THE UNIVERSITY OF CHICAGO

EXPANDING HIGH-THROUGHPUT SEQUENCING TO INVESTIGATE RNA BIOLOGY

A DISSERTATION SUBMITTED TO

THE FACULTY OF THE DIVISION OF THE BIOLOGICAL SCIENCES

AND THE PRITZKER SCHOOL OF MEDICINE

IN CANDIDACY FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

GRADUATE PROGRAM IN BIOCHEMISTRY AND MOLECULAR BIOPHYSICS

BY

MOLLY ELIZABETH EVANS

CHICAGO, ILLINOIS

AUGUST 2017

TABLE OF CONTENTS

LIST OF FIGURES

CHAPTER 1

CHAPTER 2

CHAPTER 3

LIST OF TABLES

CHAPTER 2

CHAPTER 3

CHAPTER 4

ACKNOWLEDGMENTS

I would be remiss not to thank all the people that have helped me complete this thesis. Although you may feel alone sometimes [or often] during graduate school, as they say, it certainly takes a village to complete a thesis. Or at least I think that's how the saying goes.

First, I need to thank all those that helped me perform experiments. Guanqun Zheng patiently taught me DM-tRNA-seq and passed on charged tRNA-seq to me. Żaneta Matuszek taught me everything I know about northern blots. Qing Dai performed MALDI experiments contained in Chapter 3 and also synthesized many of the oligos used in Chapter 4, with novel synthesis for m^3C and $m^2{}_2G$ phosphoramidites. Finally, and most importantly, this thesis would not have been possible without Wes Clark, who performed all sequencing, data analysis, and machine learning contained herein.

Second, I need to thank all the members of the Pan Clan who have participated in scientific [and non-scientific] discussions with me. You made the years I spent here tolerable, at times even enjoyable, and I am grateful for your scientific knowledge, life knowledge, and welcome distractions from science.

To all of my friends at UChicago, thanks for the advice and support, both scientifically and generally. You've made it fun, and we'll figure out what we want to do with our lives eventually. Whatever it is, I'm sure we'll be doing great things, science or not. Well, probably.

Finally, I need to thank my friends and family. The support I received from everyone over the past 5 years was overwhelming, and you are the best neighbors in my village. I don't say it often enough, but I feel so lucky to have you all in my life, and I will always be grateful for your blind belief in me as I climbed the graduate-school-mountain. I only hope that I can one day support you all as much as you have supported me.

ABSTRACT

High-throughput RNA sequencing was developed almost a decade ago. Used originally to map the transcriptome to single-nucleotide resolution, advancements have expanded this powerful method to investigate many different aspects of cellular RNA biology. Along with changes in transcriptome in response to a multitude of different conditions, the DNA-RNA, RNA-RNA, and RNA-protein interactions have been defined. Additionally, RNA high throughput sequencing has been used to study differential splicing, structural changes in cellular RNA, induction of transcription, RNA editing, rates of translation, and many other biological parameters.

Here, three additional methods are presented that expand the functionalities of RNA high-throughput sequencing. First, we developed a method to determine tRNA aminoacylation levels via high-throughput sequencing. Previously, chemistry was used to distinguish between charged and uncharged tRNA using microarray methods. We harness the previously established chemistry to develop a one-pot sequencing method to determine the tRNA aminoacylation levels in mammalian cells. Next, we report a method to detect modifications in tRNAs using deep sequencing. By defining the combination of mutation rate and stop rate at each site as the 'modification index', we identify likely sites of modification. Then, based on the effect of treatment with a demethylase and the contributions of the stop and mutation rate to the modification index, the identity of the modification can be discerned. Finally, we work towards developing a method to predict sites of methylation in tRNA and mRNA by high-throughput sequencing and machine learning. Using model oligonucleotides to explore the effect of sequence context on the modification index of a given RNA methylation, we define the 'modification signature' for six different modifications. This modification signature can be used

to identify the position, identity, and abundance of modifications in tRNA with high accuracy, with the possible expansion into mRNA modifications. These three methods add to the already expansive list of methods to interrogate RNA biology and can be used to further our knowledge of the importance of RNA modifications in cellular biology.

CHAPTER 1:

EXPANDING METHODS IN RNA HIGH-THROUGHPUT SEQUENCING

The development of RNA high-throughput sequencing (RNA-seq) almost a decade ago has led to appreciable advancements in RNA biology. From simply mapping the transcriptome of a cell-type or single celled organism to defining the life span of a mRNA and its interacting DNA, RNA, and protein, the methods to interrogate RNA biology by RNA-seq have steadily advanced to query the roles, regulation, and structure of RNA *in vitro* and *in vivo*.

1.1 Methods to define the transcriptome prior to RNA high-throughput sequencing

While all somatic cells contain identical genetic information encoded in the DNA, the RNA transcribed from DNA varies from cell to cell. Known as the transcriptome, these RNAs dictate the function of the cell. Prior to the development of RNA high-throughput sequencing, methods to interrogate a cell's transcriptome primarily relied on hybridization-based approaches. In order to determine RNA expression, DNA microarrays use hybridization of labeled RNA to DNA probes that are covalently attached to glass slides [1,2]. This technology allows for investigation of a large number of expressed RNAs in parallel. Advances in design and technology over the next decade allowed for more precise mapping of transcripts and the determination of alternative splice sites even in more complex organisms [3-7]. However, due to the reliance of his method on hybridization, the sequence of the originating genome must be known in order to design the microarray. Additionally, microarrays suffer from high background due to cross-hybridization of the target with probes, and microarrays cannot detect low

abundance transcripts due to the weak signal. Also, due to saturation of the signal, a large range of transcript abundance can also not be detected.

Sequence-based methods are used to directly define the transcriptome by sequencing complementary DNA (cDNA) produced from the transcriptome. A reverse transcriptase (RT) is used to convert the labile RNA into a more stable, more easily manipulated cDNA through reverse transcription. Prior to the development of high-throughput sequencing, many different methods were developed to sequence cDNAs by Sanger sequencing. The expressed sequence tag (EST) method was used to sequence cDNAs cloned into DNA vectors [8]. Serial Analysis of Gene Expression (SAGE) also relies on cloning cDNA sequences into bacteria [9]. In this method, small fragments of cDNA are cloned into the vector as concatemers with a small tag to identify the start of each cDNA, thus increasing the throughput of the reaction. However, this method is biased to 3' end of mRNA due to the poly(T) primer used to reverse transcribe the cDNA. Similar to SAGE, Cap Analysis Gene Expression (CAGE) was developed to map the 5' end of transcripts [10]. However, these methods were limited due to the high cost and relatively low throughput of Sanger sequencing.

1.2 RNA high-throughput sequencing

RNA high-throughput sequencing has revolutionized the study of RNA biology. Although different platforms for high-throughput sequencing have been developed, we will focus on Illumina technology (Figure 1.1) [11]. First, the library must be prepared with adapters on the 5' and 3' end (library prep for RNA-seq is discussed in 1.2.1-1.2.6). Then, the library is loaded into a flow-cell containg a lawn of fixed surface-bound primers that are complementary to the library adapters. Bound fragments are then amplified by bridge amplification to create a cluster

of genetically identical fragments. Next, sequencing of the RNA is performed by synthesis. Each nucleotide has a different fluorescent label, and every nucleotide serves as a terminator; thus, only one nucleotide will be incorporated at a time, and the fluorescent label identifies the nucleotide that was added at each cluster. After incorporation and imaging, the terminator is reversed and the cycle is repeated. After sequencing, the reads are processed, the adapters are removed *in silico*, and the sequences are mapped to the genome.

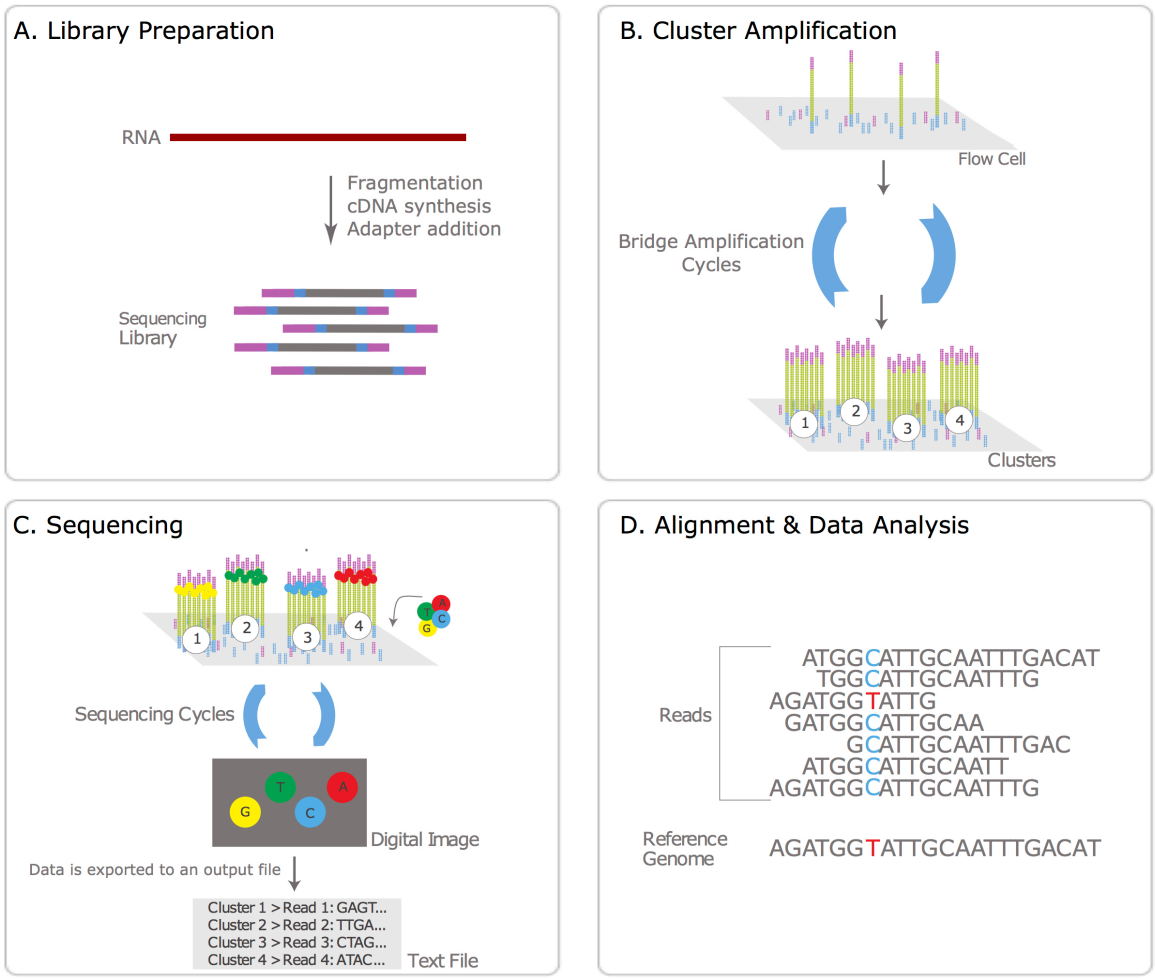


Figure 1.1: RNA high-throughput sequencing overview (adapted from 11). A) Adapter labeled cDNA is prepared from an RNA sample (details in Figure 1.2). B) The library is loaded into a flow cell, anneals to primers attached to the flow cell surface, and is PCR amplified to create clonal clusters. C) Sequencing reagents are added to the flow cell, and the first base is incorporated. Since each base has a different fluorescent label, the digital image can identify the first base of the sequence. After the image is recorded, the terminator is removed and the cycle is repeated. D) Reads are aligned to the reference genome. For RNA-seq experiments, this allows for identification of not only expressed mRNAs but also splice sites and RNA editing events.

RNA high-throughput sequencing was originally used to map the transcriptome of complex organisms at a much lower cost than both Sanger sequencing and DNA microarrays [12-18]. Additionally, this method allows mapping at single-nucleotide resolution and is quantitative. RNA-seq covers of most of the transcriptome to allow for the definition of alternative splice forms, transcription start sites, and RNA editing in a high-throughput manner.

RNA high-throughput sequencing requires sequencing cDNA prepared from an RNA sample of interest. While many variations of methods exist for preparation of cDNA libraries for high-throughput sequencing (Figure 1.2), the principles are common between all methods [19]. Depending on the experiment and the information desired, different combinations of the below methods can be deployed to avoid bias when needed.

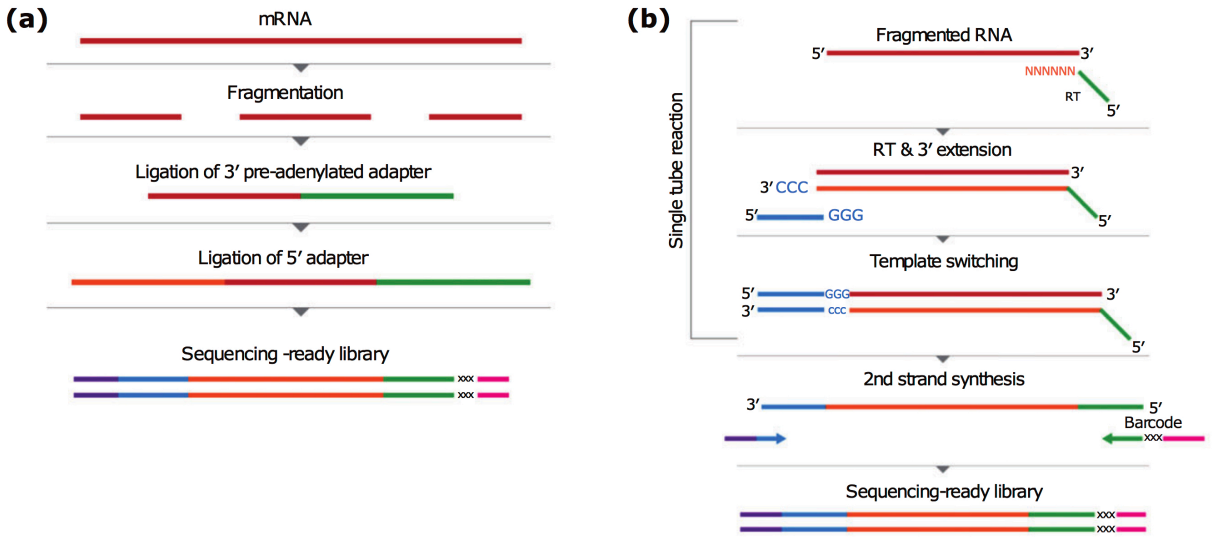


Figure 1.2: Examples of cDNA library prep for RNA high-throughput sequencing (adapted from 19). A) In this example, RNA is first fragmented. Sequential ligation of a 3' and 5' adapter is used to add the primer sequences required for Illumina sequencing. Next, these adapters are used for first and second strand cDNA synthesis and PCR amplification to create the sequencing library. B). In this method, RNA fragments are randomly primed. The random primer also contains the 3' adapter. During first strand cDNA synthesis, untemplated C nucleotides are added at the 3' end of the first strand cDNA. This overhang serves a primer binding site for second strand synthesis. Then PCR amplification is used to add a barcode and create the sequencing library.

1.2.1 RNA isolation or selective amplification of an RNA of interest

First, the RNA of interest must either be selected from total RNA or specifically targeted for cDNA synthesis. For example, if only poly(A) transcripts are the target, poly(A) RNA can be selected using hybridization to biotinylated oligo-dT. Alternatively, oligo-dT primers can be used to initiate cDNA synthesis. In each method, poly(A) mRNAs are selectively converted into cDNA.

1.2.2 Fragmentation

Second, since high-throughput sequencing cannot sequence full-length cDNAs, a fragmentation step must also be included. Typically, RNA is fragmented using chemical fragmentation by divalent cations under basic conditions, although enzymes can also be used. cDNA can also be fragmented by enzymes or sonication, but this biases reads towards the 3' end of RNAs [20].

1.2.3 First strand cDNA synthesis

Third, the RNA is used as a template for first strand cDNA synthesis. Most commercially available reverse transcriptases (RTs) require a primer to initiate synthesis. If a specific transcript is being investigated and the sequence is known, primers complementary to the RNA of interest can be added. However, to investigate the complete transcriptome, three different methods exist to prime cDNA synthesis. First, oligo-dT primers can be used. These primers anneal to the poly(A) tail of mRNAs, and RTs will extend into the full body of the mRNA. However, this biases reads to the 3' end of the RNA and can be complicated by hybridization to A-rich regions in the mRNA outside of the poly(A) tail. Next, random primers can be used. A library of random

hexamers is typically used to prime cDNA synthesis across the genome. However, even 'random' priming introduces bias into the cDNAs amplified [21,22]. Finally, an RNA adapter can be ligated directly to the 3' end of RNA so that the sequence of the 3' end of every RNA is known. A DNA primer can then be added to initiate second strand cDNA synthesis. However, this method has a strong sequence bias due to the biases inherent in the ligation enzymes. Alternatively, a novel reverse transcriptase has recently been identified that can perform a template-switching reaction [23-25]. This requires only an RNA-DNA hybrid with a single nucleotide 3' DNA overhang of any nucleotide (N). The 3' DNA overhang will anneal to the 3' end of the RNA template, and the RT is able to switch from the RNA-DNA hybrid to the RNA template to begin reverse transcription. This eliminates the need for priming of the RNA template before cDNA synthesis, thus reducing bias.

1.3.4 Second strand cDNA synthesis

Fourth, second strand cDNA synthesis must be completed before library preparation. As with first strand synthesis, RTs require a binding site that can be provided by random priming, ligation to the 5' end of the template RNA, or ligation to the 3' end of the first strand cDNA. In addition to random priming and ligation, tailing of the first strand can also be used. This allows for the addition of a known sequence to the end of the first strand, to which a primer can be bound to initiate second strand cDNA synthesis.

1.3.5 Addition of 5' and 3' adapters

Fifth, 5' and 3' adapters must be added in order for the cDNA to be sequenced on an Illumina platform. These adapters serve not only as primer binding sites for PCR amplification

of the library prior to sequencing but also as primer binding sites for bridge amplification and sequence determination during high-throughput sequencing. Adaptor addition can be done by direct ligation to the RNA, direct ligation to cDNA, or incorporation into cDNA primers. Direct ligation, as mentioned previously, has sequence biases inherent to the enzymes. Alternatively, if random cDNA primers are used, a 5' extension of the cDNA primer that incorporates the high-throughput sequencing primer sequences can be utilized. This avoids the need for ligation reactions. If the template-switching RT is used, the RNA-DNA hybrid can be designed to contain both 3' and 5' adapters. Once the RNA has been reverse transcribed, the cDNA can be circularized which allows for PCR priming in each direction around the circle to create linear cDNA with adapters on both the 3' and 5' ends.

1.2.6 Addition of a sequencing barcode

If multiple samples are to be sequenced in the same lane of a flow-cell, a barcode can be added. This allows for sequencing reads to be assigned to the originating sample. Typically, barcodes are 6-8 nucleotides and are added after second strand cDNA synthesis. During PCR amplification, primers with barcodes 5' of the adapter binding site are added, which allows for incorporation of the barcode into the final sequencing library.

1.3 Expansion of RNA high-throughput sequencing to query RNA biology

RNA high-throughput sequencing provides a powerful tool to define the transcriptome in the most complex organisms. The transcriptome of diverse cell types and organisms have been mapped, and this has led to a better understanding of how exactly the transcriptome allows cells with identical genetic information to carry out their unique functions. Additionally, the

transcriptome of cells under various stress conditions have been mapped, defining the transcriptional response to cellular perturbations and stimulations. However, high-throughput sequencing provides an opportunity to query much more than simply the transcriptome of a cell.

This powerful method has been expanded to query many more aspects of RNA biology. While there are far too many methods to be covered in detail here, a brief synopsis of the state of the field is presented as well as one method recently developed in our lab.

1.3.1 A brief synopsis of RNA high-throughput sequencing methods

In addition to defining the transcriptome, RNA high-throughput sequencing methods have been developed to follow the life span of an mRNA. By marking nascent transcripts (e.g. GRO-seq [26]), the transcription rate of each RNA in the transcriptome can be measured. Additionally, ribosome profiling has been used extensively to annotate translated sequences and quantify the amount of translation as well as the efficiency of translation of the transcriptome [27]. Methods to annotate binding partners of RNA have also been developed. By affinity-purifying an RNA binding protein after cross-linking, the RNAs bound to the protein can be annotated (e.g. PAR-CLIP [28]). RNA-RNA interactions have been detected by cross-linking of RNA duplexes followed ligation (LIGR-seq [29]). Similarly, DNA-RNA interactions can also be profiled by cross-linking and ligation (MARGI [30]). Additionally, sequencing of a single cell's transcriptome has also been developed, allowing for the analysis of differences between the transcriptome of single cells in the same tissue [31]. A method to investigate both *in vitro* and *in vivo* structural determination has also been developed [32]. While this list is nowhere near exhaustive, it is clear that RNA high-throughput sequencing has been expanded to query many different RNA parameters.

1.3.2 tRNA sequencing

Until recently, tRNA has remained resistant to sequencing due to the stable secondary structure and the high number of modified nucleotides. Recently, our lab developed a method to sequence tRNAs using demethylase treatment and a thermostable RT to overcome those two obstacles [33]. A similar method was developed in the Lowe lab that used demethylase treatment with traditional RT [34]. These methods allowed, for the first time, the definition of all tRNAs expressed in mammalian cells as well as the relative levels of each tRNA species.

1.4 Further expansion of RNA high-throughput sequencing contained herein

This thesis provides an explanation of three expansions of RNA high-throughput sequencing to determine 1) tRNA aminoacylation levels in mammalian cells, 2) modification identification and quantification in tRNA, and 3) modification identification and quantification in tRNA and mRNA by machine learning. While methods exist to quantify these biological parameters, most of the methods suffer from low-throughput or low-resolution.

1.4.1 Previous methods to determine tRNA aminoacylation

tRNA aminoacylation is an important biological parameter as the rate of translation is affect by both the amount and charged fraction of tRNA. tRNA aminoacylation levels change in response to stress, and cells sense and respond to these levels. Thus, it is important to be able to accurately measure the levels of tRNA aminoacylation.

Previously, northern blots and tRNA microarrays were employed to determine tRNA aminoacylation levels. First, northern blots employ the different mobility of charged versus uncharged tRNAs in polyacrylamide gel electrophoresis (PAGE) [35-37]. Total RNA is

separated by acidic denaturing PAGE. After transfer of the RNA to a membrane, the tRNA of interest is probed by hybridization of a radiolabeled complementary DNA probe. In this way, the charged fraction can be determined by quantifying the amount of charged (slower mobility) versus uncharged (faster mobility) tRNA. This method can only investigate a single tRNA species at a time, and not all tRNAs are efficiently separated by PAGE. Additionally, tRNAs that have very similar sequences cannot be investigated as the DNA probes can cross-hybridize.

Second, tRNA microarrays have been used to investigate tRNA aminoacylation levels [38-40]. First, uncharged tRNAs are chemically inactivated by periodate oxidation. While uncharged tRNAs have both a 2' and 3' hydroxyl that can be oxidized, charged tRNAs are protected by the amino acid covalently attached to the 3' end and thus cannot be oxidized. The tRNAs are then deacylated and labeled by ligation to fluorescent probes. Oxidation of the 3' end prevents this ligation, so in periodate treated samples, only charged tRNAs will be labeled. By comparing the periodate treated sample to buffer treated sample where all tRNAs are labeled, the charged fraction of all tRNAs can be determined by annealing to a tRNA microarray. However, as this is a hybridization-based method, tRNAs that are very similar cannot be distinguished. While this is not problematic for organisms with low tRNA sequence diversity such as bacteria and yeast, the tRNAs of organisms with high tRNA diversity such as mammals cannot be resolved.

Thus, a sequencing method is required to determine the tRNA aminoacylation fraction of organisms with high sequence diversity in a high-throughput manner. Chapter 2 describes the development of charged tRNA-seq to determine the tRNA aminoacylation levels in mammalian cells. This method shows that although most tRNA isodecoders are highly charged (>80%), $tRNA^{Ser}$ and $tRNA^{Thr}$ isodecoders have generally lower charging levels (60%-80%).

1.4.2 Previous methods to map and quantify tRNA modifications

tRNAs are the most highly modified RNA with almost 1 in 5 nucleotides modified, and their diverse modifications have been studied for decades. These modifications are distributed throughout the body of the tRNA and range from simple methylations to complex adducts that require multiple enzymes for the biosynthesis (Figure 1.3). Modifications have been shown to affect both the structure and function of tRNA (recently reviewed in [41]). Modifications in the anticodon loop affect translational efficiency and fidelity while modifications outside of the anticodon loop affect tRNA structure and quality control.

Despite the importance of tRNA modifications, the systematic mapping of modified residues in individual tRNAs remains challenging. The amount of a given modification can be determined by isolating tRNA, digesting into single nucleosides, and identification and quantification by LC-MS/MS [42]. However, this method does not give any positional information and simply reports on the abundance of a given modification in tRNA. Alternatively, individual species of tRNAs can be isolated by chromatography. Then, positional information about the modification status can be determine by partial digestion and thin layer chromatography (TLC) [43] or mass spectrometry (MS) [44]. Modifications impart different solvent mobility to a nucleoside allowing for separation by TLC. Alternatively, since all modifications (except Ψ) add mass to the RNA molecule, unmodified and modified fragments can be identified by the mass difference as measured by MS. Additionally, Ψ can be tagged with specific chemical treatment to identify the site of modification by MS. However, isolation of a large amount of sufficiently pure sample makes this method impractical to map tRNA modifications in all tRNA species.

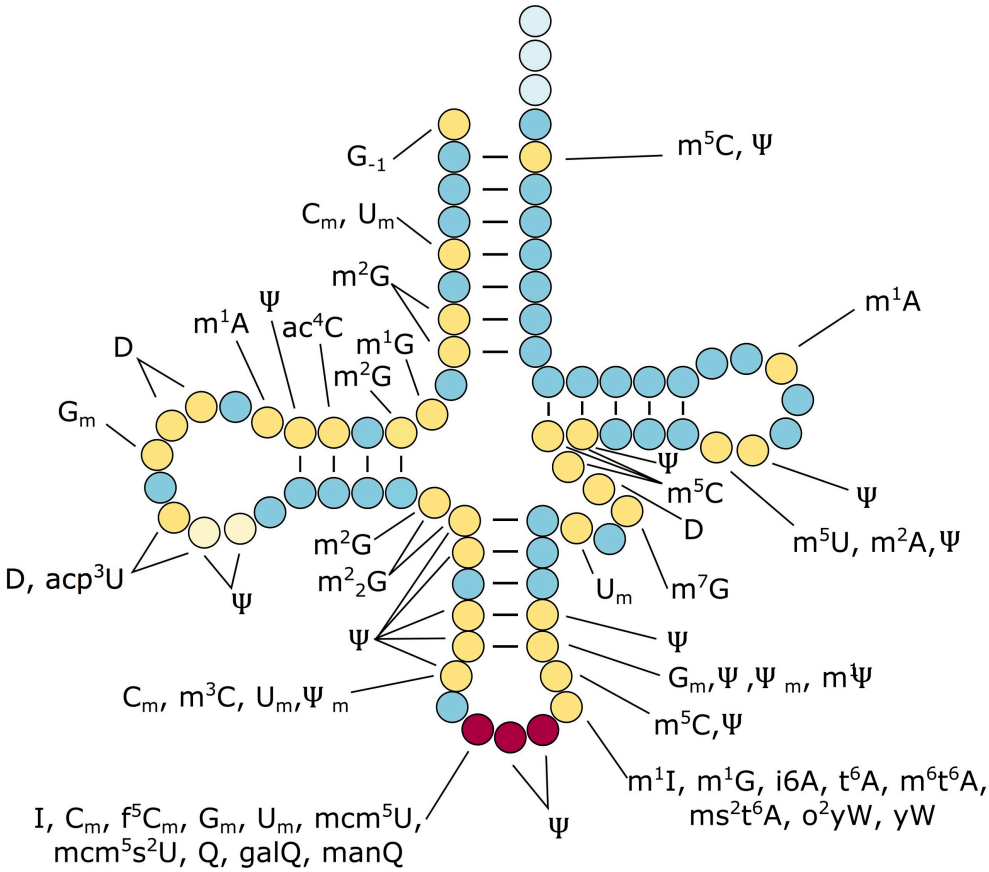


Figure 1.3: Landscape of mammalian tRNA modifications (from 41). All annotated mammalian tRNA modifications were pulled from Modomics (http://modomics.genesilico.pl/). Blue residues have no modification annotated while yellow residues are modified in at least one mammalian cytosolic tRNA (*Bos taurus*, *Homo sapiens*, *Mus musculus*, *Oryctolagus cuniculus*, *Ovis aries*, *Rattus norvegicus*). The anticodon, a hot spot of modification, is highlighted in red. Modifications that occur in the variable loop were left out for simplicity. m^2G - N2-methylguanosine; $m^2{}_2G$ - N2,N2-dimethylguanosine; m^1G - 1-methylguanosine; ac^4C - N4-acetylcytidine; m^1A - 1-methyladenosine; D - dihydrouridine; acp^3U - 3-(3-amino-3-carboxypropyl)uridine; m^3C - 3-methylcytidine; I - inosine; m^1I - 1-methylinosine; mcm^5U - 5-methoxycarbonylmethyluridine; mcm^5s^2U - 5-methoxycarbonylmethyl-2-thiouridine; Q - queuosine; galQ - galactosyl-queuosine; manQ - mannosyl-queuosine; f^5C_m - 5-formyl-2'-O-methylcytidine; t^6A- N6-threonylcarbamoyladenosine; ms^2t^6A - 2-methylthio-N6-threonylcarbamoyladenosine; m^6t^6A - N6-methyl-N6-threonylcarbamoyladenosine; i^6A - N6-isopentenyladenosine; o^2yW - peroxywybutosine; yW - wybutosine; $m^1\Psi$ - 1-methylpseudouridine; Ψ_m - 2'-O-methylpseudouridine; m^7G - 7-methylguanosine; m^5C - 5-methylcytidine; m^2A - 2-methyladenosine; m^5U - 5-methyluridine.

Reverse transcription can detect certain modifications by a specific RT signature. Modifications such as m^1A and m^1G cause RT stops while other modifications can be

specifically targeted with chemicals to form adducts that cause RT stop [45]. However, until recently, only single modifications in tRNA have been targeted.

Thus, a high-throughput sequencing analysis of cDNAs produced from tRNAs is needed to determine the location of modifications across all tRNAs. Chapter 3 presents a method to use the RT signatures of modifications in tRNA to determine the position of many modifications in tRNA. Additionally, this method can be used to determine the fraction of modified nucleotide at each site in a semiquantitative manner. This method shows, contrary to previous conjecture, many sites in tRNAs are only fractionally modified.

1.4.3 Previous methods to map/quantify mRNA modifications

mRNA modifications have been studied for many decades, but the extent and diversity of RNA modification is only beginning to be appreciated. Compared to tRNAs, only a few modifications have been identified in eukaryotic mRNA (Figure 1.4). However, some of these modifications map uniquely to particular regions of the transcript. Additionally, these modifications have been shown to affect the lifetime, translation, and localization of mRNA (recently reviewed in [41]).

Similar to tRNA modifications, in order to identify the types of modifications present in mRNA, digestion of isolated mRNA followed by LC-MS/MS has been used. Again, this does not provide sequence information. Additionally, as mRNA modifications are less abundant and many sites are only partially modified, the detection must be highly sensitive in order to identify rare modifications.

Due to the greater sequence diversity and length of mRNA compared to tRNA, partial digestion of a purified transcript followed by MS, although possible, is quite difficult to

implement. Identification and mapping of a sequence is much more difficult given the greater number of possible fragments. While a specific part of the mRNA can be protected from RNase digestion to make identification of fragments simpler, this approach is very low throughput.

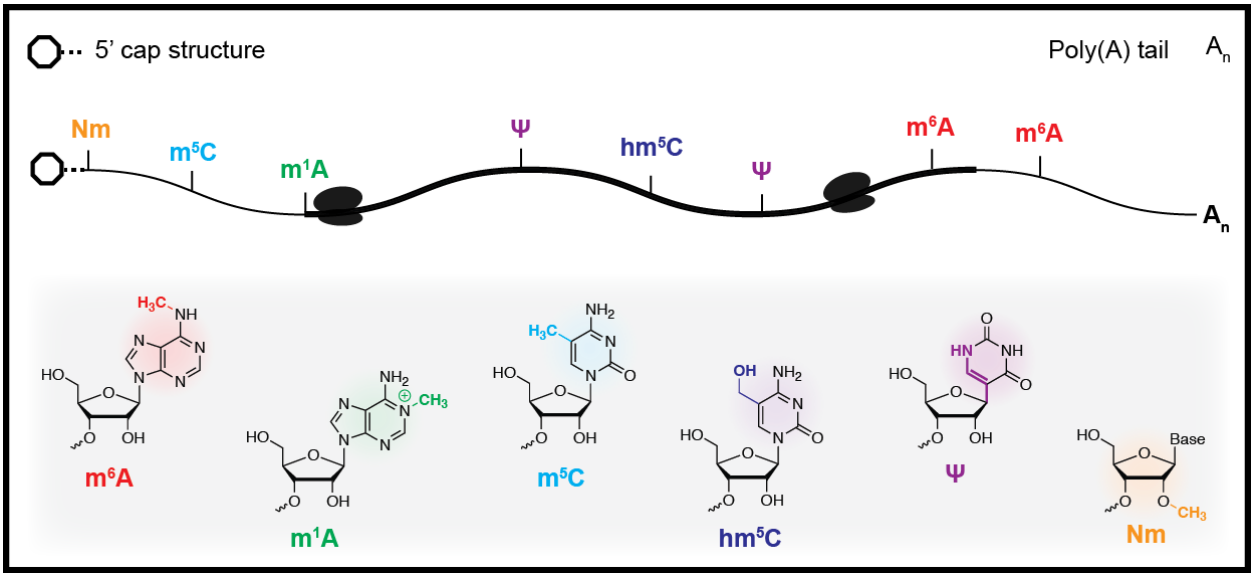


Figure 1.4: Chemical modifications in eukaryotic mRNA (from 41). A schematic representation of common chemical modifications in eukaryotic mRNA transcripts. Several of these modifications map uniquely to the mRNA cap structure, 5' or 3' untranslated regions, or the coding region (bold) of the transcript.

Since the advent of RNA high-throughput sequencing, a few methods have been developed to interrogate the position of modification in mRNA. First, chemical methods have been employed that specifically mark a specific RNA modification at its site. By targeting the specific chemical properties of a modification (e.g. Ψ reactivity with CMCT, N_m induction of RT stops), the modification can be mapped to the genome by specific signatures in the sequencing library. Affinity-based sequencing has also been used. Antibodies that specifically recognize modifications or enzymes that install modifications can be used to enrich for RNA fragments that contain modifications. After sequencing the enriched library and comparing to the total RNA

sample, positions of modifications can be identified. However, in both of these methods, not all modifications can be specifically targeted and only one modification can be interrogated at a time. Additionally, sequence context may affect the reactivity or ability of a protein to bind to the modification, so these methods are also biased.

Methods that rely on the RT-signature of modifications have also been developed. Many modifications are known to cause stops or mutations during cDNA synthesis. Searching the sequencing for an RT signature for a particular modification can identify sites of modification [46-48]. These methods can also be semiquantitative to give some information about the fraction of modification at any position. However, these methods are imperfect at identifying sites of modification even in tRNA and are not robust enough to be used in mRNA.

Here, we aim to improve the predictive power of the methods that rely on RT signature to ascertain the identity and fraction of modification at sites in mRNA. While recent experiments have shown the influence of the +1 nucleotide on the RT signature, Chapter 4 investigates the effects of the +1, +2, -1, and -2 nucleotides. Our method investigates 6 base methylations that affect Watson-Crick base pairing, and shows that the +1, +2, and -1 nucleotides influence the RT signatures. We are correctly able to identify and distinguish between m^1G and $m^2{}_2G$ sites in tRNA, and are optimistic that the modification signature can be incorporated into existing methods to better predict the identity, location, and fraction of modification in mRNA.

1.5 Original methods described in this thesis

Here are presented three methods that extend the functionalities of RNA high-throughput sequencing. These advances will allow for investigation of biological parameters that have previously used low-throughput, low-resolution, or limited approaches. Thus, we are confident

this work will contribute to future research into defining the importance of tRNA aminoacylation, tRNA modification, and mRNA modification in RNA biology.

CHAPTER 2:

DETERMINATION OF TRNA AMINOACYLATION LEVELS BY HIGH-THROUGHPUT SEQUENCING

Here we present a method to determine tRNA aminoacylation levels via high-throughput sequencing. Previously, chemistry was used to distinguish between charged and uncharged tRNA using microarray-based methods. We harness the previously established chemistry to develop a one-pot sequencing method to determine the tRNA aminoacylation levels in mammalian cells.

2.1 Introduction

tRNAs are used by the ribosome to translate mRNA into proteins, and the fraction of tRNAs that are aminoacylated (or charged) is an important biological parameter. Both the amount of tRNA and the fraction of charged tRNA affect the speed and efficiency of translation. In *E. coli*, charging levels of different tRNA isoacceptors change upon amino acid starvation [38] which is useful in the translational regulation of stress response proteins in a codon-dependent manner [49]. tRNA charging levels also vary during *E. coli* growth where they can act as sensors of cellular metabolism [50,51]. Uncharged tRNAs can also trigger the stringent response in bacteria that enables high level synthesis of the alarmone ppGpp [52]. In eukaryotes, uncharged tRNA is a well-known activator of the protein kinase GCN2 which phosphorylates eIF2α to regulate global translation activity in response to stress [53]. In human cells, tRNA charging levels are affected by proteasome inhibition, which may reflect changes in cellular metabolism when amino acid recycling is ineffective [39].

Two methods have been used in previous attempts to measure tRNA charging levels. Acid-denaturing polyacrylamide gels can, in most cases, separate charged tRNA from uncharged tRNA. After separation, a northern blot is used to determine the charged fraction of a given tRNA species. This method is sensitive but low-throughput: only one tRNA species can be queried at a time (e.g. [35-37]).

tRNA microarrays can also be used to determine charging fraction. Periodate oxidation of the 3' end is specific to uncharged tRNAs, effectively inactivating the 3' end for any further enzymatic reaction. After deacylation and subsequent ligation to fluorescent probes, the level of tRNA charging can be ascertained. By comparing untreated sample (total tRNA) to periodate-treated sample (charged tRNA), the charged fraction of tRNA can be determined [38-40]. While this method can examine the charging of all tRNAs in parallel, the resolution is quite low: an approximately eight nucleotide difference is required between the DNA probe and tRNA sequence to prevent cross-hybridization. While this specification is not problematic for organisms with low tRNA sequence diversity such as bacteria and yeast, multi-cellular organisms such as mammals have tRNA isodecoders (same anticodon, different body sequence) with only one or two divergent nucleotides which cannot be resolved on an array [54,55].

Until recently, tRNAs have been resistant to efficient and quantitative high-throughput sequencing because of the large number of modifications and rigid structure. Our lab recently developed DM-tRNA-seq, a method to quantitatively and efficiently sequence tRNA [33]. This method employs wild-type and mutant *E. coli* AlkB demethylases to remove common Watson-Crick face modifications (m^1G, m^1A, m^3C) in tRNA that commonly cause reverse transcriptase (RT) stops. Additionally, a thermostable reverse transcriptase (TGIRT) is employed to overcome the stable secondary structure. These two features allowed for quantitative high-throughput

sequencing of human tRNA from mammalian cells [33]. A similar, AlkB-based method was also developed to profile tRNA fragments but does not employ the thermostable TGIRT [34].

Here, we report an extension of the DM-tRNA-seq method to determine the charging fraction of tRNA isodecoders in mammalian cells (Figure 1). Using periodate oxidation followed by treatment at high pH, the 3' nucleotide of uncharged tRNA is removed through β-elimination while the 3' nucleotide of charged tRNA is not affected. Thus, tRNAs that were uncharged will end in 3'-CC while tRNAs that were charged will end in 3'-CCA, and a one-pot sequencing reaction can be used to precisely quantify tRNA charging levels of mammalian isodecoders at single-base resolution.

2.2 Materials and Methods

2.2.1 Cell culture and RNA isolation

Human embryonic kidney HEK293T (CRL-11268) cells from ATCC were grown in DMEM supplemented with 10% FBS, 1% 100× penicillin-streptomycin and passaged 7 times. RNA was extracted once the cells reached 70-90% confluency at P7. RNA was isolated using TRIzol, and the pH was maintained at < 5 throughout the isolation to prevent hydrolysis of the aminoacyl bond. After ethanol precipitation, total RNA was resuspended in 10 mM NaOAc/HOAc pH 4.8, 1 mM EDTA.

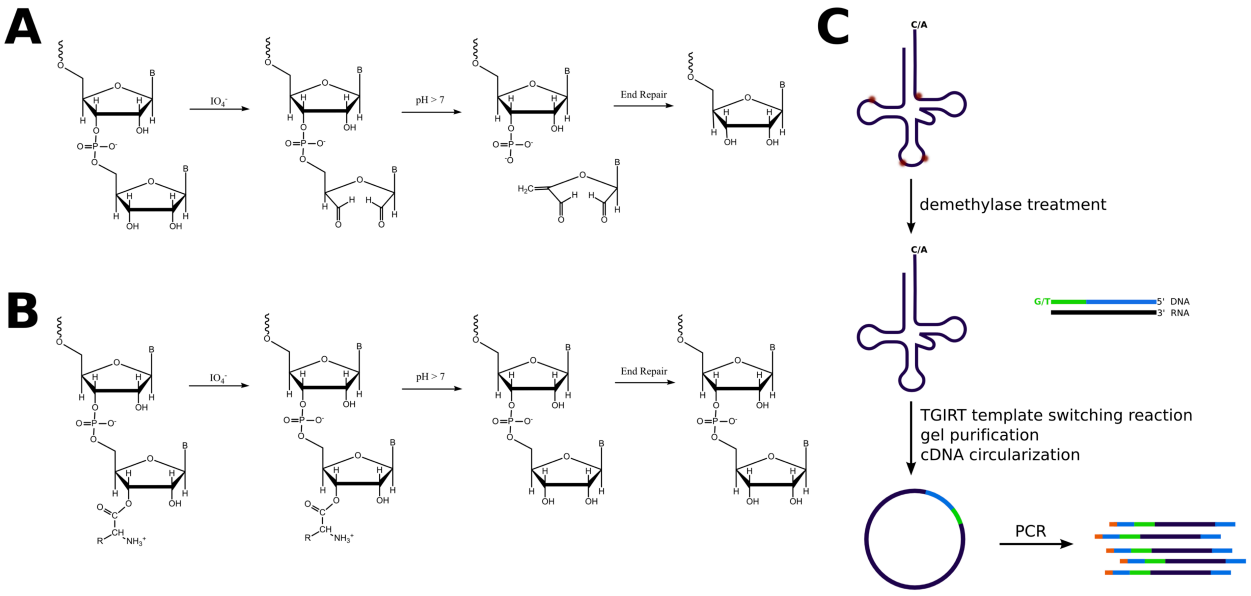


Figure 2.1: Charged DM-tRNA-seq method. A,B) Periodate oxidation and β-elimination can differentiate between charged and uncharged tRNAs prior to sequencing. Periodate selectively oxidizes the 3' end of uncharged tRNA (A), while the 3' end of charged tRNA is protected (B). Treatment with high pH causes β-elimination of the oxidized nucleotide and deacylation of charged tRNA. Thus, after end repair with T4 PNK, charged tRNAs will end in –CCA while uncharged tRNAs will end in –CC. C) Modified DM-tRNA-seq to determine charging fractions. tRNA is first treated with demethylase to remove common tRNA modifications (m^1G, m^1A, m^3C) that impair reverse transcription. Then, a DNA/RNA hybrid with a 1 nucleotide DNA overhang (T/G) is added to act as a primer for the TGIRT template switching reaction. This primer contains binding sites for Illumina PCR/library prep (light blue). For this method, we extended the DNA/RNA hybrid (green) to prevent incorrect assignment of –C and –A ending tRNAs due to *in silico* trimming. After reverse transcription and cDNA purification and circularization, the cDNA is PCR amplified to create a barcoded library for Illumina sequencing.

2.2.2 CCA radiolabeling for tRNA standards

tRNA was radiolabeled at the final bridging phosphate with α-^{32}P-ATP and *E. coli* CCA adding enzyme as previously described [56]. Briefly, 40 pmol *E. coli* $tRNA^{Tyr}$ (Sigma-Aldrich) was renatured in 20 mM Tris-HCl, pH 7.0 by heating to 85°C for 2min followed by cooling at room temperature for 3min. $MgCl_2$ was added to 10 mM, and the tRNA was incubated at 37°C for 5min. Renatured tRNA was then treated with recombinant, purified *E. coli* CCA adding

enzyme (final concentration = 30 μg/mL) at 37°C for 10min in 50 mM glycine, pH, 9.0, 13 mM $MgCl_2$, 50 μM sodium pyrophosphate, 20 μM CTP, 0.33 μM α-^{32}P-ATP. Labeled tRNA was purified by denaturing PAGE (7M Urea, 1X TBE).

2.2.3 tRNA standard preparation

Yeast $tRNA^{Phe}$ and *E. coli* $tRNA^{Lys}$ were obtained from Sigma-Aldrich and deacylated before addition to total RNA. *E. coli* $tRNA^{Tyr}$ was *in vitro* charged with total *E. coli* synthetase mix (Sigma-Aldrich). Trace amounts of 3'[^{32}P]A-labeled $tRNA^{Tyr}$ was added to 100 pmol cold $tRNA^{Tyr}$ for charging. The tRNAs were renatured as described above, then treated with total aminoacyl-tRNA synthetase mix from *E. coli* (final concentration = 5.3U/μL) in 40 mM Tris-HCl, pH 7.5, 24 mM KCl, 6 mM DTT, 3 mM ATP, 0.2 mM Tyrosine at 37°C for 15min. NaOAc/HOAc, pH 4.8 was added to 300 mM, and the reaction was phenol/chloroform extracted before ethanol precipitation. Charged tRNA was resuspended in 50 mM NaOAc/HOAc, pH 4.8, 1 mM EDTA and stored at -80°C for up to one month. To determine charging levels of the $tRNA^{Tyr}$ standard, aminoacylated tRNA was digested with P1 nuclease (final concentration = 0.15 U/μL) in 150 mM NH_4OAc pH 4.8 at room temp 20min. The reaction was resolved on a PEI cellulose TLC plate in 5% acetic acid, 100 mM NH_4Cl. Radiolabeled [^{32}P]pA-Tyr and [^{32}P]pA nucleotides were visualized by autoradiography to determine charging level (Figure 2.2).

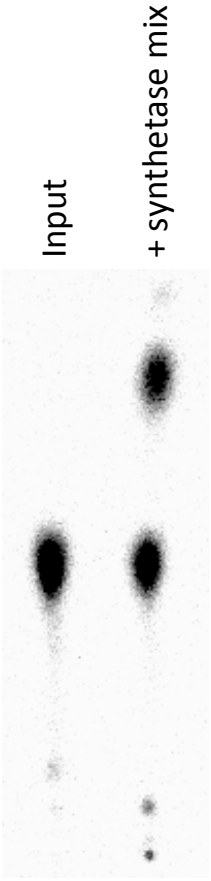


Figure 2.2: TLC quantification of *E. coli* tRNA^{Tyr} charging levels. The radiolabeled charged tRNA standard (tRNA^{Tyr}) was digested with P1 nuclease and resolved on a PEI cellulose TLC plate, and the amount of 3' [^{32}P]-A-Tyr (A*Tyr) was compared to the total A* nucleotide (A* + A*Tyr) to determine the charged fraction.

2.2.4 Periodate oxidation and β-elimination

Total RNA isolated above was mixed with prepared tRNA standards (final concentration = 0.03 pmol/μg total RNA). Periodate oxidation of total RNA (final concentration = 1 μg/μL) was carried out in 100 mM NaOAc/HOAc, pH 4.8, and freshly prepared 50 mM $NaIO_4$ at room temperature for 30 minutes. The reaction was quenched with 100 mM glucose for 5 minutes at room temperature. Any unquenched periodate was removed by Micro Bio-Spin P-6 Columns (Bio-Rad) and two ethanol precipitations. For β-elimination and deacylation, total RNA (final concentration = 1 μg/μL) was treated in 60 mM sodium borate, pH 9.5 at 45°C for 90 min. RNA was purified by Micro Bio-Spin P-6 Columns before ethanol precipitation.

To purify tRNA, total RNA was then run on an 8% denaturing polyacrylamide gel (7M Urea, 1X TBE), and tRNA was excised and eluted from the gel in 50 mM KOAc, 200 mM KCl. After ethanol precipitation, short RNA oligo standards were added in different proportions (final concentration = 0.4 pmol total / μg tRNA) to ensure that TGIRT extension from T- and G-ending primers are equivalent. The sequences of the oligo pairs are: first pair: 5' GUAAUUAUACUCAUAAAUUCGUUGUACGUGAUGCCUAAUUCCUCCA, 5' GUAAUUAUACUCAUAAAUUCGUUGUACGUGAUGCCUAAUUCCUCC; second pair: 5' GCGGACUAGGUCCUGUGUUCGAUCCACAGAGUUCGCACCA, 5' GCGGACUAGGUCCUGUGUUCGAUCCACAGAGUUCGCACC. In the first biological replicate, the oligos were added in a ratio of 75% 3'-CCA, 25% 3'-CC. In the second biological replicate, oligos were added in equal ratios (50% 3'-CCA, 50% 3'-CC). In the third biological replicate, oligos were added in a ratio of 25% 3'-CCA, 75% 3'-CC. The tRNA and model oligo mixture was used as input for DM-tRNA-seq.

2.2.5 DM-tRNA-seq

This was carried out as previously described [33] with minor modifications. First, tRNA was demethylated under previously optimized conditions. 60 pmol of total tRNA was treated with 120 pmol WT AlkB and 240 pmol D135S AlkB in 25 mM MES, pH 5.0, 300 mM KCl, 2 mM $MgCl_2$, 2 mM ascorbic acid, 300 μM α-ketoglutarate, 50 μM $(NH_4)_2Fe(SO_4)_2$ in the presence of RNasin at room temperature for 2 hours. A Zymo RNA Clean and Concentrator kit was used to remove the demethylases and clean up the reaction. Next, end repair (3' phosphate removal) was performed. 1 μM total tRNA was treated with T4 polynucleotide kinase (PNK,

Affymetrix) (final concentration = 0.2 U/μL) at 37°C for 30 min. A Zymo RNA Clean and Concentrator kit was again used to clean up the reaction.

For cDNA synthesis, TGIRT primer was 5' labeled with T4 PNK. 4 pmol of each 5' labeled TGIRT primer (T-ending: 5' GATCGTCGGACTGTAGAACTAGACGTGTGCTCTTCCGATCTTTCAGGCATTAGGCTC AAAGT, G-ending: 5' GATCGTCGGACTGTAGAACTAGACGTGTGCTCTTCCGATCTTTCAGGCATTAGGCTC AAAGG) was annealed to 8 pmol complementary RNA (5' CUUUGAGCCUAAUGCCUGAAAGAUCGGAAGAGCACACGUCUAGUUCUACAGUCCG ACGAUC/3SpC3/) in 100 mM Tris-HCl, pH 7.5, 0.5 mM EDTA at 82°C for 2min, then slow cooled to room temp. 4 pmol tRNA was then added. The tRNA/primer mixture (200 nM tRNA, 200 nM each primer) was pre-incubated at room temp for 30min in 100 mM Tris-HCl, pH 7.5, 450 mM NaCl, 5 mM $MgCl_2$, 5 mM DTT with 500 nM TGIRT (InGex, Inc.). dNTPs were added to a final concentration of 1 mM to initiate the reaction. The reverse transcription was performed at 60°C for 60min. The reactions were terminated with additions of NaOH to 0.25 M and incubation at 95°C for 3min. The reaction was neutralized with 0.25 M HCl. An equal volume of 50% Formamide, 4.5M Urea, 50mM EDTA, 0.05% Bromophenol blue, 0.05% xylene cyanol was added, and the mixture was heated at 95°C for 15 min.

cDNAs were then purified by denaturing 10% PAGE (7M Urea, 1X TBE), and extended products were cut and eluted from the gel overnight in 50 mM KOAc, 200 mM KCl. Purified cDNA was ethanol precipitated with addition of linear acrylamide (Thermo) to 20 μg/mL.

Purified cDNA was then circularized using CircLigase II (Epicentre) at 60°C overnight. After inactivation at 80°C for 10 min, samples were phenol/chloroform extracted and ethanol

precipitated. PCR library preparation for Illumina sequencing was performed using Phusion Master Mix (Thermo) for 12 PCR cycles (98°C 5s, 60°C 10s, 72°C 10s). AMPure XP Beads (Beckman-Coulter) were used to clean up the libraries before Illumina sequencing.

2.2.6 Sequencing analysis

All libraries were sequenced on an Illumina HiSeq 2000 with paired-end mapping using read lengths of 100 base pairs. Standard quality control via FastQC was performed after sequencing and also after read processing. Reads were processed using Trimmomatic v0.32 to remove the standard Illumina adapter sequence followed by subsequent trimming using custom Python scripts to remove demultiplexing artifacts, primers, and trim the extended adapter. This second trimming step ensures that reads are not over-trimmed by Trimmomatic to ensure fidelity of the 3' end of the raw reads. The resultant trimmed sequences were then aligned to the library using Bowtie 1.0 with sensitive options (--k 1 –v 3 –best –strata). Sequencing reads were aligned to a modified tRNA hg19 genome file, as in (13), containing nuclear-encoded tRNAs, mitochondrial-encoded tRNAs, and specific *E. coli* and yeast tRNAs used as standards.

tRNA isodecoder abundance was determined by raw mapped read count for each isodecoder sequence. Because we condensed same-scoring tRNAs from the genomic tRNA database [57] into one sequence, we used only mapped read count for each isodecoder without normalization to the number of genes with the same sequence. Each read was mapped to a single isodecoder based on sequence identity. If a short read could potentially be mapped to multiple isodecoders, it is thrown out due to the mapping ambiguity.

Because Bowtie 1.0 uses end-to-end alignment, the determination for charge ratio is as follows: an aligned read is considered A-ending/charged if it aligns with no mismatches to the

tRNA's 3' 15 nucleotides including the –CCA. Consequently, if the read aligns to the 3' nucleotides but only ends in –CC, it is considered C-ending/uncharged. Ratios were determined as individual fractional components over the sum of the A-ending and C-ending aligned reads.

2.2.7 Northern blot analysis

Northern probes were 5' radiolabeled with T4 PNK and gel purified (Trp: 5' TGACCCCGACGTGATTTGAACACGCAACCTTCTGATCTGGAGTCAGACGCGCTACCG TTGCGCCACGAGGTC, mt-Trp: 5' CAGAAATTAAGTATTGCAACTTACTGAGGGCTTTGAAGGCTCTTGGTCTGTATTTAA CCTAAATTTCT, SerGCT: 5'GACGAGGRTGGGATTCGAACCCACGYGTGCAGAGCACAATGGATTAGCAGTCCAT CGCCTTAACCACTCGGCCACCTCGT, ThrTGT: 5'AGGCCCCAGCGAGATTYGAACTCGCGACCCCTGGTTTACAAGACCAGTGCTCTAA CCMCTGAGCTATGGAGCC). 2.5 µg of Total RNA in 10 mM NaOAc/HOAc, pH 4.8 was mixed with an equal volume 8 M Urea, 0.1 M NaOAc/HOAc, pH 4.8, 0.05% Bromophenol Blue, 0.05% xylene cyanol. 2.5 µg of deacylated total RNA (treated in Tris-HCl, pH 9.0 at 37°C for 45min) was included as a control. Samples were run on a 6.5% PAGE sequencing gel (8M Urea, 0.1M NaOAc/HOAc pH=5.0) at 500V for 24 hours at 4°C. RNA was transferred and fixed to Hybond-XL membrane (GE-Healthcare) using a gel dryer at 80°C for 2hrs. The membrane was pre-hybridized twice at room temperature for 30min in 20 mM sodium phosphate, pH 7.0, 300 mM NaCl, 1% SDS. Hybridization of radiolabeled oligo (7 pmol) was performed in 20 mM sodium phosphate, pH 7.0, 300 mM NaCl, 1% SDS at 60°C for 16 hours. Membranes were washed twice in 20 mM sodium phosphate, pH 7.0, 300 mM NaCl, 0.1%SDS, 2 mM EDTA at

60°C for 20min. The dried membranes were exposed to imaging plates and quantified using a phosphorimager.

2.3 Results and discussion

2.3.1 Charged DM-tRNA-seq scheme

Determining charging fraction by tRNA microarrays relies on selective periodate oxidation of the 3' end of uncharged tRNAs, essentially inactivating uncharged tRNAs for downstream tRNA labeling with fluorophores [38]. However, this approach requires a two-pot reaction: a mock-treated sample measures the level of total tRNA while a periodate-treated sample measures the level of charged tRNAs. We also attempted this approach using DM-tRNA-seq, but the results displayed wide variances due to precise counting of sequencing reads from two separate sequencing libraries (not shown). We therefore devised a new approach by sequencing both charged and uncharged tRNAs in the same library. Instead of only inactivating the 3' end, we use periodate oxidation coupled to β-elimination of the 3' oxidized nucleotide to distinguish between charged and uncharged tRNA prior to reverse transcription (Figure 2.1). In this way, sequencing reads that end in 3'CCA are derived from charged tRNAs while reads that end in 3'CC are derived from uncharged tRNAs.

The procedure starts with total RNA isolated from cells under mildly acidic conditions to prevent hydrolysis of the aminoacyl bond. All mature tRNA ends with 3'CCA. Periodate will only oxidize uncharged tRNAs as the charged tRNAs are protected from oxidation by the covalently attached amino acid. After removal of periodate, β-elimination at slightly basic pH is employed to selectively remove the oxidized 3'A residue, leaving a 3' phosphate at the terminal 3'C residue (Figure 2.1A). This 3' phosphate is then removed using T4 polynucleotide kinase,

resulting in 3'C-OH for these uncharged tRNAs. The β-elimination step also deacylates all charged tRNAs, resulting in 3'A-OH for these charged tRNAs (Figure 2.1B). Demethylase treatment is then carried out to remove Watson-Crick methylations, followed by cDNA synthesis using TGIRT. cDNA is excised from denaturing gels by size, and the library is constructed after circularization and PCR amplification (Figure 2.1C).

2.3.2 Verifying the logic of charged DM-tRNA-seq

We performed several controls to ensure the completeness of each treatment step (Fig. 2.3). Using a 5'-radiolabeled model oligo, we show that the periodate oxidation followed by β-elimination completely removes the 3' nucleotide (Figure 2.3A). Mock treatment with NaCl does not affect the length of the RNA oligo while treatment with $NaIO_4$ shortens the length of the model oligo by one nucleotide. This reaction is complete even in the presence of 1 μg/μL total RNA, the condition in which the first steps of the charged tRNA-seq are performed. To ensure complete removal of the 3' phosphate, yeast tRNA was radiolabeled at the final bridging phosphate using α-^{32}P-ATP and *E. coli* CCA adding enzyme. After periodate treatment, β-elimination, and polynucleotide kinase/phosphatase treatment, the 3' phosphate is completely removed, resulting in the loss of the ^{32}P-signal (Figure 2.3B). This indicates that the resulting 3' end of the tRNA is a 3' hydroxyl group that is required for template switching reaction by TGIRT.

In order to more precisely map the 3' end of tRNA in the DM-tRNA-seq reaction, we extended the tRNA-seq primer previously used by 20 residues towards the 3' end of the DNA strand from the Illumina amplification primer binding sites. The TGIRT reaction requires an RNA oligo and its complementary DNA primer [23]. Previously, the DNA primer included only

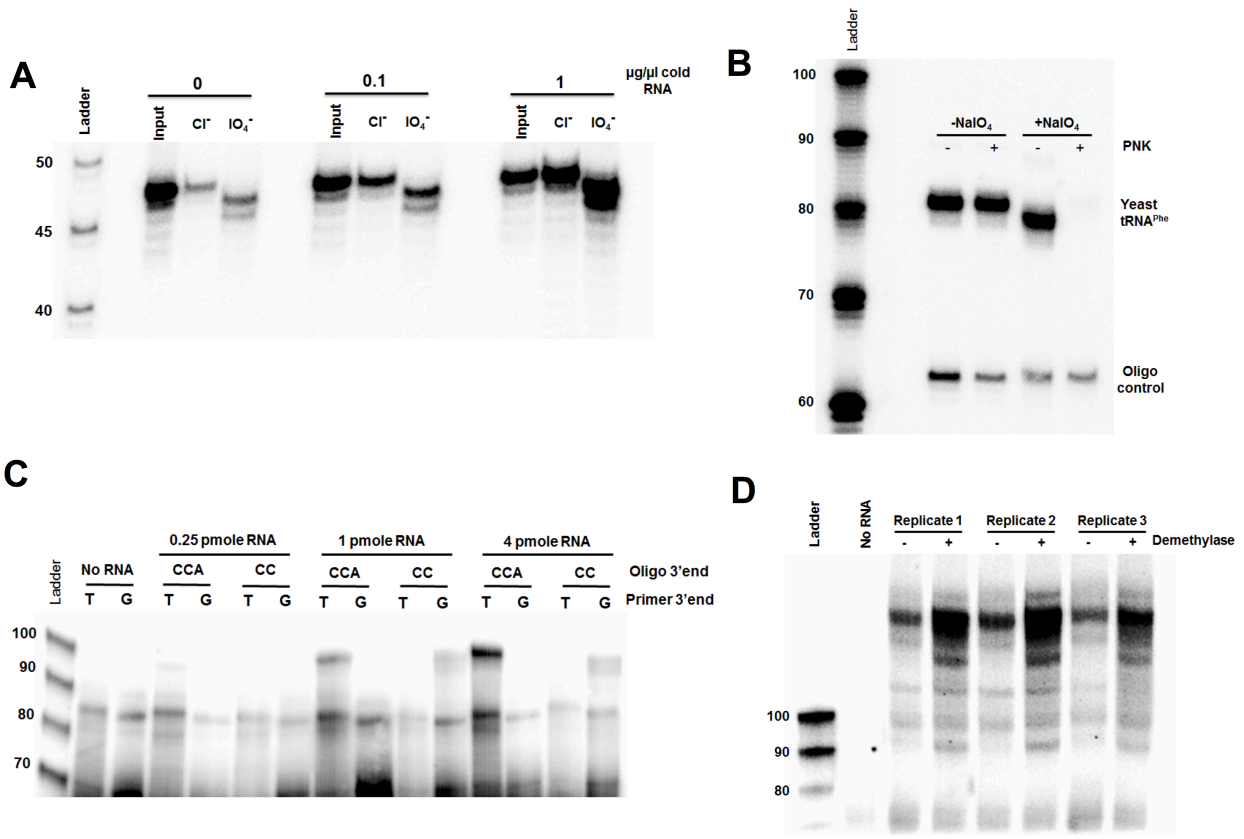


Figure 2.3: Charged tRNA-seq optimization. A) β-elimination of the final nucleotide is 100%. Using a 5'-radiolabeled model oligo, β-elimination was shown to be complete. After mock treatment with NaCl, the length of the RNA oligo remains the same. However, after treatment with $NaIO_4$, the RNA is one nucleotide shorter, even in the presence of 1 µg/µL total RNA. B) Phosphate removal from the 3' nucleotide is 100%. By labeling the final bridging phosphate, the status of the 3' end can be monitored. After mock treatment with NaCl, the bridging phosphate remains intact, even with PNK treatment. To remove this phosphate, both $NaIO_4$ and PNK treatment is required. A radiolabeled DNA oligo is included as a reference. C) There is no cross-hybridization of T- and G-ending primers with 3'A- and 3'C-ending oligos. Model oligos were used in order to assess extension of 3'A- and 3'C-ending oligos from T- and G-ending primers. TGIRT will reverse transcribe only when the primers are complementary to the last nucleotide of the oligo (3'C-ending oligo, G-ending primer and 3'A-ending oligo, T-ending primer). The expected cDNA size is 101 nucleotides for the 3'A-ending oligo and 100 nucleotides for the 3' C-ending oligo. The minor product visible in all lanes including no RNA template added (~85 nucleotides) is likely derived from aberrant RT extension of the RNA and DNA primers in the reaction mixture. D) cDNA from TGIRT reactions that were purified for Illumina sequencing. Treatment with demethylases removes m^1A58, m^1G37 and m^3C32 which are major roadblocks for TGIRT reaction. This allows for sequencing of longer cDNA transcripts, including both full-length tRNA reads and other longer abortive cDNAs caused by TGIRT stops at other modifications and/or the low processivity of TGIRT.

Illumina primer binding sites for amplification and a single nucleotide overhang for the template-switching TGIRT reaction [33]. In order to prevent *in silico* trimming that may obscure the identity of the 3' nucleotide, we added a 20 nucleotide 'spacer'. This was designed to ensure compatibility with Illumina barcoding and amplification and prevents misannotation of charged and uncharged tRNAs (Figure 2.2C). Additionally, to ensure that there is no cross-hybridization between the single nucleotide overhang in the DNA primer and the 3'-end of the RNA, model oligos were used to show that the G-ending DNA primer only extends 3'C-ending RNA while the T-ending DNA primer only extends 3'A-ending RNA (Figure 2.2C). However, this does not ensure that extension from T- and G-ending primers are equivalent, which is required in order to accurately quantify the charging level. To measure this parameter, two different pairs of model oligos ending in 3'CC and 3'CCA were added in different ratios just before the demethylation step in Fig. 2.2B to compare the measured 3'CCA/3'CC ratio to the input ratio.

To further verify that the measured values of tRNA charging fraction correspond to the actual charging fraction, uncharged *E. coli* $tRNA^{Lys}$ and yeast $tRNA^{Phe}$, and *in vitro* charged *E. coli* $tRNA^{Tyr}$ (62% charged, Figure 2.1) were added to total RNA prior to periodate oxidation. This ensures that periodate oxidation and β-elimination were complete and that charged tRNAs were not deacylated, lending confidence to the measurement of charged tRNA fraction.

2.3.3 Implementation of charged DM-tRNA-seq

We performed charged tRNA-seq using total RNA from HEK293T cells in three biological replicates. After periodate oxidation and β-elimination, DM-tRNA-seq is performed as previously described. The same tRNA samples without demethylase treatment were also sequenced for comparison. cDNA was purified from denaturing gels (Figure 2.3D), and library construction and tRNA sequence mapping were performed as previously described [33].

Approximately 30% of reads from samples that were not treated with demethylase were mapped to genomic tRNA genes, while 45-60% of reads from demethylase treated samples were mapped to genomic tRNA genes (Table 2.1). These mapping statistics are comparable to the original DM-tRNA-seq results [33].

Sample/replicates	Total processed reads	Mapped reads	Mapped rate
Untreated-rep1	21,000,566	4,931,017	23.48%
Untreated-rep2	18,785,126	4,328,746	23.04%
Untreated-rep3	13,844,890	3,882,741	28.04%
Treated-rep1	19,948,695	11,729,044	58.80%
Treated-rep2	43,660,755	19,612,649	44.92%
Treated-rep3	19,497,232	10,578,980	54.26%

Table 2.1: Sequencing statistics for HEK293T charged tRNA-seq replicates. Approximately 30% of reads from samples that were not treated with demethylase were mapped to genomic tRNA genes, while 45-60% of reads from demethylase treated samples were mapped to genomic tRNA genes. These mapping statistics are comparable to the original DM-tRNA-seq results.

The identity for most human tRNA isodecoders can be determined using the sequence of the 30 3' most nucleotides, so full-length tRNA reads are generally not required to determine tRNA charging fraction for individual isodecoder. However, m^1A58 causes RT stops even for TGIRT and is present in most human tRNAs [58]. cDNAs that abort at m^1A58 cover only ~20 residues of tRNA and can be too short to unambiguously assign to tRNA isodecoders. Thus, in order to properly quantify isodecoder amounts for mammalian samples, demethylase treatment is required. Because of the higher number of mapped reads, demethylase treated samples were used for the remaining analysis. The charged fraction of tRNA isodecoders from demethylase treated

replicates correlated well (Figure 2.4), so averages of the three replicates were used for all further analysis.

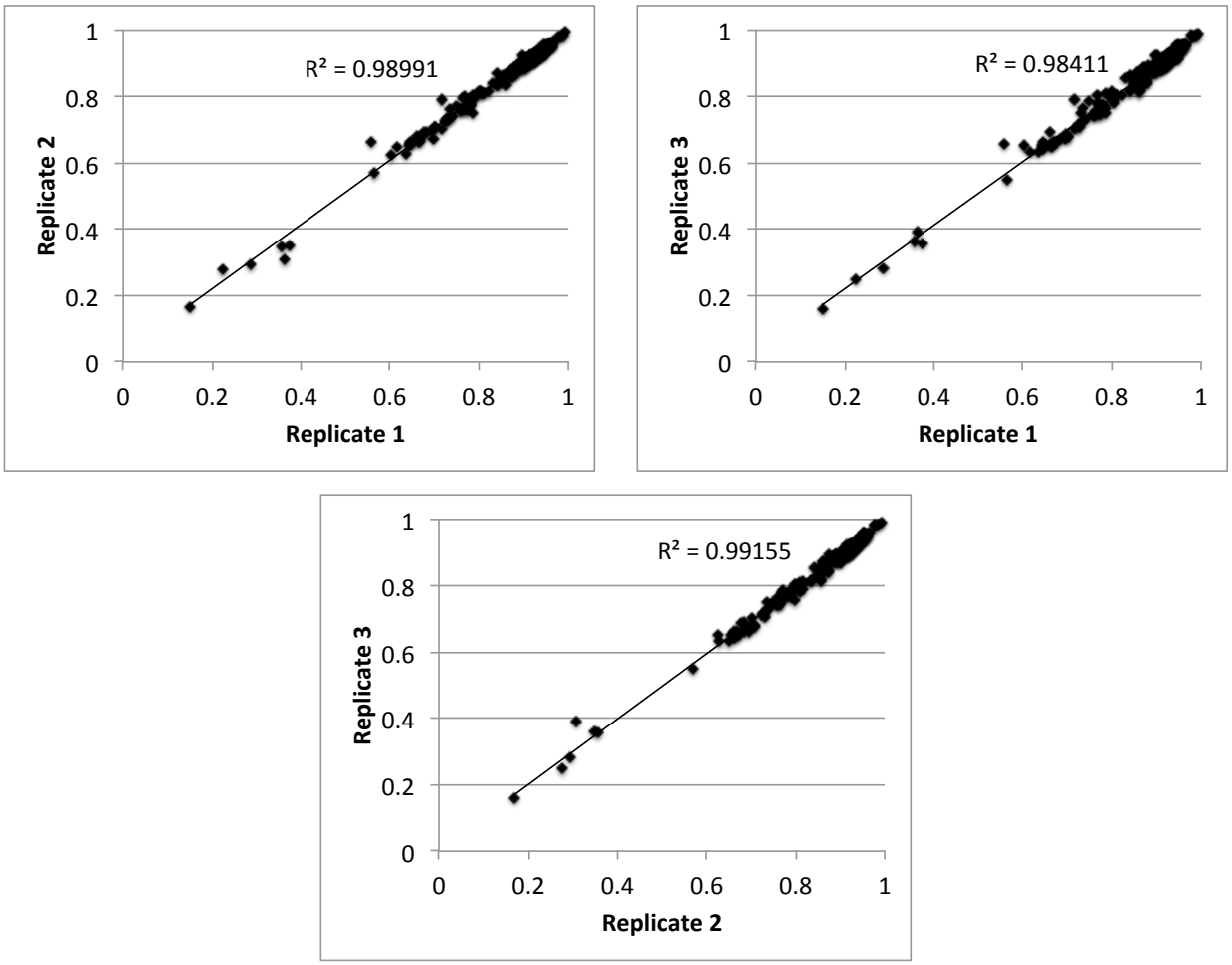


Figure 2.4: Charging levels plotted between biological replicates. All nuclear and mitochondrial-encoded tRNA isodecoders with read counts within 1,000-fold range relative to the highest abundant tRNA ($tRNAi^{Met}$ for these samples) are shown.

We first determined the quantitative nature of our experiment using the spike-in standards (Figure 2.5A). Standards are *E. coli* and yeast tRNAs with known charging fractions and RNA oligo pairs with known 3'A/3'C ratios. All data points have a linear correlation between the expected and measured values (R^2=0.999) with a slope of 0.955 and intercept of 0.06. This result indicates the fully quantitative nature of our method.

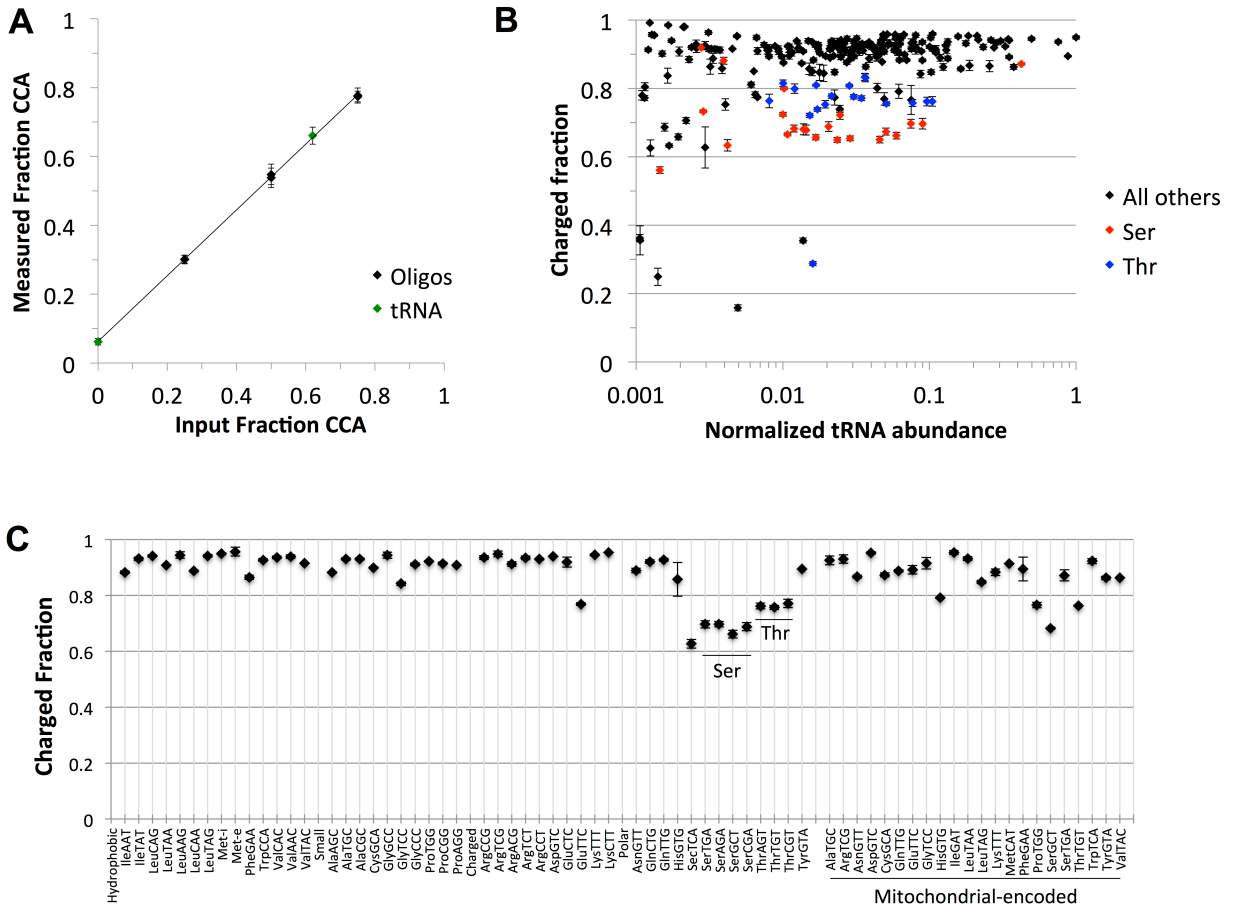


Figure 2.5: Charged tRNA-seq results. A) Spike-in model oligos and tRNA standards show that charged-tRNA-seq is quantitative. Two pairs of model oligos (black) were added in different molar ratios to the three biological replicates (see *Materials and Methods*). Additionally two uncharged tRNAs and one *in vitro* charged tRNA were also added. The 'charged fraction' measured by sequencing was plotted against the input 'charged fraction'. This calibration curve shows that the measured charged fraction is linear across different charged fractions. B) The measured charged fraction is independent of the abundance of tRNA. The abundance of nuclear and mitochondrial-encoded tRNA isodecoders was normalized to the most abundant isodecoder. tRNA isodecoders that were within 1,000-fold of the most abundant tRNA were plotted. The charged fraction is independent of the abundance, suggesting that the measured charged fraction should be accurate even for low abundance tRNA isodecoders. $tRNA^{Ser}$ and $tRNA^{Thr}$ isodecoders are highlighted. C) Most abundant tRNA isodecoders are highly charged. The charged fraction of the top abundance isodecoder for each isoacceptor was plotted. Most abundant isodecoders are highly charged (>80%) while $tRNA^{Ser}$ and $tRNA^{Thr}$ isodecoders have lower charging fraction (60-80%).

We then plotted the charging level of all individual isodecoders that are within a 1,000-fold range of the most abundant isodecoder (an initiator $tRNA^{Met}$ in this case, Figure 2.5B). Most

tRNA isodecoders are charged at a level of >80%, consistent with the expected range of charged tRNA fractions in the cell. The small amount of uncharged tRNAs likely resides in the E-site of the ribosome. There is little correlation between tRNA isodecoder abundance and charging fraction within this abundance range, suggesting that our method should be useful to measure charging fraction of all tRNAs, even including rare tRNA isodecoders. On the other hand, the most abundant tRNA isodecoders are typically charged at high levels, indicating that the highly abundant isodecoders are likely used for protein synthesis (Figure 2.5C). Unexpectedly, $tRNA^{Ser}$ and $tRNA^{Thr}$ isodecoders are charged at lower levels compared to other tRNAs (Figure 2.5B, 2.5C). This result is reminiscent of an *E. coli* study using microarrays where $tRNA^{Ser}$ and $tRNA^{Thr}$ isoacceptors have lower charging levels when grown in media where serine is also heavily used for metabolic purposes [51].

2.3.4 Validation of charged DM-tRNA-seq

We performed the standard acidic denaturing gel electrophoresis followed by Northern blot to validate the charging levels of several tRNAs, including the unexpected $tRNA^{Ser}$ and $tRNA^{Thr}$ results (Figure 2.6). High levels of tRNA charging were seen in $tRNA^{Trp}$ (94.0% ± 2.4%, 3 biological replicates) and mt-$tRNA^{Trp}$ (91.0% ± 2.8%, 3 replicates) by Northern blot (Figure 2.6A and 2.6B), agreeing well with values obtained by sequencing for the most abundant $tRNA^{Trp}$ (92.6% ± 0.8%, 3 replicates) and mt-$tRNA^{Trp}$ (92.3% ± 0.3%, 3 replicates). Similarly, we observed good correspondence between the charging levels measured by Northern blot for $tRNA^{Ser}$(GCU) (68.1 ± 6.0, 5 biological replicates; 66.2% ± 1.0% by sequencing, 3 replicates) and $tRNA^{Thr}$(UGU) (74.6% ± 12.1%, 5 replicates; 75.7% ± 1.0% by sequencing, 3 replicates) (Figure 2.6C and 2.6D).

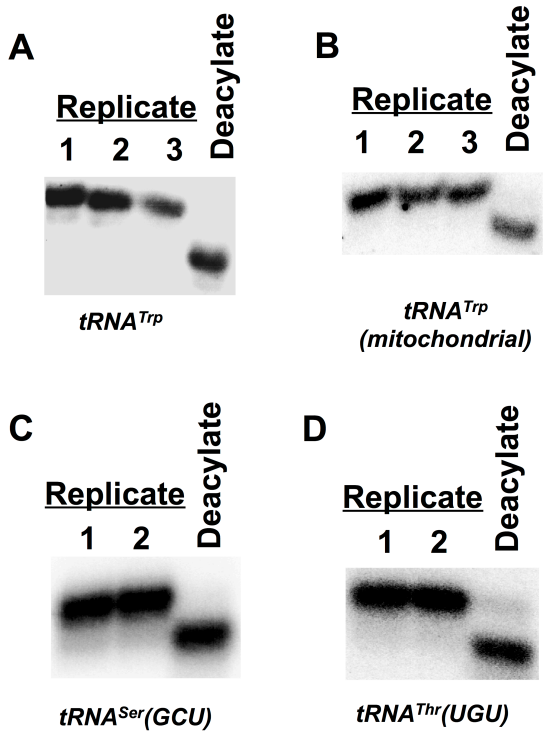


Figure 2.6: Validation of low charging fraction for tRNA$^{Ser/Thr}$. Total RNA was loaded onto a 6.5% acid denaturing gel to separate charged from uncharged tRNAs. Deacylated RNA was also run as a control. Trp (A) and mt-Trp (B) are highly charged while tRNA^{Ser}(GCU) (C) and tRNA^{Thr}(UCU) (D) have a lower charged fraction. In (C) and (D), 5 biological replicates were performed and only two are shown for simplicity.

2.3.5 Discussion

While our charged tRNA-seq method is not crucial for organisms with low tRNA sequence diversity (e.g. bacteria), determination of charging levels of tRNA isodecoders in mammals requires single-base resolution. The cause of or purpose for high tRNA diversity in mammals is not readily apparent. While tRNAs that are charged are most certainly used in translation, different isodecoders could potentially play distinct roles in translation. For example, different tRNA isodecoders of tRNA^{Ser}(UAG) show different stop codon suppression efficiency [59]. Additionally, tRNA^{Glu} isodecoders seem to have different mistranslation efficiency [60]. A tRNA^{Arg}(UCU) isodecoder is specifically expressed in central nervous system, and its expression is needed to alleviate ribosome pausing [61]. tRNA isodecoders can also have extra-translational functions. A tRNA^{Asp} isodecoder that is not charged has been shown to bind to the 3'UTR of aspartyl-tRNA synthetase (AspRS) transcript in HeLa cells to regulate polyadenylation, mRNA

stability, and AspRS translation [62]. However, the individual functions of the vast majority of tRNA isodecoders have not yet been investigated. The intriguing result that some tRNA isodecoders have a much lower charging fraction (Figure 3B) could potentially signal that these tRNA have extra-translational functions, and these isodecoders may be candidates for future studies for tRNA 'moonlighting' functions.

2.4 Conclusions

In summary, we show that the charging fractions of tRNA isodecoders can be determined in a one-pot reaction using DM-tRNA-seq. Our high-throughput, high-resolution method to determine the charged fraction of tRNA isodecoders will allow for investigation into differences in charging levels under different cellular conditions such as stress or environmental perturbation.

CHAPTER 3

TRNA BASE METHYLATION IDENTIFICATION AND QUANTIFICATION VIA HIGH-THROUGHPUT SEQUENCING

Here we report a method to detect modifications in tRNAs using deep sequencing. By defining the combination of mutation rate and stop rate at each site as the 'modification index', we identify likely sites of modification. Then, based on the effect of treatment with a demethylase and the contributions of the stop and mutation rate to the modification index, the identity of the modification can be discerned.

3.1 Introduction

Over 100 types of post-transcriptional RNA modifications have been identified in biology [63]. RNA modifications are a source of cellular and biological tuning of RNA function [64-67]. The most extensively modified cellular RNAs include rRNA and tRNA, each with multi-faceted functions. rRNA modifications affect ribosome maturation and numerous aspects of protein synthesis [68-70]. Mammalian tRNAs are the most highly modified RNA molecule in the cell, containing on average 14 modified nucleotides per molecule, or one modification for approximately every five residues [33,64,71,72]. tRNA modifications are known to affect all aspects of tRNA biology including decoding, charging efficiency, fidelity, *in vivo* stability, and intracellular localization [73-77].

The human tRNAome consists of ~500 tRNA genes distributed among 49 isoacceptors, i.e. tRNAs with different anticodon sequences [57]. Furthermore, human isoacceptor families can contain many unique sequences due to differences in the body sequence, so that ~300 distinct tRNAs are encoded in a human genome [54,57]. These tRNAs potentially have different

modification patterns depending on sequence, anticodon stem-loop context, and other factors. Although it is commonly assumed that many modification sites in tRNA are fully modified, previous literature provide scattered evidence that certain modification sites in specific tRNAs can be fractionally modified [47,78-81]. This suggests that dynamic differences in tRNA modification could be a potentially useful parameter for biological regulation.

Extensive studies have been performed to identify tRNA modifications in a site-specific manner (e.g. [43,44]). The most common method was to first isolate a tRNA of interest from total cellular RNA, followed by digestion to oligonucleotides and the determination of modification site by either thin-layer chromatography (TLC) or liquid chromatography and mass spectrometry (LC/MS). These methods have the ability to access all modification types and have been the standards in establishing the full modification patterns in individual tRNAs. Major drawbacks of these methods are the difficulty and large amount of material needed to isolate an individual tRNA from cellular RNA. Another method of identifying tRNA modification sites relies on reverse transcriptase (RT) stops and/or mutations that occur at several specific modification types at the Watson-Crick face of the nucleobase [45,46]. This "modification-detection-by-synthesis" method can be applied transcriptome-wide for tRNA. However, previous tRNA sequencing methods were inefficient and not quantitative due to the low quality of sequencing reads derived from poorly characterized RT stop and mutation efficiencies at individual modification types and sites. It was also shown that RT stop and mutation are strongly context dependent for a modification type such as N^1-methyladenosine (m^1A) [47,48]. Thus, in TLC, LC/MS and previous sequencing studies, it has not been feasible to establish quantitative information on tRNA modification fractions at individual sites.

Recent studies on N6-methyladenosine (m^6A) and N1-methyladenosine (m^1A) in mRNAs indicate that dynamic RNA modifications play important biological roles [82-85]. Dynamic mRNA modification has a cell type and cell state dependent pattern that includes both the location of the modification sites and the modification fractions at each site. On the other hand, tRNA modifications are generally present in the same locations derived from the specificity of the modification enzymes and tRNA sequence/structure. Therefore, dynamic tRNA modifications would most likely only be derived from the variations in the modification fractions at each site. Indeed, a previous study using low resolution microarrays on to determine the amount of m^1A58 modification present in tRNAs shows that m^1A58 sites in some specific tRNAs are hypomodified [80]. Despite previous efforts, transcriptome-wide method has been lacking to systematically quantify tRNA modifications at single base resolution.

Recently, our group and Lowe/Phizicky labs published Illumina sequencing methods for tRNA: DM-tRNA-Seq (demethylase tRNA sequencing) and ARM-Seq (AlkB-facilitated RNA methylation sequencing) [33,34]. The principle of both methods is to use the *E. coli* AlkB demethylase and its mutant to remove m^1A, N3-methyl-cytosine (m^3C), and N1-methyl-guanosine (m^1G) modifications at the Watson-Crick face in tRNA prior to cDNA synthesis. While the initial results mainly emphasized the increased frequency in the full-length tRNA or tRNA fragment reads due to demethylation, neither publication provided a detailed analysis on tRNA modification landscape, especially the ability of using tRNA-seq to quantify modification fractions. In our published DM-tRNA-seq results [33], we showed that the application of a thermophilic RT (TGIRT) enabled large number of reads derived from readthroughs of these modifications. Furthermore, TGIRT leaves a very strong mutation and stop signature for

different modification types and sites. However, a thorough analysis on the applicability of the DM-tRNA-seq method on tRNA modifications was not carried out.

Here, we provide a comprehensive, detailed study on applying DM-tRNA-seq to identify specific base methylations as well as to generate quantitative information of modification fractions at these sites in the human tRNA transcriptome. We were able to thoroughly analyze 6 base methylations in tRNA and rRNA (m^1A, m^3C, m^1G, N^2,N^2-dimethyl-guanosine (m^2_2G), 1-methylinosine (m^1I), N^3-methyl-uridine (m^3U), Figure 3.1).

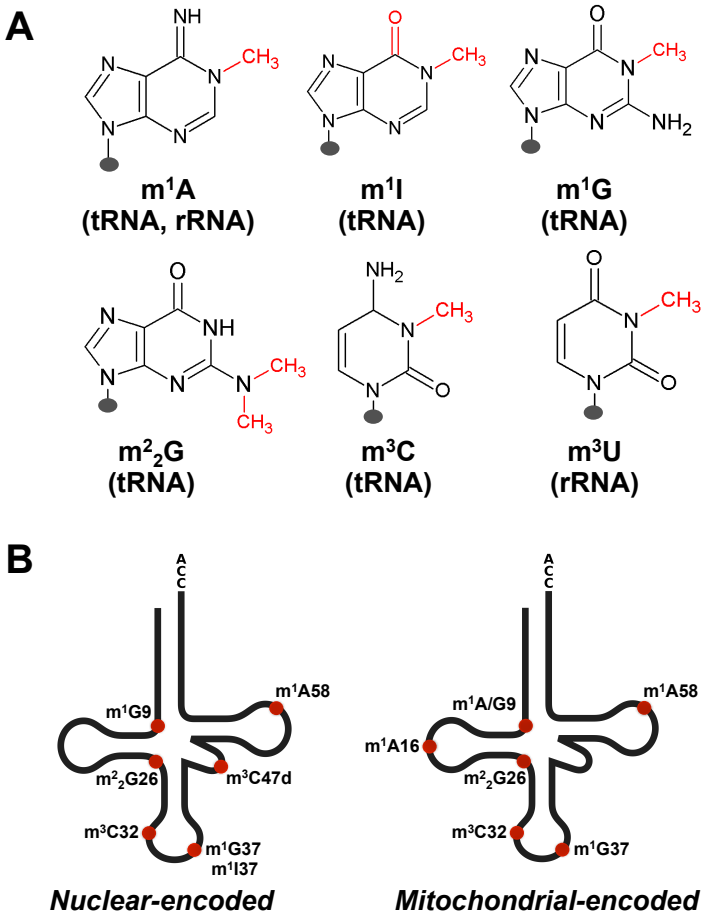


Figure 3.1 Methylations in human tRNA and rRNA investigated in this study. (A) Chemical structure of the 6 modifications where the methyl group is shown in red. The presence of each modification in tRNA and/or rRNA is listed beneath each base. (B) Schematic tRNA cloverleaf structure where the methylations for nuclear-encoded (left) and mitochondrial-encoded (right) tRNAs investigated in this work are located.

Using known positions of modification in tRNA, we can infer identification of 5 other modifications (inosine (I), dihydrouridine (D), methyl-2-thio-N6-threonylcarbamoyl-adenosine (ms^2t^6A), methyl-2-thio-N6-isopentenyl-adenosine (ms^2i^6A), 1-methyl-3-(3-amino-3-

carboxypropyl)pseudouridine ($m^1\Psi$-X)). We provide a complete map of the human tRNA transcriptome on four methylations (m^1A, m^3C, m^1G, $m^2{}_2G$) in both nuclear and mitochondrial-encoded tRNAs in HEK293T cells, many of which were unknown or not previously validated. We apply a quantitative metric, the "Modification Index" (MI) to describe each modification site that includes frequencies of mutations and stops. MI values give semi-quantitative information regarding the modification levels at individual tRNA position. We validate the quantitative nature of the MI analysis for several tRNAs by primer extension. Our results establish DM-tRNA-seq as a high throughput method to detect and quantify numerous tRNA modifications and its applicability to studies of dynamic tRNA modifications.

3.2 Materials and Methods

3.2.1 DM-tRNA-seq of HEK293T and HeLa cells

Experimental details of DM-tRNA-seq method were described in [33]. The same method was applied for HeLa rRNA sequencing except the input RNA was isolated rRNA that was subjected to chemical fragmentation before library preparation. Our DM-tRNA-seq data is composed of biological replicates of total tRNA from HEK293T cells without and with demethylase treatment. Both sets of replicates map with $r^2 \geq 0.99$. For the analysis here, we used the untreated sample 2 which has 9,002,637 mapped reads to nuclear-encoded tRNA and 2,269,204 reads to mitochondrial-encoded tRNA, and treated sample 1 with 15,740,700 mapped reads to nuclear-encoded tRNA and 2,851,034 reads to mitochondrial-encoded tRNA. The rRNA-seq data contains 7,531,718 mapped reads.

3.2.2 Sequencing alignments

All reads were sequenced on an Illumina HiSeq 2000 with paired-end mapping with read lengths of 100 base pairs. Standard quality control via FastQC was performed after sequencing and after subsequent trimming mentioned next.

Sequencing reads were aligned using Bowtie to a modified tRNA genome file containing nuclear-encoded tRNAs, mitochondrially-encoded tRNAs and human rRNAs (found at the GEO accession for this paper). Splicing of tRNAs was accounted for in the modified genome file and only mature tRNAs were used for alignment. Briefly, the tRNA library was adapted from the tRNAScan-SE library ((http://gtrnadb.ucsc.edu/Hsapi19/, [57]) by appending 3'CCA to tRNAs from the genomic tRNA database. Isodecoders with identical SEscan scores from the genomic tRNA database were consolidated for ease of identity assignment, so the total number of annotated tRNA genes and pseudogenes were reduced from 625 to 462. Prior to mapping, reads were processed using Trimmomatic v0.32 as well as further trimming using custom Python scripts to remove any further artifacts from demultiplexing and removal of primers, adapters, or any other low quality sequences. Sequences greater than 15 base pairs were then aligned to the library using Bowtie 1.0 with sensitive options using the highest allowed mismatch settings for Bowtie 1.0. Mapping to all references occurred simultaneously. Only one alignment (best with Bowtie1 or k=1 with Bowtie2, k refers to the number of allowed distinct alignments per raw read) declared as valid by the respective mapping software was reported for each read. Analysis was also performed using k = 3 to allow for up to 3 alignments per read and also v = 0,1,2,3 (how many mismatches allowed in the seed sequence per raw read) in Bowtie 1 to allow for fewer mismatches.

3.2.3 Modification Index (MI)

Mapping pipeline proceeded by conversion from the SAM output from Bowtie using custom C and Python scripts to separate isodecoders based on previous alignment. From here, further cleanup based on redundancy (any read that could inherently map due to misalignment) was also discarded. For each position in the reference file, the following was calculated: at each position = *n*, how many counts existed (total counts) = *c*, how many misincorporations or mutations = *m,* and how many aligned reads stopped at the position = *s*. This leads to a full modification index (MI) calculated at each position by (mutations at position + stops at position directly 3' to position)/(total counts at position), or *(m+s)/c*, and individual mutation and stop components to the metric can be further reduced by calculating *m/c* and *s/c*, respectively.
For high stringency detection purposes for this work, sites corresponding to a MI value of 15% or below were discarded as noise due to sequencing error or basal level of misincorporation from the RT reaction. For simplicity and ease of discussion, sites with MI ≥ 15% were categorized in the range of 15-50% as low, 50%-80% as medium and 80-100% as highly or completely modified. Positional shifts of modifications due to tRNA length (either due to variations in length of type I tRNAs or type I versus type II tRNAs) were recognized and adjusted for in calculations for heat maps. Modification index in regions of variable loop are accounted for accordingly. Modification index 3' to $m^1A57/58/59/68$ (m^1A58 in standard tRNA nomenclature) is only derived from mutations due to the lack of stop information at these positions because of the short reads needed for stops at this m^1A position.

3.2.4 Reverse Transcription Primer Extension

Total RNA was isolated from HEK293T cells in biological triplicates using TRIzol (Life technologies). tRNA was gel purified by 8% denaturing PAGE (7M Urea, 1X TBE). 700 ng of total RNA or 200 ng of total tRNA was annealed to 5 pmol of specific primer in 30 mM Tris-HCl pH 7.5, 2 mM KCl at 90ºC for 90 sec. The mixture was cooled at room temperature for 3 min. The reverse transcription reaction was carried out in 1X AMV Buffer, 0.1 mM each cold (d)dNTP, 1.25 µM radiolabeled dNTP, and 0.2U/µL AMV RT (New England BioLabs). The reaction was performed at 37ºC for 30 min. To degrade the RNA after the RT reaction, 10 U of RNaseH (Epicentre) was added, and the reaction was incubated at 37ºC for 5 min. RNA Loading dye (9 M Urea, 100 mM Na_2EDTA, Bromophenol blue, xylene cyanol) was added and the reaction was resolved by denaturing gels (7M Urea, 1X TBE).

3.2.5 Validation of m^1A9 in $tRNA^{Asp}$ (AspGTC) and m^3C47d in $tRNA^{Leu}$ (LeuCAG)

Total RNA was isolated from HEK293T cells using TRIzol (Life Technologies). Total tRNA was gel purified by 8% denaturing PAGE (7M Urea, 1X TBE).

RNase H cleavage and fragment isolation

For AspGTC, 5' ^{32}P-labeled tRNA was spiked into 20 µg of total tRNA and annealed to 20 pmol DNA oligo (5'-TGGCTCCCCGTCGGGGAATCGAACCCCGGTCTCCCGCGTGACAGGCGGGG) complementary to nucleotides 26-76. For LeuCAG, 3' ^{32}P-labeled tRNA was spiked into 20 µg of total tRNA and annealed to 20 pmol DNA oligo (5'-CTGCGACCTGAACGCAGCGCCTTAGACCGCTCGGCCATCCTGAC) complementary to

nucleotides 1-42. The tRNA/oligo was annealed in 100 mM Tris-HCl, pH 7.5, 0.5 mM EDTA at 90ºC for 90s. The reaction was cooled at room temp for 3 min. Digestion was performed by adding NaCl to 100 mM, $MgCl_2$ to 10 mM, and RNase H (Epicenter) to a final concentration of 1U/µl. The reaction was incubated at 37ºC for 30 min. 0.1U/µL DNase I (NEB) was added and the reaction was incubated at 37ºC for 30 min. The reaction was then ethanol precipitated, resuspended in 1X RNA loading dye (4.5 M Urea, 50 mM EDTA), and resolved on 10% denaturing PAGE (7 M Urea, 1XTBE). The digestion products were eluted in 200 mM NH_4Cl, 50 mM NH_4OAc overnight. Glycogen was added to aid precipitation.

MALDI-TOF

Approximately 2 pmole of eluted fragments were digested with 0.5 U/µL T1 (Ambion) in 10 mM NH_4OAc for 30 min at 37ºC. One pmole of the digested fragments was mixed with an equal amount of MALDI matrix, composed by 9:1 (v:v) ratio of 20 ,40 ,60 - trihydroxyacetophenone (THAP, 10 mg/ml in 50% CH_3CN/H_2O): diammonium citrate (50 mg/ml in H_2O). The mixture was spotted on a MALDI sample plate, dried under vacuum and analyzed by a Bruker ultrafleXtreme MALDI-TOF Mass Spectrometer in reflector, positive mode.

LC/MS-QQQ

Approximately 1 pmole of RNase T1 digested fragments was digested with Nuclease P1 in 100 mM ammonium acetate at 37ºC for 2 hours. 3µl of freshly prepared 1M ammonium carbonate was added with 1 U of Alkaline Phosphatase (Roche), and further incubated at 37ºC for 2 hours. Mixture was centrifuged through a 0.22 m PVDF syringe filter (Millex-GV) to

remove contaminants. Remaining sample was then analyzed using an Agilent 6460 Triple Quad MS-MS with attached 1290 UHPLC.

Hydrazine/aniline cleavage and tRNA microarray

3' ^{32}P-labeled tRNA was spiked into 0.2 µg/µL total tRNA and incubated with 10% Hydrazine in 3M NaCl at 4ºC for 10min, quenched with 0.3M NaOAc/HOAc, pH 5.5, and ethanol precipitated. The pellet was then resuspended in 1M aniline, 1M HOAc. The reaction was incubated at 60ºC for 10min, quenched with 0.3M NaOAc/HOAc, pH 5.5, and ethanol precipitated. The pellet was resuspended in 1X RNA loading buffer (4.5M Urea, 50mM EDTA) and resolved on 15% denaturing PAGE (7M Urea, 1X TBE). Fragments were eluted in 200 mM KCl, 50 mM KOAc overnight. Poly(A) RNA and salmon sperm DNA were added to aid in precipitation. The fragments were then hybridized to custom-made tRNA microarrays at 60ºC for 16 hours and exposed to phosphorimaging plates as previously described [86].

3.3 Results and Discussion

Next generation sequencing of tRNA was previously inefficient and not quantitative due to the presence of a large number of base modifications at the Watson-Crick face and the stable tRNA structure. Our lab and the Lowe/Phizicky labs independently developed tRNA-seq methods that first removed many base methylations using AlkB-derived enzymes before cDNA synthesis [33,34]. We also used a thermophilic reverse transcriptase (TGIRT) that could more efficiently read through base methylations than the commonly used superscript RTs. Hence, more mutation signatures are present in our sequencing data using cellular RNA from HEK293T cells. Our sequencing strategy split each sample into two, one directly sequenced and the other

first treated with the demethylase enzymes before cDNA synthesis. In this work, the untreated sequencing data are used for modification analysis, whereas the demethylase-treated sequencing data are used for validating the presence of modifications.

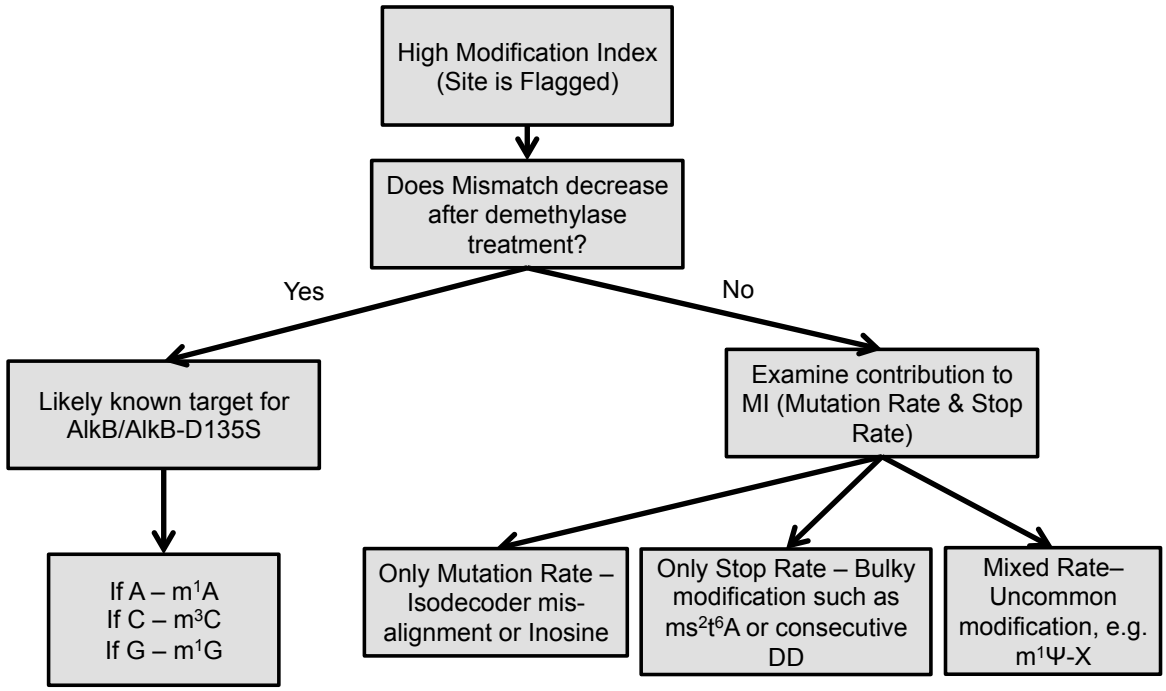


Figure 3.2: Flowchart depicting the analysis used for modification identification in tRNA and rRNA. To begin, all sites in tRNA and rRNA with a modification index ≥15% in the untreated data are flagged. Each flagged site is then compared to the corresponding site in the demethylase treated data. If MI decreases, the modification at that position is considered to belong to the substrates of the demethylases, m^1A, m^3C, m^1G, $m^2{}_2G$, and m^3U. If the site is not affected by demethylase treatment, it is further considered through the individual MI parameters of mutation and stop components. If the site has a purely mutational component, it is either identified as an inosine for the tRNA genes with an A at that position, an isodecoder misalignment, or more rarely a tRNA SNP for tRNA genes with an G/C/T at that position. If the site has a purely stop component, then it is either a bulky modification in the W-C face such as ms^2t^6A, or a modification with a largely base destabilizing motif such as dihydrouridine (D). Lastly, a site with a mixed contribution from both mutation and stop seems to represent uncommon modifications such as m^1acp^3-Ψ.

3.3.1 Identification of tRNA methylation sites.

To identify potential sites of modifications that result in RT readthrough mutations and stops, all sequencing reads were aligned to mature tRNA sequences to obtain maximal coverage of modified molecules. A decision tree is shown for the flagged modification indices ≥ 0.15 at

each position (Figure 3.2). We first look for the effect of demethylase treatment at each flagged site. m^1A, m^3C, m^1G, $m^2{}_2G$ and m^3U sites show decreased MI values upon demethylase treatment (left branch). Sites that are unaffected by demethylase treatment are further assessed based on whether the MI is composed of only mutations, only stops, or a mixture of mutations and stops (right branch). Among the right branch, only mutations corresponds to A-to-Inosine modification or isodecoder misalignment, only stops corresponds to the presence of bulky modifications in the Watson-Crick face and dihydrouridine (D), and a mixture of mutations and stops corresponds to other, uncommon modifications in the Watson-Crick face. This type of analysis enables us to identify the specifics of the modification using the MI values.

tRNAs are made of type I's which have a 4-5 nucleotide variable loop, and type II's which have a longer variable loop that folds into a short hairpin. We plotted the cumulative MI for all type I tRNAs comprised of 18 of the 20 amino acids, and type II tRNA comprised of the amino acids leucine and serine (Figure 3.3). For type I tRNAs, high MI peaks are apparent around nucleotide 58, 37, 34, 26, 20, 9, and 4-7. Upon demethylase treatment, MI values are reduced to near background levels for the peak around 58, more than 50% for the peak around 37, and moderately reduced for the peak around 9. As we have already shown previously [33], this result validates the peaks around position 58 as m^1A58, 37 as m^1G37, 9 as m^1G9 as m^1A and m^1G are known tRNA modification targets of the AlkB and AlkB-D135S enzymes used in the sequencing experiment. The same conclusion can be applied to the type II tRNAs for the peak around position 67 as m^1A at the same location as m^1A58 in the type I tRNA, 50 as m^3C located in the loop of the variable loop hairpin (m^3C47d in the standard tRNA nomenclature), 37 as m^1G37, 32 as m^3C32, and 9 as m^1G9. As expected from known tRNA modifications, some positions are not affected by demethylase treatment including the peak around position 34 as Inosine, around 20

that can be found commonly in tRNAs containing two consecutive dihydrouridines in the D loop.

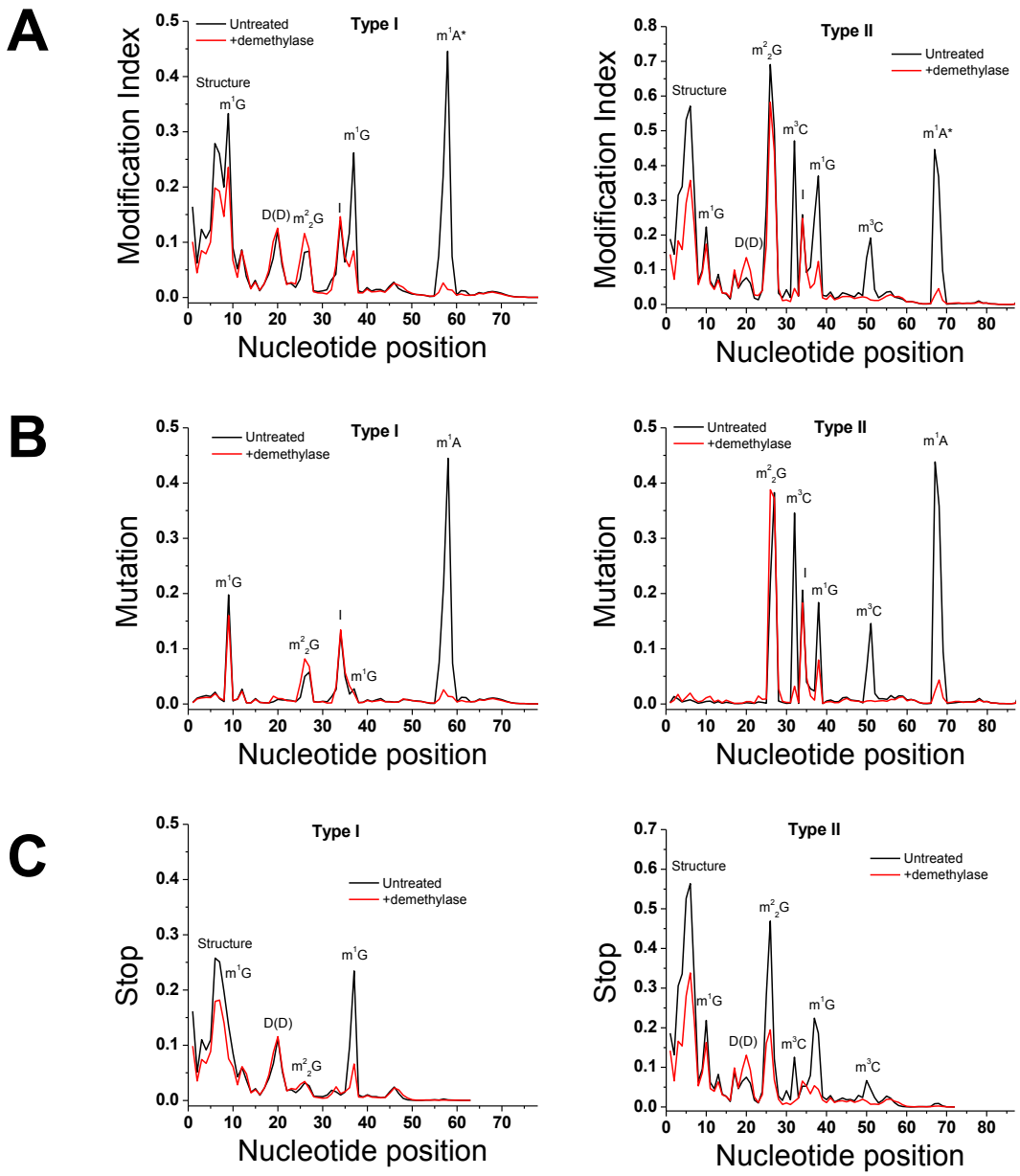


Figure 3.3: Cumulative modification index and mutation and stop plots for type I and type II tRNAs. The x-axis corresponds to the position along the tRNA. Black line: untreated, red line: demethylase treated samples. (A) MI for type I tRNA on the left and type II tRNA on the right. The modifications assigned to each peak are also indicated in the graph. Demethylase treated samples show a significant decrease for the m^1A, m^1G, and m^3C modifications. (B) Mutation fractions for type I tRNA on the left and type II tRNA on the right. (C) Stop fractions for type I tRNA on the left and type II tRNA on the right.

The demethylase treatment significantly reduced the MI values of the m^1G37 peak, but only moderately reduced the MI value of the m^1G9 peak, even though both are made of the same chemical structure (Figure 3.3A). This result can be interpreted as the tRNA structure interfering with the demethylase reaction: m^1G37 is located in the anticodon loop and easily accessible to the demethylase enzyme, whereas m^1G9 forms a base triple with the 23-12 base pair of the D stem and is more buried. Our demethylase enzyme reactions were performed under conditions where the tRNAs are still folded. The demethylase treatment requires divalent cations, so any conditions that denature tRNAs would also lead to metal-ion-dependent tRNA hydrolysis.

Interestingly, the demethylase reaction also slightly reduced the MI values of the type II tRNAs at the peak around 26, corresponding to m^2_2G26 modifications. This result suggests that our demethylase mixture is also reactive toward the m^2_2G26 modification. m^2_2G26 pairs with nucleotide 44 and this pair is sandwiched between the D and the anticodon stems. This makes it even less accessible than m^1G9, possibly explaining the poor reactivity of the demethylase toward the m^2_2G26 modification.

We further plotted the two parameters that compose the MI values, mutation fraction and stop fraction, separately for the cumulative type I and II tRNAs (Figure 3.3B, 3.3C). Prevalent sites of high mutation rate are all attributed to the four known types and locations of base methylations in tRNA: m^1A, m^1G, m^2_2G, and m^3C, as well as the well-known A34 to inosine modification. Prevalent sites of high stop rate are attributed to the known m^1G, m^2_2G, m^3C, ms^2t^6A, and consecutive dihydrouridine (DD) modifications. Strong stops are also present around nucleotides 4-7 which are moderately reduced upon demethylase treatment. This region is not known to be modified, yet the RT appears to fall off frequently during cDNA synthesis. At this time, we do not understand why these stops are present. In some cases, it is likely due to the

presence of m^1G9 in the RNA template of the RT reaction, since removal of m^1G9 by demethylase prior to cDNA synthesis simultaneously reduces the extent of stops in this region (see the graph for several mitochondrial tRNAs below). In other cases, it may be due a reduction of RT processivity after encountering many modified nucleotides in the RNA template, or some kind of RNA-cDNA structure effects. Reverse transcriptase stops at m^1A58 are not reported in this sequencing data set due to the short cDNA products generated in the sequencing reaction.

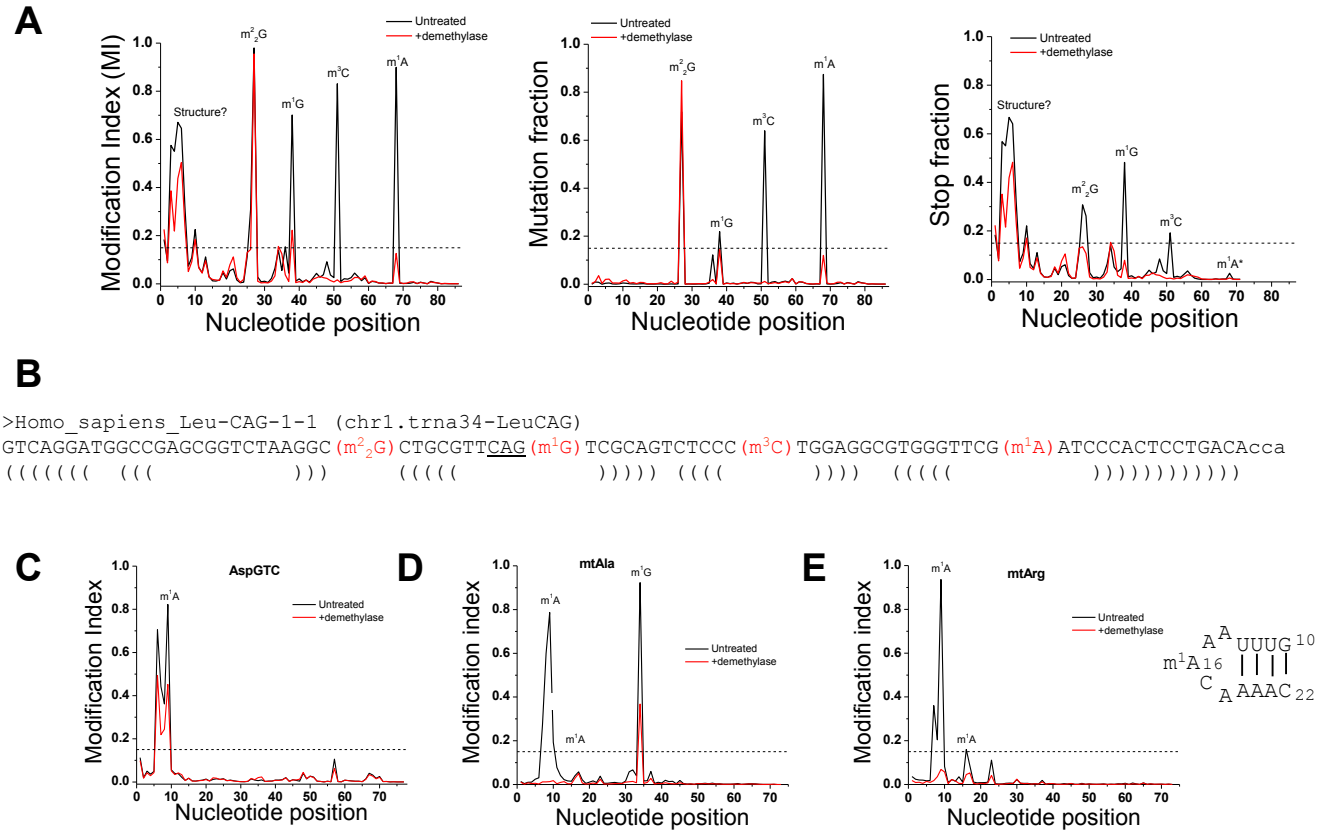


Figure 3.4: Plots for individual tRNAs with known and unknown methylations. Black line: untreated, red line: demethylase treated samples. Dashed line shows the MI value of 15% which we use here as the threshold for high confidence modification detection. (A) LeuCAG c1t34 (chromosome 1, tRNA34) isodecoder contains all four demethylase reactive modifications of m^1A, m^3C, m^1G and $m^2{}_2G$. MI plot is on the left, mutation plot in the middle, and stop plot on the right. (B) LeuCAG c1t34 sequence with annotated modifications identified in panel (A). The brackets mark the acceptor, D, anticodon, and T stems of this tRNA. (C) MI plot of AspGTC showing a newly identified m^1A9 modification. (D) MI plot of mitochondrial tRNAAla showing a newly identified m^1G37 modification. (E) MI plot of mitochondrial tRNAArg showing a newly identified m^1A modification in the D loop whose sequence is also shown.

A representative example of how MI, mutation and stop fraction plots look for a specific tRNA with multiple methylations at W-C face is shown in Figure 3.4A. In these graphs, the x-axis designates the nucleotide position and the y-axis the fraction of mutated and stopped (left panel), only mutated (middle), or only stopped (right) reads aligned to each position in this tRNA, LeuCAG isodecoder from chromosome 1, tRNA34 (CAG-c1t34) according to the hg19 genomic tRNA database [57]. Because of the background, we choose to interpret only these peaks with a MI value ≥ 0.15 as a confidence threshold (dashed line in Figure 3.4A). Three known methylations, m^1A, m^1G and $m^2{}_2G$ (Figure 3.4B) can be readily identified from these graphs as all positions have reduced MI values upon demethylase treatment. The large stop peak toward the 5' end of the tRNA is likely derived from low RT retaining processivity as discussed for the cumulative type I and type II plots (Figure 3.3C).

For LeuCAG, we found another peak located at C47d that was also removed upon demethylase treatment, consistent with a previously unidentified m^3C at this position. We also found an unknown adenosine modification at position 9 in AspGTC (Figure 3.4C). For position 9 modifications in human tRNA, m^1G9 is common among cytoplasmic tRNAs, whereas m^1A9 is common among mitochondrial tRNAs. We were surprised to find that the cytoplasmic $tRNA^{Asp}$ also contains a modification that responded to demethylase treatment, which is consistent with an m^1A modification. We performed additional experiments to independently validate these two new methylation sites (see below).

In order to validate the usefulness of our alignment strategy with respect to available aligners, we compared mapping using either Bowtie1 or Bowtie2 aligner (Figure 3.5). A major difference between Bowtie1 and Bowtie2 is that Bowtie2 allows for insertions and deletions (indels). For the highly abundant tRNAs shown in Figure 3.5, the modification sites are easily

identified with both alignment programs. The alignments using Bowtie2 seem to be a little noisier than the alignments using Bowtie1, which may be in part derived from the presence of a small amount of indels shifting the mapped positions as arbitrary misincorporations.

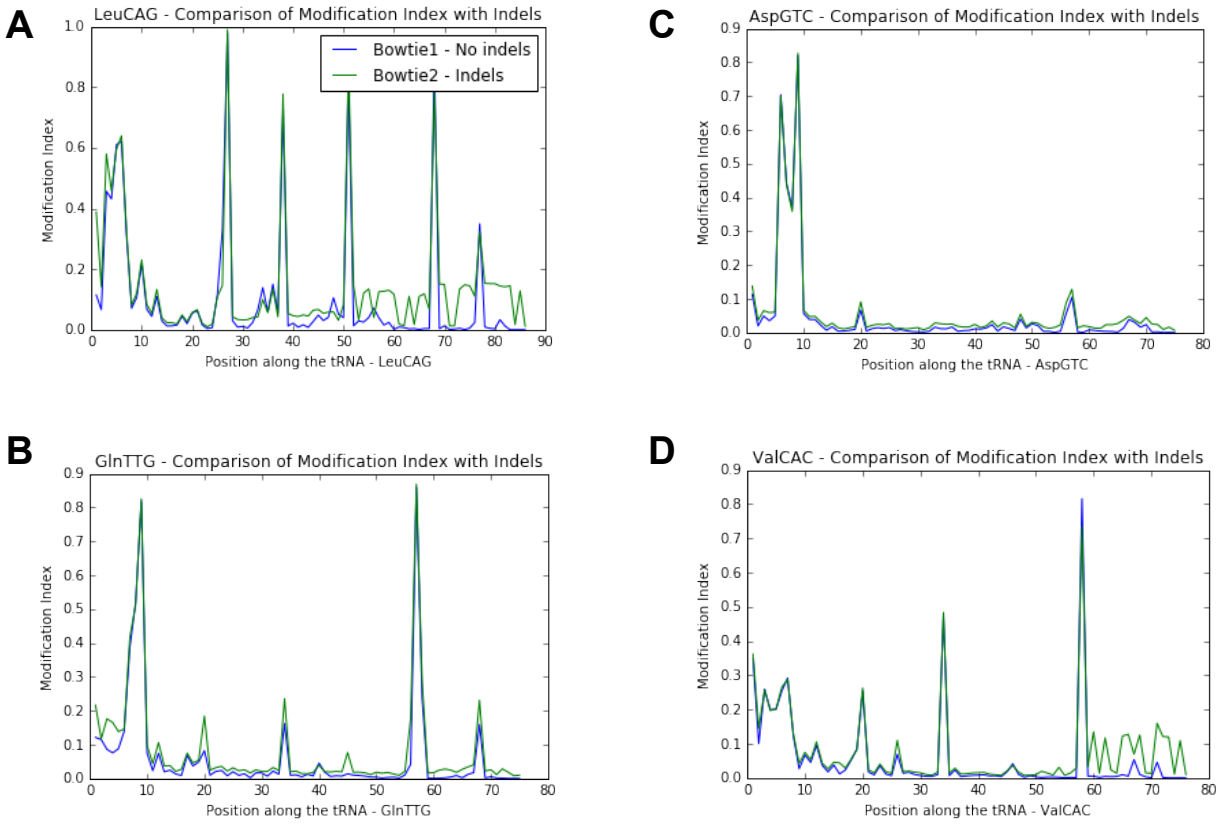


Figure 3.5: Comparison of alignment protocols for Bowtie 1 and Bowtie 2 to determine appropriate options for alignment. Shown as the plots for four highly abundant tRNAs: A) LeuCAG, B) GlnTTG, C) AspGTC, and D) ValCAC. Bowtie1 does not allow insertion and deletions, whereas Bowtie 2 does.

We also adjusted for the number of mismatches in the seed sequence, or v = 0,1,2, and 3 mismatches, a common feature in the Bowtie1 aligner (Figure 3.6). Allowing for fewer mismatches reduced the overall number of reads due to the abundant mutations located close to the 3' end of tRNAs such as m^1A58. However, even for a heavily modified tRNA such as the

LeuCAG, we were able to analyze MIs at v = 2,3 mismatches. Comparing the MI for the Bowtie1 maps relative to the indel-containing Bowtie2 maps also showed a relatively small difference in MI across all of the modifications. Average MI across these highly abundant tRNAs also did not change significantly, nor were there any substantial differences in the plots for cumulative MI across the whole tRNA. These results gave us confidence in the method of alignment and further analysis.

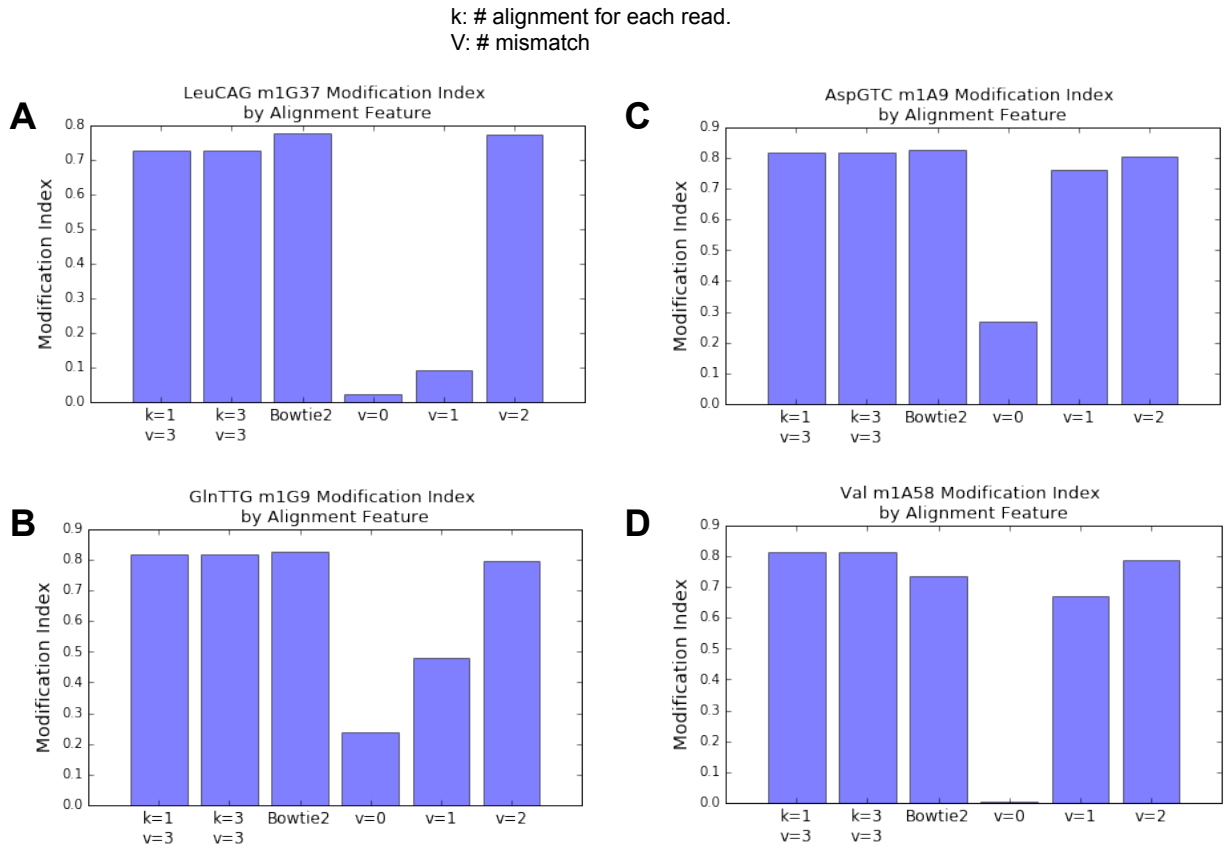


Figure 3.6: Comparison of alignment protocols to determine appropriate options for alignment. Flagged modifications for four highly abundant tRNAs A) LeuCAG, B) GlnTTG, C) AspGTC, and D) ValCAC were compared using all options for k = 1 or k = 3 (either one alignment or three alignments are reported), v = 0, 1, 2, 3 (0, 1, 2, or 3 mismatches are allowed in the seed sequence by the aligner).

3.3.2 Validation of m^1A9 in $tRNA^{Asp}$ (AspGTC) and m^3C47d in $tRNA^{Leu}$ (LeuCAG)

We performed additional experiments to validate the two new modifications at A9 in AspGTC and C47d in LeuCAG. Since the sequencing was performed using reverse transcription, we made every effort to avoid any RT reaction in our validation.

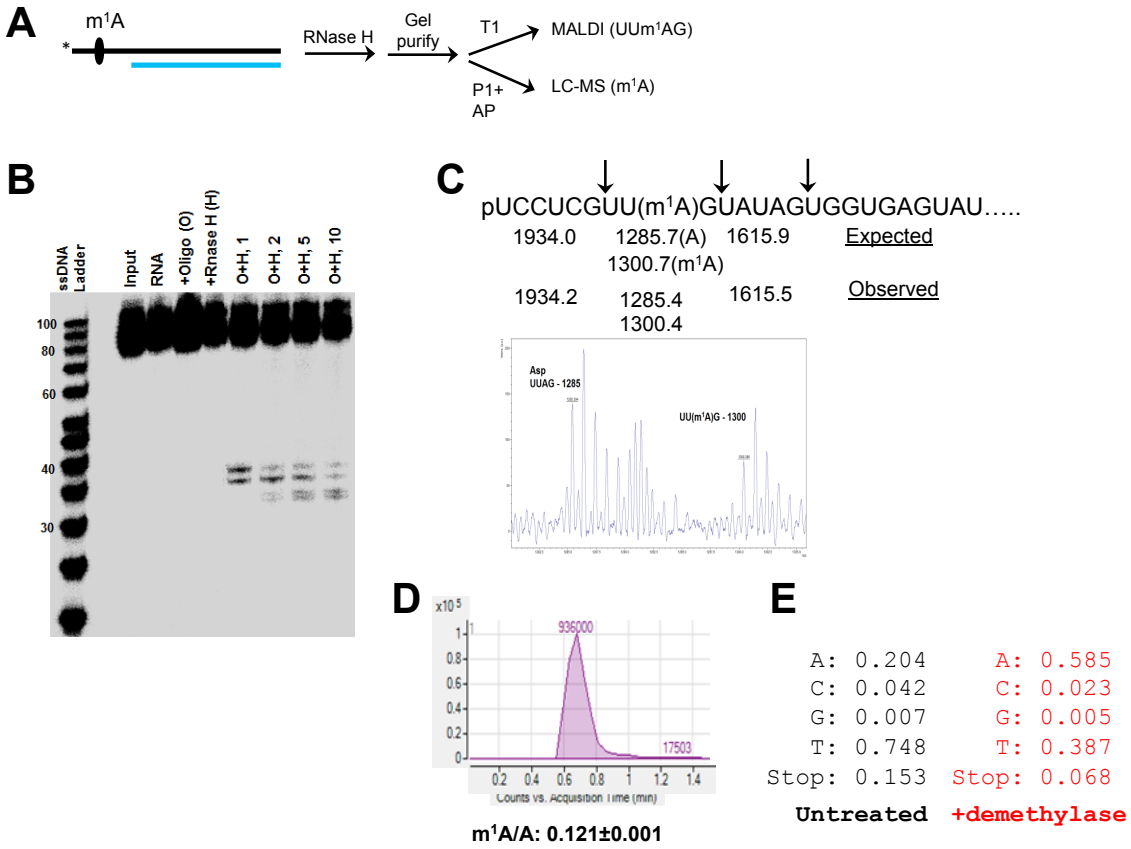


Figure 3.7: Validation of m^1A9 in AspGTC. (A) Experimental strategy. AspGTC RNA is shown as a black line, and the complementary DNA oligo is in blue. The star represents the location of the 5' 32P-label. (B) Optimization of RNase H cleavage reaction. 5' 32P-labeled total tRNA from HEK293T cells is shown. The amount of DNA oligos added (in pmole) is indicated in the 4 right lanes. The amount of cleavage product is ~5.7% which is within the expected range of the amount of AspGTC RNA in the sample. The 1pmole condition is the most stringent and used in subsequent experiments. (C) MALDI results. The sequence of the AspGTC fragment is shown with the corresponding mass of the nuclease T1 digested fragments. The mass spectrum shows the presence of the m^1A-modified fragment. (D) QQQ LC-MS result of nuclease P1 and alkaline phosphatase (AP) treated AspGTC fragment showing the assigned m1A peak. (E) The mutation and stop fractions from DM-tRNA-seq at the A9 site.

For AspGTC, our approach was to use an AspGTC specific oligo and RNase H cleavage to first isolate the 5' fragments of the AspGTC RNA from gel purified total tRNA from HEK293T cells, followed by nuclease T1 digestion to detect a modified fragment containing A9 by MALDI or nuclease P1 and alkaline phosphatase treatment to detect a methylated A residue by LC-MS (Figure 3.7A). We tested a variety of conditions and applied the most stringent condition for fragment isolation (Figure 3.7B). We detected all three expected nuclease T1 fragments in the MALDI assay, including the UUAG/UU(A*)G peaks that differ by 15 daltons, indicating a presence of a methyl group (Figure 3.7C). We also readily detected m^1A using the established QQQ LC-MS method for m^1A RNA methylation studies [84] (Figure 3.7D). These results, together with the mutation and stop signatures at this position (Figure 3.7E), indicate that the modified A9 residue in AspGTC corresponds to N1-methyladenosine.

For LeuCAG, we also used the approach of isolating RNase H cleaved LeuCAG fragments from gel purified total tRNA from HEK293T cells, followed by nuclease P1 and alkaline phosphatase treatment to specifically identify m^3C (Figure 3.8A). Again, stringent RNase H cleavage conditions were used to obtain LeuCAG fragments (Figure 3.8B), and QQQ LC-MS showed appropriate amount of m^3C as well as m^1A (from m^1A58 of LeuCAG) from the isolated fragments (Figure 3.8C). We also utilized an m^3C-specific base chemistry to validate the LeuCAG site (Figure 3.8D). Hydrazine reacts with 3-methylcytosine and renders the site abasic, which can be detected as a strand scission after aniline treatment [87,88]. Using 3' ^{32}P-labeled total tRNA from HEK293T cells, hydrazine/aniline treatment yielded three major fragments that are consistent with the cleavage products of $tRNA^{Thr}$ at m^3C32, $tRNA^{Ser}$ at m^3C32, and $tRNA^{Ser}$ at m^3C47d and the LeuCAG site at m^3C47d (Figure 3.8E). These specific tRNA products were confirmed by tRNA microarrays (Figure 3.8F) [86]. These results, together with the mutation

and stop signatures at this position (Figure 3.8G), indicate that the modified C47d residue in LeuCAG corresponds to N3-methylcytosine.

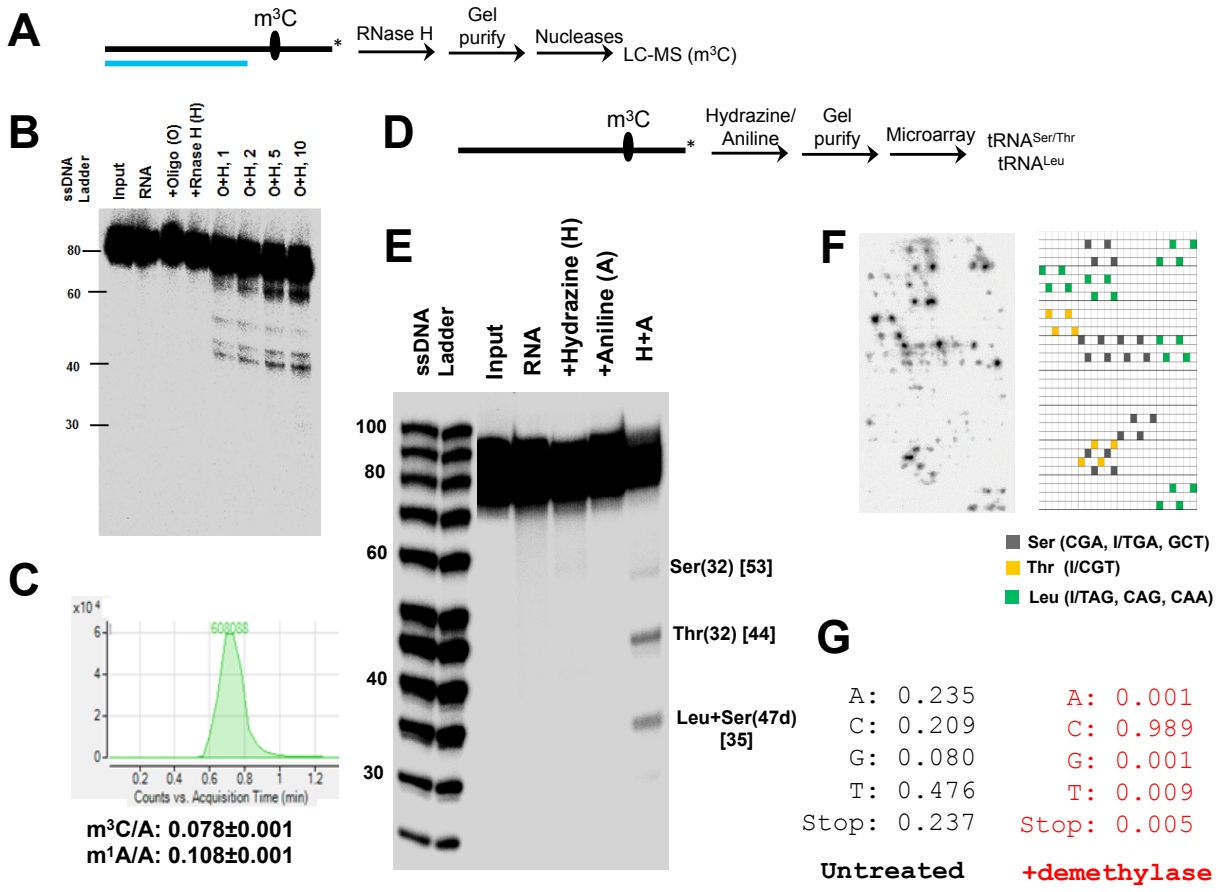


Figure 3.8: Validation of m^{3}C47d in LeuCAG. (A) Experimental strategy #1. LeuCAG RNA is shown as a black line, and the complementary DNA oligo is in blue. The star represents the location of the 3' 32P-label. (B) Optimization of RNase H cleavage reaction. 3' 32P-labeled total tRNA from HEK293T cells is shown. The amount of DNA oligos added (in pmole) is indicated in the 4 right lanes. The amount of cleavage product is ~2.1% which is within the expected range of the amount of LeuCAG RNA in the sample. The 1pmole condition is the most stringent and used in subsequent experiment. (C) QQQ LC-MS result of nuclease P1 and alkaline phosphatase (AP) treated LeuCAG fragment showing the assigned m^{3}C peak. (D) Experimental strategy #2 using hydrazine/aniline cleavage. 3' 32P-labeled total tRNA is shown as a black line. (E) Cleavage occurs only with both hydrazine and aniline treatments (right most lane). Three fragments can be readily seen. By size (nucleotides in brackets), they correspond to cleavage at known m^{3}C32 of tRNASer, m^{3}C32 of tRNAThr, m^{3}C47d of tRNASer and putative LeuCAG. (F) tRNA microarray of the cleavage products from panel (E). The array layout map is on the right. All major products are derived from tRNASer and tRNAThr and tRNALeu. For tRNALeu, the 3' sequence from 47d-3'CCA are identical for LeuCAG and LeuCAA, and differ by just 2 out of 35 nucleotides for LeuI/TAG, hence the hybridization to all three tRNALeu probes on the array. (G) The mutation and stop fractions from DM-tRNA-seq at the C47d site.

3.3.3 Assessment of quantification of methylation fractions

We calculated the MI value for all tRNA isodecoders among the top 25% expressed tRNAs according to the number of aligned reads in the untreated sample. Our DM-tRNA-seq data is of sufficient quality that the mapped reads for each of these top quartile tRNAs range from 20,000 – 700,000. Although the MI values can be precisely measured, there are several inherent caveats to interpret these as the precise modification fraction. In particular, reverse transcription based method is prone to underestimate the modification fraction to varying degrees in a context dependent manner [46-48]. In our case, this underestimation can be derived from either the inadequacy of RT stops or the ability of RT to also incorporate the correct nucleotide opposite to the modified base. The m^1A58 stop fraction is likely underestimated in our sequencing data; these RT stop reads are only 18 residues and can be easily missed in our experimental design which was aimed for sequencing longer RT products. This effect may be mitigated in future experiments by expanding the size range of RT products to be sequenced. The quantitative variation of incorporating the correct nucleotide, however, is a large obstacle in obtaining strictly quantitative information as this effect can be highly dependent on the chemical structure and the sequence context of the modified base [47,48]. For these reasons, the MI values cannot yet to be used to fully represent the precise modification fraction without further study. In this work, we choose to group the MI values in three regimes (0.15-0.50, 0.50-0.80, and 0.80 – 1.0) as a semi-quantitative metric for ease of description and discussion. Table 3.1 shows the MI groups of four methylation types in the tRNA transcriptome.

Table 3.1: Methylations identified in the human tRNAome by DM-tRNA-seq

tRNA[a]	m^1A58[b,c]	m^1G37[b]	m^2_2G26[b]	m^1G9[b]	m^3C[b]
AlaAGC	+++		+++		
AlaCGC	++[d]		+++		
AlaTGC	++[d]		+++		
CysGCA	+++[d]	+++			
AspGTC	[e]			+++ (A)	
GluCTC	+			+	
GluTTC	++			++	
PheGAA	+++, + (14)		+++		
GlyCCC	+++				
GlyGCC	+++				
GlyTCC	+++				
HisGTG	+++	++			
IleAAT	+++		+++		
IleTAT	+++[d]		+++	+++	
LysCTT	+++				
LysTTT	+++				
LeuAAG	+++	+	+++		
LeuCAG	+++	++	+++		+++ (47d)
LeuTAG	+++	+	+++		
LeuCAA	+++	+++	+++		
LeuTAA	+++	+++	+++		
Met-i	++[f]			+++	
Met-e	+++		+++		+++ (20)
AsnGTT	+++		+++	+++	
ProAGG	+++	++		+++	
ProCGG	+++	++		+++	
ProTGG	+++	+++		+++	
GlnCTG	+++			+++	
GlnTTG	+++			+++	
ArgACG	+++	+++	+++	+++	
ArgCCG	+++	+++	+++	+++	
ArgTCG	+++[d]	+++	+++	+++	
ArgCCT	+++			+++	+ (32)
ArgTCT	+++			+++	++ (32)
SecTCA	+++				
SerAGA	+++		+++		++ (32) + (47 d)
SerCGA	+++[d]		+++		++ (32) + (47 d)
SerTGA	+++		+++		+++ (32) + (47 d)
SerGCT	+++		+++		++ (32), ++ (47 d)
ThrAGT	+++		+++	++	++ (32)
ThrCGT	+++		+++	+++	++ (32)
ThrTGT	+++		+++	+++	++ (32)
ValAAC	++				
ValCAC	+++				
ValTAC	+++				
TrpCCA	+++	+++	+++	++	
TyrGTA	+++	+++	+++		

[a]Not bold, known in human or other mammalian tRNA (10). Bold, previously unknown in human and other mammalian tRNA.
[b](+) MI = 0.15–0.50; (++) MI = 0.50–0.80; (+++) MI = 0.80–1.0.
[c]MI values for m^1A58 include mutations only.
[d]Also described in Cozen et al. (2015) (32).
[e]m^1A is still present as validated by primer extension (Fig. 4B). The MI value is 0.11 in untreated and 0.06 in demethylase-treated data.
[f]MI value is 0.77, close to the boundary of the high group (0.80).

m^1A58: 40 of 46 (87%) sites are in the highly modified group, suggesting that most tRNAs are fully modified at this position. Aside from AspGTC tRNA with MI <0.15 (Figure 3.4C), the only other tRNA in the low group is GluCTC. Both Asp and Glu tRNA results are consistent with literature that these tRNAs are hypomodified at m^1A58 [78,79]. We also validate the low m^1A58 modification in $tRNA^{Asp}$ by primer extension (Figure 3.10). As described above, we were unable to obtain stop information for the m^1A58 modifications in the previous sequencing experiment.

The only other tRNA of note in this group is the initiator tRNA (Met-i) with MI value of 0.77 without the consideration of stops. In yeast, m^1A58 is known to be required for the stability of the initiator-tRNA, but not for other tRNAs [89]. It is possible that m^1A58 is also required for Met-i stability in mammalian cells. If this is the case, the MI of 0.77 is close to the high regime (≥0.80), so Met-i could still be fully modified in HEK293T cells.

m^1G37: 9 of 15 (60%) sites are highly modified. This modification is known to prevent ribosome frameshifting in specific tRNA and codon contexts [73,90], so high levels of modifications are expected. Surprisingly, 6 of 15 tRNAs are in the intermediate and low groups. This result is not fully consistent with the known function m^1G37 modification. It may be derived from the strong context dependence of MI, such that the RT is more likely to incorporate the correct nucleotide opposite the m^1G in these particular tRNAs.

$m^2{}_2G26$: Without exception, all 25 sites belong to the highly modified group. This modification is known to provide rigidity of the coaxially stacked D and anticodon stems [91] and to prevent misfolding of at least one tRNA [92,93]. Full modification at these sites would ensure that these tRNAs all have proper conformations and rigidity for translation.

m^1G9: 15 of 19 (79%) sites are highly modified. This modification is known to play a role in maintaining the structure of mitochondrial tRNAs and possibly cytosolic tRNAs as well [94]. Hence, the modification fraction is expected to be high for most of these tRNAs.

m^1A9: To our surprise, AspGTC (Figure 3.4B) has a high MI at position 9 which is reduced upon demethylase treatment, suggesting the presence of m^1A9 modification which we validated by mass spectrometry (Figure 3.7) and by primer extension below (Figure 3.10). m^1A9 is not known previously to be present in cytosolic tRNAs, but is common among mitochondrial-encoded tRNAs [44].

m^3C : These include 9 sites at C32, of which only 1 is in the highly modified group (11%); among the 5 sites at C47d located in the loop region of the hairpin in the variable arm of type II tRNAs, only one is in the highly modified group (20%); the single site at C20 in the elongator $tRNA^{Met}$, also predicted by HAMR analysis [46] is in the highly modified group. The functions of these m^3C modifications are unclear; they may perform structural roles for tRNA. However, the prevalence of many m^3C32 and m^3C47d modifications present in the low and intermediate groups suggests that m^3C modification fractions may be useful for regulatory purposes.

3.3.3.6 Mitochondrial tRNA modifications

We also analyzed the modification indexes for the 22 human mitochondrial tRNAs (Table 3.2). The complete mitochondrial tRNA modifications have been mapped in the bovine liver [44], and several human mitochondrial tRNA modifications are also known [72]. We were

able to detect all 19 expected m^1A/G9 modifications, and 16 of 19 (84%) belong to the highly modified group. m^1A/G9 in mitochondrial tRNAs is very efficiently removed by the demethylases (Figure 3.4C,D), presumably due to the lower structural stability of mitochondrial versus cytosolic tRNAs. On the basis of the bovine mitochondrial tRNA modifications, we detected m^1A58 or equivalent in mtLeu(TAA), mtLys, and mtSer(TGA), but not in mtGlu, mtIle, and mtCys. We detected all three m^1G37 modifications in mtGln, mtLeu(TAG), and mtPro, $m^2{}_2G26$ in mtIle, and m^3C32 in mtSer(TGA) and mtThr, as predicted from bovine mt-tRNAs.

Table 3.2:Methylations identified in the human mitochondrial tRNAome by DM-tRNA-seq

tRNA[a]	m^1A58[b,c]	m^1G37[b]	$m^2{}_2G26$[b]	m^1A/G9[b]	Other[b]
mtAlaTGC		+++		++	
mtArgTCG				+++	+ (m^1A16)
mtAsnGTT			+	+++	
mtAspGTC				+++	
mtCysGCA				++	
mtGlnTTG		+++		+++	
mtGluTTC				+++	
mtGlyTCC				+++	
mtHisGTG				+++	
mtIleGAT			+++	+++	
mtLeuTAA	+++			+++	
mtLeuTAG		++		+++	
mtLysTTT	+			+++	
mtMetCAT					
mtPheGAA				++	
mtProTGG		+++		+++	
mtSerGCT					
mtSerTGA	+				+ (m^3C32)
mtThrTGT				+++	++ (m^3C32)
mtTrpTCA				+++	
mtTyrGTA				+++	
mtValTAC				+++	

[a](Not bold) Known in human mitochondrial tRNA (10). (Bold) Previously unknown in human and other mammalian mitochondrial tRNA, but can be inferred from the modification maps of bovine mitochondrial tRNAs (25). (Blue) New modifications in human mitochondrial tRNA.
[b](+) MI=0.15–0.50; (++) MI=0.50–0.80; (+++) MI=0.80–1.0.
[c]MI values for m^1A58 include mutations only.

We also found three new mitochondrial tRNA modifications that could not be inferred from the bovine mitochondrial tRNAs. Human mtAla has a G after the anticodon UGC, and this G is highly modified to m^1G (Figure 3.4D). We also observe above threshold MI values for G26 in mtAsn that could be modified to $m^2{}_2G$ and A16 in mtArg that could be modified to m^1A (Figure 3.4E).

3.3.4 MI value correlations by primer extension

We applied the standard primer extension method to correlate the quantitative nature of the MI values in assessing tRNA methylations (Figure 3.10). Primer extension relies on using retroviral reverse transcriptases such as AMV RT that stop at methylated nucleotides such as m^1A. Our experimental design involves a primer that ends one nucleotide away from the modification site such as m^1A58 (Figure 3.9). The primer extension reaction is visualized by using the appropriate α-^{32}P-dNTP incorporated at position 59. When the reaction mixture contains 3 dNTPs and one specific ddNTP, the short reaction product corresponds to the stops at the modified nucleotide, and the long reaction product corresponds to the amount of unmodified tRNA. In this way, we validated the presence and the correlation of MI values from sequencing of m^1A9 in AspGTC tRNA (Figure 3.10A), as well as the low fraction of m^1A58 in AspGTC (Figure 3.10B). We also correlated the MI values for m^1A58 of GlyGCC tRNA, a tRNA known to have m^1A58 in the Modomics database and of ArgTCG tRNA, a tRNA not present in the Modomics database.

Figure 3.9: Experimental design for validation of m^1A by primer extension stops using AMV RT. The tRNA sequences for AspGTC, GlyGCC and ArgTCG are shown on top. Primer sequences are complementary to tRNA sequences and are in black. The specific mixture of one α-^{32}P-dNTP, one ddNTP and two dNTPs for each modification site is indicated above the tRNA sequence. The extended product is shown in green for non-radioactive nucleotides and in red for the location of ^{32}P nucleotide incorporated.

In all four cases, the modification indices obtained by primer extension and sequencing are similar (Figure 3.10C). The MI values from DM-tRNA-seq seem to slightly underestimate the MI values from AMV RT primer extension by up to ~10%. For m^1A58, this result may be derived from the under-counting of reads derived from RT stops at this position in our sequencing data as discussed above. Our results clearly confirm the very low m^1A58 modification levels of AspGTC at ~12%, as well as the presence of m^1A9 in AspGTC. These

results indicate that the MI values from DM-tRNA-seq are excellent parameters for estimating modification fractions.

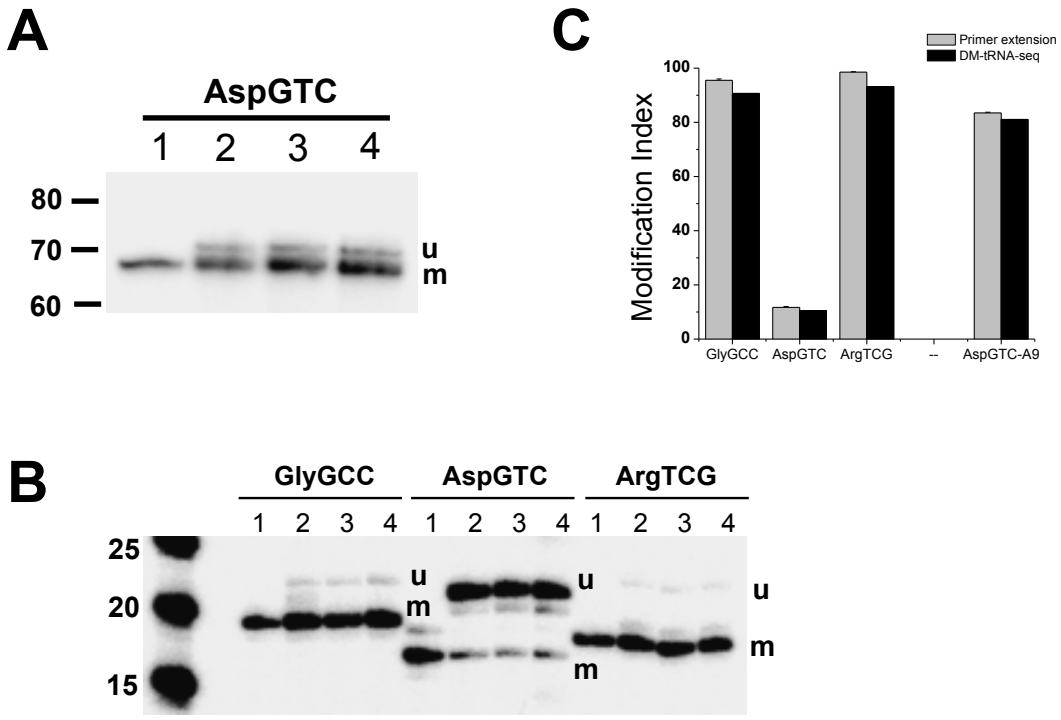


Figure 3.10: Validation of modification index by primer extension stops with AMV RT. All experiments were designed in such a way that the lower band corresponds to the amount of modified tRNA (m) and the upper band the amount of unmodified tRNA (u), see Fig. S4. Lane 1: extension using only the respective α-^{32}P-dNTP to indicate the product location. Lanes 2-4: extension from biological triplicates. (A) AspGTC for position 9, a putative m^1A identified in DM-tRNA-seq. Lane 1: α-^{32}P-dCTP only; Lanes 2-4: dATP, α-^{32}P-dCTP, ddGTP, dTTP. (B) GlyGCC, AspGTC, ArgTCG for position m^1A58. Lanes 2-4 for GlyGCC and for AspGTC: α-^{32}P-dATP, dCTP, ddGTP, dTTP. Lanes 2-4 for ArgTCG: ddATP, α-^{32}P-dCTP, dGTP, dTTP. (C) Quantitative comparison of modification fraction determined by primer extension (gray) and by DM-tRNA-seq (black).

3.3.5 Other modifications

We also looked for peaks with >15% MI values that do not change upon demethylase treatment (Figure 3.11). AlaAGC is known to contain two inosine modifications at 34 (first anticodon nucleotide) and 37, and three W-C face methylations at m^1A58, m^1I37, and $m^2{}_2G26$

[72,95]. I34 is readily apparent in the mutation graph as essentially 100% of this residue has been converted from A to I, which reads as G, and it does not respond to demethylase treatment (Figure 3.11B). m^1A58 and $m^2{}_2G26$ can be easily picked out due to the reduction in peak height upon demethylase treatment. m^1I37 leads to mostly stops in the untreated sample; demethylase fully removes the methyl group so that the RT stop is fully eliminated; at the same time, the mutation fraction goes up to nearly 100% because of the A37 to I conversion. AsnGTT is known to contain two consecutive dihydrouridines in the D-loop [72]. Each D modification is known to weaken W-C base pairing by ~1 kcal/mol [96], so that the presence of two consecutive Ds leads to strong RT stops that do not respond to demethylase treatment (Figure 3.11C). LysTTT is known to contain a bulky modification, 2-methylthio-6-threonylcarbamoyl-A (ms^2t^6A) at A37 [72]. This modification leads to a ~98% stop in the RT reaction which does not respond to the demethylase treatment (Figure 3.11D). Because we are starting out with ~300,000 reads for this tRNA (Figure 3.11D inset), we are still able to obtain ~6,000 reads that process past the modification, and we can assess potential modifications upstream of ms^2t^6A37 after this strong stop. Mitochondrial $tRNA^{Tyr}$ is known to contain another bulky modification, 2-methylthio-6-isopentenyl-A (ms^2i^6A) at A37 [72]. This modification leads to a ~95% stop in the RT reaction which does not respond to the demethylase treatment (Figure 3.11E). Similarly, the high number of read counts for this tRNA (Figure 3.11E inset) still enables the analysis of m^1G9 modification in mtTyr.

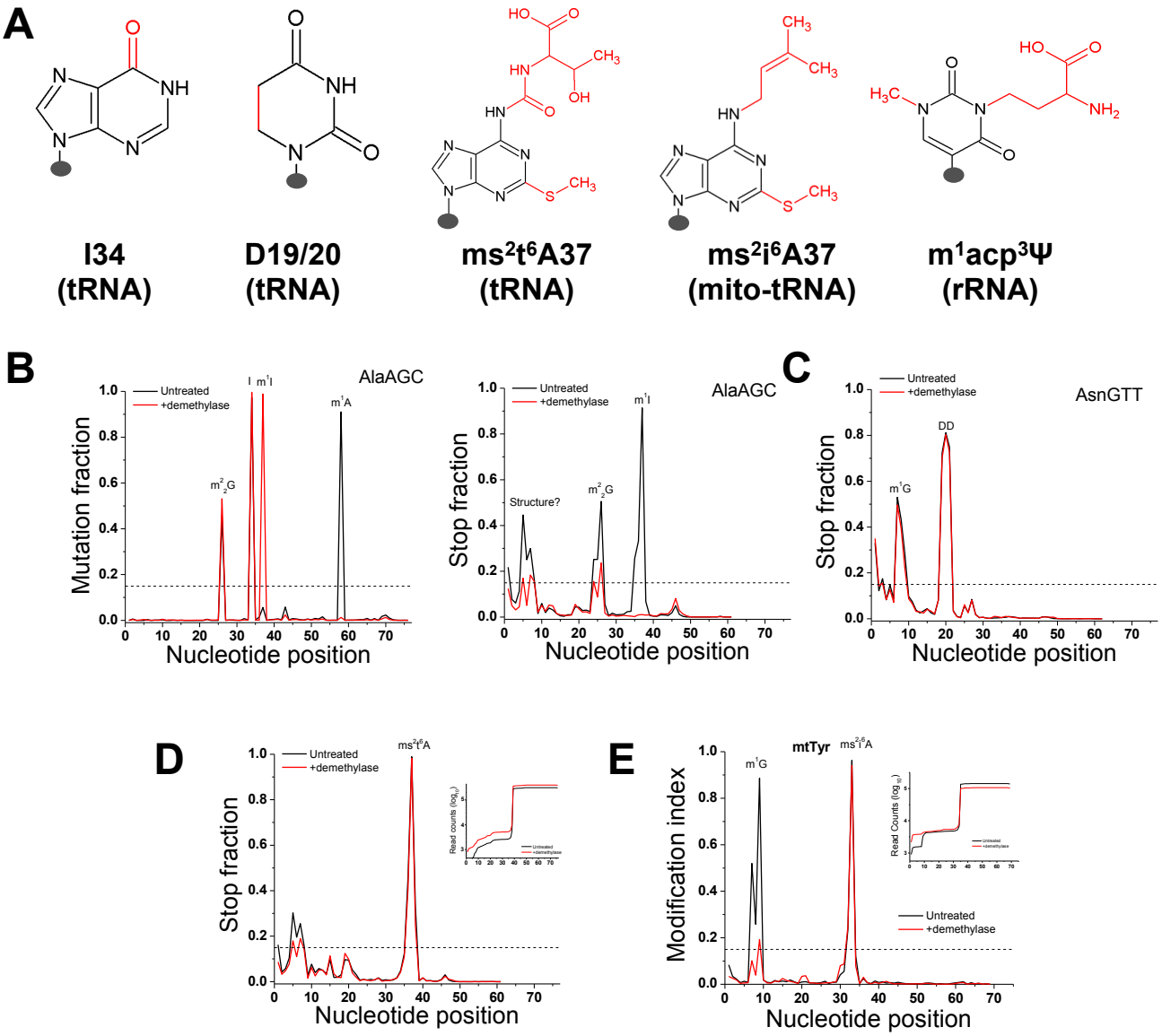


Figure 3.11: Plots for individual tRNAs with other methylations from DM-tRNA-seq. The MI value of these modifications does not change upon demethylase treatment. (A) Chemical structure of the 5 modifications where the chemical group is shown in red. The presence of each modification in tRNA and/or rRNA is listed beneath each base. (B) Mutation and stop plots showing AlaAGC with the known m^1I37 and I34 modifications. (C) Stop plot showing the consecutive dihydrouridine modifications (DD) in AsnGTT. (D) Stop plot for LysTTT showing the ~98% stop at ms^2t^6A37. Inset (y-axis = log10 counts along the LysTTT tRNA) shows that even with such strong stop, sufficient number of counts is still present for the analysis of modifications in this tRNA upstream of this stop. (E) MI plot for mitochondrial tRNATyr showing a ~95% stop at ms^2i^6A37. Inset (y-axis = log10 counts along the mtTyr tRNA) shows that even with such strong stop, sufficient number of counts is still present for the analysis of the m^1G9 modifications in this tRNA upstream of this stop.

3.3.6 Modification index heat maps for all abundant tRNAs

We also show mutation and stop fractions across all tRNA isodecoders that are present in the top 25% abundant tRNAs as heat maps ($tRNA^{Leu}$ isodecoders in Figure 3.12; tRNA isodecoders for other amino acids not shown). To enhance presentation, all nucleotide positions in the heat map (x-axis) are converted to standard tRNA nomenclature so that the anticodon nucleotides are always numbered 34-36, D loop nucleotides 14-20, and variable loop nucleotides 44-48. Human $tRNA^{Leu}$ contains five distinct isoacceptors with the anticodons of AAG, CAG, TAG, CAA and TAA. In the mutation graph, the untreated samples show high mutation values at m^1A58, m^1G37, m^2_2G26 for all isodecoders, m^3C47d for the two CAG-isodecoders, and I for the two AAG-isodecoders. As expected, the m^1A58, m^1G37, and m3C47d modifications are significantly reduced upon demethylase treatment. m^1G37 and m^2_2G26 also show significant stops in the untreated sample, but the stops are substantially reduced for m^1G37, and moderately reduced for m^2_2G26.

Applying the above criteria, all identified m^1A, m^1G, m^2_2G, and m^3C sites in the tRNA transcriptome for 47 isoacceptors from HEK293T cells are shown in Table 3.1 (the remaining 2 annotated isoacceptors in the genomic tRNA database are excluded here because they are present at very low levels). They include 12/46 new m^1A58 (26%), 4/15 new m^1G37 (27%), 10/25 new m^2_2G (40%), 6/21 new m^1G9 (29%), and 7/15 new m^3C (47%) sites not present in the human and other mammalian tRNAs in the Modomics database [72].

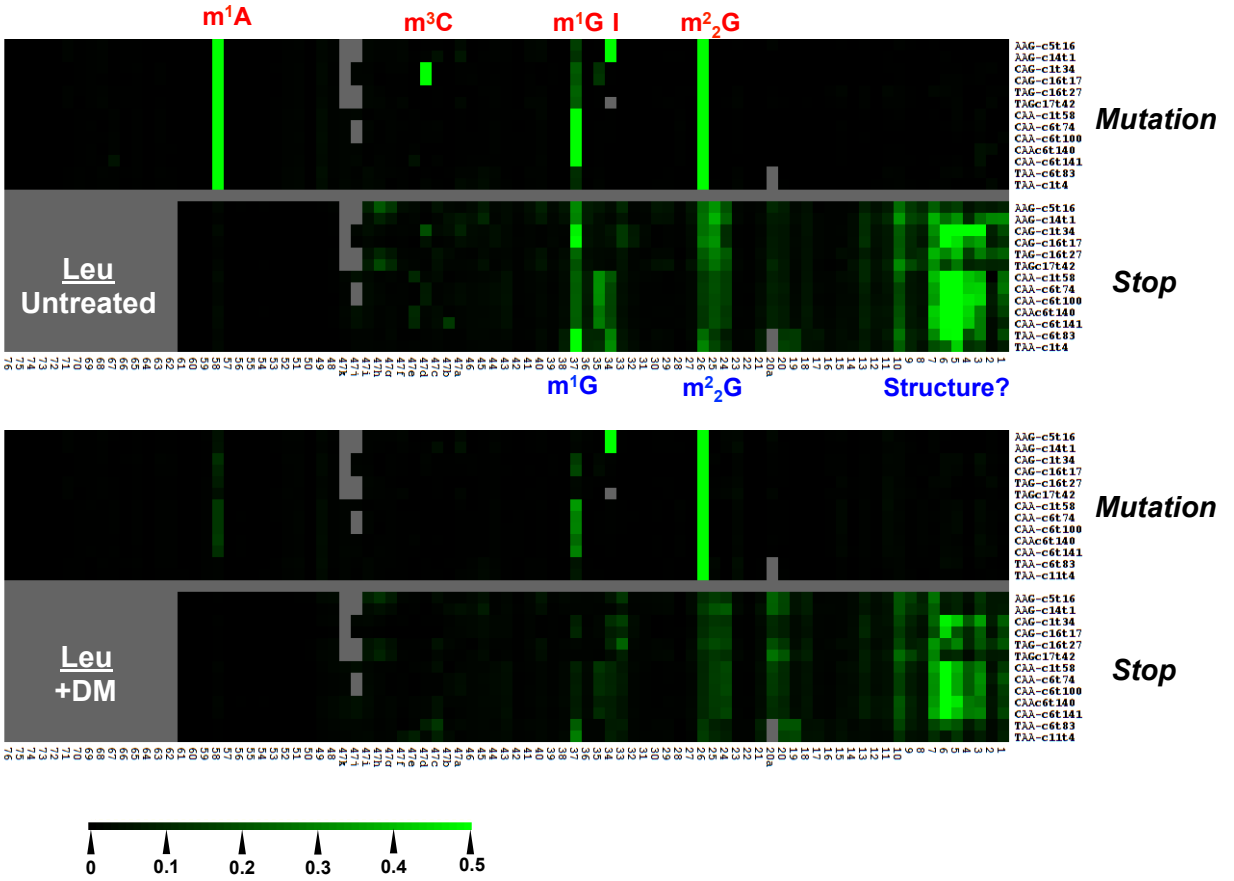


Figure 3.12: Mutation and stop heat maps for all abundant isodecoders in the tRNALeu family. The x-axis indicates the tRNA position according to the standard tRNA nomenclature so that the anticodons are always numbered 34-36, and the last nucleotide is always 76. Nucleotides not present according to this nomenclature are shown in gray. Stops can only be analyzed up to nucleotide 61 so nucleotides 62-76 are shown in gray. The scale of the heat map is shown on the lower left. Modifications are annotated in orange or in blue according to their identification through mutation or stop component, respectively. Only the isodecoders that are among the top 25% most abundant tRNA species are shown.

3.3.7 rRNA modifications

To extend the application of DM-RNA-seq, we applied it to sequence human rRNAs from HeLa cells. Human rRNA has four known modifications present at the Watson-Crick face of nucleobases: 1-methyl-3-(3-amino-3-carboxypropyl) pseudouridine ($m^1acp^3\Psi$, Figure 3.11A) and N6,N6-dimethyladenosine ($m^6{}_2A$) in the 18S rRNA, m^1A and 3-methyl uridine (m^3U, Figure

3.1A) in the 28S rRNA. The DM-RNA-seq method was able to detect three of these modifications (Figure 3.13). The $m^6{}_2A$ site was not accessible in our present study because it is located 19 nucleotides away from the 3' end of 18S rRNA, such that very short sequencing reads will be needed to obtain stop fraction. The m^1A site shows a very high MI value of 0.976 contributed about equally from mutations and stop; this high MI value is reduced to background levels upon demethylase treatment as expected (Figure 3.13A). The rRNA m^1A result indicates that both mutations and stop can contribute significantly to the m^1A modification signature. The m^3U site shows a MI value of 0.713 mostly derived from mutations; it is also reduced to background levels upon demethylase treatment (Figure 3.13B), indicating that our demethylase mixture also acts on m^3U. The $m^1acp^3\Psi$ site shows a MI value of 0.885 contributed about equally from mutations and stop; it does not respond to demethylase treatment (Figure 3.13 C).

Both the m^1A site in the 28S rRNA and the $m^1acp^3\Psi$ site in the 18S rRNA are previously known to be fully modified [97,98]. The very high MI values of 0.885-0.976 at these two sites nicely reflect these high modification fractions.

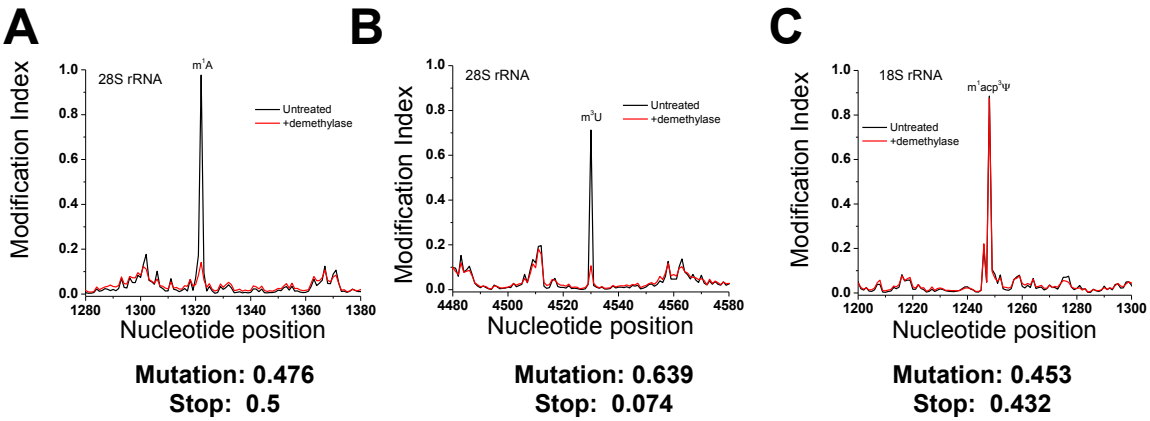


Figure 3.13: MI plots for rRNA modifications. Black line: untreated, red line: demethylase treated samples. The mutation and stop component of each modification is shown below the graph. (A) The 100 nucleotide region in the 28S rRNA around m^1A1322. (B) The 100 nucleotide region in the 28S rRNA around m^3U4530. For both (A) and (B), demethylase treatment reduces the MI to the background levels. (C) The 100 nucleotide region in the 18S rRNA around the m^1acp^3-Ψ modification at 1248. The black and red lines overlap, indicating that the demethylases do not act on this modification.

3.3.8 Discussion

This work demonstrates that RNA methylations at the Watson-Crick face can be precisely identified and, importantly, their modification fractions assessed by the newly developed DM-tRNA-seq method. We show that the high extent of readthrough of the modifications by the thermophilic group II intron RT (TGIRT) generates adequate amount of reads to enable transcriptome-wide analysis of 5 tRNA and 2 rRNA methylations. The tRNA methylations we detect and assess quantitatively cover approximately one third of all human tRNA modification sites. We were able to identify many previously known sites in the widely used RNA modification databases as well as new sites for which no prior experimental data exist. Examples of completely novel sites include the m^3C47d site in LeuCAG and the m^1A9 in AspGTC that we validated by sequencing-independent methods (Figures 3.7 and 3.8).

Comparing the parallel sequenced, demethylase treated and untreated data lends high confidence regarding the m^1A, m^3C, m^1G and $m^2{}_2G$ methylations that are widespread in the tRNA transcriptome. Interestingly, the particular mutation and stop patterns can also allow for initial assessment of the modification types for the same base such as m^1G and $m^2{}_2G$: m^1G is manifested by more significant stop than mutation fractions, whereas $m^2{}_2G$ is manifested by more mutation than stop signatures (Figure 3.3B, 3.3C).

Of course, vast prior knowledge of these methylations in defined positions in many tRNAs significantly aided the validation and strengthening of our method. On the other hand, human tRNA modifications are still inadequately represented in the widely used modification databases such as Modomics and the RNA Modification Database [71,72]. Of the 47 human tRNA isoacceptors shown in Table 3.1, 11 do not have any data among the human and mammalian tRNAs and 20 have data only among other mammalian tRNAs. Of the 22 human

mitochondrial tRNAs shown in Table 3.2, 17 do not have data among the human mitochondrial tRNAs. Our results clearly help fill some of these large gaps.

A unique and new feature of our method is the use of the modification index (MI) to assess the quantitative nature of each detectable modification site. The MI values from DM-tRNA-seq also correlate well with the classical primer extension stops using AMV RT for several m^1A sites (Figure 3.9). The slightly lower modification fraction obtained in sequencing compared to primer extension stops may be explained by TGIRT inserting the "correct" base some of the time at the methylation sites. Furthermore, the MI values for m^1A58 in tRNAs likely underestimates the m^1A58 modification fraction in this study due to the lack of stop information in our experimental design. As we have seen for the m^1A9 sites for mitochondrial tRNAs and the m^1A1322 site in rRNA, TGIRT does not read through m^1A all the time. In the future, the m^1A58 stop fraction should be obtained by using a longer TGIRT sequencing template for the short RNA fragments derived from m^1A58 stops.

Surprisingly, we found that numerous methylation sites seem to fall into low and medium modified categories in the HEK293T tRNA transcriptome (Tables 3.1, 3.2). This result seems to suggest that dynamic ranges exist for tRNA methylations to have a cell type and cell state dependent pattern. However, a confounding factor not investigated in detail here is the effect of sequence context of the modification site on the MI value. Recent work by the Helm lab [47,48] shows clear differences in mutation and stop fraction for m^1A with a different +1 nucleotide sequence, indicating that sequence context is an important parameter in determining the ratio of mutations and stops at each site. Without additional studies of context dependence for each modification type and site, MI values cannot yet be fully used to precisely identify modification

fractions. Despite these potential caveats, MI values can be interpreted as providing a lower bound on the modification fraction for each methylation site.

Biologically, quantitative differences in specific tRNA modifications have been well documented for the 5-methoxycarbonylmethyl (mcm^5) and 5-methoxy-carbonyl-methyl-2-thio (mcm^5s^2) U34 modifications in $tRNA^{Arg}$(UCU) and $tRNA^{Glu}$(UUC) and 5-methyl (m^5C)-C34 in $tRNA^{Leu}$(CAA) to enhance stress response [99-102]. In yeast, global levels of many modification types such as m^5C, $m^2{}_2G$, and 2' O-methyl-C (Cm) can change significantly when cells are exposed to distinct types of chemical stressors [42,103].

3.4 Conclusions

In summary, using high throughput DM-tRNA-seq data, we identified and quantified 6 base methylations and evaluated 5 other modifications in the human tRNA transcriptome and rRNA. Using the modification index metric, we assessed quantitative information on methylations at the Watson-Crick face among all isoacceptors in nuclear-encoded tRNAs and all mitochondrial-encoded tRNAs. While the functional investigation on the consequence of potential differences in modification fractions between cells is beyond the scope of this work, our result demonstrates the feasibility of using DM-tRNA-seq to investigate dynamic tRNA modification patterns.

CHAPTER 4

PREDICTING WATSON-CRICK BASE METHYLATIONS USING SEQUENCE SIGNATURES AND MACHINE LEARNING

Here we work towards developing a method to predict sites of methylation in tRNA and mRNA by high-throughput sequencing and machine learning. Using model oligonucleotides to explore the effect of sequence context on the modification index of a given RNA methylation, we define the 'modification signature' for six different modifications. This modification signature can be used to identify the position, identity, and abundance of modifications in tRNA with high accuracy, with the possible expansion into mRNA modifications.

4.1 Introduction

RNA modifications play a central role in biology. RNA modifications are present in all domains of life, and all the various types of RNAs contain modifications. These modifications are often regulatory, affecting the lifetime, localization, and function of the RNA. Although tRNAs and rRNAs contain the largest number and diversity of modifications, mRNAs, lncRNAs, and snRNAs are also modified. These modifications are sometimes dynamic and often serve regulatory roles in RNA processing, translation, and turnover (recently reviewed in [41]).

Previous methods to determine the location of RNA modifications genome-wide have relied on chemical-based methods and/or enrichment of modified RNA fragments (recently reviewed in [104,105]). Generally, chemical-based methods rely on the specific properties of the modified nucleotide to distinguish it from different nucleotide modifications and unmodified bases (e.g. Ψ reactivity with CMCT [106-109], m^5C insensitivity to bisulfite [110], N_m induction

of RT stops under certain reaction conditions [111], and N_m resistance to alkaline hydrolysis [112,113]). Generally, this method can precisely identify positions of modifications genome-wide and, when carefully implemented, can give quantitative information about modification fraction at a given genomic location. However, each modification must be looked at individually, and some modifications cannot be specifically targeted. Additionally, reagents generally cannot target 100% of modified residues while also avoiding off target effects, so the identification of the complete set of sites of modification remains difficult.

Alternatively, proteins that specifically recognize a modification can be used to enrich for modified RNA. Both antibodies that directly recognize the modification (m^5C [114], m^6A [82,83], m^1A [84,85], hm^5C [115]) and enzymes that install the modification (m^5C [116,117]) can be used to specifically pull down RNA fragments that contain the modification of interest. After sequencing of the 'input' and 'enriched' fraction, RNA fragments that are more prevalent in the enriched fraction are annotated as modified. Recently, this method has been improved to give single nucleotide resolution at sites of modification [118-120] and information about percent modification [121]. However, antibodies are only available for m^1A, m^6A, m^5C, and hm^5C, so this technique cannot be widely applied. Additionally, antibodies can often only recognize the modification in certain contexts such that antibody-enrichment will not identify all sites of modification. Antibodies can also cross react with different modified or unmodified nucleotides, leading to false positives.

As not all modifications are tractable for chemical modification, and truly specific binders only exist for a few RNA modifications, new methods must be developed in order to evaluate fractional modification across the RNA landscape. Modifications are known to cause reverse transcription stops and mutations, creating an 'RT signature' in sequencing data. This

signature includes reads that are truncated at the modification and reads that cannot be mapped to the original sequence due to 'mutations', or incorporation of the non-complementary nucleotide at the position of modification. Previous methods have used deep sequencing data to identify sites of modification through mutational RT signatures [46,122-124]. As these methods only consider the misincorporation of nucleotides in order to identify sites of modification, the methods miss the added information of RT stops.

More recently, the full RT signature (including both stop and mutation information) was used to identify sites of modification. The Helm lab has used stop and mutation data to look at m^1A in tRNA and rRNA from bacteria and human samples [47,48]. This work has shown that the RT signature for m^1A varies with the sequence context surrounding the modification. Both the stop rate and the fidelity of the RT are affected by the sequence context 3' of an m^1A modification. However, this research has focused only on m^1A, neglecting other RNA modifications. Our lab reported a method to identify new sites of many different RNA modification in tRNA by looking at the full RT signature [58]. However, we did not consider the contest surrounding the modification in our analysis.

Here, we aim to use both the RT signature as well as the sequence context to predict sites of 6 modifications using deep sequencing data by defining a 'modification signature' for each modification. This modification signature is the aggregate of the RT signatures across different sequence contexts. In order to define the modification signature, we designed model oligos to be able to query the effect of the -2, -1, +1, and +2 nucleotides on the RT signature of 6 Watson-Crick methylations: m^1A, $m^6{}_2A$, m^3C, m^1G, $m^2{}_2G$, and m^3U. In order to use this method for precise quantification, we additionally performed calibration curves with unmodified oligos.

Upon compiling RT signature information for all 256 context combinations, we were able to discern that the nucleotides surrounding the modification assert different effects on the resulting sequencing output. We found significant effects on both stop rate and mutation rate due to the +1 and +2 positions in all 6 modifications interrogated. Subsequently, in order to parse out these effects, we used pre-existing tRNA-sequencing data to query $m^2{}_2G$ and m^1G. Using a random forest machine learning algorithm, we were able to computationally identify and separate $m^2{}_2G$ and m^1G with 97% certainty across 101 G, m^1G, and $m^2{}_2G$ instances in previous sequencing data. Using this method, we can identify not only sites of modification, but also the type of modification (i.e. can distinguish between m^1G and $m^2{}_2G$), suggesting that this will be a useful tool to identify position, type, and fraction of modification in mRNA.

4.2 Material and Methods

4.2.1 Synthesis of m^3C phosphoramidite

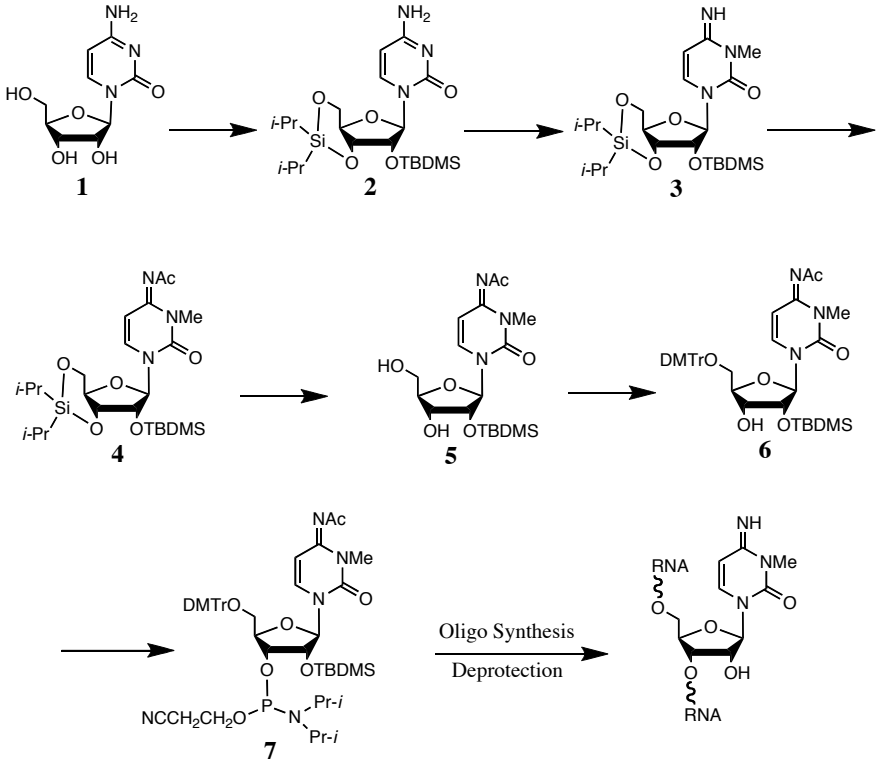


Figure 4.1: Synthesis of m^3C phosphoramidite. Once synthesized, the phosphoramidite was used in solid-phase phosphoramidite synthesis to produce the oligo listed in Table 4.1.

4.2.2 Synthesis of $m^2{}_2G$ phosphoramidite

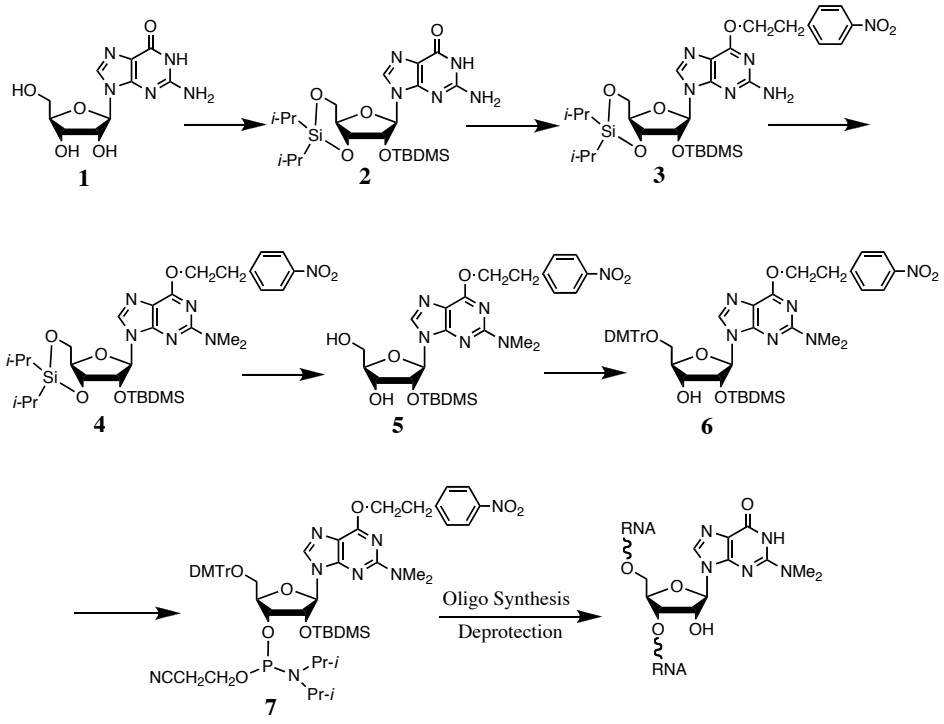


Figure 4.2: Synthesis of $m^2{}_2G$ phosphoramidite. Once synthesized, the phosphoramidite was used in solid-phase phosphoramidite synthesis to produce the oligo listed in Table 4.1.

4.2.3 Synthesis of model oligonucleotides

Table 4.1 lists the sequences of the model oligonucleotides used for this experiment. In order to test the effect of sequence context on the RT signature for the six modifications, the four nucleotides surrounding the modification (**bold**) were randomized using equal ratios of phosphoramidites during solid-phase synthesis. Each oligo was designed with a molecular barcode (*italicized*) in order to map reads directly back to the original sequence. Outside of this molecular barcode, the sequences are identical apart from the modification. Unmodified oligos and the $m^6{}_6A$ oligo were ordered from Dharmacon. The remaining modified oligos synthesized by solid-phase phosphoramidite sequencing.

Oligos were gel purified by denaturing 8% PAGE (7M Urea, 1X TBE), eluted from the gel overnight in 50 mM KOAc, 200 mM KCl. Purified oligos were ethanol precipitated, resuspended in ddH_2O, and the concentration was measured by Nanodrop.

Table 4.1: Sequences of modified oligonucleotides used in this experiment

Modification	Sequence
A	GUAAUUAUACN**N**(A)**NN**AUUCGUUGUA*ACCUAC*GCCUAAUGCCUGAA
m^1A	GUAAUUAUACN**N**(m^1A)**NN**AUUCGUUGUA*CGUGAU*GCCUAAUGCCUGAA
$m^6{}_2A$	GUAAUUAUACN**N**($m^6{}_2A$)**NN**AUUCGUUGUA*ACAAAU*GCCUAAUGCCUGAA
C	GUAAUUAUACN**N**(C)**NN**AUUCGUUGUA*CACUCU*GCCUAAUGCCUGAA
m^3C	GUAAUUAUACN**N**(m^3C)**NN**AUUCGUUGUA*CAAGAA*GCCUAAUGCCUGAA
G	GUAAUUAUACN**N**(G)**NN**AUUCGUUGUA*AUGUAA*GCCUAAUGCCUGAA
m^1G	GUAAUUAUACN**N**(m^1G)**NN**AUUCGUUGUA*CUUUGA*GCCUAAUGCCUGAA
$m^2{}_2G$	GUAAUUAUACN**N**($m^2{}_2G$)**NN**AUUCGUUGUA*AGAAUC*GCCUAAUGCCUGAA
U	GUAAUUAUACN**N**(U)**NN**AUUCGUUGUA*CCUCCU*GCCUAAUGCCUGAA
m^3U	GUAAUUAUACN**N**(m^3U)**NN**AUUCGUUGUA*GAUUAA*GCCUAAUGCCUGAA

4.2.4 TGIRT reverse transcription reactions

TGIRT primer was 5' labeled with T4 PNK. 4 pmol of 5' labeled TGIRT primer (5'GATCGTCGGACTGTAGAACTAGACGTGTGCTCTTCCGATCTT) was annealed to 4 pmol complementary RNA (5'AGAUCGGAAGAGCACACGUCUAGUUCUACAGUCCGACGAUC/3SpC3/) in 100 mM Tris-HCl, pH 7.5, 0.5 mM EDTA at 82°C for 2min, then slow cooled to room temp. 4 pmol model oligo was then added. The oligo/primer mixture (200 nM oligo, 200 nM primer) was pre-incubated at room temp for 30min in 100 mM Tris-HCl, pH 7.5, 450 mM NaCl, 5 mM $MgCl_2$, 5 mM DTT with 500 nM TGIRT (InGex, Inc.). dNTPs were added to a final concentration of 1 mM to initiate the reaction. The reverse transcription was performed at 60°C for 60min. The reactions were terminated with additions of NaOH to 0.25 M and incubation at 95°C for 3min.

The reaction was neutralized with 0.25 M HCl. At this point, the reaction was spilt into analytical (4µL) and preparative fractions (16µL).

4.2.5 SSIII reverse transcription reactions

The SSIII primer used in this experiment was identical to the TGIRT primer except instead of the 3'T DNA overhang, the primer included 14 nucleotides complementary to the final 14 nucleotides of the model oligo (5' GATCGTCGGACTGTAGAACTAGACGTGTGCTCTTCCGATCTTTCAGGCATTAGGC). The primer was 5' labeled with T4 PNK. 2pmol of 5' labeled SSIII primer was annealed to 7.4 fmol of model oligo in the presence of 0.5mM each dNTP and 25ng/µL poly(A) RNA in 14µL. The reaction was heated at 65°C for 5min and cooled on ice for 1min. Then, 4µL of 5X First Strand buffer, 1µL 100mM DTT, and 1µL 200U/µL SSIII (Thermo) was added to a volume of 20µL. The reaction was incubated at 50°C for 30 min. The enzyme was inactivated at 70°C for 15min, followed by treatment with 10U RNaseH (Epicentre) at 37°C for 20min to degrade the RNA oligo. At this point, the reaction was spilt into analytical (4µL) and preparative fraction (16µL).

4.2.6 Analysis of RT reactions

To the analytical fraction (4µL), an equal volume of 50% Formamide, 4.5M Urea, 50mM EDTA, 0.05% Bromophenol blue, 0.05% xylene cyanol was added, and the mixture was heated at 95°C for 15 min. The products were resolved by denaturing 10% PAGE (7M Urea, 1X TBE) (Figure 4.5). The gel was dried using a gel dryer at 80°C for 2hrs and was exposed to imaging plates and imaged using a phosphorimager.

4.2.7 Library prep

The preparative fractions for each modification were combined and ethanol precipitated. For the initial experiments, cDNA containing 0% modification and 100% modification were combined into a single tube – due to the molecular barcode, the reads that come from each sequence can be sorted during sequencing analysis. For the calibration curve, the reactions were combined into 6 different fractions (Table 4.2) in order to be able to map the reads back to the original oligos. Specifically, reactions that both contained the same unmodified oligo (e.g. 75% m^1A, 75% $m^6{}_6A$) were prepared separately.

Fraction	RT reactions
1	75% m^1A, 75% m^3C, 75% $m^2{}_2G$
2	75% $m^6{}_6A$, 75% m^1G, 75% m^3U
3	50% m^1A, 50% m^3C, 50% $m^2{}_2G$
4	50% $m^6{}_6A$, 50% m^1G, 50% m^3U
5	25% m^1A, 25% m^3C, 25% $m^2{}_2G$
6	25% $m^6{}_6A$, 25% m^1G, 25% m^3U

Table 4.2: Reactions contained in each of six sequencing fractions. The RT reactions for the calibration curve were partitioned such that reactions that contained the same unmodified oligo would undergo library prep in separate fractions.

Precipitated reactions were then resuspended in an equal volume of ddH_2O and 50% Formamide, 4.5M Urea, 50mM EDTA, 0.05% Bromophenol blue, 0.05% xylene cyanol and purified by denaturing 10% PAGE (7M Urea, 1 X TBE). The gel was exposed to an imaging plate for 30min-2hrs, and extended products were cut and eluted from the gel overnight in 50 mM KOAc, 200 mM KCl. Purified cDNA was ethanol precipitated with addition of linear acrylamide (Thermo) to 20 μg/mL.

Purified cDNA was then circularized using CircLigase II (Epicentre) at 60°C overnight. After inactivation at 80°C for 10 min, samples were phenol/chloroform extracted and ethanol precipitated. PCR library preparation for Illumina sequencing was performed using Phusion

Master Mix (Thermo) for 12 PCR cycles (98°C 5s, 60°C 10s, 72°C 10s). AMPure XP Beads (Beckman-Coulter) were used to clean up the libraries before Illumina sequencing.

4.2.8 Data Analysis

In order to separate out the various raw sequences by modification, MIGEC was used to characterize each set of raw sequences by the molecular barcode. Subsequent trimming was performed by Cutadapt 1.9 to remove the reverse transcriptase-specific adapters. Alignment was then performed using a library comprised of the 256 sequence permutations per modification and aligned with 1 mismatch allowed in the seed sequence. Subsequently, analysis was performed by combining specific combinations of contexts (for instance, the 16 +1/+2 nucleotide sequences) in order to analyze the stop and mutation component of respective signatures.

4.2.9 Machine learning

Code can be found at (github.com/wescclark/Modification-Signature/Manuscript_Code/). Briefly, the supervised learning algorithm of random forest ensemble was used from the sklearn libraries. The training sets of data included known m^2_2G/m^1G data previously described in [33] using the respective mutation and stop information, as well as unmodified G nucleotides from the same datasets. The training sets included equal instances of G, m^2_2G, and m^1G data. In addition, a separate validation set sites for at the 5' end was included since these sites often have a high modification index due to TGIRT's poor processivity on tRNA. The information used to train was the mutation rate, stop rate, as well as the respective ratio of inserted nucleotide (fraction of As, Cs, or Ts at a site). The forests were built upon 10 classifiers, and the ensemble information was averaged over 5 sets of training.

4.3 Results and Discussion

Previous attempts to identify modifications *de novo* from sequencing data have suffered from lack of incorporation of arrest rate as well as lack of accounting for the surrounding nucleotides [46,58,122-124]. Most recently, it was shown that the nucleotide 3' of an m^1A modification affects both the probability that the RT will stop and also the proportions of mutations [47]. In this study, model oligos were used with different nucleotides +1 to the m^1A modification but identical elsewhere. Further investigation showed a minimal effect of the -1 and +2 nucleotides, but these were only investigated in a limited context [48].

Here, we seek to define the effect of the nucleotide context surrounding different modifications. RT signatures (misincorporation and mutation rate) were queried across 256 different sequence contexts for six different Watson-Crick methylations: m^1A, $m^6{}_2A$, m^3C, m^1G, $m^2{}_2G$, and m^3U (Figure 4.3). The aggregate of the 256 RT signatures for a given modification is referred to as the 'modification signature' and is used to predict sites of modification from sequencing data. By incorporating two different modifications for A and G, we also aim to be able to identify, in addition to the position of modification, the identity of the modification.

4.3.1 Experimental design

We designed 10 model oligonucleotides (4 unmodified, 6 modified) to define the modification signatures for these modifications (Figure 4.4A). The model oligos are 45nt long and synthesized by solid-phase phosphoramidite synthesis. Novel synthesis was required for m^3C and $m^2{}_2G$ phosphoramidites (Figures 4.1, 4.2). The sequence of the RNA oligo was optimized to be mostly unstructured and compatible with Illumina barcoding and sequencing. Each oligo has a unique molecular barcode (green), so that reads can be sorted to the originating oligo. The two

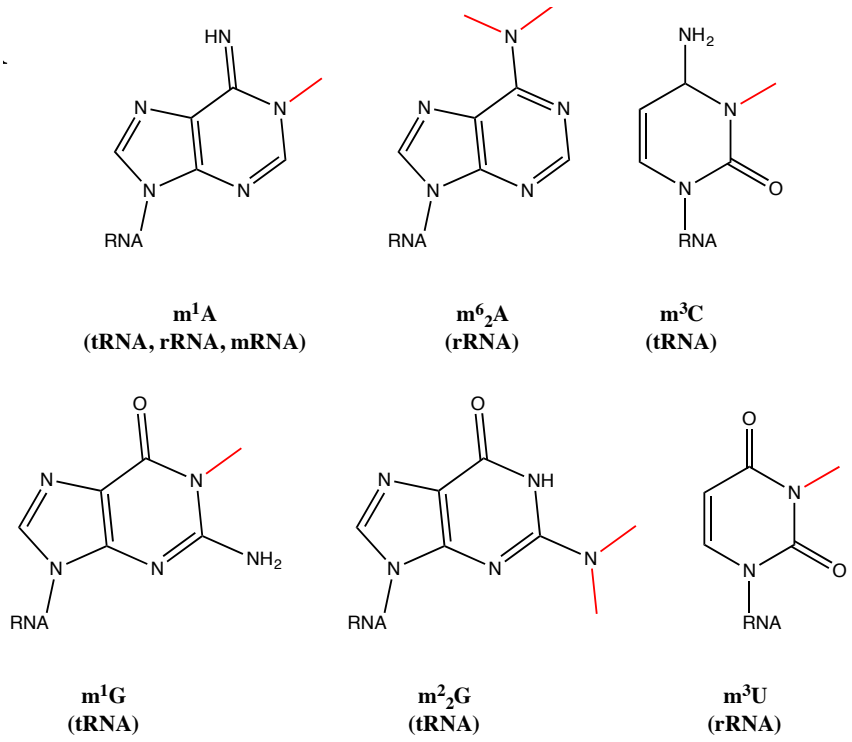


Figure 4.3: Modifications investigated in this experiment. Methylations investigated in this experiment. Six Watson-Crick face methylations have been identified in cellular RNA: m^1A, $m^6{}_2A$, m^3C, m^1G, $m^2{}_2G$, m^3U. m^1A has already been identified in mRNA, while the other five exist in tRNA and rRNA. Methylations are highlighted in red.

nucleotides 5' and 3' of the modification (X) were randomized (N) so that 256 different contexts can be interrogated.

Once the model RNAs are reverse transcribed, the reactions are combined and sequenced (Figure 4.4B). After sequencing, the reads can be sorted by the molecular barcode to assign the reads to the correct model oligo. After sorting, the reads are analyzed for fraction of stopped and mismatched reads, and the context dependence of each signature can be defined. Subsequent analysis of 16 (+1/+2) and 256 (-2/-1/+1/+2) signatures allow us to parse apart and determine the specific contributions of positional factors on misincorporation and stop to the ensemble modification index. The signatures can then be used to predict sites of modification from existing sequencing date. High stop or mutation rates in sequencing data can indicate highly structured regions and SNPs respectively. Alternatively, the high stop/mutation could indicate a position of

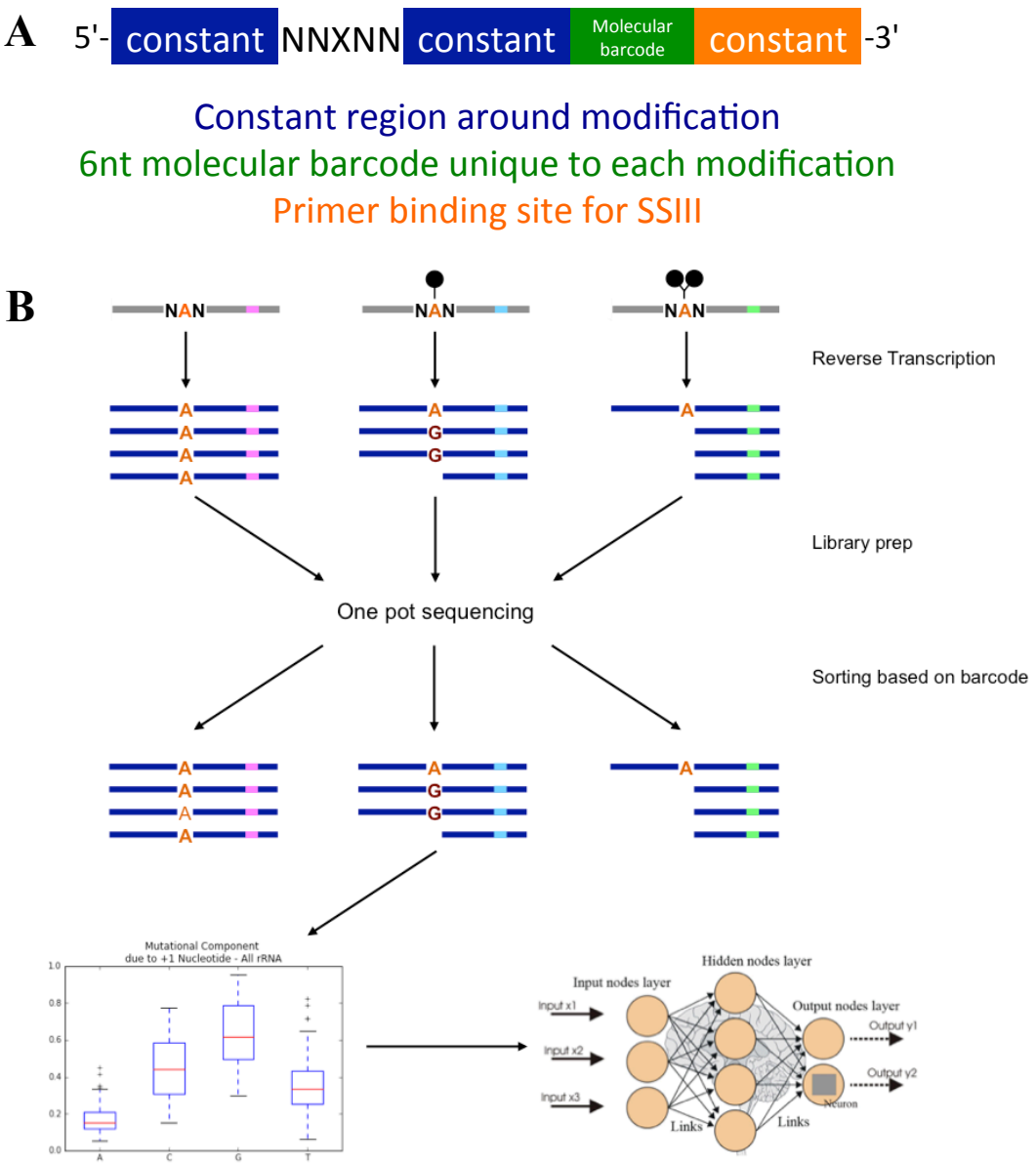


Figure 4.4: Oligonucleotide design and experimental outline. A) RNA oligonucleotide design. The 45mer includes one modified nucleotide (X) surrounded on each side by two random nucleotides (N). Each modification has a unique 6-nucleotide 'molecular barcode' that allows for sorting of reads back to the original sequence. Outside of the modification and the molecular barcode, all oligos are identical. The design was also optimized to decrease structure and to be compatible with Illumina barcoding and sequencing. The 3' end is used as a primer binding site to allow for SSIII reverse transcription. B) Experimental outline. Each RNA oligonucleotide is reverse transcribed in a separate reaction. The reactions are then combined and prepared for Illumina sequencing. Because each RNA has a 'molecular barcode', the sequencing reads can be sorted and assigned to the RNA oligo from which they are derived. This allows for the identification of the context dependent stop and mutation rate for each modification, here termed the modification signature. Using this signature and machine learning, analysis of sequencing data can yield the position and identity of modified nucleotides.

RNA modification. Utilizing modification signature information as features for supervised learning predictors, we can attempt to classify previously unknown sites as positions of modification or otherwise aberrant sequencing output. Further, if different nucleoside modifications have different signatures, the identity of the modification could also be discerned.

4.3.2 Defining the modification signature

First, the gel-purified model oligos were reverse transcribed with SSIII and TGIRT in separate reverse transcription reactions (0% modified, 100% modified) (Figure 4.5). For unmodified oligos, no stop products are visible. For 100% modified oligos, both stop and read-through products are visible in both the SSIII and TGIRT reactions. However, a higher fraction of read-through/stop is seen with TGIRT. Since the nucleotide -1/-2 context cannot be discerned from reads that are truncated at the modification, read-through cDNA gives more context-dependent information than stop products. Thus, further analysis used only TGIRT reverse transcription.

TGIRT reverse transcription followed by Illumina sequencing reveals the effect of the surrounding nucleotides on the stop and mutation rate (Figure 4.6, 4.7, 4.8, 4.9, 4.10 and 4.11). Modification index (MI) is a measure of the total effect of the modification on reverse transcription and is a simple sum of stop rate and mutation rate [58]. This metric can only be defined for 16 contexts – the sequence information for the -1 and -2 nucleotides are lost with stopped reads. For m^1G, all 16 contexts have a similarly high modification index (>0.8) but the contributions from stop rate and mutation rate vary widely (Figure 4.6A). This indicates that both the +1 and +2 positions can affect the modification signature. The same high modification index with varying contributions from stop and mutation was also seen in m^1A, m^3C, $m^2{}_2G$, and m^3U

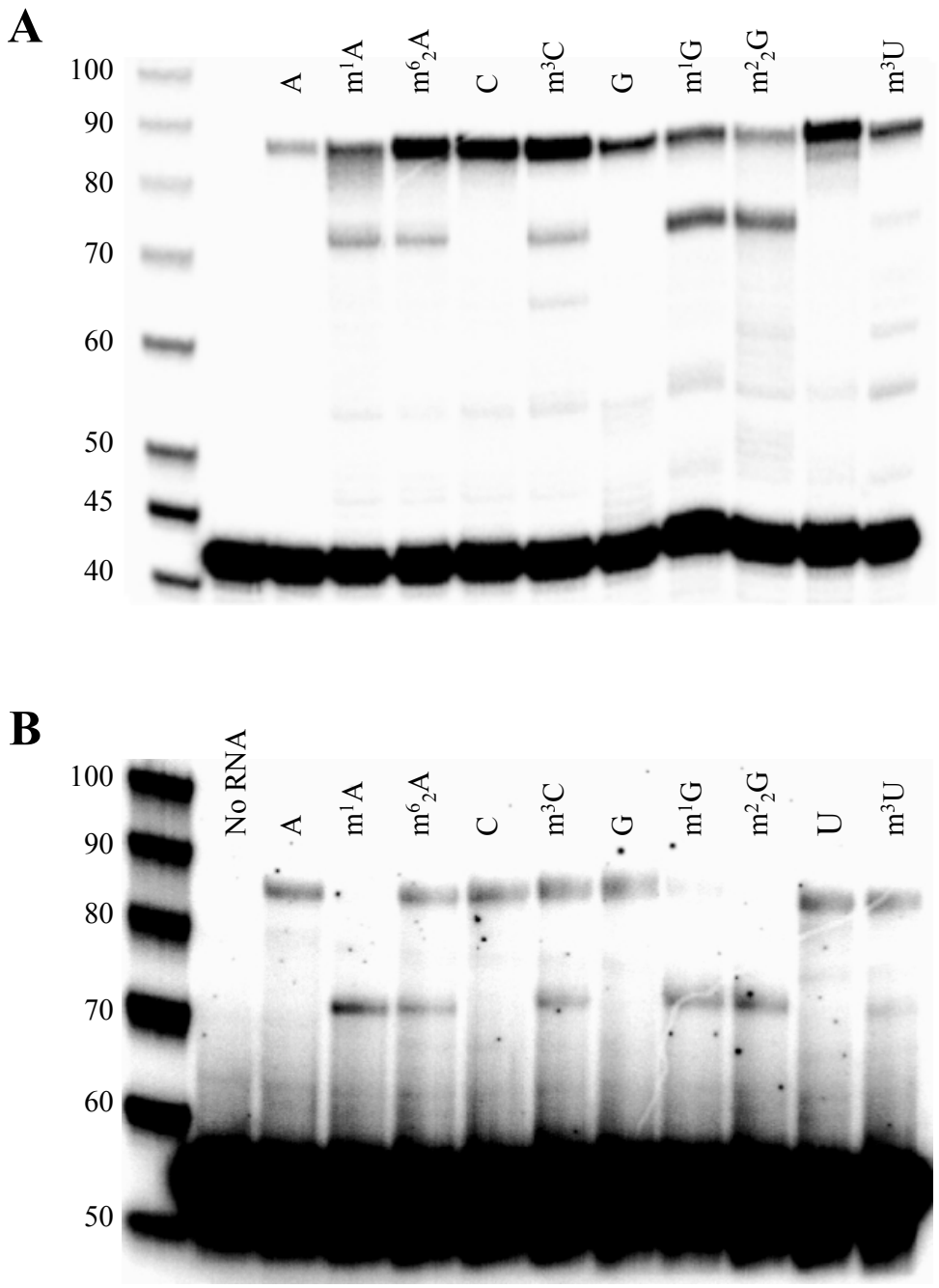


Figure 4.5: Reverse transcription of model oligonucleotides. A) TGIRT was used to reverse transcribe each RNA oligonucleotide. The stop product is 73 nucleotides and the readthrough product is 86 nucleotides. While a significant amount of modification-induced stops are observed, many full-length products are also observed, suggesting that TGIRT is efficient at reading through many Watson-Crick methylations. B) SSIII was used to reverse transcribe each RNA oligonucleotide. The stop product is 73 nucleotides and the readthrough product is 86 nucleotides. Here, most reads are modification-induced stops, as SSIII cannot efficiently read through modifications.

(Figures 4.7A, 4.8A, 4.9A, 4.10A). This data contradicts the previously reported data that the +2 position does not affect the reverse transcription signature of m^1A [48]. This previous experiment was limited by oligo synthesis, and the results do not explore the full possible contextual information. In our experiment, we are querying in depth all potential sequence contexts around the modification, so we may be able to detect effects that were not seen in their limited library. $m^6{}_2A$ did not show a high MI across contexts, and the stop rate was the main contributor (Figure 4.11A).

Next we analyzed the mutation rate for all 256 possible contexts. This metric looks only at full length reads, and the mutation rate is defined as the fraction of misincorporation of a non-complementary dNTP. The mutation rate for m^1G varied between 0.2-1.0 across the 265 different contexts (Figure 4.6B). Mutation rate can further be broken down into its nucleotide components (Figure 4.6C). From this metric, it is clear that the identity of the non-complementary incorporated nucleotide varies with the surrounding context and is influenced not only by the +1 nucleotide. m^1A, m^3C, $m^2{}_2G$, and m^3U also show large variations in mutation rate (Figures 4.7B, 4.8B, 4.9B, 4.10B) and mutation context (Figures 4.7C, 4.8C, 4.9C, 4.10C) while $m^6{}_2A$ has generally much lower mutation rate for all contexts (Figure 4.11B, 4.11C).

4.3.3 Calibration curves to determine the quantitative nature of the method

In order to determine if our method is quantitative, we performed TGIRT reverse transcription and sequencing on libraries containing fractional modification (25%, 50%, 75%) to create a calibration curve. For this, we combined and barcoded libraries in order to be able to map reads back to the original oligo (e.g. no libraries containing the unmodified A oligo were combined into one barcoding reaction) (Table 4.2). Good correlation was found between the

input fraction of m^1G and the measured fraction of m^1G (Figure 4.6D). The correlation between the input and measured fraction of the other modifications ranged from good (m^3C, $m^2{}_2G$, and m^3U) to poor (m^1A, and $m^6{}_2A$) (Figures 4.7D, 4.8D, 4.9D, 4.10D, 4.11D). However, as sequencing can be biased towards short reads, our method is only semiquantitative; precise quantification likely requires interrogation of individual sites by a more low-throughput method.

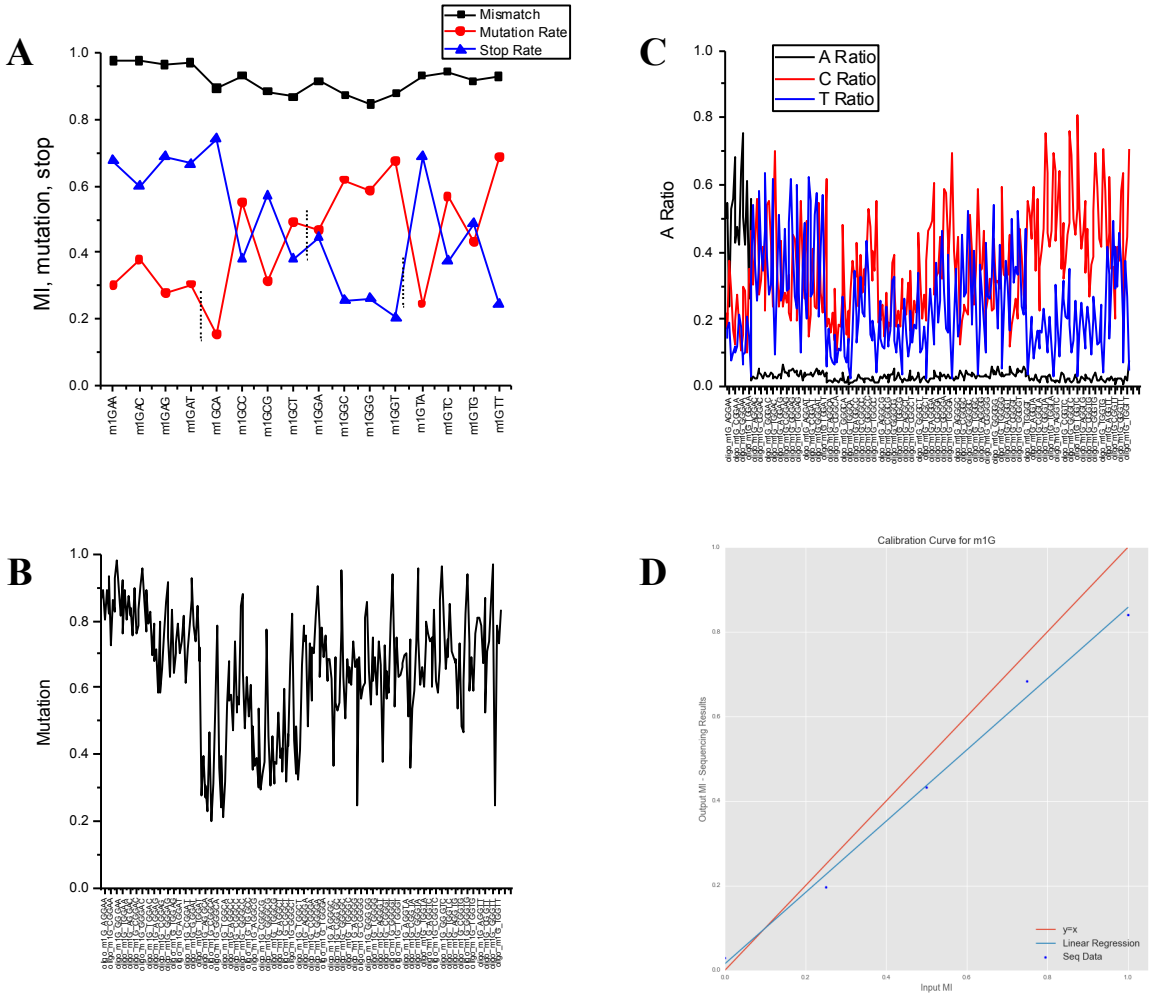


Figure 4.6: m^{1}G modification signature. A) Modification index at each of the 16 contexts. Modification index (black) is plotted for each of the 16 contexts for which it can be calculated. The two components [mutation (red) and stop (blue)] are also plotted. While the modification index is uniformly high across all contexts (>0.8), the contribution from RT stops and mutation varies widely. B). The mutation rate at each of the 256 contexts. For full-length reads, the fraction of mutations (TGIRT incorporates a non-complementary nucleotide) is plotted. The mutation rate varies widely across all contexts, from 0.2-1.0. C) The identity of the mutations is plotted for each of the 256 contexts. A, C, and T are all preferably incorporated dependent on the sequence context of the modification. D) Calibration curve. The red line is a perfect correlation (x=y). For m^{1}G, the calibration curve trends well.

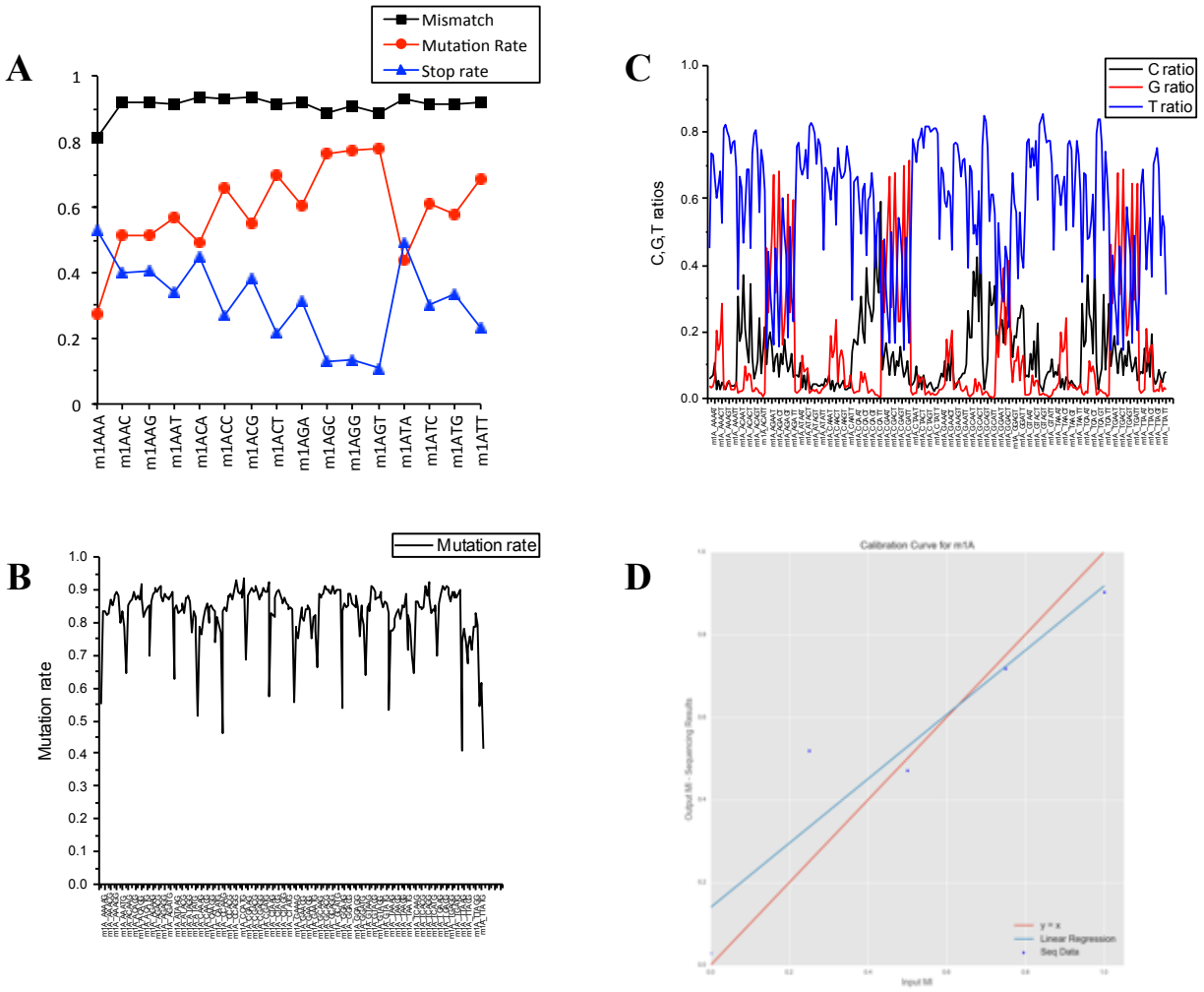


Figure 4.7: m[1]A modification signature. A) Modification index at each of the 16 contexts. While the modification index is uniformly high across all contexts (>0.8), the contribution from RT stops and mutation varies widely. B). The mutation rate at each of the 256 contexts. The mutation rate varies widely across all contexts, from 0.4-09. C) The identity of the mutations is plotted for each of the 256 contexts. G and T are both preferably incorporated dependent on the sequence context of the modification, while C is generally incorporated at lower levels. D) Calibration curve. For m[1]A, the calibration curve does not trend well, but it appears that 25% modified sample might be an outlier.

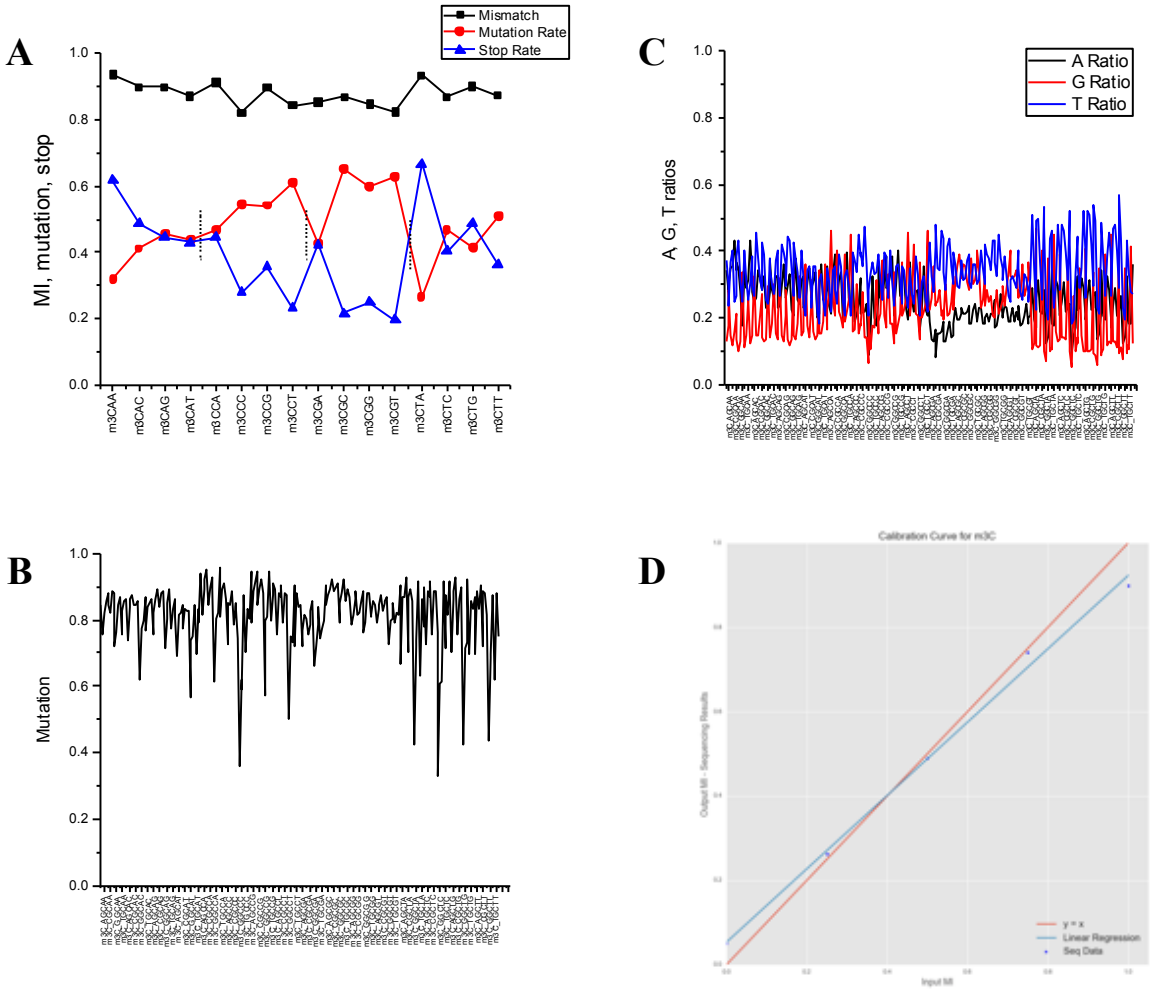


Figure 4.8: m^3C modification signature. A) Modification index at each of the 16 contexts. While the modification index is uniformly high across all contexts (>0.8), the contribution from RT stops and mutation varies widely. B). The mutation rate at each of the 256 contexts. The mutation rate varies widely across all contexts, from 0.3-09. C) The identity of the mutations is plotted for each of the 256 contexts. Here, the three nucleotides are fairly evenly incorporated, although the exact fractions do vary between contexts. D) Calibration curve. For m^3C, the calibration curve trends well.

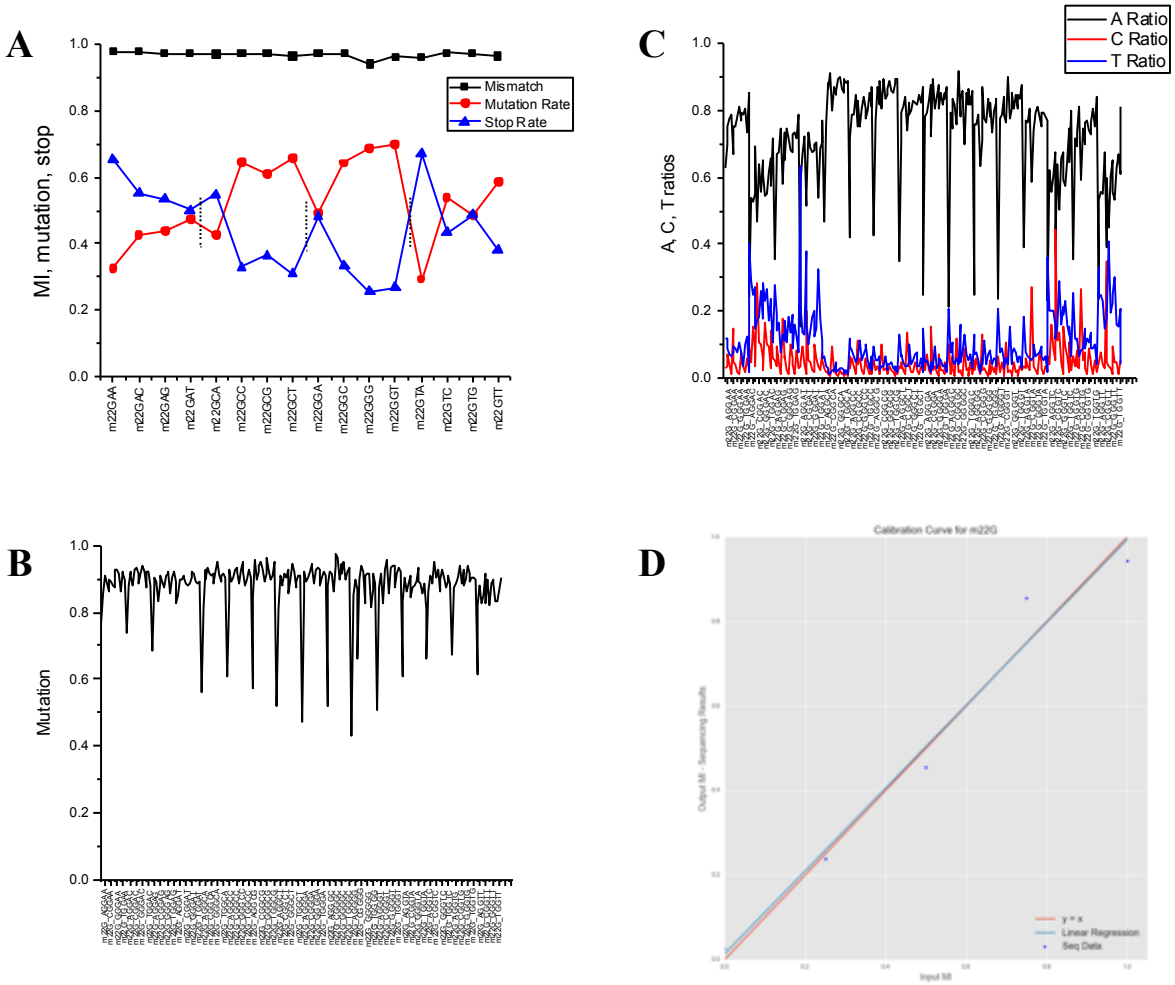


Figure 4.9: $m^2{}_2G$ modification signature. A) Modification index at each of the 16 contexts. While the modification index is uniformly high across all contexts (>0.9), the contribution from RT stops and mutation varies widely. B). The mutation rate at each of the 256 contexts. The mutation rate varies widely across all contexts, from 0.4-09. C) The identity of the mutations is plotted for each of the 256 contexts. A is preferentially incorporated independent of the context, while the amount of C and T incorporation varies between contexts. D) Calibration curve. For $m^2{}_2G$, the calibration curve trends well.

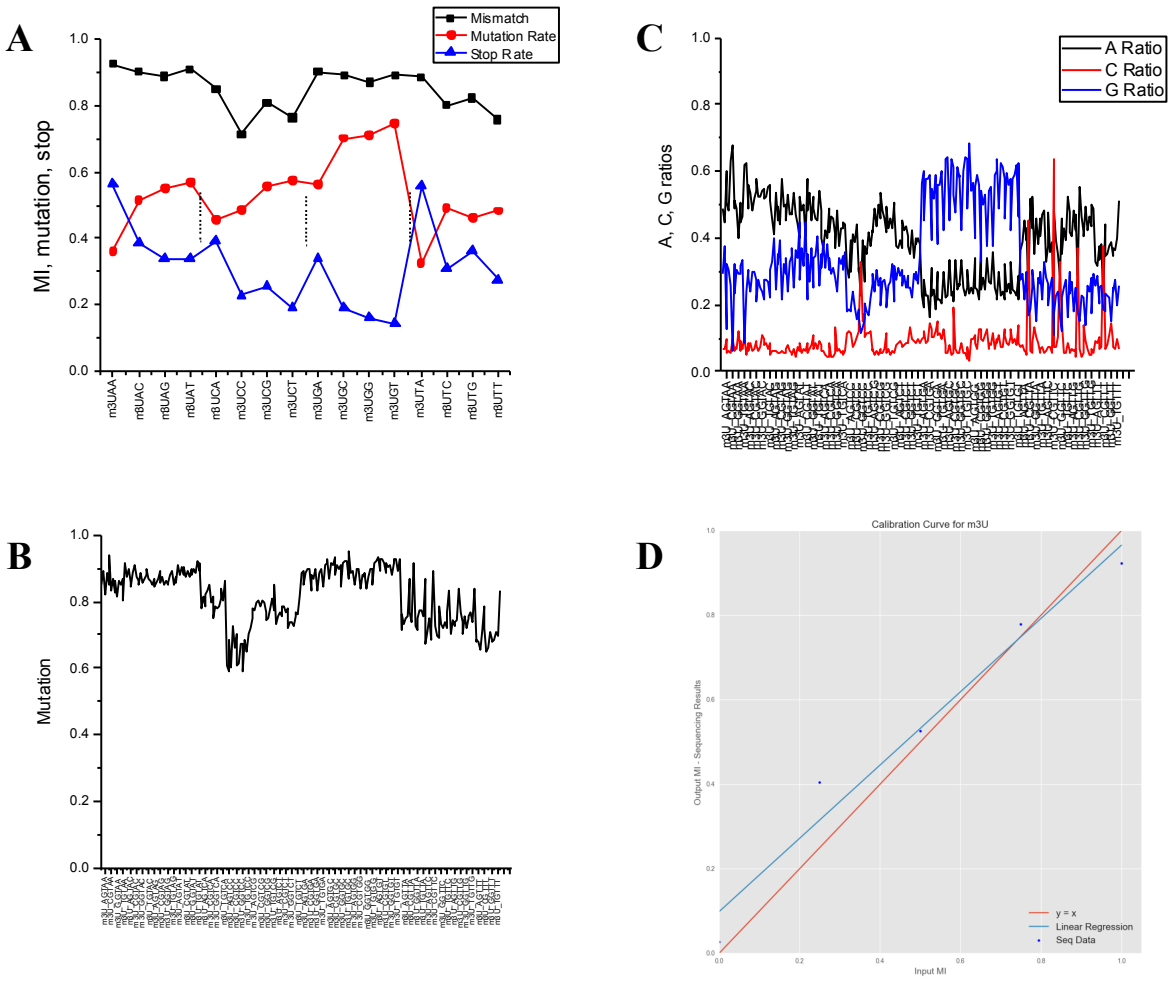


Figure 4.10: m^{3}U modification signature. A) Modification index at each of the 16 contexts. While the modification index is quite high across all contexts (>0.7), the contribution from RT stops and mutation varies widely. B). The mutation rate at each of the 256 contexts. The mutation rate varies across all contexts, from 0.6-09. C) The identity of the mutations is plotted for each of the 256 contexts. A, C, and G are all preferably incorporated dependent on the sequence context of the modification. D) Calibration curve. For m^{3}G, the calibration curve trends well.

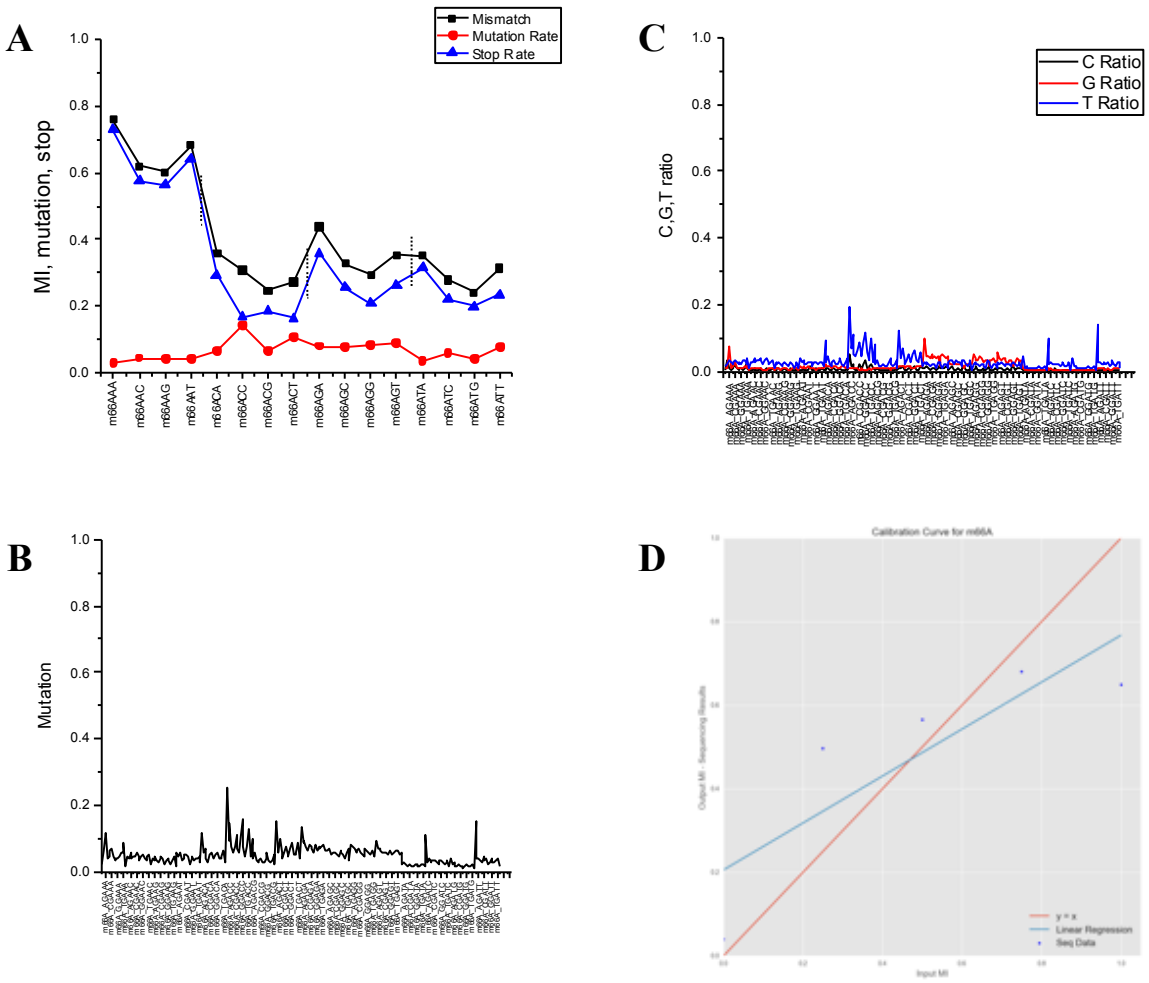


Figure 4.11: $m^6{}_2A$ modification signature. A) Modification index at each of the 16 contexts. The modification index is significantly lower than the other modifications across most contexts. Additionally, most of the contribution to modification index is from the stop component. B). The mutation rate at each of the 256 contexts. The mutation rate is quite low compared to the other modifications investigated (<0.3). C) The identity of the mutations is plotted for each of the 256 contexts. G and T are both preferably incorporated dependent on the sequence context of the modification while C is less efficiently incorporated. D) Calibration curve. For $m^6{}_2A$, the calibration curve does not trend well due to the high variability in the stop data between the 16 contexts.

4.3.4 K-means clustering of 256 m^1A contexts based on misincorporation rates

Next, to ascertain the context dependent nucleotide incorporation at the site of modification, all 256 contexts were plotted for m^1A using the non-A incorporation ratios (C, G, and T incorporations at site 0, Figure 4.12). Subsequently, K-means clustering algorithms were utilized, with grid search tuning in order to properly adjust our learning estimator. For m^1A modification, six separate clusters were found to best separate the data based only on non-complementary nucleotide incorporation at the site of modification. One of the best clustering cases finds that the -1 nucleotide for m^1A directly influences non-complementary incorporation in two cases. One cluster shows that a large fraction of -1G sites (21 out of 64 possible

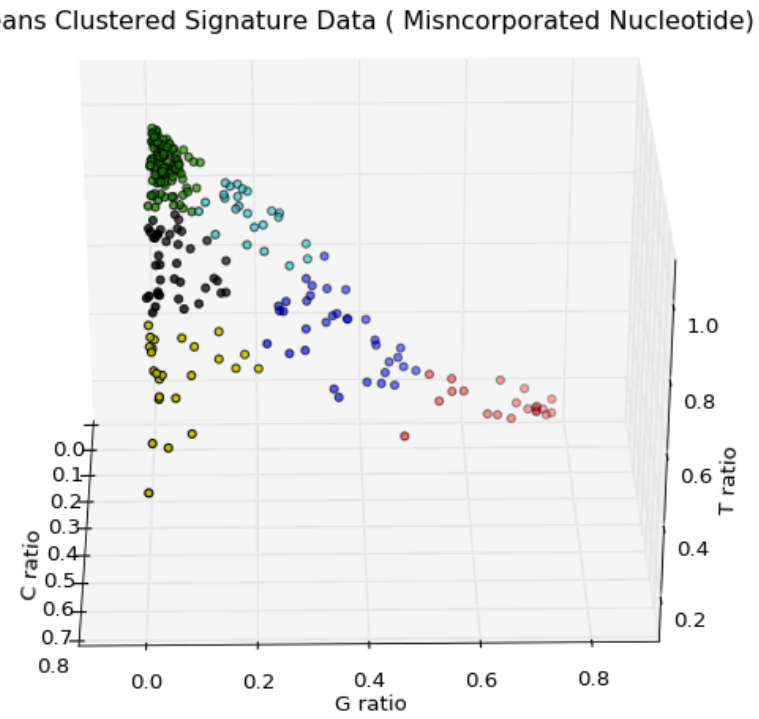


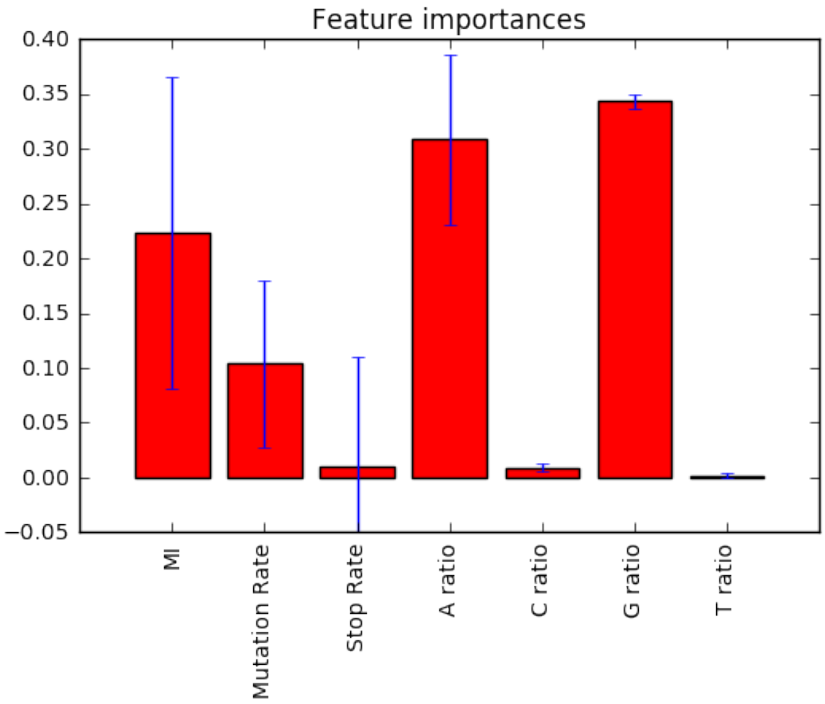
Figure 4.12: K-means clustering reveals importance of the -1 nucleotide identity on misincorporation at m^1A sites. The amount of each misincorporation is plotted on a 3D graph for all 256 contexts. Clustering was used to search for potential influences of sequence context on the misincorporation. The red cluster includes 21/64 possible NG(m^1A)NN contexts that have a high G incorporation at position 0 with low C and T incorporation. The yellow cluster includes 22/64 NC(m^1A)NN contexts that have high C incorporation at position 0 with low G and T incorporation. These two clusters suggest that the context of the -1 nucleotide can affect the misincorporation of non-complementary nucleotides at m^1A modifications.

NG(m^1A)NN contexts) promote G incorporation by TGIRT at the site of m^1A modification (Figure 4.12, red cluster). Another cluster shows that a large fraction of -1C sites (22 out of 64 possible NC(m^1A)NN contexts) promote C incorporation by TGIRT at the site of m^1A modification (Figure 4.12, yellow cluster). Thus, the -1 nucleotide can influence the misincorporation of nucleotides at the site of m^1A modification, contrary to previous studies.

4.3.5 Machine learning to predict identity of tRNA G, m^1G, and m^2_2G nucleotides

Finally, to see if we could utilize the parameters from our analysis to train machine learning, we used known sites of m^2_2G and m^1G modification from our previously published tRNA-sequencing data. We extended previously published supervised learning methods [47,48,125] to determine more than just presence or absence of a modification. We hoped that with modification signatures for multiple modifications, we could determine not only presence/absence of a modification but also the identity of the modification (e.g. m^2_2G or m^1G). We tuned our random forest estimator using the training parameters of non-G incorporated nucleotides, mutation ratio, stop ratio, and full modification index – the last metric, while seemingly redundant, allows for further separation (Figure 4.13A) and was a statistically significant contributor in generating decision branches. For separating unmodified sites from m^1G and m^2_2G, our composite accuracy was 97%, only mislabeling one m^1G site as a m^2_2G modification (Figure 4.13B). Finally, TGIRT has been shown in our data to wildly misincorporate toward the 5' end of tRNA due to an assumed decreased processivity caused by the tRNA structure. This leads to high MI sites near the 5' end of tRNA due to either the RT falling off before completely transcribing the tRNA or misincorporating non-complementary nucleotides. So, in order to challenge our random forest model, we included equal parts

A



B

	Predicted Identity		
Actual Identity	Unmodified	m^1G	$m^2{}_2G$
Unmodified	34	0	0
m^1G	0	38	1
$m^2{}_2G$	0	0	28

C

	Predicted Identity			
Actual Identity	Erroneous	Unmodified	m^1G	$m^2{}_2G$
Erroneous	29	0	3	4
Unmodified	0	29	0	0
m^1G	1	0	28	0
$m^2{}_2G$	0	0	0	21

Figure 4.13: Using machine learning to predict modification identity in tRNA. A) Contributions of different parameters to the classification of positions as unmodified G, m^1G, or $m^2{}_2G$. Although both stop rate and mutation rate are included as training parameters, modification index still contributes to the overall classification of positions. B) Results of the machine learning classification of positions in tRNA. The accuracy was very high, only misclassifying one m^1G as $m^2{}_2G$. C) Results of the machine learning classification of positions in tRNA with the addition of erroneous sequences caused by decreases in TGIRT processivity. The algorithm is correctly able to identify 29/36 erroneous sites correctly while still maintaining high accuracy of G/m^1G/$m^2{}_2G$ classification.

unmodified, m^1G, $m^2{}_2G$, and these erroneously flagged 5' tRNA sites. Our model still predicted m^1G and $m^2{}_2G$ from either unmodified or erroneous sites with 93% accuracy (Figure 4.13C), showing that the features generated from our sequence analysis were sufficient for modification classification.

4.3.6 Discussion

As the significance of RNA modifications in biology becomes increasingly obvious, the need for genome-wide identification of sites of modifications is critical. Current methods rely on selective enrichment or chemical-targeting of the modification. These methods can only identify one modification at a time and suffer from weaknesses (recently discussed in [105]). Both selective enrichment and chemical-targeting can suffer from imperfect specificity (false positives). Since modifications are rare compared to unmodified nucleotides, if a reagent or protein-binder can recognize an off-target (but more abundant) nucleotide, the amount of false positives can be quite high. Additionally, both methods can also suffer from incompleteness (false negatives). If proteins only recognize a certain context of modification, many sites of modification will not be identified. Similarly, ensuring that chemical targeting reaches 100% (without off-target effects) regardless of the structure or context is essential for identifying the complete set of modified sites. Finally, selective enrichment and chemical-targeting protocols have only been developed for certain modifications.

A sequencing-based method that relies on the modification signature presents distinct advantages. Many different modifications can be queried in the same sequencing reaction, given that there are distinct signatures for each modification. In addition, if a signature can be defined (through similar work to the current paper), the modification can be investigated. Here we show

that the 256 different contexts surrounding the modification can sufficiently define the signature for that modification for 6 different Watson Crick face methylations. Here, we use our model to correctly separate unmodified sites from m^1G and $m^2{}_2G$ with 97% accuracy.

While tRNA modifications are easy targets for identification, modifications within mRNA present new challenges for sequence based-identification of the presence of modifications. tRNA modifications are abundant, and many modifications are near 100%. Further, the sequence diversity of tRNAs is low compared to mRNA, so the number of reads needed to sufficiently cover tRNAs for confident identification of modifications is lower than mRNA. In addition to the deep sequencing required to sufficiently cover mRNA, modifications in mRNA are much less prevalent, and the modification fraction at a given site is typically less than 50% [121]. However, with detailed modification signatures, the identification of mRNA modifications with high quality sequencing data can likely be achieved.

4.4 Conclusions

In summary, we work towards a method to determine the position, identity, and fraction of modified residues in mRNA. By determining the RT signature for 256 different contexts surrounding 6 Watson-Crick methylations, we define six different modification signatures. We show that there are large variations in both the mutation rate and stop rate in cDNA due to the surrounding nucleotide identity for all six modifications. Furthermore, we validate our results with the tRNA modification sites previously described and demonstrate the capability of our machine learning prediction algorithm to recognize not only positions of modifications but also the identity of the modification. Moving forward, we aim to use these 'modification signatures' to

evaluate existing sequencing data sets in order to identify new sites of RNA modifications in mRNA and lncRNA.

CHAPTER 5

CONCLUSIONS, CHALLENGES, AND FUTURE DIRECTIONS

5.1 Conclusions

Here, we present three methods that extend the functionality of RNA high-throughput sequencing. First, a method was developed to determine the tRNA aminoacylation level by high-throughput sequencing. Previous microarray methods used chemistry to distinguish between charged and uncharged tRNAs. We harnessed and extended this chemistry to develop a one-pot sequencing method to determine tRNA aminoacylation levels in mammalian cells. Second, a method to identify and quantify tRNA modifications by high-throughput sequencing was developed. As modifications cause both cDNA stops and mutations, a 'modification index' that combined the two parameters was defined at every position for each tRNA. Using this index and sensitivity to demethylase treatment, modifications were identified and quantified. Finally, we report a method towards predicting modifications in mRNA. Using oligonucleotide libraries, we queried the effect of the surrounding context on the modification index for six different modifications. We defined the 'modification signature' as the context dependent modification index. This, in turn, was used to identify the position, identity, and abundance of modifications in tRNA, with the possible expansion to mRNA.

These three methods contribute to the functionality of RNA-seq to answer important biological questions. tRNA aminoacylation levels in isodecoders and tRNA and mRNA modification position, identity, and quantification are three biological parameters that effect many aspects of cellular biology. By accurately defining these parameters, we move towards being able to better understand biological phenomenon and predict cellular responses.

5.2 Future directions

A few questions arise from the small amount of data collected from the initial experiments performed in the development of these three methods. The first is why $tRNA^{Ser}$ and $tRNA^{Thr}$ have a lower overall charging than other tRNA species. Similar results were recently reported for *E. coli* [51], but it is unclear if the molecular details would be conserved. Additionally, some specific tRNA isodecoders had particularly low charging levels. It would be interesting to investigate whether these isodecoders are involved in 'moonlighting' functions outside of translation. Development of the method to identify and quantify tRNA modifications showed many partially modified sites in tRNA. It would be interesting to determine if partial modification is functional. Additionally, we would like to know if incomplete installation and/or partial removal of the modification are responsible for the fractional modifications. The first tRNA demethylase was only recently identified [126], opening the door to the possibility that many of these modifications are reversible and regulatory.

In addition to the specific questions raised by the initial experiments, these methods are available for immediate use to answer biological questions. For instance, although the changes in tRNA aminoacylation level upon starvation in human cells are known at isoacceptor resolution, certain isodecoders are potentially preferentially charged under starvation conditions. Charged tRNA-seq can now answer this question. Additionally, there could be large differences in aminoacylation levels or modifications of specific tRNA isodecoders between tissues in the same organism. While these differences may or may not be regulatory, they can be easily addressed with the methods presented here. Finally, if our modification signature method can be scaled to identify modifications in mRNA, we should be able to define the modification landscape of mRNAs across cell types without having to perform individual sequencing reactions for each

modification. Again, while differences in the modification landscape between tissue types may or may not be regulatory, this will likely be a useful biological parameter to define.

5.3 Challenges

RNA high-throughput sequencing has become a staple in biochemical research as the method has been extended to look at a plethora of biological parameters. Unfortunately, the answer to most scientific questions today has been to design an RNA-seq experiment. While these experiments do provide a wealth of information to researchers, the biases inherent in library preparation and data analysis have yet to be precisely defined. Hence, while scientists feel confident that their results represent biologically relevant phenomenon, we are blind to the accurate conclusions of these experiments.

For example, although ligation bias is well known, it is rarely accounted for in data analysis. Library prep also requires PCR amplification before sequencing, and while only a few rounds of PCR are performed in order to reduce bias, short sequences and sequences with low GC-content are preferentially enriched. Additionally, rare cDNAs can be lost after only a few cycles of PCR. When precise counting is used to analyze this wealth of data, small biases over only a few cycles of PCR will cause distortion of the input ratios of cDNA and can lead to false conclusions. Additionally, when scientists are looking for a certain result in their sequencing data, they can manipulate the mapping methods in order to find what they are looking for. Although 'blind' analysis isn't realistic, we should take care to ensure that our expectations about the outcome of the experiment do not influence the final sequencing analysis.

RNA-seq also suffers from statistical uncertainty, despite the amount of data produced from each experiment. When considering small changes in gene expression, only a few replicates

are insufficient to determine differential expression. For example, it has been suggested that *at least 12 biological replicates per condition* be performed in order to define most of the differentially expressed genes between conditions [127]. While this is likely both time and cost prohibitive, we would be wise to consider more rigorous statistical analyses to define to the False Discovery Rate (FDR) for every experiment in order to properly scrutinize the conclusions drawn from that experiment. Additionally, validation of multiple targets would add to confidence in experimental results.

Additionally, RNA-seq experiments suffer from a lack of transparency about sample preparation and quality as well as sequence mapping and analysis. For instance, the quality of the cDNA library submitted for sequencing is rarely reported. Additionally, reads uploaded to the Gene Expression Omnibus (GEO) have often already been processed, and methods to map reads and statistical analysis of the mapping are added to the experimental methods section as an afterthought.

Literacy in sequencing methods and data analysis is also too low among biologists. When scrutinizing a paper, many reviewers and readers simply accept the sequencing method and analysis without being critical. Although this is problematic, a lack of scientific literacy in scientists preparing the cDNA libraries and analyzing the data is more troubling. While kits for RNA-seq library preparation have increased efficiency and availability of the method, understanding the principles of library prep and the limitations of the method are often lacking. Similarly, oftentimes a scientist with no background in computer science is simply coloring-by-numbers when analyzing data. While this might be sufficient for simple experiments, more complex methods require a deep understanding of data processing and computational methods.

As we move into the future, more standardized methods of sequence mapping and statistical analysis must be defined in order to gain confidence in our methods. Additionally, transparency in the quality of libraries prepared and in the sequence mapping, trimming, and analysis is key. Finally, improving the sequencing literacy should be a top priority as more and more scientists incorporate these methods into their experiments. Since experiments build upon one another, we must ensure that our experimental methods and analysis are of sufficient quality to keep science moving forward.

REFERENCES:

1. Schena M, Shalon D, Davis RW, Brown PO (1995) Quantitative monitoring of gene expression patterns with a complementary DNA microarray. Science 270: 467–470.

2. Lockhart DJ, Dong H, Byrne MC, Follettie MT, Gallo MV, et al. (1996) Expression monitoring by hybridization to high-density oligonucleotide arrays. Nature Biotechnology 14: 1675–1680. doi:10.1038/nbt1296-1675.

3. Clark TA, Sugnet CW, Ares M (2002) Genomewide analysis of mRNA processing in yeast using splicing-specific microarrays. Science 296: 907–910. doi:10.1126/science.1069415.

4. Yamada K, Lim J, Dale JM, Chen H, Shinn P, et al. (2003) Empirical analysis of transcriptional activity in the Arabidopsis genome. Science 302: 842–846. doi:10.1126/science.1088305.

5. Bertone P, Stolc V, Royce TE, Rozowsky JS, Urban AE, et al. (2004) Global identification of human transcribed sequences with genome tiling arrays. Science 306: 2242–2246. doi:10.1126/science.1103388.

6. Cheng J, Kapranov P, Drenkow J, Dike S, Brubaker S, et al. (2005) Transcriptional maps of 10 human chromosomes at 5-nucleotide resolution. Science 308: 1149–1154. doi:10.1126/science.1108625.

7. David L, Huber W, Granovskaia M, Toedling J, Palm CJ, et al. (2006) A high-resolution map of transcription in the yeast genome. Proc Natl Acad Sci USA 103: 5320–5325. doi:10.1073/pnas.0601091103.

8. Adams MD, Kelley JM, Gocayne JD, Dubnick M, Polymeropoulos MH, et al. (1991) Complementary DNA sequencing: expressed sequence tags and human genome project. Science 252: 1651–1656.

9. Velculescu VE, Zhang L, Vogelstein B, Kinzler KW (1995) Serial analysis of gene expression. Science 270: 484–487.

10. Shiraki T, Kondo S, Katayama S, Waki K, Kasukawa T, et al. (2003) Cap analysis gene expression for high-throughput analysis of transcriptional starting point and identification of promoter usage. Proc Natl Acad Sci USA 100: 15776–15781. doi:10.1073/pnas.2136655100.

11. An Introduction to Next-Generation Sequencing Technology (2016) An Introduction to Next-Generation Sequencing Technology: 1–16. Available: https://www.illumina.com/content/dam/illumina-marketing/documents/products/illumina_sequencing_introduction.pdf. Accessed 2 July 2017.

12. Lister R, O'Malley RC, Tonti-Filippini J, Gregory BD, Berry CC, et al. (2008) Highly

Integrated Single-Base Resolution Maps of the Epigenome in Arabidopsis. Cell 133: 523–536. doi:10.1016/j.cell.2008.03.029.

13. Wilhelm BT, Marguerat S, Watt S, Schubert F, Wood V, et al. (2008) Dynamic repertoire of a eukaryotic transcriptome surveyed at single-nucleotide resolution. Nature 453: 1239–1243. doi:10.1091/mbc.01-10-0507.

14. Nagalakshmi U, Wang Z, Waern K, Shou C, Raha D, et al. (2008) The transcriptional landscape of the yeast genome defined by RNA sequencing. Science 320: 1344–1349. doi:10.1126/science.1158441.

15. Cloonan N, Forrest ARR, Kolle G, Gardiner BBA, Faulkner GJ, et al. (2008) Stem cell transcriptome profiling via massive-scale mRNA sequencing. Nat Meth 5: 613–619. doi:10.1038/nmeth.1223.

16. Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B (2008) Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nat Meth 5: 621–628. doi:10.1038/nmeth.1226.

17. Marioni JC, Mason CE, Mane SM, Stephens M, Gilad Y (2008) RNA-seq: An assessment of technical reproducibility and comparison with gene expression arrays. Genome Research 18: 1509–1517. doi:10.1101/gr.079558.108.

18. Morin R, Bainbridge M, Fejes A, Hirst M, Krzywinski M, et al. (2008) Profiling the HeLa S3 transcriptome using randomly primed cDNA and massively parallel short-read sequencing. Biotech 45: 81–94. doi:10.2144/000112900.

19. Hrdlickova R, Toloue M, Tian B (2017) RNA-Seq methods for transcriptome analysis. WIREs RNA 8. doi:10.1002/wrna.1364.

20. Wang Z, Gerstein M, Snyder M (2009) RNA-Seq: a revolutionary tool for transcriptomics. Nature Reviews Genetics 10: 57–63. doi:10.1038/nrg2484.

21. Hansen KD, Brenner SE, Dudoit S (2010) Biases in Illumina transcriptome sequencing caused by random hexamer priming. Nucleic Acids Research 38: e131–e131. doi:10.1093/nar/gkq224.

22. Roberts A, Trapnell C, Donaghey J, Rinn JL, Pachter L (2011) Improving RNA-Seq expression estimates by correcting for fragment bias. Genome Biol 12: R22. doi:10.1186/gb-2011-12-3-r22.

23. Mohr S, Ghanem E, Smith W, Sheeter D, Qin Y, et al. (2013) Thermostable group II intron reverse transcriptase fusion proteins and their use in cDNA synthesis and next-generation RNA sequencing. RNA 19: 958–970. doi:10.1261/rna.039743.113.

24. Qin Y, Yao J, Wu DC, Nottingham RM, Mohr S, et al. (2015) High-throughput sequencing of human plasma RNA by using thermostable group II intron reverse transcriptases. RNA 22: 111–128. doi:10.1261/rna.054809.115.

25. Nottingham RM, Wu DC, Qin Y, Yao J, Hunicke-Smith S, et al. (2016) RNA-seq of human reference RNA samples using a thermostable group II intron reverse transcriptase. RNA 22: 597–613. doi:10.1261/rna.055558.115.

26. Core LJ, Waterfall JJ, Lis JT (2008) Nascent RNA sequencing reveals widespread pausing and divergent initiation at human promoters. Science 322: 1845–1848. doi:10.1126/science.1162228.

27. Ingolia NT, Ghaemmaghami S, Newman JRS, Weissman JS (2009) Genome-wide analysis in vivo of translation with nucleotide resolution using ribosome profiling. Science 324: 218–223. doi:10.1126/science.1168978.

28. Hafner M, Landthaler M, Burger L, Khorshid M, Hausser J, et al. (2010) Transcriptome-wide Identification of RNA-Binding Protein and MicroRNA Target Sites by PAR-CLIP. Cell 141: 129–141. doi:10.1016/j.cell.2010.03.009.

29. Sharma E, Sterne-Weiler T, O'Hanlon D, Blencowe BJ (2016) Global Mapping of Human RNA-RNA Interactions. Mol Cell 62: 618–626. doi:10.1016/j.molcel.2016.04.030.

30. Sridhar B, Rivas-Astroza M, Nguyen TC, Chen W, Yan Z, et al. (2017) Systematic Mapping of RNA-Chromatin Interactions In Vivo. Curr Biol 27: 602–609. doi:10.1016/j.cub.2017.01.011.

31. Tang F, Barbacioru C, Wang Y, Nordman E, Lee C, et al. (2009) mRNA-Seq whole-transcriptome analysis of a single cell. Nat Meth 6: 377–382. doi:10.1016/j.tig.2007.12.007.

32. Lucks JB, Mortimer SA, Trapnell C, Luo S, Aviran S, et al. (2011) Multiplexed RNA structure characterization with selective 2'-hydroxyl acylation analyzed by primer extension sequencing (SHAPE-Seq). Proc Natl Acad Sci USA 108: 11063–11068. doi:10.1073/pnas.1106501108.

33. Zheng G, Qin Y, Clark WC, Dai Q, Yi C, et al. (2015) Efficient and quantitative high-throughput tRNA sequencing. Nat Meth 12: 835–837. doi:10.1038/nmeth.3478.

34. Cozen AE, Quartley E, Holmes AD, Hrabeta-Robinson E, Phizicky EM, et al. (2015) ARM-seq: AlkB-facilitated RNA methylation sequencing reveals a complex landscape of modified tRNA fragments. Nat Meth 12: 879–884. doi:10.1038/nmeth.3508.

35. Varshney U, Lee CP, RajBhandary UL (1991) Direct analysis of aminoacylation levels of tRNAs in vivo. Application to studying recognition of Escherichia coli initiator tRNA mutants by glutaminyl-tRNA synthetase. J Biol Chem 266: 24712–24718.

36. McClain WH, Jou YY, Bhattacharya S, Gabriel K, Schneider J (1999) The reliability of in vivo structure-function analysis of tRNA aminoacylation. Journal of Molecular Biology 290: 391–409. doi:10.1006/jmbi.1999.2884.

37. Sørensen MA (2001) Charging levels of four tRNA species in Escherichia coli Rel+ and Rel− strains during amino acid starvation: a simple model for the effect of ppGpp on translational accuracy11Edited by D. E. Draper. Journal of Molecular Biology 307: 785–798. doi:10.1006/jmbi.2001.4525.

38. Dittmar KA, Sørensen MA, Elf J, Ehrenberg M, Pan T (2005) Selective charging of tRNA isoacceptors induced by amino-acid starvation. EMBO Reports 6: 151–157. doi:10.1038/sj.embor.7400341.

39. Zhou Y, Goodenbour JM, Godley LA, Wickrema A, Pan T (2009) High levels of tRNA abundance and alteration of tRNA charging by bortezomib in multiple myeloma. Biochemical and Biophysical Research Communications 385: 160–164. doi:10.1016/j.bbrc.2009.05.031.

40. Zaborske JM, Narasimhan J, Jiang L, Wek SA, Dittmar KA, et al. (2009) Genome-wide analysis of tRNA charging and activation of the eIF2 kinase Gcn2p. Journal of Biological Chemistry 284: 25254–25267. doi:10.1074/jbc.M109.000877.

41. Roundtree IA, Evans ME, Pan T, He C (2017) Dynamic RNA Modifications in Gene Expression Regulation. Cell 169: 1187–1200. doi:10.1016/j.cell.2017.05.045.

42. Chan CTY, Dyavaiah M, DeMott MS, Taghizadeh K, Dedon PC, et al. (2010) A Quantitative Systems Approach Reveals Dynamic Control of tRNA Modifications during Cellular Stress. PLoS Genet 6: e1001247. doi:10.1371/journal.pgen.1001247.s010.

43. Kuchino Y, Hanyu N, Nishimura S (1987) Analysis of modified nucleosides and nucleotide sequence of tRNA. Meth Enzymol 155: 379–396.

44. Suzuki T, Suzuki T (2014) A complete landscape of post-transcriptional modifications in mammalian mitochondrial tRNAs. Nucleic Acids Research 42: 7346–7357. Available: https://academic.oup.com/nar/article-lookup/doi/10.1093/nar/gku390.

45. Motorin Y, Muller S, Behm Ansmant I, Branlant C (2007) Identification of Modified Residues in RNAs by Reverse Transcription-Based Methods. Methods in Enzymology. Methods in Enzymology. Elsevier, Vol. 425. pp. 21–53. doi:10.1016/S0076-6879(07)25002-5.

46. Ryvkin P, Leung YY, Silverman IM, Childress M, Valladares O, et al. (2013) HAMR: high-throughput annotation of modified ribonucleotides. RNA 19: 1684–1692. doi:10.1261/rna.036806.112.

47. Hauenschild R, Tserovski L, Schmid K, Thüring K, Winz M-L, et al. (2015) The reverse transcription signature of N-1-methyladenosine in RNA-Seq is sequence dependent. Nucleic Acids Research 43: 9950–9964. doi:10.1093/nar/gkv895.

48. Tserovski L, Marchand V, Hauenschild R, Blanloeil-Oillo F, Helm M, et al. (2016) High-throughput sequencing for 1-methyladenosine (m1A) mapping in RNA. Methods

107: 110–121. doi:10.1016/j.ymeth.2016.02.012.

49. Elf J, Nilsson D, Tenson T, Ehrenberg M (2003) Selective charging of tRNA isoacceptors explains patterns of codon usage. Science 300: 1718–1722. doi:10.1126/science.1083811.

50. Subramaniam AR, Pan T, Cluzel P (2013) Environmental perturbations lift the degeneracy of the genetic code to regulate protein levels in bacteria. Proc Natl Acad Sci USA 110: 2419–2424. doi:10.1073/pnas.1211077110.

51. Avcilar-Kucukgoze I, Bartholomäus A, Cordero Varela JA, Kaml RF-X, Neubauer P, et al. (2016) Discharging tRNAs: a tug of war between translation and detoxification in Escherichia coli. Nucleic Acids Research 44: 8324–8334. doi:10.1093/nar/gkw697.

52. Haseltine WA, Block R (1973) Synthesis of guanosine tetra- and pentaphosphate requires the presence of a codon-specific, uncharged transfer ribonucleic acid in the acceptor site of ribosomes. Proc Natl Acad Sci USA 70: 1564–1568.

53. Dever TE, Feng L, Wek RC, Cigan AM, Donahue TF, et al. (1992) Phosphorylation of initiation factor 2α by protein kinase GCN2 mediates gene-specific translational control of GCN4 in yeast. Cell 68: 585–596.

54. Goodenbour JM, Pan T (2006) Diversity of tRNA genes in eukaryotes. Nucleic Acids Research 34: 6137–6146. doi:10.1093/nar/gkl725.

55. Parisien M, Wang X, Pan T (2014) Diversity of human tRNA genes from the 1000-genomes project. rnabiology 10: 1853–1867. doi:10.1016/j.ymeth.2007.08.001.

56. Wolfson AD, Uhlenbeck OC (2002) Modulation of tRNAAla identity by inorganic pyrophosphatase. Proc Natl Acad Sci USA 99: 5965–5970. doi:10.1073/pnas.092152799.

57. Chan PP, Lowe TM (2009) GtRNAdb: a database of transfer RNA genes detected in genomic sequence. Nucleic Acids Research 37: D93–D97. doi:10.1093/nar/gkn787.

58. Clark WC, Evans ME, Dominissini D, Zheng G, Pan T (2016) tRNA base methylation identification and quantification via high-throughput sequencing. RNA 22: 1771–1784. doi:10.1261/rna.056531.116.

59. Geslain R, Pan T (2010) Functional Analysis of Human tRNA Isodecoders. Journal of Molecular Biology 396: 821–831. doi:10.1016/j.jmb.2009.12.018.

60. Gomes AC, Kordala AJ, Strack R, Wang X, Geslain R, et al. (2016) A dual fluorescent reporter for the investigation of methionine mistranslation in live cells. RNA 22: 467–476. doi:10.1261/rna.054163.115.

61. Ishimura R, Nagy G, Dotu I, Zhou H, Yang X-L, et al. (2014) Ribosome stalling induced by mutation of a CNS-specific tRNA causes neurodegeneration. Science 345: 455–459.

doi:10.1126/science.1249749.

62. Rudinger-Thirion J, Lescure A, Paulus C, Frugier M (2011) Misfolded human tRNA isodecoder binds and neutralizes a 3' UTR-embedded Alu element. Proc Natl Acad Sci USA 108: E794–E802. doi:10.1073/pnas.1103698108.

63. Grosjean H (2005) Modification and editing of RNA: historical overview and important facts to remember. Topics in Current Genetics. Topics in Current Genetics. Berlin, Heidelberg: Springer Berlin Heidelberg, Vol. 12. pp. 1–22. doi:10.1007/b106848.

64. Phizicky EM, Hopper AK (2010) tRNA biology charges to the front. Genes & Development 24: 1832–1860. doi:10.1101/gad.1956510.

65. Wei F-Y, Suzuki T, Watanabe S, Kimura S, Kaitsuka T, et al. (2011) Deficit of tRNA(Lys) modification by Cdkal1 causes the development of type 2 diabetes in mice. J Clin Invest 121: 3598–3608. doi:10.1172/JCI58056.

66. Fu Y, Dominissini D, Rechavi G, He C (2014) Gene expression regulation mediated through reversible m6A RNA methylation. Nature Reviews Genetics 15: 293–306. doi:10.1038/nrg3724.

67. Li S, Mason CE (2014) The Pivotal Regulatory Landscape of RNA Modifications. Annu Rev Genom Hum Genet 15: 127–150. doi:10.1146/annurev-genom-090413-025405.

68. Decatur WA, Fournier MJ (2002) rRNA modifications and ribosome function. Trends Biochem Sci 27: 344–351.

69. Liang X-H, Liu Q, Fournier MJ (2009) Loss of rRNA modifications in the decoding center of the ribosome impairs translation and strongly delays pre-rRNA processing. RNA 15: 1716–1728. doi:10.1261/rna.1724409.

70. Higa-Nakamine S, Suzuki T, Uechi T, Chakraborty A, Nakajima Y, et al. (2012) Loss of ribosomal RNA modification causes developmental defects in zebrafish. Nucleic Acids Research 40: 391–398. doi:10.1093/nar/gkr700.

71. Cantara WA, Crain PF, Rozenski J, McCloskey JA, Harris KA, et al. (2011) The RNA Modification Database, RNAMDB: 2011 update. Nucleic Acids Research 39: D195–D201. doi:10.1093/nar/gkq1028.

72. Machnicka MA, Milanowska K, Osman Oglou O, Purta E, Kurkowska M, et al. (2013) MODOMICS: a database of RNA modification pathways--2013 update. Nucleic Acids Research 41: D262–D267. doi:10.1093/nar/gks1007.

73. Bjork GR, Wikström PM, Bystrom AS (1989) Prevention of translational frameshifting by the modified nucleoside 1-methylguanosine. Science 244: 986–989.

74. Gerber AP, Keller W (1999) An adenosine deaminase that generates inosine at the wobble position of tRNAs. Science 286: 1146–1149.

75. Ohira T, Suzuki T (2011) Retrograde nuclear import of tRNA precursors is required for modified base biogenesis in yeast. Proc Natl Acad Sci USA 108: 10502–10507. doi:10.1073/pnas.1105645108.

76. Whipple JM, Lane EA, Chernyakov I, D'Silva S, Phizicky EM (2011) The yeast rapid tRNA decay pathway primarily monitors the structural integrity of the acceptor and T-stems of mature tRNA. Genes & Development 25: 1173–1184. doi:10.1101/gad.2050711.

77. Zaborske JM, Bauer DuMont VL, Wallace EWJ, Pan T, Aquadro CF, et al. (2014) A Nutrient-Driven tRNA Modification Alters Translational Fidelity and Genome-wide Protein Coding across an Animal Genus. PLoS Biol 12: e1002015. doi:10.1371/journal.pbio.1002015.s006.

78. Kuchino Y, Shindo-Okada N, Ando N, Watanabe S, Nishimura S (1981) Nucleotide sequences of two aspartic acid tRNAs from rat liver and rat ascites hepatoma. J Biol Chem 256: 9059–9062.

79. Chan JC, Yang JA, Dunn MJ, Agris PF, Wong TW (1982) The nucleotide sequence of a glutamate tRNA from rat liver. Nucleic Acids Research 10: 4605–4608.

80. Saikia M, Fu Y, Pavon-Eternod M, He C, Pan T (2010) Genome-wide analysis of N1-methyl-adenosine modification in human tRNAs. RNA 16: 1317–1327. doi:10.1261/rna.2057810.

81. Vinayak M, Pathak C (2009) Queuosine modification of tRNA: its divergent role in cellular machinery. Biosci Rep 30: 135–148. doi:10.1042/BSR20090057.

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

83. Meyer KD, Saletore Y, Zumbo P, Elemento O, Mason CE, et al. (2012) Comprehensive Analysis of mRNA Methylation Reveals Enrichment in 3′ UTRs and near Stop Codons. Cell 149: 1635–1646. doi:10.1016/j.cell.2012.05.003.

84. Dominissini D, Nachtergaele S, Moshitch-Moshkovitz S, Peer E, Kol N, et al. (2016) The dynamic N1-methyladenosine methylome in eukaryotic messenger RNA. Nature 530: 441–446. doi:10.1038/nature16998.

85. Li X, Xiong X, Wang K, Wang L, Shu X, et al. (2016) Transcriptome-wide mapping reveals reversible and dynamic N(1)-methyladenosine methylome. Nat Chem Biol 12: 311–316. doi:10.1038/nchembio.2040.

86. Parisien M, Wang X, Perdrizet G II, Lamphear C, Fierke CA, et al. (2013) Discovering RNA-Protein Interactome by Using Chemical Context Profiling of the RNA-Protein Interface. CellReports 3: 1703–1713. doi:10.1016/j.celrep.2013.04.010.

87. Peattie DA (1979) Direct chemical method for sequencing RNA. Proc Natl Acad Sci USA 76: 1760–1764.

88. Behm Ansmant I, Helm M, Motorin Y (2011) Use of specific chemical reagents for detection of modified nucleotides in RNA. Journal of Nucleic Acids 2011: 408053. doi:10.4061/2011/408053.

89. Kadaba S, Krueger A, Trice T, Krecic AM, Hinnebusch AG, et al. (2004) Nuclear surveillance and degradation of hypomodified initiator tRNAMet in S. cerevisiae. Genes & Development 18: 1227–1240. doi:10.1101/gad.1183804.

90. Gamper HB, Masuda I, Frenkel-Morgenstern M, Hou Y-M (2015) Maintenance of protein synthesis reading frame by EF-P and m1G37-tRNA. Nature Communications 6: 1–13. doi:10.1038/ncomms8226.

91. Steinberg S, Cedergren R (1995) A correlation between N2-dimethylguanosine presence and alternate tRNA conformers. RNA 1: 886–891.

92. Edqvist J, Stråby KB, Grosjean H (1995) Enzymatic formation of N2,N2-dimethylguanosine in eukaryotic tRNA: importance of the tRNA architecture. Biochimie 77: 54–61. doi:10.1016/0300-9084(96)88104-1.

93. Pallan PS, Kreutz C, Bosio S, Micura R, Egli M (2008) Effects of N2,N2-dimethylguanosine on RNA structure and stability: crystal structure of an RNA duplex with tandem m2 2G:A pairs. RNA 14: 2125–2135. doi:10.1261/rna.1078508.

94. Helm M, Giegé R, Florentz C (1999) A Watson−Crick Base-Pair-Disrupting Methyl Group (m 1A9) Is Sufficient for Cloverleaf Folding of Human Mitochondrial tRNA Lys †. Biochemistry 38: 13338–13346. doi:10.1021/bi991061g.

95. Grosjean H, Auxilien S, Constantinesco F, Simon C, Corda Y, et al. (1996) Enzymatic conversion of adenosine to inosine and to N1-methylinosine in transfer RNAs: a review. Biochimie 78: 488–501.

96. Dalluge JJ, Hashizume T, Sopchik AE, McCloskey JA, Davis DR (1996) Conformational flexibility in RNA: the role of dihydrouridine. Nucleic Acids Research 24: 1073–1079.

97. Liang XH, Liu Q, Fournier MJ (2009) Loss of rRNA modifications in the decoding center of the ribosome impairs translation and strongly delays pre-rRNA processing. RNA 15: 1716–1728. doi:10.1261/rna.1724409.

98. Meyer B, Wurm JP, Kötter P, Leisegang MS, Schilling V, et al. (2011) The Bowen-Conradi syndrome protein Nep1 (Emg1) has a dual role in eukaryotic ribosome biogenesis, as an essential assembly factor and in the methylation of Ψ1191 in yeast 18S rRNA. Nucleic Acids Research 39: 1526–1537. doi:10.1093/nar/gkq931.

99. Begley U, Dyavaiah M, Patil A, Rooney JP, DiRenzo D, et al. (2007) Trm9-catalyzed

tRNA modifications link translation to the DNA damage response. Mol Cell 28: 860–870. doi:10.1016/j.molcel.2007.09.021.

100. Chan CTY, Pang YLJ, Deng W, Babu IR, Dyavaiah M, et al. (2012) Reprogramming of tRNA modifications controls the oxidative stress response by codon-biased translation of proteins. Nature Communications 3: 937. doi:10.1038/ncomms1938.

101. Patil A, Dyavaiah M, Joseph F, Rooney JP, Chan CTY, et al. (2012) Increased tRNA modification and gene-specific codon usage regulate cell cycle progression during the DNA damage response. Cell Cycle 11: 3656–3665. doi:10.4161/cc.21919.

102. Patil A, Chan CTY, Dyavaiah M, Rooney JP, Dedon PC, et al. (2012) Translational infidelity-induced protein stress results from a deficiency in Trm9-catalyzed tRNA modifications. rnabiology 9: 990–1001. doi:10.4161/rna.20531.

103. Gu C, Begley TJ, Dedon PC (2014) tRNA modifications regulate translation during cellular stress. FEBS Lett 588: 4287–4296. doi:10.1016/j.febslet.2014.09.038.

104. Li X, Xiong X, Yi C (2016) Epitranscriptome sequencing technologies: decoding RNA modifications. Nat Meth 14: 23–31. doi:10.1038/nmeth.4110.

105. Helm M, Motorin Y (2017) Detecting RNA modifications in the epitranscriptome: predict and validate. Nature Reviews Genetics 18: 275–291. doi:10.1038/nrg.2016.169.

106. Schwartz S, Bernstein DA, Mumbach MR, Jovanovic M, Herbst RH, et al. (2014) Transcriptome-wide Mapping Reveals Widespread Dynamic-Regulated Pseudouridylation of ncRNA and mRNA. Cell 159: 148–162. doi:10.1016/j.cell.2014.08.028.

107. Carlile TM, Rojas-Duran MF, Zinshteyn B, Shin H, Bartoli KM, et al. (2014) Pseudouridine profiling reveals regulated mRNA pseudouridylation in yeast and human cells. Nature 515: 143–146. doi:10.1038/nature13802.

108. Lovejoy AF, Riordan DP, Brown PO (2014) Transcriptome-Wide Mapping of Pseudouridines: Pseudouridine Synthases Modify Specific mRNAs in S. cerevisiae. PLoS ONE 9: e110799. doi:10.1371/journal.pone.0110799.s005.

109. Li X, Zhu P, Ma S, Song J, Bai J, et al. (2015) Chemical pulldown reveals dynamic pseudouridylation of the mammalian transcriptome. Nat Chem Biol 11: 592–597. doi:10.1038/nchembio.1836.

110. Squires JE, Patel HR, Nousch M, Sibbritt T, Humphreys DT, et al. (2012) Widespread occurrence of 5-methylcytosine in human coding and non-coding RNA. Nucleic Acids Research 40: 5023–5033. doi:10.1093/nar/gks144.

111. Incarnato D, Anselmi F, Morandi E, Neri F, Maldotti M, et al. (2017) High-throughput single-base resolution mapping of RNA 2′-O-methylated residues. Nucleic Acids Research 45: 1433–1441. doi:10.1093/nar/gkw810.

112. Marchand V, Blanloeil-Oillo F, Helm M, Motorin Y (2016) Illumina-based RiboMethSeq approach for mapping of 2′-O-Me residues in RNA. Nucleic Acids Research 44: e135–e135. doi:10.1093/nar/gkw547.

113. Dai Q, Moshitch-Moshkovitz S, Han D, Kol N, Amariglio N, et al. (2017) Nm-seq maps 2′-O-methylation sites in human mRNA with base precision. Nat Meth: 1–6. doi:10.1038/nmeth.4294.

114. Edelheit S, Schwartz S, Mumbach MR, Wurtzel O, Sorek R (2013) Transcriptome-Wide Mapping of 5-methylcytidine RNA Modifications in Bacteria, Archaea, and Yeast Reveals m5C within Archaeal mRNAs. PLoS Genet 9: e1003602. doi:10.1371/journal.pgen.1003602.s001.

115. Delatte B, Wang F, Ngoc LV, Collignon E, Bonvin E, et al. (2016) Transcriptome-wide distribution and function of RNA hydroxymethylcytosine. Science 351: 282–285. doi:10.1126/science.aac5253.

116. Hussain S, Sajini AA, Blanco S, Dietmann S, Lombard P, et al. (2013) NSun2-mediated cytosine-5 methylation of vault noncoding RNA determines its processing into regulatory small RNAs. CellReports 4: 255–261. doi:10.1016/j.celrep.2013.06.029.

117. Khoddami V, Cairns BR (2013) Identification of direct targets and modified bases of RNA cytosine methyltransferases. Nature Biotechnology 31: 458–464. doi:10.1038/nbt.2566.

118. Linder B, Grozhik AV, Olarerin-George AO, Meydan C, Mason CE, et al. (2015) Single-nucleotide-resolution mapping of m6A and m6Am throughout the transcriptome. Nat Meth 12: 767–772. doi:10.1016/j.molcel.2010.05.004.

119. Chen K, Lu Z, Wang X, Fu Y, Luo G-Z, et al. (2014) High-Resolution N6-Methyladenosine (m 6A) Map Using Photo-Crosslinking-Assisted m 6A Sequencing. Angew Chem Int Ed 54: 1587–1590. doi:10.1002/anie.201410647.

120. Ke S, Alemu EA, Mertens C, Gantman EC, Fak JJ, et al. (2015) A majority of m 6A residues are in the last exons, allowing the potential for 3? UTR regulation. Genes & Development 29: 2037–2053. doi:10.1101/gad.269415.115.

121. Molinie B, Wang J, Lim KS, Hillebrand R, Lu Z-X, et al. (2016) m6A-LAIC-seq reveals the census and complexity of the m6A epitranscriptome. Nat Meth 13: 692–698. doi:10.1261/rna.032912.112.

122. Ebhardt HA, Tsang HH, Dai DC, Liu Y, Bostan B, et al. (2009) Meta-analysis of small RNA-sequencing errors reveals ubiquitous post-transcriptional RNA modifications. Nucleic Acids Research 37: 2461–2470. doi:10.1093/nar/gkp093.

123. Iida K, Jin H, Zhu J-K (2009) Bioinformatics analysis suggests base modifications of tRNAs and miRNAs in Arabidopsis thaliana. BMC Genomics 10: 155. doi:10.1186/1471-2164-10-155.

124. Findeiss S, Langenberger D, Stadler PF, Hoffmann S (2011) Traces of post-transcriptional RNA modifications in deep sequencing data. Biol Chem 392: 305–313. doi:10.1515/BC.2011.043.

125. Pedregosa F, Varoquaux GEL, Gramfort A, Michel V, Thirion B, et al. (2011) Scikit-learn: Machine Learning in Python. J Mach Learn Res 12: 2825–2830.

126. Liu F, Clark W, Luo G, Wang X, Fu Y, et al. (2016) ALKBH1-Mediated tRNA Demethylation Regulates Translation. Cell 167: 816–818.e816. doi:10.1016/j.cell.2016.09.038.

127. Schurch NJ, Schofield P, Gierliński M, Cole C, Sherstnev A, et al. (2016) How many biological replicates are needed in an RNA-seq experiment and which differential expression tool should you use? RNA 22: 839–851. doi:10.1261/rna.053959.115.