detect m^2G_6 within these particular tRNA isodecoders. To ensure accurate analysis of m^2G_6 within the specific tRNA isodecoders, we designed specific primers for the respective decoders (Figure 4-3a). THUMPD3 knockdown on human lung cell line A549 and RNA extraction were performed by the Kouzarides lab followed by PhOxi treatment and sequencing performed by me. YAMAT-seq[70] confirmed accurate targeting of the relevant isodecoders (Figure 4-3b). Our data clearly show that PhOxi-seq detects m^2G_6 within tRNA-Gly-GCC-2 (Figure 4-4a, WT) and tRNA-Gly-CCC-2 (Figure 4-4c, WT).

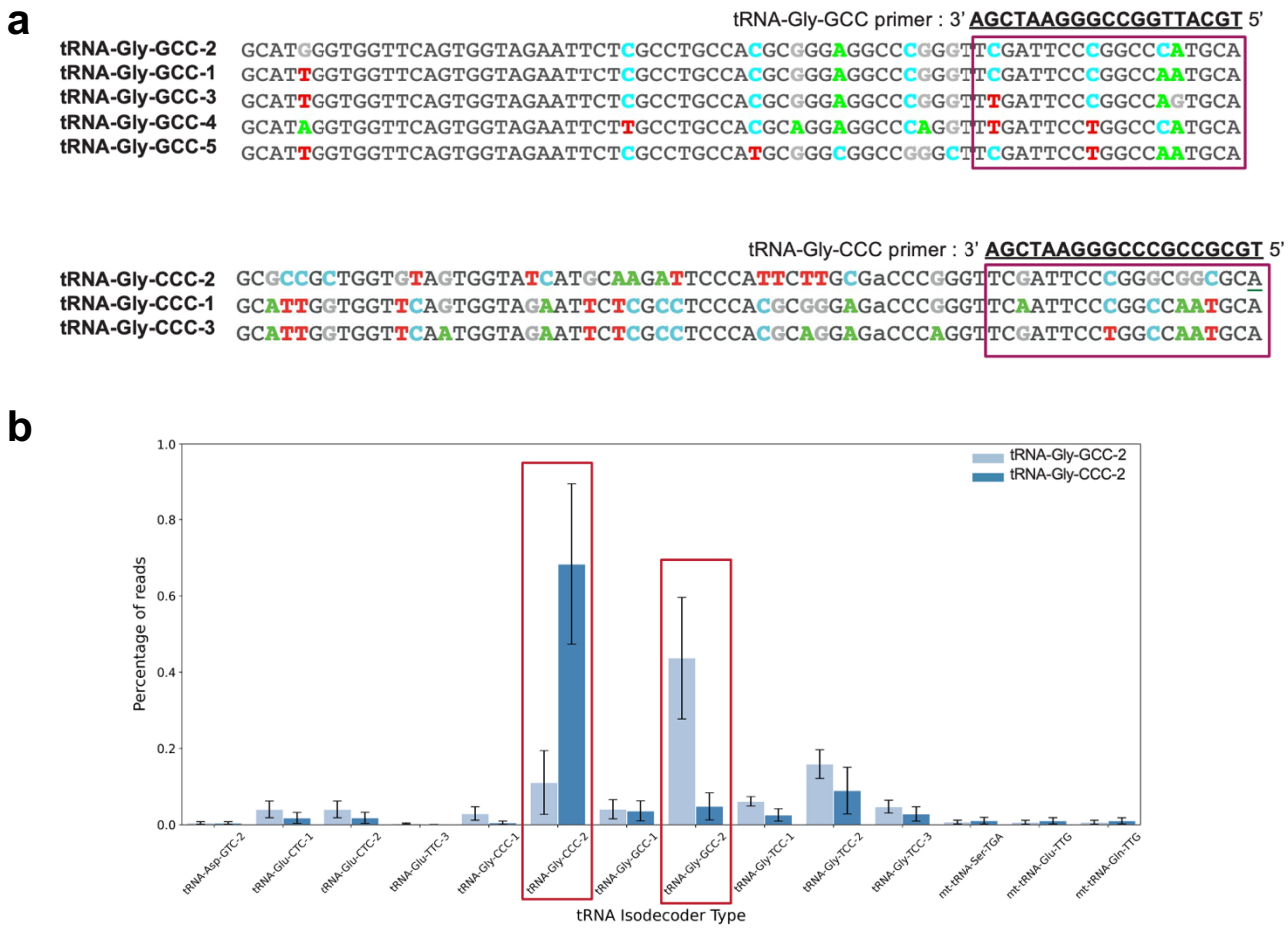


Figure 4-3: a) Annealing of unique tRNA-Gly-GCC-2 RT-PCR primer to five tRNA-Gly-GCC isodecoders and annealing of unique tRNA-Gly-CCC-2 RT-PCR primer to three tRNA-Gly-CCC isodecoders. b) Reads acquired from YAMAT-seq were aligned to a comprehensive set of nuclear-encoded cytoplasmic tRNA and mitochondrial tRNA sequences. The top tRNAs that the reads were predominantly aligned to are depicted. The data is illustrated as the proportion of reads, where 0 represents 0% and 1.0 signifies 100%. Error bars represent the mean ± SD of 3 independent replicates (n=3).

Next, we asked whether PhOxi-seq could identify THUMPD3-dependent sites of m^2G within the same tRNA isodecoders. We compared the sequencing error rates, following blue light treatment (+BL) of RNA samples harvested from WT cells and THUMPD3-depleted (ΔD3k.d.) cells. The data reveal significantly reduced sequencing error rates at G6 of tRNA-Gly-GCC-2 (Figure 4-4a) and tRNA-Gly-CCC-2 (Figure 4-4c) in tRNA samples isolated from THUMPD3-depleted cells, indicating that THUMPD3 catalyzes a significant proportion of m^2G at this site. Remaining m^2G signal could result from modifications installed by other guanine methylases. As expected with guanosine oxidation pathways, we observed T and C substitutions as significant nucleotide sequencing errors at known m^2G positions (Figure 4-4b and 4-4d).

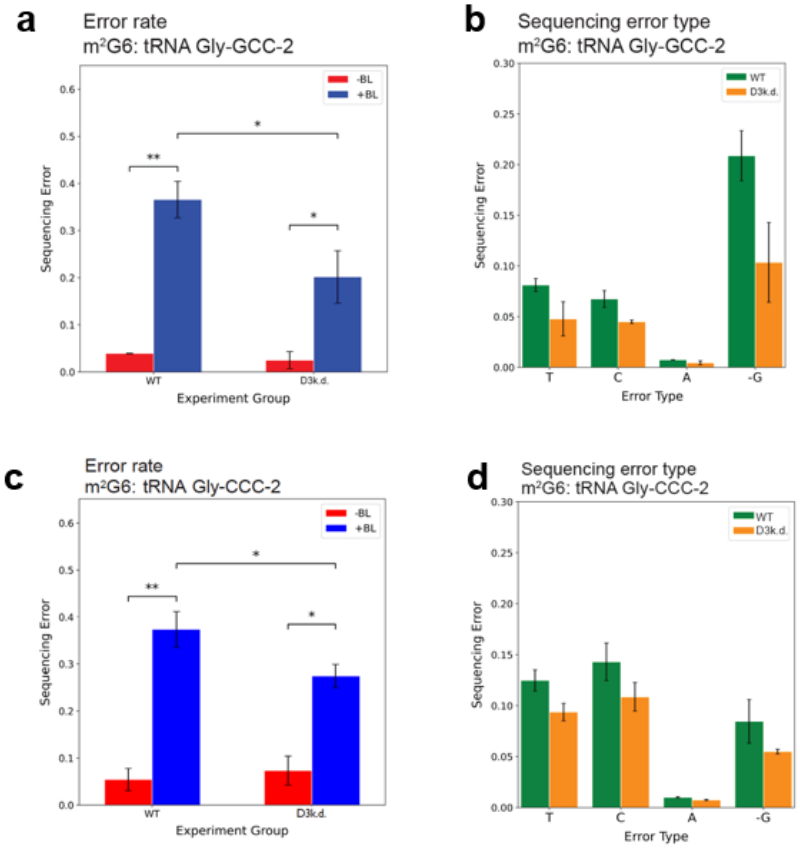


Figure 4-4: THUMPD3-depletion in A549 lung cancer cells results in reduced m^2G_6 in tRNA, detected via PhOxi-seq. a) & c) Plot representing sequencing error at G6 of tRNA-Gly-GCC-2 and tRNA-Gly-CCC-2 with (+BL; blue) and without (-BL; red) blue-light treatment. Two experimental groups represent control (WT) and THUMPD3-depleted cell lines (D3k.d.). Error bars represent the mean ± SD of 3 independent replicates (n=3). Welch's t-test was used to assess the significance of changes in error rates at G6 on tRNA-Gly-GCC-2. ns - $P > 0.05$, * - $P \leq 0.05$, ** - $P \leq 0.01$, *** - $P \leq 0.001$, **** - $P \leq 0.0001$. b) & d) Sequencing error data represented by the blue bars (+BL samples) were deconvoluted into individual nucleotides. Respective proportion of reads corresponding to main types of misincorporation – G → T substitution, G → C substitution, G → A substitution, and single base deletion were plotted. Green and orange bars represent WT and D3k.d. data, respectively. Error bars represent the mean ± SD of 3 independent replicates (n=3).

Interestingly, in addition to the expected substitutions, another error type, G skipping, was prevalent. This identified a novel type of sequencing error that can be used in future analysis to detect genome-wide m^2G distribution. Taken together, these findings highlight PhOxi-seq as a technique capable of detecting enzyme-dependent sites of m^2G within tRNA. Furthermore, the unexpected error rate profile may prove helpful in identifying other potential THUMPD3-dependent sites of m^2G. Based on these findings, we decided to explore whether PhOxi-seq could be used to detect m^2G sites within the wider transcriptome.

4.3.2. Utilising PhOxi-seq to detect m^2G in other RNA species.

Our initial findings validate PhOxi-seq as an effective method of identifying enzyme-dependent sites of m^2G within specific tRNAs. However, its application and relevance to mRNA and other non-coding RNAs have not been explored. Based on my prior success with tRNA and rRNA, we reasoned that blue light-induced photo-oxidation of m^2G is independent of RNA type. Therefore, to identify sites of m^2G within other types of RNA, I treated polyA+ enriched RNA provided by the Kouzarides lab with the same photooxidation condition. Then, the Kouzarides lab subjected the resulting RNA libraries to high-throughput sequencing. However, as PhOxi-seq has not previously been tested on this class of RNA, the Kouzarides lab had to generate a unique bioinformatic pipeline described in the experimental detail section.

We leveraged PhOxi-seq and our novel bioinformatic pipeline to identify sites of m^2G within a human polyA+ enriched transcriptome isolated from A549 lung cancer cells. Analysis of our tRNA data revealed that each m^2G site treated with photooxidation mutates to any base (although with different frequencies, Figure 4-4) or gets deleted. Applying these characteristics where any change or deletion of Gs is scored within our pipeline, we identified 381 potential THUMPD3-dependent sites of m^2G within 298 distinct polyA+ enriched RNAs.

Since multiple base substitutions and deletion signatures were observed at m^2G_6 in tRNAs, we applied to the pipeline a requirement to further filter our hits for at least two photooxidation induced mutational changes at the sites identified above in Figure 4-4. This revealed 62 high confidence hits (Figure 4-5a). These m^2G sites were distributed across various mRNA regions, including exons (15%), introns (61%), 3' UTRs (11%), and 5' UTR (2%) (Figure 4-5b). Interestingly, m^2G sites were enriched within intronic regions. This aligns well with the previously reported role of THUMPD3 in the regulation of RNA splicing.[86]

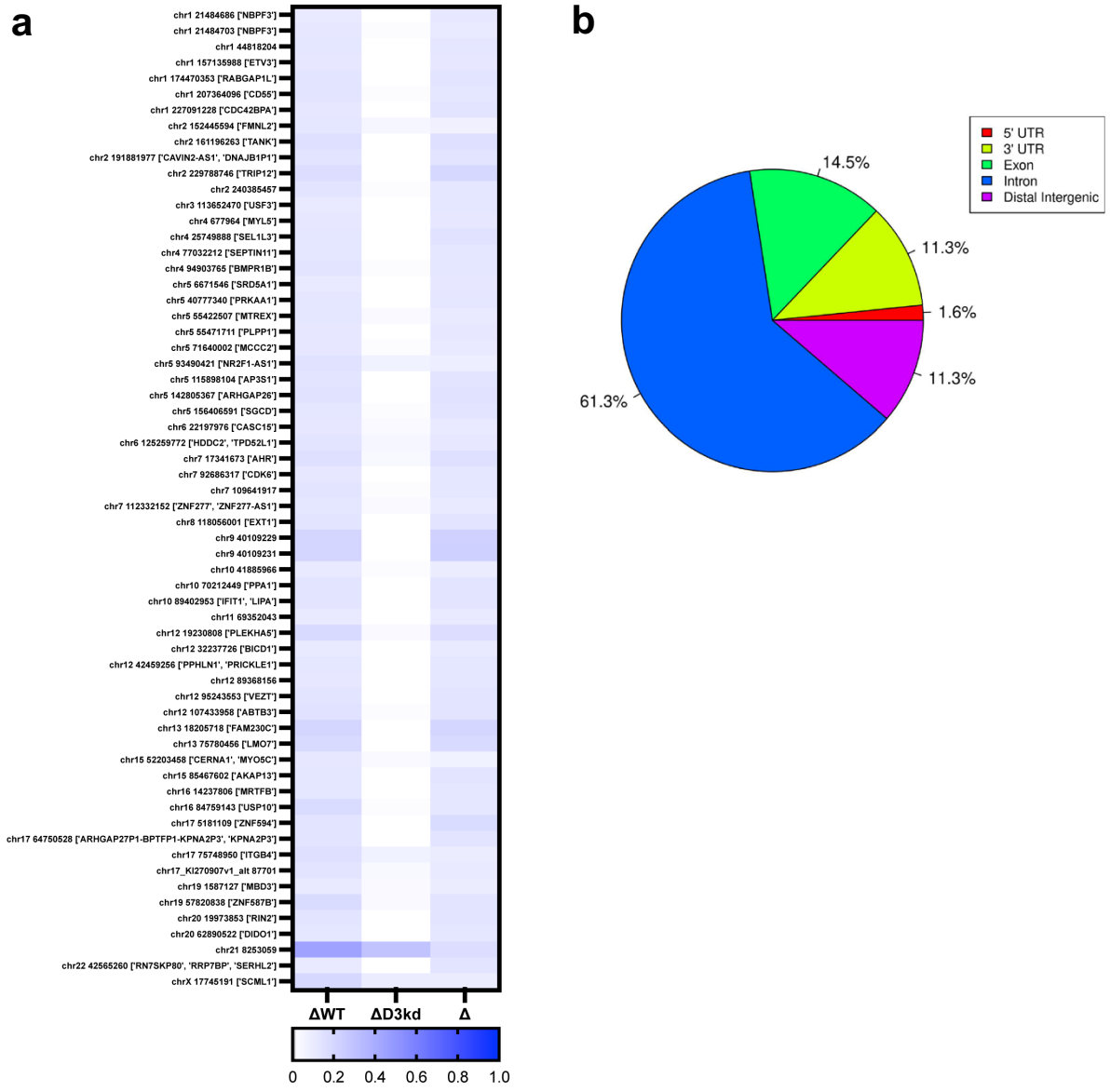


Figure 4-5: a) Heatmap of high confidence THUMPD3-dependent m^2G sites. Transcript and location are plotted with difference in PhOxi-seq response in wild type A549 lung cancer cells (ΔWT), difference in PhOxi-seq response in THUMPD3 knockdown A549 lung cancer cells (ΔD3k.d.), and the difference between the ΔWT and ΔD3 experiments (Δ). The sequencing error is illustrated in percentage, where 0 is 0% and 1 represents 100%. b) Pie chart of the percentages of 62 top potential m^2G sites that exhibit

THUMPD3 dependency located across different genomic regions (exon, intron, intergenic, 5' UTR, 3' UTR).

Consistent with previous observations that elevated THUMPD3 is associated with cancer phenotypes,[86] several potential m^2G sites were located on oncogenic transcripts. A notable example is integrin β4 subunit (ITGB4), which has been associated with several cancer types, including lung carcinoma,[89] and showed a marked response in PhOxi-seq for a potential m^2G site in an exon location (Figure 4-6a). This signal was significantly decreased for THUMPD3-depleted samples (Figure 4-6b). Another potential m^2G site was identified in the intron of A-kinase anchor protein 13 (AKAP13), which has been involved with breast cancer.[90] An increase in error was observed after PhOxi treatment, which was largely resolved upon THUMPD3 knockdown (Figure 4-6c-d).

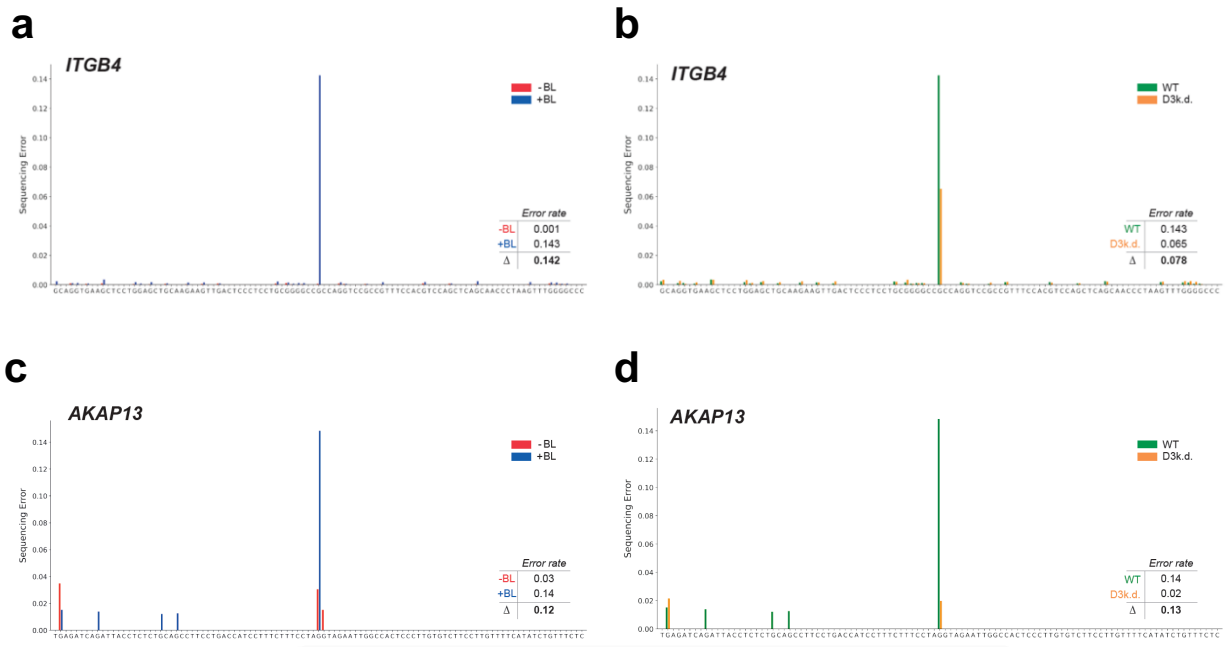


Figure 4-6: Bar plot of guanosine error rates in WT cells with (+BL, depicted in blue) and without (-BL, depicted in red) in THUMPD3-dependent sites a) ITGB4, c) AKAP13. Plot representing guanosine error for the photooxidized samples (+BL) in WT and THUMPD3 knockdown (D3k.d.) cells in b) ITGB4, d) AKAP13. Green bars indicate WT cell lines and orange bars indicate THUMPD3 depleted cell line D3k.d. The sequencing error is illustrated in percentage, where 0 is 0% and 1 represents 100%. Insert in graph: tables indicating error rates at given m^2G sites.

Several potential m^2G sites appeared to be highly dependent on THUMPD3, such as that observed in the intron of the E3 ligase Thyroid hormone receptor interactor 12 (TRIP12), which experienced complete loss of photooxidation-mediated mutation in THUMPD3-depleted cells (Figure 4-5a). Interestingly, TRIP12 is often mutated in lung adenocarcinoma patients,[91] and is known to regulate the extracellular matrix.[92] Notably, our analysis also uncovered m^2G sites within a well-characterised non-coding RNA,

pseudogene of snRNA 7SK (RN7SKP80) (Figure 4-7a-b). On the other hand, potential m^2G sites were observed in the 3' UTR of Zinc Finger Protein 587B (ZNF587B) (Figure 4-7c-d). This gene is believed to have tumor suppressor activity in ovarian cancer.[93]

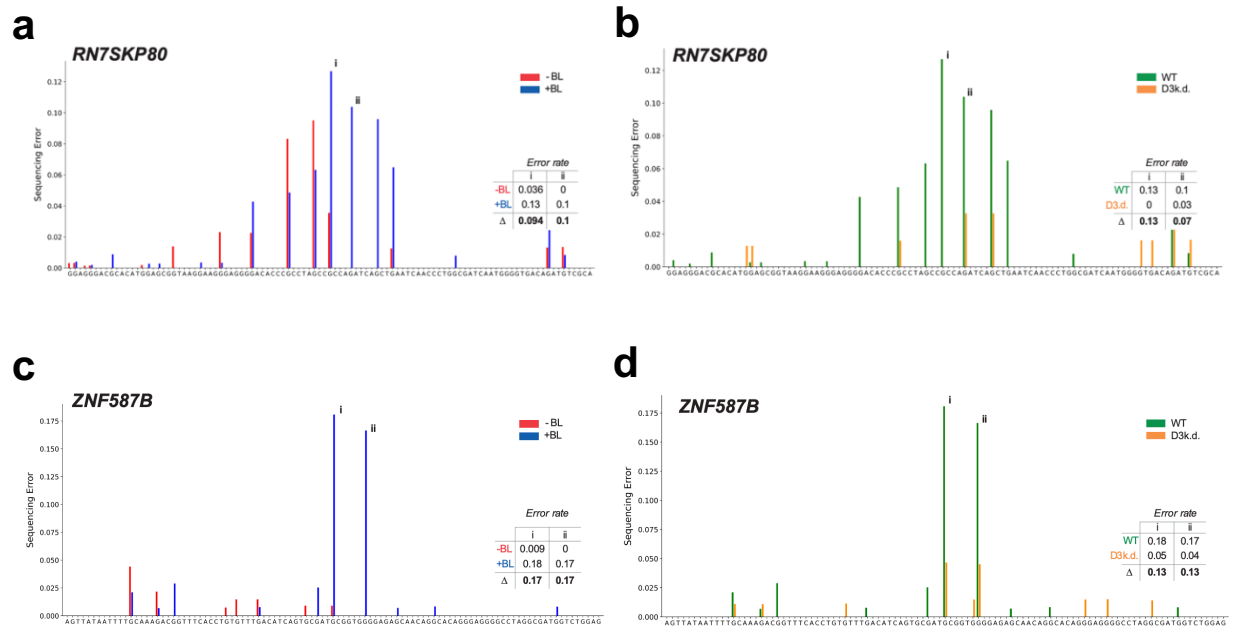


Figure 4-7: Bar plot of guanosine error rates in WT cells with (+BL, depicted in blue) and without (-BL, depicted in red) in THUMPD3-dependent sites a) RN7SKP80, c) ZNF587B. Plot representing guanosine error for the photooxidized samples (+BL) in WT and THUMPD3 knockdown (D3k.d.) cells in b) RN7SKP80, d) ZNF587B. Green bars indicate WT cell lines and orange bars indicate THUMPD3 depleted cell line D3k.d. The sequencing error is illustrated in percentage, where 0 is 0% and 1 represents 100%. Insert in graph: tables indicating error rates at given m^2G sites.

Furthermore, we identified 2978 or 362 blue-light sensitive G sites (lower and higher confidence hits, respectively), which are based on only WT samples. Additional sites that were unaffected by THUMPD3 depletion suggest the presence of a THUMPD3-

independent modification at these sites. For example, a potential THUMPD3-independent m^2G site is located in the mannose receptor C-type 2 transcript (MRC2), which does not resolve photooxidation-induced error in THUMPD3-depleted samples (Figure 4-8a-b). Additionally, a THUMPD3-independent m^2G modification potentially exists in contactin associated protein-like 3 pseudogene 2 (CNTNAP3P2) shown by an increase in error after PhOxi treatment, which is unaffected when THUMPD3 is targeted (Figure 4-8c-d).

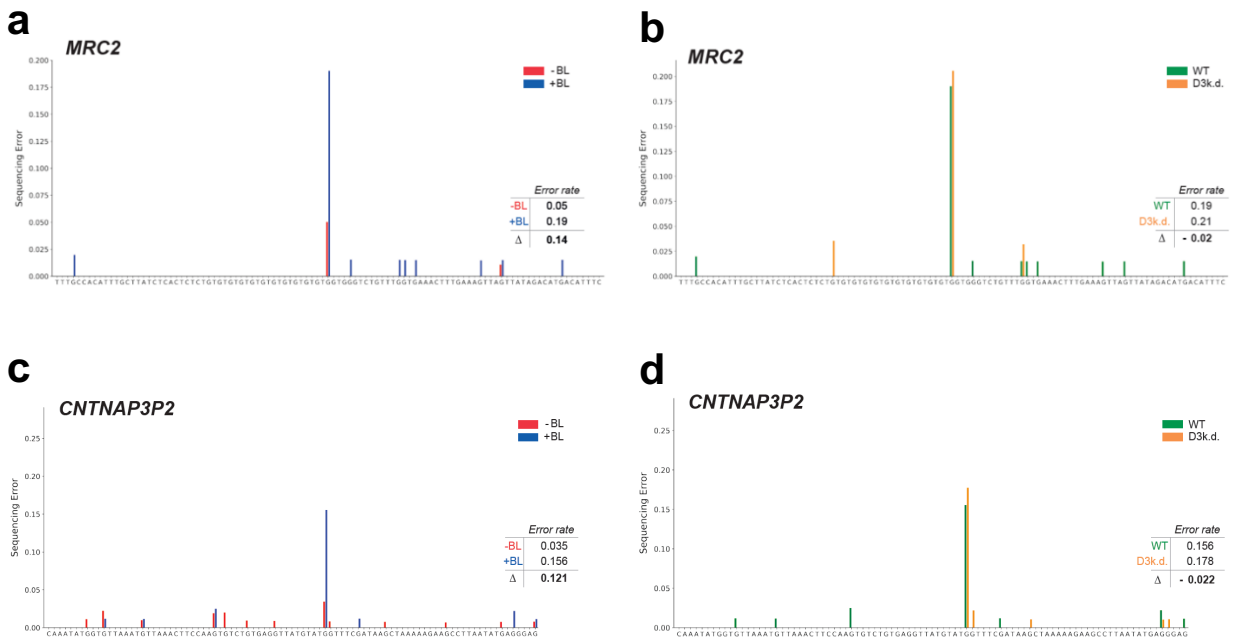


Figure 4-8: a) Plot representing guanosine error for the photooxidized samples (+BL) in THUMPD3-independent site, MRC2. b) Bar plot of guanosine error rates in MRC2 for the photo-oxidised samples (+BL) in WT and THUMPD3-depleted (ΔD3k.d.) cells. c) Plot of guanosine error in THUMPD3-independent site, CNTNAP3P2, for samples with (+BL) and without (-BL) blue-light treatment. d) Guanosine error rates in CNTNAP3P2 for the photo-oxidised samples (+BL) in WT and THUMPD3-depleted (ΔD3k.d.) cells. Green bars indicate WT cell lines and orange bars indicate THUMPD3 depleted cell line

(D3k.d). The sequencing error is illustrated in percentage, where 0 is 0% and 1 represents 100%. Insert in graph: tables indicating error rates at given m^2G sites.

Finally, potential modifications not installed by THUMPD3 were identified in the intron of tumor suppressor genes of Zinc finger homeobox 3 (ZFHX3) and CKLF like MARVEL transmembrane domain containing 8 (CMTM8). Their bar plots demonstrate largely unresolved photooxidation-induced error when THUMPD3 is depleted (Figure 4-9). Both ZFHX3 and CMTM8 are associated with various cancers including NSCLC.[94-95] Such sites may be under the control of another m^2G methylases, such as THUMPD2 or TRMT11.

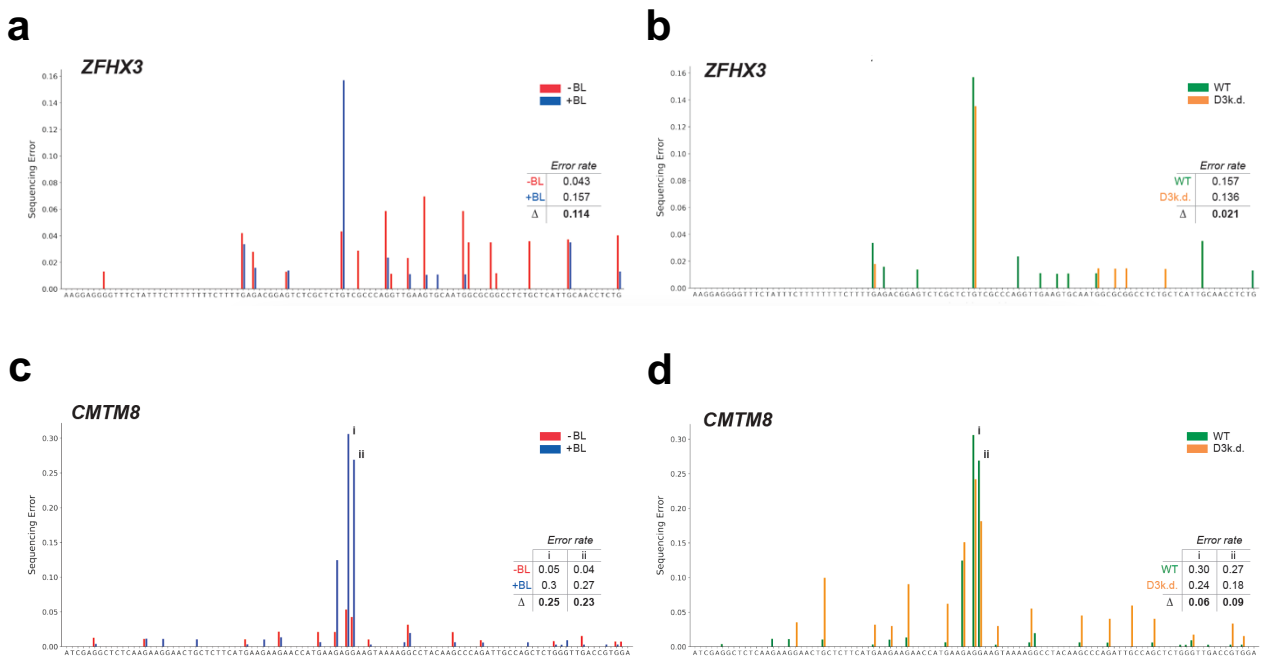


Figure 4-9: a) Plot representing guanosine error for the photooxidized samples (+BL) in THUMPD3-independent site, ZFHX3. b) Bar plot of guanosine error rates in ZFHX3 for

the photo-oxidised samples (+BL) in WT and THUMPD3-depleted (ΔD3k.d.) cells. c) Plot of guanosine error in THUMPD3-independent site, CMTM8, for samples with (+BL) and without (-BL) blue-light treatment. d) Guanosine error rates in CMTM8 for the photo-oxidised samples (+BL) in WT and THUMPD3-depleted (ΔD3k.d.) cells. Green bars indicate WT cell lines and orange bars indicate THUMPD3 depleted cell line (D3k.d). The sequencing error is illustrated in percentage, where 0 is 0% and 1 represents 100%. Insert in graph: tables indicating error rates at given m^2G sites.

4.4. Conclusion

Given the lack of tools to study m^2G, the Kouzarides lab and I sought to establish high confidence m^2G sites by knocking down THUMPD3, a known methyltransferase for m^2G installation in RNA. Our analysis of THUMPD3-dependent sites of m^2G within tRNA indicated that PhOxi-seq could detect such sites with a good signal-to-noise ratio. Developing an advanced bioinformatic pipeline to identify enzyme-dependent sites of modification was essential for analyzing other types of RNA. At present, the identification of specific sites of a given modification within less abundant RNAs remains a significant challenge within the RNA modification community. Consequently, we applied PhOxi-seq coupled with our novel bioinformatic pipeline to provide the first map of numerous potential sites of m^2G, including THUMPD3-dependent and independent targets, within polyA+ enriched RNA from A549 lung cancer cells. It is likely that our initial list of m^2G-modified sites within transcripts represents a significant underestimate of the true number. This is largely because of inherent limitations in our ability to differentiate signal-to-noise in our analyses. To increase this ratio, future approaches could include depletion of other

m^2G methyltransferases, either singularly or in combination. In summary, these findings emphasize the diverse landscape of m^2G within the transcriptome. Many of these transcripts are involved in human diseases, especially cancer, which is of particular note due to the association of THUMPD3 with human cancers. Additionally, we anticipate that these data will provide a valuable resource for other researchers in the field and support seminal studies on the role of m^2G in human health and disease.

4.5. Experimental details

4.5.1. Cell culture media, reagents, and maintenance

All cell culture was carried out under sterile conditions in a standard laminar flow hood. Human cell lines were maintained in filter top flasks or culture dishes and cultured in DMEM (Gibco), 10% [v/v] FBS growth media. A549 cell line was grown in DMEM supplemented with 1× penicillin-streptomycin (pen-strep) antibiotics and 10% [v/v] fetal bovine serum (FBS). Cells were grown at 37 °C, 5% [v/v] CO_2, and were maintained at > 80% viability. Cells were passaged every 3-5 days and seeded at 1 x 10^5 live cells per mL. Cells were passaged using conventional cell culture techniques. The maximum number of passages was 20 before a new vial of cells was revived. Tests for mycoplasma contamination were carried out every month or when a new vial of cells was revived.

4.5.2. Lentiviral production and infection

The LentiCRISPR v2 system was used to generate CRISPR/Cas9 knockout cell lines of RNA methyltransferase THUMPD3.[96] The 'all-in-one' pLKO-Tet-On system was used to generate inducible shRNA-mediated THUMPD3 knockdowns.[97-98] Constructs carrying sgRNAs and shRNAs are listed in Table 4-1. A549 cells were seeded onto L-polylysine (1 µg/mL) coated 10 cm dishes 24 hours prior to transfection in antibiotic-free DMEM media with 10% [v/v] FCS. Cells were 80% confluent on the day of transfection. To produce lentiviral particles, cells were transfected using 42 µL FuGENE® 4K Transfection Reagent (Promega, E5911) with 6 µg of lentiCRISPR v2 or pLKO-Tet-On constructs together with 3.5 µg psPAX2 (Addgene, 12260) and 4 µg pCMV-VSV-G (Addgene, 8454). 16 hours later, the media was exchanged for fresh media. 48 hours post-infection, virus particles were harvested and sterile filtered using 0.45 µm syringe filters (Millipore). Aliquots of lentiviral supernatants were stored at -70 °C.

Target cells (A549) were then transduced with the virus with polybrene (8 µg/mL) to increase the efficiency of transduction. 48 hours after infection transduced cells were selected for by treatment with puromycin (1 µg/mL). To induce expression of shRNAs, 10-100 ng/mL doxycycline was added to the media and the cells were incubated for 4-8 days to allow adequate expression of the shRNA.

4.5.3. Total RNA extraction with DNase treatment

Approximately 10^6 cells were homogenized in 1 mL QIAzol lysis reagent (QIAGEN). Homogenates were incubated for 5 minutes at room temperature. RNA was supplemented with chloroform (0.2 mL per 1 mL of QIAzol) and vigorously shaken for 20 seconds. After 3 minutes incubation at room temperature, homogenates were centrifuged at 12000 g for 15 minutes, at 4 °C. The aqueous (top) phase was carefully removed and transferred to a new RNase-free tube. 1.5 volumes of 100% EtOH were added. RNA was purified using a RNeasy Mini Kit (QIAGEN, 74104) with an additional step of DNase digestion after the first wash following the RNase-Free DNase Set protocol (QIAGEN, 79254). After the washes, columns were centrifuged for an additional 2 minutes to remove residual washing buffer. RNA was eluted with 30-50 µL of RNase-free water. RNA concentrations were measured using a Nanodrop spectrophotometer (Thermo Fisher) or Qubit BR RNA (Thermo Fisher, Q10211). RNA was stored at -70 °C until further use.

Table 4-1: Plasmids used to generate CRISPR KO and shRNA-inducible knockdown cells.

Name	Vector	Insert
sgscr	lentiCRISPR v2	**fwd** - CACCGGCGAGGTATTCGGCTCCGCG **rev** - AAACCGCGGAGCCGAATACCTCGCC
sgD3 #3	lentiCRISPR v2	**fwd** - CACCGCGCAAGAGGCCAATGGAGTG **rev** - AAACCACTCCATTGGCCTCTTGCGC
shscr	pLKO-Tet-On	**fwd** - CCGGCCTAAGGTTAAGTCGCCCTCGCTC GAGCGAGGGCGACTTAACCTTAGGTTTTT **rev** - AATTAAAAACCTAAGGTTAAGTCGCCCTCG CTCGAGCGAGGGCGACTTAACCTTAGG
shD3 #1	pLKO-Tet-On	**fwd** - CCGGCCTGAAGGAATGAAGCATTAT CTCGAGATAATGCTTCATTCCTTCAGGTTTTT **rev** - AATTAAAAACCTGAAGGAATGAAGCATTAT CTCGAGATAATGCTTCATTCCTTCAGG

4.5.4. Purification of small and large RNA fractions

RNA Clean & Concentrator kits (ZYMO RESEARCH, R1013 or R1017) were used to separate small (< 200nt) and large (> 200nt) RNA fractions, following the manufacturer's instructions. RNA concentration was determined using a QubitTM RNA HS Assay Kit (Thermo Fisher, Q32855). RNA was stored at -70 °C until needed.

4.5.5. PhOxi-seq on tRNA

A549 CRISPR/Cas9 targeted KO cell lines (sgscr and sgD3 #3, Table 1) were grown to 70-80% confluency. Cells were then harvested by scraping and pellets. Total RNA was then extracted from the cells as described above (Section 4.5.3.). Next, the small RNA fraction was purified (Section 4.5.4.) and subjected to blue light induced photo-oxidation described in chapter 2. Primers that specifically amplify tRNA-Gly-GCC-2 and tRNA-Gly-CCC-2 were used prior to YAMAT-seq[8]. Three biological replicates were included for each treatment. Bioinformatic analysis was performed as described in Section 4.5.7.

4.5.6. PhOxi-seq using mRNA as a template

A549 cell lines harboring shRNA vectors (shscr and shD3 #1, Table 1) were seeded into 15 cm dishes. shRNA expression was induced by doxycycline for 8 days. Cell pellets were collected, and total RNA was extracted as described above. Photooxidation was performed as described in chapter 2. rRNA HMR Removal Kit (Qiagen, 334386) was used in conjunction with NEXTFLEX® Rapid Directional RNA-Seq Kit 2.0 (PerkinElmer, NOVA-

5198-02) to deplete rRNA and perform whole transcriptome analysis according to the manufacturer's instructions. The RNA quality was assessed using an RNA Screen Tape (Agilent, 5067-5576) analyzed on a 4200 TapeStation System (Agilent, G2991BA). The NEXTFLEX® Rapid Directional RNA-Seq Kit 2.0 protocol was followed to perform all the remaining steps for library construction steps. The final multiplexed library, with an average size of 350-550 bp (at 10-20 nM) was submitted for 150 bp, paired-end sequencing on an Illumina NovaSeq® instrument. To ensure deep sequencing, a maximum of six samples was sequenced at each time. 120 GB of raw data were obtained per multiplexed library. Bioinformatic analyses was performed as describe in Section 4.5.7.

4.5.7. Data analysis

4.5.7.1. tRNA sequencing analysis

4.5.7.1.1. Data processing

Read quality was assessed using FastQC (v0.11.9)[99]. Reads were then filtered to include reads between 103 and 123 base pair (bp) long (113± 10 bp) using custom bash (v5.0.17) scripts. Then reads were filtered to include mature tRNA reads that end in 'GGA' or CCA' using custom bash scripts.

Filtered reads were then aligned to the tRNA reference sequence using bwa-mem (v0.7.17)[100] with parameters -t 32 -w 20. Then, aligned reads were filtered to include only

primarily mapped reads using samtools (v1.10)[101] view with parameter -F 260. Hard-clipped reads were also removed using samtools. The filtered reads were then sorted and indexed using samtools. The proportion of different bases or indels at each nucleotide position of the tRNA sequence for aligned reads of each tRNA isodecoder type were summarised using bam-readcount (v1.0.1)[102].

4.5.7.2. tRNA reference sequence

A comprehensive and non-repetitive set of reference sequences of tRNAs was constructed. It consists of isodecoder sequences of high confidence mature nuclear-encoded cytoplasmic tRNA genes (n=432) from GtRNAdb[103] and mitochondrial tRNA genes (n=28) from tRNAdb[104].

4.5.7.3. PhOxi-seq pipeline based on transcriptome-wide analysis

4.5.7.3.1. Data processing

Adapter sequences were trimmed from 150bp paired-end reads using Trimmomatic (v0.40-rc1)[105]. Filtered reads were aligned using STAR (v2.7.10a)[106] two-pass mode with alignment parameters -- outFilterMultimapNmax 20 --alignSJoverhangMin 5 --alignSJDBoverhangMin 3 --outFilterMismatchNmax 999 --outFilterMismatchNoverReadLmax 1.0 --alignIntronMin 21 --sjdbOverhang 149 --alignIntronMax 1000000 --alignMatesGapMax 1000000. Aligned reads were then filtered to include only primarily mapped reads using samtools (v1.10) view with parameter -F

260. Biological replicates (n=3) for treated and untreated samples were merged respectively using samtools merge. Reads spanning introns was reformatted using SplitNCigarReads from GATK (v4.3.0.0)[107]. The variant calling was performed for merged reads using bcftools (v1.16) mpileup with parameters --max-depth 1000000 and bcftools call under multiallelic mode[108] with parameters –multiallelic-caller –variants-only. The resulting vcf file containing variant information was processed using bedtools (v2.30.0)[109] intersect program to obtain bed file of genomic regions with variant calls. bam-readcount (v1.0.1). The subsequent analysis and plotting were performed using custom Python (v3.8) scripts.

4.5.7.3.2. Data exploration

As a relatively crude overview of the data to determine if THUMPD3 depletion affected error rates in a subset of reads, we calculated density plots of variant allele frequency (VAF) data distribution for each sample (Figure 4-10). The overall distribution of variant allele frequency (VAF) data of variant sites from bcftools (with non-zero depth in all replicates) (n=1192693) was investigated. VAF data encompasses single-nucleotide polymorphism (SNPs) and other single-nucleotide variations from the reference genome (i.e., sequencing error rates). The plots clearly show reduction of variant sites with VAF values between 0.1 and 0.5 upon THUMPD3 depletion. These data are consistent with loss of blue-light induced changes at sites of m^2G when THUMPD3 is targeted.

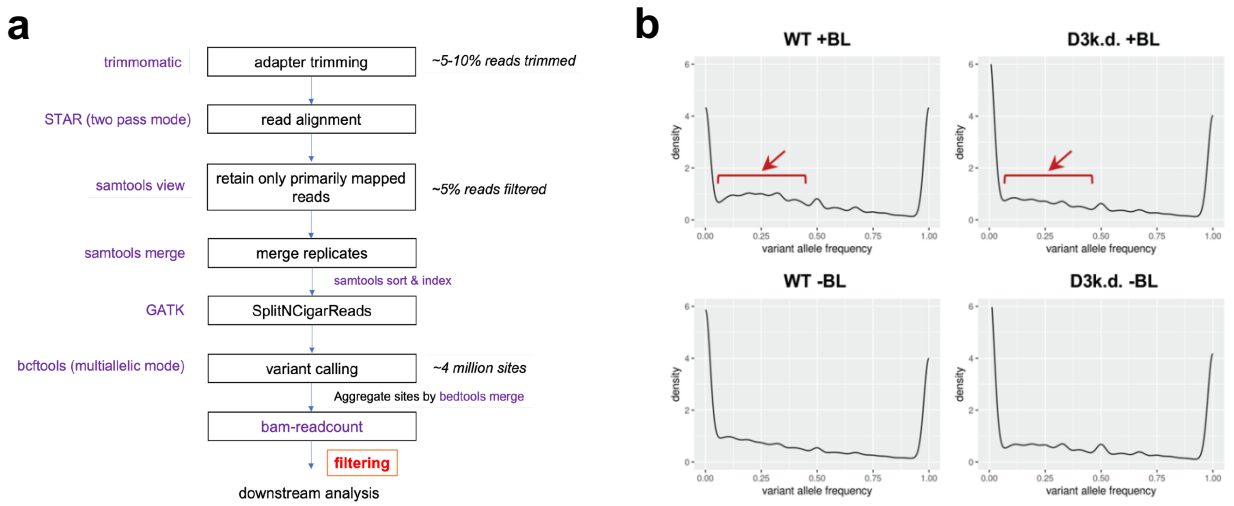


Figure 4-10: a) Overview of PhOxi-seq pipeline on the whole transcriptome. Main steps of the pipeline were presented in flowchart, involving steps of read processing, alignment, variant calling, and downstream analysis. Software involved at each step were labelled on the left and coloured purple, percentage of reads processed at each step and number of variant sites were labelled on the right. The conditions of filtering step coloured in red could be found in Figure 4-12b. The detailed description, rationale and quality control steps of the pipeline could be found in experimental detail section. b) Representative density plots from one biological replicate illustrating overall distribution of variant allele frequency (VAF) data of variant sites of representative replicate from each experiment group (control cell line (WT) with and without BL (±BL) treatment, THUMPD3 depleted cell line (D3k.d.) ±BL treatment). The y-axis of the density curve is density and the area under the curve gives the percentage. The x-axis ranges from 0-1, representing the VAF data, which here indicates the proportion of reads that do not match the reference genome at each variant site. The consistent pattern in the plots with two peaks corresponds to specific sets of data: one peak with VAF values (x-axis) approaching 1 represents mostly SNPs; one with VAF values approaching 0 could be mostly noises. The VAF data in this plot covers variant sites from the variant calling step that show non-zero depth in all replicates (n=1192693).

We next assessed the relationship between VAF of each sample. Our analysis revealed that blue light-treated RNA from WT cells exhibited the greatest differences from all other samples (Figure 4-11a). When only the top high confidence targets were considered, this difference was even more pronounced (Figure 4-11b), which is consistent with the expected higher sequencing error rate in samples with intact m^2G prior to THUMPD3 depletion. Overall, our findings suggest that PhOxi-treatment induces sequencing error rates that are largely resolved upon THUMPD3 depletion, underscoring the utility and effectiveness of PhOxi-seq in identifying enzyme-dependent m^2G.

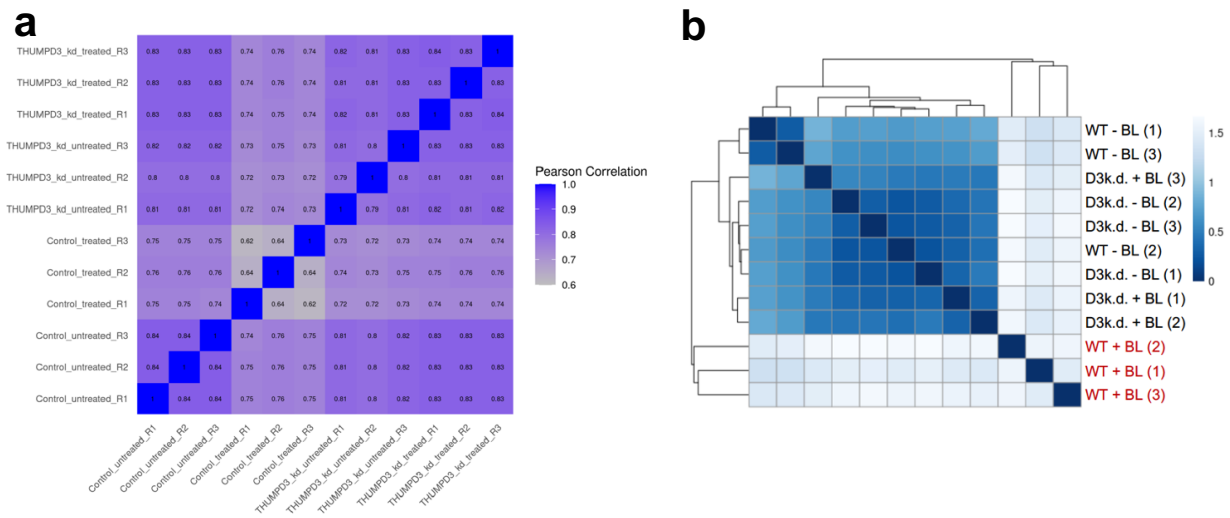


Figure 4-11: a) Heatmap based on Pearson Correlation coefficient of VAF data between samples. The VAF data here covers variant sites that show non-zero depth in all replicates (n=1192693). b) Heatmap of the sample-to-sample distances based on Euclidean distance measure for sequencing error rates of top potential m^2G sites (n=62) of all samples. A hierarchical clustering was also performed based on the sample distances and presented in the plot.

4.5.7.3.3. Filtering conditions for potential m^2G sites

At the end of the bioinformatic pipeline, prior to the target identification, we analyzed whether low sequencing depth leads to more uncertainty of the quantification of low abundance genes, as might be expected from results in both bulk and single-cell sequencing technologies. We therefore calculated the average sequencing depth of all SNPs across different samples. These were then sorted, in ascending order, by coverage and we employed a non-overlapping sliding window approach to group the sorted sites into 'windows'. We then calculated Spearman's rank correlation coefficient and plotted the data as shown in Figure 4-12a. This shows that correlations increase with higher sequencing depth, providing a rationale to filter out low read depth regions (indicated by dotted line, Figure 4-12a; read depth ≥ 50). This cut-off was included in our downstream analysis of the pipeline data (Figure 4-12b).

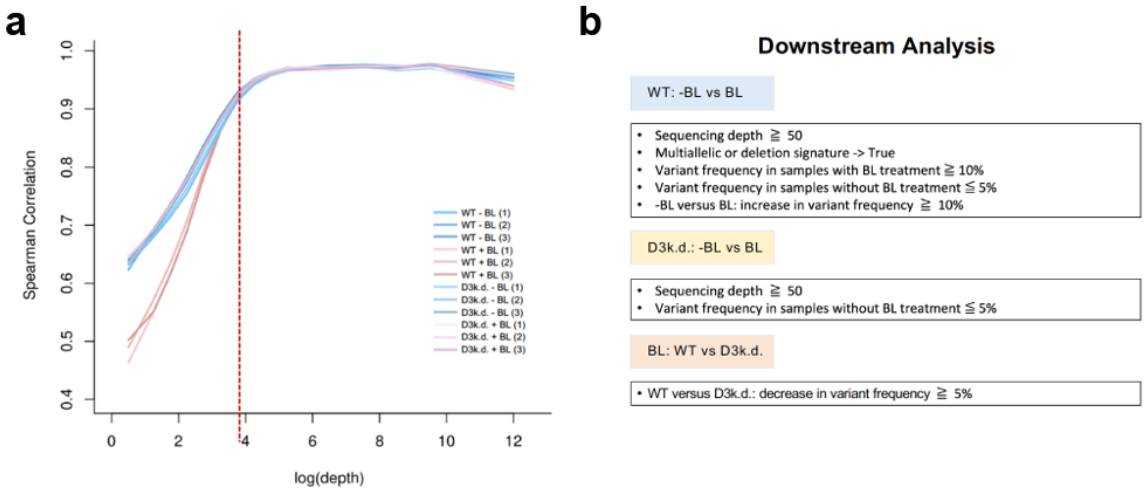


Figure 4-12: a) Line plot of average Spearman's rank correlation coefficient (y-axis) of VAF data for each replicate with the rest of the replicates of increasing windows of

average sequencing coverage (x-axis). The y-axis ranges from 0.4 – 1. The closer the coefficient values to 1, the stronger the correlation, and vice versa. The average sequencing depth of all 'SNP' sites across different samples were calculated. These sites are then sorted, in ascending order, by coverage. A non-overlapping sliding window approach is used to group the sorted sites where sites with similar coverage in a specific range are grouped into 'windows'. For each window (e.g. log(depth): 2-3, 3-4, log here represents natural log), the Spearman's rank correlation of the VAF data of the sites from the sample of interest and all other samples is calculated and averaged, and plotted. A vertical dotted line indicates the depth threshold of 50 at x-axis = 3.912 (log(50)). b) Table illustrating filtering conditions for potential m^2G modified sites during downstream analysis. The conditions include threshold on sequencing coverage, levels of sequencing error changes in different groups and presence of multiallelic signature.

The filtering conditions to select top potential m^2G modified sites focused on these aspects:

1. Sufficient sequencing coverage
2. Multiallelic or deletion signature
3. Levels of sequencing errors increase upon BL treatment
4. Levels of sequencing errors decrease upon enzyme depletion

The conditions are based on the observation of positive samples in tRNA sequence analysis. The cutoff of levels of sequencing errors it follows the assumptions that the levels of sequencing errors in control samples should be low since m^2G is non base pair-

disrupting modification. Also, the cutoff of sequencing error rate is non-stringent allowing more potential sites to be included.

4.5.7.3.4. Analysis of sequencing error rates of target sites

For selected top targets of the whole transcriptome, sequencing error data of genomic region 50bp upstream to 50bp downstream of the target site were extracted and guanosine errors were plotted. In terms of tRNAs, sequencing error data of complete sequence positions of tRNAs were extracted and guanosine errors were plotted. Specifically, sequencing error changes upon BL treatment in control samples and sequencing error changes upon enzyme depletion in BL treated samples were plotted in bar plots.

In tRNA sequencing analysis, levels of main error types (G → T substitution, G → C substitution, G → A substitution, and single base deletion) with and without BL treatment in control samples were plotted respectively for two tRNA isodecoder types (tRNA-Gly-GCC-2 and tRNA-Gly-CCC-2).

Replicate information was also included when analysing sequencing error levels at nucleotide level. In this case, reads from individual replicate were processed through the pipeline instead of merging the replicates for deeper sequencing coverage and used for downstream analysis.

4.5.7.4. Statistical analysis

Welch's t-test was used to compare sequencing error levels (n=3) between different samples given that the data follows the normality assumption but with different variances. Pearson's correlation coefficients were calculated to compare the similarity of overall VAF data between samples. Spearman's rank correlation coefficient was calculated for similarity between data of different samples in different depth ranges.

4.5.7.5. Code availability

The computational methods and custom scripts are available in the following Github repository: https://github.com/hanzhang2000/PhOxi_seq_transcriptome_code.

Chapter 5: Conclusions and future directions

Since their discovery in the early 1950s, over 160 post-transcriptional RNA modifications have been identified across all life.[50] These chemical modifications play a crucial role in regulating a wide range of RNA functions and governing key cellular processes such as splicing, translation, and degradation.[47] Unsurprisingly, increasing evidence highlights their critical involvement in the development of human diseases, especially cancer. Sequencing technologies capable of mapping these RNA modifications at single-nucleotide resolution across the transcriptome have been essential in advancing functional studies.[49]

Throughout my research journey, I developed a photo-oxidative sequencing method (PhOxi-seq) to identify modifications such as N^2-methylated guanosines at the single-nucleotide level. Using the chemoselectivity of photoredox chemistry, this technique can selectively photooxidize N^2-methylguanosine (m^2G) and N^2, N^2-dimethylguanosine ($m^2{}_2G$) in the presence of other RNA bases. I demonstrated that read-through errors significantly increase for m^2G due to the proposed formation of m^2-Iz and m^2-OG, which pair with G and A, respectively. In contrast, errors were reduced for $m^2{}_2G$, with a 6-fold increase in read-through resulting from the formation of $m^2{}_2$-Iz. The method proved compatible with a variety of sequence contexts in both tRNA and rRNA.

I further assessed the limitations of PhOxi-seq as a potential tool to detect N^1-methylated guanosine (m^1G). Although m^1G sites in RNA are sensitive to photooxidation and can be cleanly detected with PhOxi-seq, the same mutations occur at m^2G and the variability in read-through errors in the no-photooxidation control posed a potential confounding

factor. To address this limitation in the future, the use of an m^1G demethylase to fully convert m^1G into G will be studied.

In collaboration with the Kouzarides lab at the University of Cambridge (UK), we adapted PhOxi-seq to provide evidence for potential sites of THUMPD3-dependent m^2G in various RNA species isolated from a human lung cancer cell line (A549). In combination with the knockdown of an m^2G methyltransferase, THUMPD3, PhOxi-seq confirmed known sites of THUMPD3-dependent m^2G in tRNA. The scope of PhOxi-seq was further extended to detect THUMPD3-dependent m^2G sites in other RNA species. This method serves as a starting point for further studies to validate the presence of m^2G in other human RNA types. In addition, our approach identified many other potential m^2G sites within the transcriptome, which appear not to be installed by THUMPD3. These sites represent a useful database for researchers in the RNA modification field, particularly those analysing THUMPD2 and TRMT11.

In summary, PhOxi-seq detects methylated guanosine such as m^1G, m^2G, and $m^2{}_2G$ in various RNA types using photoredox chemistry. In collaboration with the Kouzarides lab, PhOxi-seq was optimized to unveil numerous potential sites of m^2G, including THUMPD3-dependent and independent targets, across mRNA regions and non-coding RNAs. These findings highlight the diverse landscape of m^2G within the transcriptome. Additionally, the data will support further exploration into the functional implications of these modified sites in RNA biology.

References

1) Minchin, S.; Lodge, J. Understanding Biochemistry: Structure and Function of Nucleic Acids. *Essays Biochem.* **2019**, *63* (4), 433–456.

2) Aubert, M.; O'Donohue, M.-F.; Lebaron, S.; Gleizes, P.-E. Pre-Ribosomal RNA Processing in Human Cells: From Mechanisms to Congenital Diseases. *Biomolecules* **2018**, *8* (4), 123.

3) Hocine, S.; Singer, R. H.; Grunwald, D. RNA Processing and Export. *Cold Spring Harb. Perspec. Biol.* **2010**, *2* (12), a000752.

4) Oberbauer, V.; Schaefer, M. tRNA-Derived Small RNAs: Biogenesis, Modification, Function and Potential Impact on Human Disease Development. *Genes* **2018**, *9* (12), 607.

5) O'Brien, J.; Hayder, H.; Zayed, Y.; Peng, C. Overview of MicroRNA Biogenesis, Mechanisms of Actions, and Circulation. *Front. Endocrinol.* **2018**, *9*, 402.

6) Statello, L.; Guo, C.-J.; Chen, L.-L.; Huarte, M. Gene Regulation by Long Non-Coding RNAs and Its Biological Functions. *Nat. Rev. Mol. Cell Biol.* **2020**, *22* (2), 96–118.

7) Tang, Q.; Li, L.; Wang, Y.; Wu, P.; Hou, X.; Ouyang, J.; Fan, C.; Li, Z.; Wang, F.; Guo, C.; Zhou, M.; Liao, Q.; Wang, H.; Xiang, B.; Jiang, W.; Li, G.; Zeng, Z.; Xiong, W. RNA Modifications in Cancer. *Br. J. Cancer* **2023**, *129* (2), 204–221.

8) Davis, F. F.; Allen, F. W. Ribonucleic Acids from Yeast Which Contain a Fifth Nucleotide. *J. Biol. Chem.* **1957**, *227* (2), 907–915.

9) Liu, W.-W.; Zheng, S.-Q.; Li, T.; Fei, Y.-F.; Wang, C.; Zhang, S.; Wang, F.; Jiang, G.-M.; Wang, H. RNA Modifications in Cellular Metabolism: Implications for

Metabolism-Targeted Therapy and Immunotherapy. *Signal Transduct. Target. Ther.* **2024**, *9*, No. 70.

10) Boccaletto, P.; Stefaniak, F.; Ray, A.; Cappannini, A.; Mukherjee, S.; Purta, E.; Kurkowska, M.; Shirvanizadeh, N.; Destefanis, E.; Groza, P.; Avşar, G.; Romitelli, A.; Pir, P.; Dassi, E.; Conticello, S. G.; Aguilo, F.; Bujnicki, J. M. MODOMICS: A Database of RNA Modification Pathways. 2021 Update. *Nucleic Acids Res.* **2021**, *50* (D1), D231-D235.

11) Arzumanian, V. A.; Dolgalev, G. V.; Kurbatov, I. Y.; Kiseleva, O. I.; Poverennaya, E. V. Epitranscriptome: Review of Top 25 Most-Studied RNA Modifications. *Int. J. Mol. Sci*, **2022**, *23* (22), 13851.

12) Esteve-Puig, R.; Bueno-Costa, A.; Esteller, M. Writers, Readers and Erasers of RNA Modifications in Cancer. *Cancer Lett.* **2020**, *474*, 127–137.

13) Abdelraheem, E.; Thair, B.; Varela, R. F.; Jockmann, E.; Popadić, D.; Hailes, H. C.; Ward, J. M.; Iribarren, A. M.; Lewkowicz, E. S.; Andexer, J. N.; Hagedoorn, P.; Hanefeld, U. Methyltransferases: Functions and Applications. *ChemBioChem* **2022**, *23* (18), e202200212.

14) Yang, Y.; Hsu, P. J.; Chen, Y.-S.; Yang, Y.-G. Dynamic Transcriptomic m^6A Decoration: Writers, Erasers, Readers and Functions in RNA Metabolism. *Cell Res.* **2018**, *28* (6), 616–624.

15) Zeng, C.; Huang, W.; Li, Y.; Weng, H. Roles of METTL3 in Cancer: Mechanisms and Therapeutic Targeting. *J. Hematol. Oncol.* **2020**, *13* (1), 117.

16) Yang, Z.; Yu, G.; Zhu, X.; Peng, T.; Lv, Y. Critical Roles of FTO-Mediated mRNA m6A Demethylation in Regulating Adipogenesis and Lipid Metabolism: Implications in Lipid Metabolic Disorders. *Genes Dis.* **2021**, *9* (1), 51–61.

17) Wei, W.; Ji, X.; Guo, X.; Ji, S. Regulatory Role of N^6-Methyladenosine (m^6A) Methylation in RNA Processing and Human Diseases. *J. Cell. Biochem.* **2017**, *118* (9), 2534–2543.

18) Wang, J.; Wang, J.; Gu, Q.; Ma, Y.; Yang, Y.; Zhu, J.; Zhang, Q. The Biological Function of m6A Demethylase ALKBH5 and Its Role in Human Disease. *Cancer Cell Int.* **2020**, *20* (1), 347.

19) Cui, L.; Ma, R.; Cai, J.; Guo, C.; Chen, Z.; Yao, L.; Wang, Y.; Fan, R.; Wang, X.; Shi, Y. RNA Modifications: Importance in Immune Cell Biology and Related Diseases. *Signal Transduct. Target. Ther.* **2022**, *7* (1), 334.

20) Wilkinson, E.; Cui, Y.-H.; He, Y.-Y. Roles of RNA Modifications in Diverse Cellular Functions. *Front. Cell Dev. Biol.* **2022**, *10*, 828683.

21) Horowitz, S.; Horowitz, A.; Nilsen, T. W.; Munns, T. W.; Rottman, F. M. Mapping of N6-Methyladenosine Residues in Bovine Prolactin mRNA. *Proc. Natl. Acad. Sci.* **1984**, *81* (18), 5667–5671.

22) Dominissini, D.; Moshitch-Moshkovitz, S.; Amariglio, N.; Rechavi, G. Transcriptome-Wide Mapping of N^6-Methyladenosine by m^6A-Seq. *Methods Enzymol.* **2015**, *560*, 131–147.

23) Hussain, S.; Sajini, A. A.; Blanco, S.; Dietmann, S.; Lombard, P.; Sugimoto, Y.; Paramor, M.; Gleeson, J. G.; Odom, D. T.; Ule, J.; Frye, M. NSun2-Mediated Cytosine-5 Methylation of Vault Noncoding RNA Determines Its Processing into Regulatory Small RNAs. *Cell Rep.* **2013**, *4* (2), 255–261.

24) Motorin, Y.; Helm, M. Methods for RNA Modification Mapping Using Deep Sequencing: Established and New Emerging Technologies. *Genes* **2019**, *10* (1), 35.

25) Zhang, Y.; Lu, L.; Li, X. Detection Technologies for RNA Modifications. *Exp. Mol. Med*. **2022**, *54* (10), 1601–1616.

26) Tang, J.; Chen, S.; Jia, G. Detection, Regulation, and Functions of RNA N^6-Methyladenosine Modification in Plants. *Plant Commun.* **2023**, *4* (3), 100546.

27) Ryvkin, P.; Leung, Y. Y.; Silverman, I. M.; Childress, M.; Valladares, O.; Dragomir, I.; Gregory, B. D.; Wang, L.-S. HAMR: High-Throughput Annotation of Modified Ribonucleotides. *RNA* **2013**, *19* (12), 1684–1692.

28) Hong, T.; Yuan, Y.; Chen, Z.; Xi, K.; Wang, T.; Xie, Y.; He, Z.; Su, H.; Zhou, Y.; Tan, Z.-J.; Weng, X.; Zhou, X. Precise Antibody-Independent m6A Identification via 4SedTTP-Involved and FTO-Assisted Strategy at Single-Nucleotide Resolution. *J. Am. Chem. Soc.* **2018**, *140* (18), 5886–5889.

29) Hauenschild, R.; Tserovski, L.; Schmid, K.; Thüring, K.; Winz, M.-L.; Sharma, S.; Entian, K.-D.; Wacheul, L.; Lafontaine, D. L. J.; Anderson, J.; Alfonzo, J.; Hildebrandt, A.; Jäschke, A.; Motorin, Y.; Helm, M. The Reverse Transcription Signature of N-1-Methyladenosine in RNA-Seq is Sequence Dependent. *Nucleic Acids Res.* **2015**, *43* (20), 9950-9964.

30) Dong, Z.-W.; Shao, P.; Diao, L.-T.; Zhou, H.; Yu, C.-H.; Qu, L.-H. RTL-P: A Sensitive Approach for Detecting Sites of 2′-O-Methylation in RNA Molecules. *Nucleic Acids Res.* **2012**, *40* (20), e157.

31) Aschenbrenner, J.; Werner, S.; Marchand, V.; Adam, M.; Motorin, Y.; Helm, M.; Marx, A. Engineering of a DNA Polymerase for Direct m6A Sequencing. *Angew. Chem. Int. Ed. Engl.* **2017**, *57* (2), 417–421.

32) Cozen, A. E.; Quartley, E.; Holmes, A. D.; Hrabeta-Robinson, E.; Phizicky, E. M.; Lowe, T. M. ARM-Seq: AlkB-Facilitated RNA Methylation Sequencing Reveals a Complex Landscape of Modified tRNA Fragments. *Nat. Methods* **2015**, *12* (9), 879–884.

33) Zheng, G.; Qin, Y.; Clark, W. C.; Dai, Q.; Yi, C.; He, C.; Lambowitz, A. M.; Pan, T. Efficient and Quantitative High-Throughput tRNA Sequencing. *Nat. Methods* **2015**, *12* (9), 835–837.

34) Dai, Q.; Zheng, G.; Schwartz, M. H.; Clark, W. C.; Pan, T. Selective Enzymatic Demethylation of N^2,N^2-dimethylguanosine in RNA and Its Application in High-throughput tRNA Sequencing. *Angew. Chem. Int. Ed. Engl.* **2017**, *56* (18), 5017–5020.

35) Meyer, K. D. DART-Seq: An Antibody-Free Method for Global m^6A Detection. *Nat. Methods* **2019**, *16* (12), 1275–1280.

36) Knutson, S. D.; Arthur, R. A.; Johnston, H. R.; Heemstra, J. M. Selective Enrichment of A-to-I Edited Transcripts from Cellular RNA Using Endonuclease V. *J. Am. Chem. Soc.* **2020**, *142* (11), 5241–5251.

37) Imanishi, M.; Tsuji, S.; Suda, A.; Futaki, S. Detection of N^6-Methyladenosine Based on the Methyl-Sensitivity of Mazf RNA Endonuclease. *Chem. Commun.* **2017**, *53* (96), 12930–12933.

38) Schaefer, M.; Pollex, T.; Hanna, K.; Lyko, F. RNA Cytosine Methylation Analysis by Bisulfite Sequencing. *Nucleic Acids Res.* **2008**, *37* (2), e12.

39) Mahdavi-Amiri, Y.; Chung Kim Chung, K.; Hili, R. Single-Nucleotide Resolution of N^6-Adenine Methylation Sites in DNA and RNA by Nitrite Sequencing. *Chem. Sci.* **2021**, *12* (2), 606–612.

40) Liu, C.; Sun, H.; Yi, Y.; Shen, W.; Li, K.; Xiao, Y.; Li, F.; Li, Y.; Hou, Y.; Lu, B.; Liu, W.; Meng, H.; Peng, J.; Yi, C.; Wang, J. Absolute Quantification of Single-Base m^6A Methylation in the Mammalian Transcriptome Using GLORI. *Nat. Biotechnol.* **2022** *41* (3), 355–366.

41) Herzog, V. A.; Reichholf, B.; Neumann, T.; Rescheneder, P.; Bhat, P.; Burkard, T. R.; Wlotzka, W.; von Haeseler, A.; Zuber, J.; Ameres, S. L. Thiol-Linked Alkylation of RNA to Assess Expression Dynamics. *Nat. Methods* **2017**, *14* (12), 1198–1204.

42) Suzuki, T.; Ueda, H.; Okada, S.; Sakurai, M. Transcriptome-Wide Identification of Adenosine-to-Inosine Editing Using the Ice-Seq Method. *Nat. Protoc.* **2015**, *10* (5), 715–732.

43) Lin, S.; Liu, Q.; Jiang, Y.-Z.; Gregory, R. I. Nucleotide Resolution Profiling of m^7G tRNA Modification by TRAC-Seq. *Nat. Protoc.* **2019**, *14* (11), 3220–3242.

44) Orsolic, I.; Carrier, A.; Esteller, M. Genetic and Epigenetic Defects of the RNA Modification Machinery in Cancer. *Trends Genet.* **2023**, *39* (1), 74–88.

45) Li, M.; Zhu, J.; Lv, Z.; Qin, H.; Wang, X.; Shi, H. Recent Advances in RNA-targeted Cancer Therapy. *ChemBioChem* **2023**, *25* (4), e202300633.

46) Qiu, L.; Jing, Q.; Li, Y.; Han, J. RNA Modification: Mechanisms and Therapeutic Targets. *Mol. Biomed*. **2023**, *4* (1), 25.

47) Zaccara, S.; Ries, R. J.; Jaffrey, S. R. Reading, Writing and Erasing mRNA Methylation. *Nat. Rev. Mol. Cell Biol.* **2019**, *20* (10), 608−624.

48) Jonkhout, N.; Tran, J.; Smith, M. A.; Schonrock, N.; Mattick, J.S.; Novoa, E. M. The RNA Modification Landscape in Human Disease. *RNA* **2017**, *23* (12), 1754−1769.

49) Barbieri, I.; Kouzarides, T. Role of RNA Modifications in Cancer. *Nat. Rev. Cancer* **2020**, *20* (6), 303−322.

50) Boccaletto, P.; Machnicka, M. A.; Purta, E.; Piątkowski, P.; Bagiński, B.; Wirecki, T. K.; de Crécy-Lagard, V.; Ross, R.; Limbach, P. A.; Kotter, A.; Helm, M.; Bujnicki, J. M. MODOMICS: A Database of RNA Modification Pathways. 2017 Update. *Nucleic Acids Res.* **2018**, *46* (D1), D303−D307.

51) Bartee, D.; Gamage, S. T.; Link, C. N.; Meier, J. L. Arrow Pushing in RNA Modification Sequencing. *Chem. Soc. Rev.* **2021**, *50* (17), 9482−9502.

52) Li, X.; Xiong, X.; Yi, C. Epitranscriptome Sequencing Technologies: Decoding RNA Modifications. *Nat. Methods* **2017**, *14* (1), 23−31.

53) Jurga, S.; Barciszewski, J. *Epitranscriptomics*, 1st ed.; Springer, 2021.

54) Sergiev, P. V.; Bogdanov, A. A.; Dontsova, O. A. Ribosomal RNA Guanine-(N2)-Methyltransferases and Their Targets. *Nucleic Acids Res.* **2007**, *35* (7), 2295−2301.

55) Chen, Q.; Yan, M.; Cao, Z.; Li, X.; Zhang, Y.; Shi, J.; Feng, G.; Peng, H.; Zhang, X.; Zhang, Y.; Qian, J.; Duan, E.; Zhai, Q.; Zhou, Q. Sperm TsRNAs Contribute to Intergenerational Inheritance of an Acquired Metabolic Disorder. *Science* **2016**, *351* (6271), 397–400.

56) Zhang, Y.; Zhang, X.; Shi, J.; Tuorto, F.; Li, X.; Liu, Y.; Liebers, R.; Zhang, L.; Qu, Y.; Qian, J.; Pahima, M.; Liu, Y.; Yan, M.; Cao, Z.; Lei, X.; Cao, Y.; Peng, H.; Liu, S.; Wang, Y.; Zheng, H.; Woolsey, R.; Quilici, D.; Zhai, Q.; Li, L.; Zhou, T.; Yan, W.; Lyko, F.; Zhang, Y.; Zhou, Q.; Duan, E.; Chen, Q. Dnmt2 Mediates Intergenerational Transmission of Paternally Acquired Metabolic Disorders through Sperm Small Non-Coding RNAs. *Nat. Cell Biol.* **2018**, *20* (5), 535–540.

57) Blaesius, K.; Abbasi, A. A.; Tahir, T. H.; Tietze, A.; Picker-Minh, S.; Ali, G.; Farooq, S.; Hu, H.; Latif, Z.; Khan, M. N.; Kaindl, A. Mutations in the tRNA Methyltransferase 1 Gene TRMT1 Cause Congenital Microcephaly, Isolated

Inferior Vermian Hypoplasia and Cystic Leukomalacia in Addition to Intellectual Disability. *Am. J. Med. Genet. A* **2018**, *176* (11), 2517–2521.

58) Najmabadi, H.; Hu, H.; Garshasbi, M.; Zemojtel, T.; Abedini, S. S.; Chen, W.; Hosseini, M.; Behjati, F.; Haas, S.; Jamali, P.; Zecha, A.; Mohseni, M.; Püttmann, L.; Vahid, L. N.; Jensen, C.; Moheb, L. A.; Bienek, M.; Larti, F.; Mueller, I.; Weissmann, R.; Darvish, H.; Wrogemann, K.; Hadavi, V.; Lipkowitz, B.; Esmaeeli-Nieh, S.; Wieczorek, D.; Kariminejad, R.; Firouzabadi, S. G.; Cohen, M.; Fattahi, Z.; Rost, I.; Mojahedi, F.; Hertzberg, C.; Dehghan, A.; Rajab, A.; Banavandi, M. J.; Hoffer, J.; Falah, M.; Musante, L.; Kalscheuer, V.; Ullmann, R.; Kuss, A. W.; Tzschach, A.; Kahrizi, K.; Ropers, H. H. Deep Sequencing Reveals 50 Novel Genes for Recessive Cognitive Disorders. *Nature* **2011**, *478* (7367), 57–63.

59) Dai, Q.; Zheng, G.; Schwartz, M. H.; Clark, W. C.; Pan, T. Selective Enzymatic Demethylation of N^2, N^2-Dimethylguanosine in RNA and Its Application in High-Throughput tRNA Sequencing. *Angew. Chem., Int. Ed.* **2017**, *56* (18), 5017−5020.

60) Chen, W.; Song, X.; Lv, H.; Lin, H. iRNA-m2G: Identifying N^2-methylguanosine Sites Based on Sequence-Derived Information. *Mol. Ther. Nucleic Acids* **2019**, *18*, 253−258.

61) Ao, C.; Zou, Q.; Yu, L. RFhy-M2G: Identification of RNA N2-Methylguanosine Modification Sites Based on Random Forest and Hybrid Features. *Methods* **2022**, *203*, 32–39.

62) Ryvkin, P.; Leung, Y. Y.; Silverman, I. M.; Childress, M.; Valladares, O.; Dragomir, I.; Gregory, B. D.; Wang, L.-S. HAMR: High-Throughput Annotation of Modified Ribonucleotides. *RNA* **2013**, *19* (12), 1684−1692.

63) Kobayashi, K.; Tagawa, S. Direct Observation of Guanine Radical Cation Deprotonation in Duplex DNA Using Pulse Radiolysis. *J. Am. Chem. Soc.* **2003**, *125* (34), 10213−10218.

64) Morikawa, M.; Kino, K.; Oyoshi, T.; Suzuki, M.; Kobayashi, T.; Miyazawa, H. Product Analysis of Photooxidation in Isolated Quadruplex DNA; 8-Oxo-7,8-Dihydroguanine and Its Oxidation Product at 3′-G Are Formed Instead of 2,5-Diamino-4h-Imidazol-4-One. *RSC Adv.* **2013**, *3* (48), 25694.

65) Kino, K.; Sugiyama, H. Possible cause of G-C→C-G Transversion Mutation by Guanine Oxidation Product, Imidazolone. *Chem. Biol.* **2001**, *8* (4), 369−378.

66) Xie, L.-I.; Yang, Y.-T.; Wang, R.-L.; Cheng, H.-P.; Li, Z.-Y.; Liu, L.; Mao, L.; Wang, M.; Cheng, L. Identification of Flavin Mononucleotide as a Cell-Active Artificial N^6-Methyladenosine RNA Demethylase. *Angew. Chemie. Int. Ed.* **2019**, *58* (15), 5028−5032.

67) Nappi, M.; Hofer, A.; Balasubramanian, S.; Gaunt, M. J. Selective Chemical Functionalization at N6-Methyladenosine Residues in DNA Enabled by Visible-Light-Mediated Photoredox Catalysis. *J. Am. Chem. Soc.* **2020**, *142* (51), 21484−21492.

68) Patel, S.; Badir, S. O.; Molander, G. A. Developments in Photoredox-Mediated Alkylation for DNA-Encoded Libraries. *Trends Chem.* **2021**, *3* (3), 161−175.

69) Kino, K.; Kobayashi, T.; Arima, E.; Komori, R.; Kobayashi, T.; Miyazawa, H. Photoirradiation Products of Flavin Derivatives, and the Effects of Photooxidation on Guanine. *Bioorg. Med. Chem. Lett.* **2009**, *19* (17), 2070−2074.

70) Shigematsu, M.; Honda, S.; Loher, P.; Telonis, A. G.; Rigoutsos, I.; Kirino, Y. YAMAT-seq: an efficient method for high-throughput sequencing of mature transfer RNAs. *Nucleic Acids Res.* **2017**, *45* (9), No. e70.

71) Langmead, B.; Trapnell, C.; Pop, M.; Salzberg, S. L. Ultrafast and Memory-Efficient Alignment of Short DNA Sequences to the Human Genome. *Genome Biol.* **2009**, *10* (3), R25.

72) Korfmann, C. snp-counter, GitHub. https://github.com/ckorfmann/snp-counter (accessed 2021-05-07)

73) Wang, Y.; Katanski, C. D.; Watkins, C.; Pan, J. N.; Dai, Q.; Jiang, Z.; Pan, T. A High-Throughput Screening Method for Evolving a Demethylase Enzyme with Improved and New Functionalities. *Nucleic Acids Res.* **2021**, *49* (5), No. e30.

74) Dai, Q.; Zheng, G.; Schwartz, M. H.; Clark, W. C.; Pan, T. Selective Enzymatic Demethylation of N^2, N^2-Dimethylguanosine in RNA and Its Application in High-Throughput tRNA Sequencing. *Angew. Chemie. Int. Ed.* **2017**, *56* (18), 5017−5020.

75) Björk, G. R; Jacobsson K; Nilsson K; Johansson M. J.; Byström A. S.; Persson O. P. A Primordial tRNA Modification Required for the Evolution of Life? *EMBO J.* **2001**, *20* (1-2), 231–239.

76) Powell, C. A.; Kopajtich, R.; D'Souza, A. R.; Rorbach, J.; Kremer, L. S.; Husain, R. A.; Dallabona, C.; Donnini, C.; Alston, C. L.; Griffin, H.; Pyle, A.; Chinnery, P. F.; Strom, T. M.; Meitinger, T.; Rodenburg, R. J.; Schottmann, G.; Schuelke, M.; Romain, N.; Haller, R. G.; Ferrero, I.; Haack, T. B.; Taylor, R. W.; Prokisch, H.; Minczuk, M. TRMT5 Mutations Cause a Defect in Post-Transcriptional Modification of Mitochondrial tRNA Associated with Multiple Respiratory-Chain Deficiencies. *Am. J. Hum. Genet.* **2015**, *97* (2), 319–328.

77) Bhatta, A.; Dienemann, C.; Cramer, P.; Hillen, H. S. Structural Basis of RNA Processing by Human Mitochondrial RNase P. *Nat. Struct. Mol. Biol.* **2021**, *28* (9), 713–723.

78) Yew, T. W.; McCreight, L.; Colclough, K.; Ellard, S.; Pearson, E. R. tRNA Methyltransferase Homologue Gene TRMT10A Mutation in Young Adult-Onset Diabetes with Intellectual Disability, Microcephaly and Epilepsy. *Diabet. Med.* **2016**, *33* (9), e21-e25.

79) Seidel, A.; Brunner, S.; Seidel, P.; Fritz, G. I.; Herbarth, O.; Modified Nucleosides: An Accurate Tumour Marker for Clinical Diagnosis of Cancer, Early Detection and Therapy Control. *Br J Cancer.* **2006**, *94* (11):1726-33.

80) Chung Kim Chung, K.; Mahdavi-Amiri, Y.; Korfmann, C.; Hili, R. PhOxi-Seq: Single-Nucleotide Resolution Sequencing of N^2-methylation at Guanosine in RNA by Photoredox Catalysis. *J. Am. Chem. Soc.* **2022**, *144* (13), 5723–5727.

81) Zuker, M. Mfold Web Server for Nucleic Acid Folding and Hybridization Prediction. *Nucleic Acids Res.* **2003**, *31* (13), 3406–3415.

82) Jones, J. D.; Franco, M.; Smith, T.; Snyder, L. R.; Anders, A. G.; Ruotolo, B. T.; Kennedy, R. T.; Koutmou, K. S. Methylated Guanosine and Uridine Modifications in S. cerevisiae mRNAs Modulate Translation Elongation. *RSC Chem. Biol.* **2023**, *4* (5), 363–378.

83) Yang, W.-Q.; Xiong, Q.-P.; Ge, J.-Y.; Li, H.; Zhu, W.-Y.; Nie, Y.; Lin, X.; Lv, D.; Li, J.; Lin, H.; Liu, R.-J. THUMPD3-TRMT112 Is a m^2G Methyltransferase Working on a Broad Range of tRNA Substrates. *Nucleic Acids Res.* **2021**, *49* (20), 11900-11919.

84) Wang, C.; Ulryck, N.; Herzel, L.; Pythoud, N.; Kleiber, N.; Guérineau, V.; Jactel, V.; Moritz, C.; Bohnsack, M. T.; Carapito, C.; Touboul, D.; Bohnsack, K. E.; Graille, M. N^2-Methylguanosine Modifications on Human tRNAs and SnRNA U6 Are Important for Cell Proliferation, Protein Translation and Pre-mRNA Splicing. *Nucleic Acids Res.* **2023**, *51* (14), 7496–7519.

85) Yang, W.-Q.; Ge, J.-Y.; Zhang, X.; Zhu, W.-Y.; Lin, L.; Shi, Y.; Xu, B.; Liu, R.-J. THUMPD2 Catalyzes the N2-Methylation of U6 SnRNA of the Spliceosome Catalytic Center and Regulates Pre-mRNA Splicing and Retinal Degeneration. *Nucleic Acids Res.* **2024**, *52* (6), 3291–3309.

86) Klimontova, M.; Zhang, H.; Campos-Laborie, F.; Webster, N.; Andrews, B.; Chung Kim Chung, K.; Hili, R.; Kouzarides, K.; Bannister, A. J. THUMPD3 Regulates Alternative Splicing of ECM Transcripts in Human Lung Cancer Cells and Promotes Proliferation and Migration. Submitted to PLOS One.

87) Yu, Y. P.; Ding, Y.; Chen, Z.; Liu, S.; Michalopoulos, A.; Chen, R.; Gulzar, Z. G.; Yang, B.; Cieply, K. M.; Luvison, A.; Ren, B.-G.; Brooks, J. D.; Jarrard, D.; Nelson, J. B.; Michalopoulos, G. K.; Tseng, G. C.; Luo, J.-H. Novel Fusion Transcripts Associate with Progressive Prostate Cancer. *Am. J. Pathol.* **2014**, *184* (10), 2840–2849.

88) Zheng, H.-X.; Zhang, X.-S.; Sui, N. Advances in The Profiling of N^6-Methyladenosine (m^6A) Modifications. *Biotechnol. Adv.* **2020**, *45*, 107656.

89) Huang, W.; Fan, L.; Tang, Y.; Chi, Y.; Li, J. A Pan-Cancer Analysis of the Oncogenic Role of Integrin Beta4 (ITGB4) in Human Tumors. *Int. J. Gen. Med.* **2021**, 14, 9629–9645.

90) Wirtenberger, M.; Tchatchou, S.; Hemminki, K.; Klaes, R.; Schmutzler, R. K.; Bermejo, J. L.; Chen, B.; Wappenschmidt, B.; Meindl, A.; Bartram, C. R.; Burwinkel, B. Association of Genetic Variants in the Rho Guanine Nucleotide Exchange Factor AKAP13 with Familial Breast Cancer. *Carcinogenesis* **2005**, *27* (3), 593–598.

91) Li, G.; Yi, S.; Yang, F.; Zhou, Y.; Ji, Q.; Cai, J.; Mei, Y. Identification of Mutant Genes with High-frequency, High-risk, and High-expression in Lung Adenocarcinoma. *Thorac. Cancer* **2014**, *5* (3), 211–218

92) Lee, K. K.; Rajagopalan, D.; Bhatia, S.S.; Tirado-Magallanes, R.; Chng, W. J.; Jha, S. The Oncogenic E3 Ligase TRIP12 Suppresses Epithelial–Mesenchymal Transition (EMT) and Mesenchymal Traits through ZEB1/2. *Cell Death Discov.* **2021**, *7* (1), 95

93) Liu, Y.; Ouyang, Q.; Sun, Z.; Tan, J.; Huang, W.; Liu, J.; Liu, Z.; Zhou, H.; Zeng, F.; Liu, Y. The Novel Zinc Finger Protein 587B Gene, ZNF587B, Regulates Cell Proliferation and Metastasis in Ovarian Cancer Cells in vivo and in vitro. *Cancer Manag. Res.* **2020**, *12*, 5119–5130.

94) Zhang, J.; Zhou, N.; Lin, A.; Luo, P.; Chen, X.; Deng, H.; Kang, S.; Guo, L.; Zhu, W.; Zhang, J. ZFHX3 Mutation as a Protective Biomarker for Immune Checkpoint Blockade in Non-Small Cell Lung Cancer. *Cancer Immunol. Immunother.* **2021**, *70* (1), 137–151.

95) Zhang, W.; Qi, H.; Mo, X.; Sun, Q.; Li, T.; Song, Q.; Xu, K.; Hu, H.; Ma, D.; Wang, Y. CMTM8 is Frequently Downregulated in Multiple Solid Tumors. *Appl. Immunohistochem. Mol. Morphol.* **2017**, *25* (2), 122–128.

96) Shalem, O.; Sanjana, N. E.; Hartenian, E.; Shi, X.; Scott, D. A.; Mikkelsen, T. S.; Heckl, D.; Ebert, B. L.; Root, D. E.; Doench, J. G.; Zhang, F. Genome-Scale CRISPR-Cas9 Knockout Screening in Human Cells. *Science* **2014**, *343* (6166), 84–87.

97) Wiederschain, D.; Susan, W.; Chen, L.; Loo, A.; Yang, G.; Huang, A.; Chen, Y.; Caponigro, G.; Yao, Y.; Lengauer, C.; Sellers, W. R.; Benson, J. D. Single-Vector Inducible Lentiviral RNAi System for Oncology Target Validation. *Cell Cycle* **2009**, *8* (3), 498–504.

98) Wee, S.; Wiederschain, D.; Maira, S.-M.; Loo, A.; Miller, C.; deBeaumont, R.; Stegmeier, F.; Yao, Y.-M.; Lengauer, C. PTEN-Deficient Cancers Depend on PIK3CB. *Proc. Natl. Acad. Sci. U.S.A* **2008**, *105* (35), 13057–13062.

99) Andrews S. (2010). FastQC: a quality control tool for high throughput sequence data. Available online at: http://www.bioinformatics.babraham.ac.uk/projects/fastqc

100) Li H. (2013) Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. arXiv:1303.3997v2 [q-bio.GN]

101) Danecek, P.; Bonfield, J. K.; Liddle, J.; Marshall, J.; Ohan, V.; Pollard, M. O.; Whitwham, A.; Keane, T.; McCarthy, S. A.; Davies, R. M.; Li, H. Twelve Years of SAMtools and BCFtools, *GigaScience* **2021**, *10* (2), giab008.

102) Khanna, A.; Larson, D. E.; Srivatsan, S. N.; Mosior, M.; Abbott, T. E.; Kiwala, S.; et al. Bam-Readcount - Rapid Generation of Basepair-Resolution Sequence Metrics. *J. Open Source Softw.* **2022**, *7* (69), 3722.

103) Chan, P. P.; Lowe, T. M. GtRNAdb 2.0: An Expanded Database of Transfer RNA Genes Identified in Complete and Draft Genomes. *Nucleic Acids Res.* **2016**, *44* (D1), D184–189.

104) Jühling, F.; Mörl, M.; Hartmann, R. K.; Sprinzl, M.; Stadler, P. F.; Pütz, J. tRNAdb 2009: Compilation of tRNA Sequences and tRNA Genes. *Nucleic Acids Res.* **2009**, *37* (Database issue), D159–162.

105) Bolger, A. M.; Lohse, M.; Usadel, B. Trimmomatic: A Flexible Trimmer for Illumina Sequence Data. *Bioinformatics* **2014**, *30* (15), 2114–20.

106) Dobin, A.; Davis, C. A.; Schlesinger, F.; Drenkow, J.; Zaleski, C.; Jha, S.; Batut, P.; Chaisson, M.; Gingeras, T. R. Star: Ultrafast Universal RNA-Seq Aligner. *Bioinformatics* **2012**, *29* (1), 15–21.

107) McKenna, A.; Hanna, M.; Banks, E.; Sivachenko, A.; Cibulskis, K.; Kernytsky, A.; Garimella, K.; Altshuler, D.; Gabriel, S.; Daly, M.; DePristo, M. A. The Genome Analysis Toolkit: A Mapreduce Framework for Analyzing Next-Generation DNA Sequencing Data. *Genome Res.* **2010**, *20* (9), 1297–1303.

108) Danecek, P.; Schiffels, S.; Durbin, R. Multiallelic calling model in bcftools (-m). https://samtools.github.io/bcftools/call-m.pdf.

109) Quinlan, A. R.; Hall, I. M. BEDTools: A Flexible Suite of Utilities for Comparing Genomic Features. *Bioinformatics* **2010**, *26* (6), 841–842.