TRANSCRIPTOME-WIDE RNA MODIFICATION PROFILING

by

Vahid Khoddami Vishteh

A dissertation submitted to the faculty of
The University of Utah
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

Department of Oncological Sciences

The University of Utah

August 2013

The University of Utah Graduate School

STATEMENT OF DISSERTATION APPROVAL

The dissertation of **Vahid Khoddami Vishteh**

has been approved by the following supervisory committee members:

Bradley R. Cairns	, Chair	**05 / 21 / 2013** Date Approved
Brenda L. Bass	, Member	**05 / 21 / 2013** Date Approved
Cynthia J. Burrows	, Member	**05 / 21 / 2013** Date Approved
David A. Jones	, Member	**05 / 21 / 2013** Date Approved
Stephen L. Lessnick	, Member	**05 / 21 / 2013** Date Approved

and by **Bradley R. Cairns** , Chair of

the Department of **Oncological Sciences**

and by Donna M. White, Interim Dean of The Graduate School.

ABSTRACT

Post-transcriptional RNA modifications provide new structural and functional features to modified RNA molecules. Extensive research in the past has resulted in isolation of over 100 distinct nucleotide modifications from different organisms and in different RNA species. These modified nucleotides are distributed within the entire transcriptome comprising the cellular epitranscriptome. The ultimate goal of the research in the field is to address what the specific functions of specific modifications are, and also the impact of each on cellular physiology. However, the first question to be addressed is how these > 100 modified nucleotides are distributed within the transcriptome.

RNA modification profiling using conventional techniques has provided a great body of knowledge about the distribution of many modifications in RNAs. However, these findings remained limited mostly to tRNAs and rRNAs, the two most abundant and also highly modified RNA species in different organisms. This is partly because of the lower sensitivity of applied classical technologies.

Here in this dissertation, in Chapter 2, we are reporting an optimized new RNA bisulfite protocol suitable for high-throughput RNA cytosine methylation profiling. We present the results of application of this technique for 5-methyl-cytosine (m^5C) profiling in mouse embryonic fibroblasts (MEFs) RNAs, isolated from wt and dnmt2-/- mice to explore the target specificity of DNA methyltransferase 2 (DNMT2) enzyme.

In Chapter 3, we present a substantially novel technique: Aza-IP, for enrichment and identification of the direct targets of RNA cytosine methyltransferases (m^5C-RMTs) as well as

determination of the exact modified bases in the same experiment. We provide the results of the Aza-IP technique for two human m^5C-RMTs; DNMT2 and NSUN2, representing their known and novel RNA targets/modified bases.

In Chapter 4 we discuss how similar technologies to both of the RNA bisulfite sequencing and Aza-IP techniques as well as other methodologies can be applied and extended for transcriptome-wide profiling of RNA modifications other than m^5C.

In Chapter 5 we present the future directions of the work focused on cataloguing the direct targets of all human m^5C-RMTs in human cultured cells in mouse and fish model systems, to elucidate the functions of cytosine methylation in RNA molecules.

To my parents

TABLE OF CONTENTS

LIST OF FIGURES

Figure

LIST OF TABLES

Table

ACKNOWLEDGMENTS

First of all I would like to thank my mentor; Dr. Bradley Cairns, for giving me the opportunity to join his lab and experience the cutting edge science under his supervision. Joining Cairns lab was an honor and full of lessons for me. Cairns lab is a wonderful place to live, learn, experience, flourish and prosper. By this I would like to thank Dr. Cairns for his efforts to lead such a productive research lab.

I would like to thank all members of the Cairns lab for making such a friendly helpful environment to live and work, especially Alisha Schlichter, the lab manager, for her continuous support for our daily experiments in the lab. I would like to thank the senior scientists in the cairns lab: Margaret Kasten, Tim Parnell, and Cedric Clapier, as well as the former lab members: Jacqueline Wittmeyer, Kunal Rai, Andrew Oler, and Kaede Hinata for their helpful and often vital comments for many of the experiments I have presented in different chapters of this dissertation. I would also like to thank all the graduate students in Cairns lab, first for being my close friends here in Utah, and second for offering their help and comments for my projects.

I would like to thank the members of my thesis committee: Dr. Brenda Bass, Dr. Cynthia Burrows, Dr. David Jones, and Dr. Steve Lessnick for all their support and also their constructive suggestions. Also special thanks to Dr. Cynthia Burrows for letting me to do part of my project in her lab.

A number of other people have also contributed directly or indirectly during my thesis, including Xiaoyun Xu, Aaron Fleming and James Muller (from Dr. Burrows' lab), Somaye

Dehghanizadeh (Jones lab), Dr. Vicente Planelles and members of his lab, especially Carlos Maximiliano Rêgo Monteiro Filho, Brian Dalley and Nicole Moss (Sequencing core), David Nix, Brett Milash, Ying Sun and Tim Mosbruger (Bioinformatics Core), and many others here at Huntsman Cancer Institute, as well as people in the Department of Biochemistry and Department of Medicinal Chemistry at the University of Utah. I would like to thank all of them here. I also would like to thank Jessica Askin and Dee DalPonte for all their administrative support.

Mentioning all these names and many others I may have missed to name here just reminds me that what I have achieved during my PhD - here at Huntsman Cancer Institute/University of Utah - belongs to many people not just me. So, THANKS EVERYBODY.

CHAPTER 1

INTRODUCTION

Modifications and flow of genetic information

The three fundamental informative polymers of life - DNA, RNA and proteins - are polymerized from nucleotides, ribonucleotides and amino acid building blocks, respectively. DNA, in most organisms, exists in double-stranded form and serves as the information storage polymer consisting of only four standard nucleotides: A, C, G and T. To decode and utilize the stored information in DNA, different types of RNA polymerases transcribe the DNA at specific defined locations to produce variety of single-stranded RNA molecules by accurately copying the sequences from DNA to RNA transcripts. The RNA molecules, again, are composed of only four standard ribonucleotides: A, C, G and U. A portion of RNA transcripts – mRNAs - code for protein synthesis in ribosomes to translate every triplet codons within mRNAs, into one amino acid in the polypeptide chains. The protein molecules are composed of 21 standard amino acids.

Although the limited number of aforementioned building blocks in each polymer type seems to be necessary for the natural flow of genetic information from DNA to RNA to protein, they are not sufficient to fulfill all expected functions of such polymers in living organisms. Thus, although DNA, RNA and proteins are built from only few standard building blocks, some of the monomers at specific defined locations may undergo some targeted enzymatic changes to produce new modified entities. These postreplicational, -transcriptional or – translational modifications provide new structural and/or functional features for these polymers making them suitable to fulfill more elaborated functions and through more controlled/regulatable ways. Thus, understanding the full spectrum of modified building blocks, their exact positions within the polymers, their actual functions - as well as studying the effector proteins/enzymes producing (writing), interpreting (reading) or removing (erasing) those modified entities - is fundamental for understanding the logic of life.

Diversity of RNA modifications

Naturally modified nucleotides in DNA molecules are limited (6-methyladenosine (m^6A), 5-methylcytosine (m^5C), 5-hydroxymethylcytosine (hm^5C), and the Base J[1]), whereas more than 100 distinct modified structures have been identified so far in RNA, and the list is still growing[2] (Figure 1.1). From a chemical standpoint, the DNA and RNA polymers are substantially similar in structure with two exceptions: presence of hydroxyl group at 2' position of the ribose in RNA and replacement of T in DNA with U in RNA which differs only in one methyl group. Also, DNA is mostly double-stranded while RNA is largely single-stranded (although it can make local double-stranded structures as well). The majority of modifications taking place in RNA are chemically possible to take place in DNA as well, but the observed number of modifications in DNA is much below the expected numbers. This significant difference in the numbers may reflect the resistance of DNA during evolution, as an information storage entity, to accommodate more modifications in order to maintain the fidelity and integrity of information during DNA replication. This will become more clear when considering that some modified nucleotides can change the A:T or G:C base pairing rules, and therefore can cause mutations, and some others may stop the polymerases during replication, and therefore can cause interruption in DNA synthesis and duplication, impairing the normal cell cycle and division.

RNA in contrast is not the information storage polymer, except in some RNA viruses, and once it is transcribed it no longer serves as the polymerization template, except in the case of telomere maintenance and in retroviruses and retrotransposons. This has probably allowed the RNA molecules to accommodate such a diverse pool of modified entities. In addition, RNA has diverse responsibilities in living organisms including coding, transfer, catalytic, structural, regulatory or guide functions which all can become enhanced or regulated by post-transcriptional modifications.

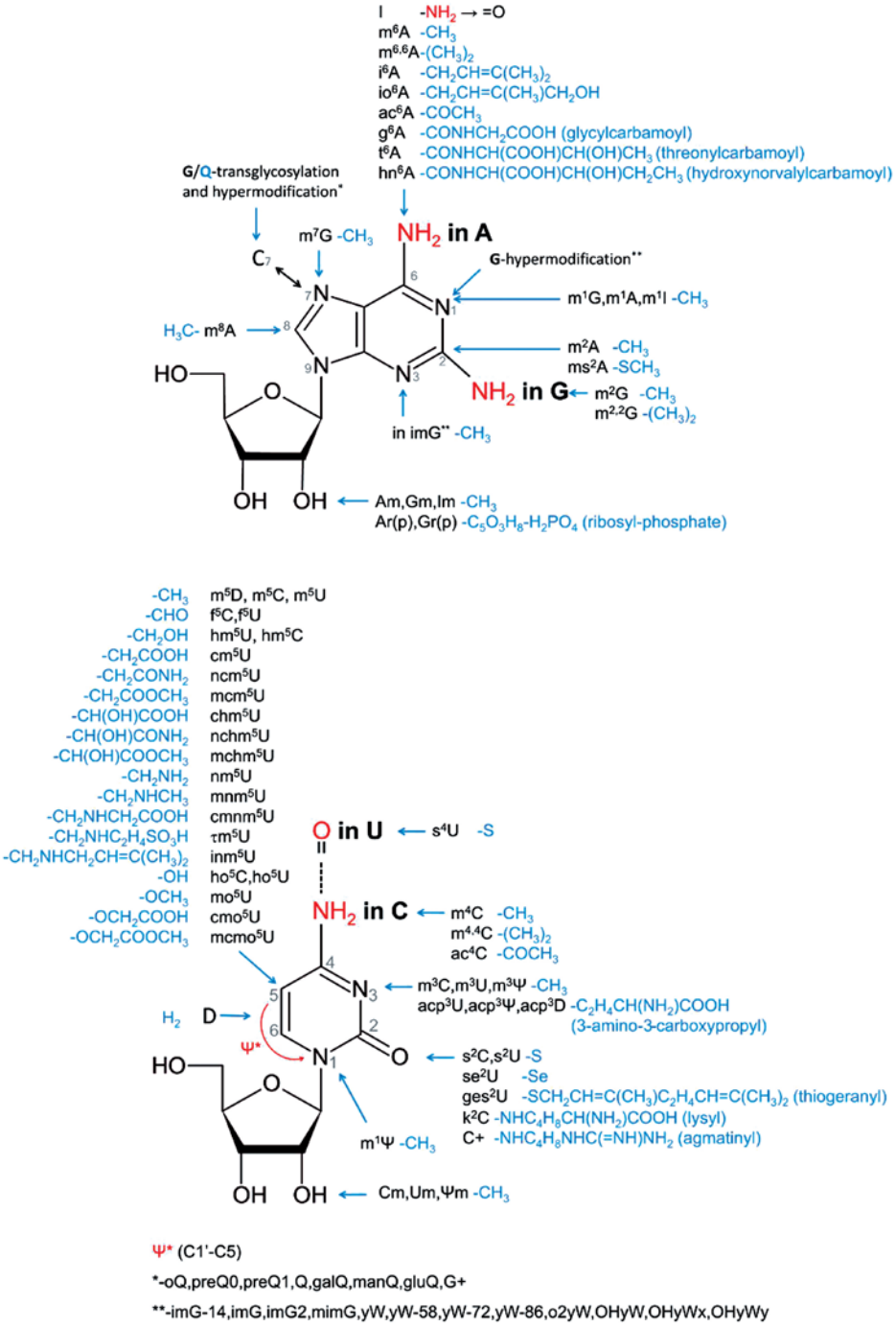


Figure 1.1 | Diversity of nucleotide modifications in RNA[2]. This schematic, from reference,[2] shows that all accessible portions of nucleotides within an RNA molecule can be a subject for modification.

The emerging concept of "epitranscriptome"

Although modified nucleotides have been observed in almost all RNA species, tRNAs and rRNAs harbor majority of known RNA modifications regarding both of the number and diversity. Results obtained over the past five decades, from classical RNA profiling approaches, suggest that most known RNA modifications reside in tRNAs and rRNAs but not other RNA species. mRNAs, snRNAs, snoRNAs etc., are believed to show less diverse and lower numbers of modifications, in comparison to tRNAs and rRNAs. Since tRNAs and rRNAs are both highly abundant and highly modified in cells, it has been easier to isolate the RNAs and characterize these modifications. However, these advantages for studying modification in these two RNA species may be considered disadvantages for studying modifications in other RNA species using classical methodologies. This is because preparing pure RNA samples of low-copy RNA species from cell extracts without contamination by highly abundant tRNAs and rRNAs (which are also highly modified), is almost impossible. This will result in higher background, lower resolution and difficulties in interpretation of RNA modification profiling, which limits the detection of less abundant and slightly-modified RNA species. On the other hand, in recent years, several new noncoding RNA (ncRNA) species have been characterized in different organisms, which are considered new targets of RNA modification profiling as modified nucleotides have already been isolated in some of them. These challenges prompt the need for development of sensitive discovery tools to unravel the real scope of RNA modifications in all RNA species to define the "epitranscriptome" of living organisms. This may be possible by recruiting and/or modifying the recently developed and still evolving high-throughput modification profiling techniques.

Here in this dissertation we explain and exemplify the cytosine methylation/methyltransferases and cytosine methylation profiling methodologies extendable for studying other types of RNA modifications / modifiers.

5-methylcytosine in DNA

5-methylcytosine (m^5C) is the first modified nucleotide described in 1948 in DNA.[3] Since then the functions and enzymology of m^5C-methylation have been the focus of extensive research in many laboratories worldwide. This endeavor has resulted in generation of a massive body of knowledge about m^5C in DNA, the DNA methyltransferase enzymes (DNMTs), the m^5C binders (MBD proteins), the de-methylase complexes as well as physiological attributes and pathological implications of disruptions or misregulation of DNA methylation machinery.[4]

5-methylcytosine in RNA

m^5C has been found in many RNA species.[5, 6] Majority of m^5C bases in total cellular RNAs are present in tRNAs as they are both abundant molecules and bear more m^5C sites considering their short length of approximately 75 bp on average. tRNAs of archaeal and eukaryotic origin, not eubacterial organisms, contain m^5C at specific conserved locations.[5] In humans over 200 m^5C distinct sites have been identified in different tRNAs and their isoacceptors (tRNAs with different anticodons that accept the same amino acid) or isodecoders (tRNAs with the same anticodons that accept the same amino acid but have different body sequences).[6] However, the most conserved and frequently modified sites within various tRNAs of different species are C48 and C49 positions at the junction of variable region with TΨC-stem of tRNAs.[5, 6]

Prokaryotic and eukaryotic rRNAs also show conserved m^5C sites indicating functional importance. However, compared to tRNAs, rRNAs, with much longer length, show much lower number of m^5Cs. For example there are only three m^5Cs reported in *Escherichia coli* rRNAs, two in human 28S and none in the many eukaryotic 18S rRNAs that have been identified so far.[5] The scope of ribosomal m^5C methylation in different species, however, remained to become fully

explored as it is possible that only a subset of rRNAs receive this modification, which may be missed using available analysis tools.

m^5C has also been reported previously in mRNAs and also some viral RNAs such as sindbis virus, adenovirus and Turnip Yellow Mosaic Virus (TYMV) RNAs.[7-11] Finally, a recent high-throughput RNA methylation profiling by bisulfite sequencing in HeLa cells verified and extended m^5C presence in tRNAs and rRNAs, and also showed m^5C in other noncoding RNAs as well as a limited number of mRNAs.[6] These recent results in HeLa cells motivate a more thorough examination of the scope (cell types and developmental contexts) and functions of this modification.

RNA cytosine methyltransferases (m^5C-RMTs)

All m^5C sites in all bacterial, archeal and eukaryotic organisms are laid down enzymatically by the action of RNA cytosine methyltransferases (m^5C-RMTs).[5] Different types of m^5C-RMTs have been characterized so far that all catalyze the methyl transfer from the methyl donor S-Adenosyl Methionine (SAM) cofactor to the fifth position (C5) of the target base cytosine producing methylated RNA and S-Adenosyl-L-Homocysteine (SAH) as the product and byproduct of this reaction, respectively.[5] All m^5C-RMTs tested so far, similar to DNMTs, form a covalent enzyme-substrate intermediate with their target cytosine through the sulfur atom of the cysteine residue in their catalytic domain to the C6 position of the base in the target RNA. Next enamine methylation of the C5 position of the target cytosine is catalyzed by the enzyme using SAM as the methyl donor, followed by beta-elimination to complete the methylation cycle and releasing the enzyme[5] (Figure 1.2). This catalytic mechanism is the basis for design and construction of DNMT and m^5C-RMT suicide inhibitors such as 5-azacytidine (5-aza-C).[12] RNA polymerases can incorporate the 5-aza-C in the growing RNA molecule, randomly in place of

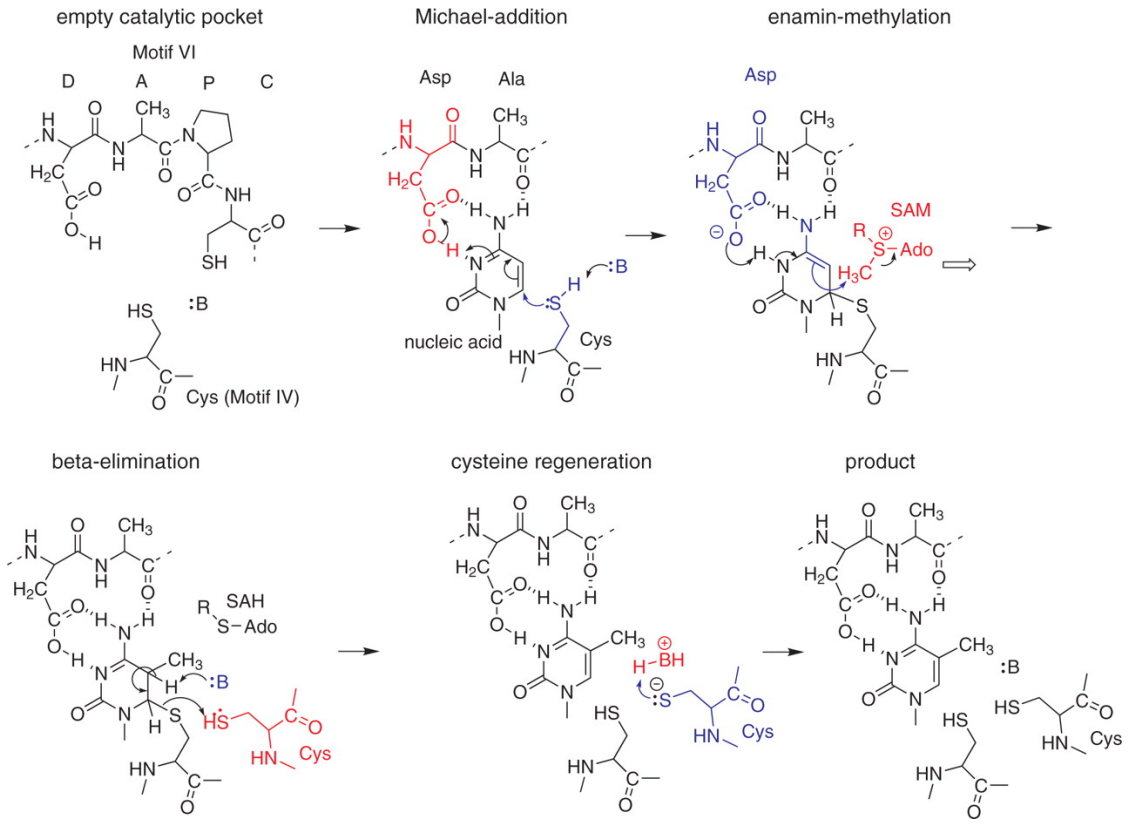


Figure 1.2 | Mechanism of RNA cytosine methylation by m^{5}C-RMTs.[5] This schematic, from reference,[5] represents the detailed mechanism of RNA cytosine methylation by m^{5}C-RMTs. Upon recognition of the exact target cytosine within an RNA molecule a conserved cysteine residue from the catalytic site of the m^{5}C-RMT makes a covalent connection to the 6th position of the target cytosine through Michael addition. Next a methyl group is transferred from the methyl-donor; SAM, to the 5th position of the target base through enamine methylation. Finally the cysteine residue is released from the base through beta elimination resulting in methyl-cytosine and SAH as product and byproduct of the methylation reaction, respectively.

cytosines. 5-aza-C, once incorporated at the target site, can block the m^5C-RMTs. Due to substitution of carbon atom at position 5 of the base with nitrogen in 5-aza-C, the methylation cycle cannot be completed and the enzyme remains covalently attached to the base. This RMT-RNA adduct formation depletes the cells from m^5C-RMTs resulting in RNA hypomethylation.[12-15]

Based on the exact catalytic mechanism there are two types of m^5C-RMTs.[5] Members of the first enzyme type, exemplified by DNMT2 enzyme, for methylation use a single cysteine in their catalytic site similar to the action of DNMTs.[16] In contrast members of the second type, exemplified by the members of NSUN family of RNA methyltransferases, utilize two different cysteine residues from separate domains in their catalytic site for the catalysis. In this second enzyme type, one cysteine form the covalent connection with the cytosine base and after the completion of methyl transfer from SAM to the C5 of the base the second cysteine residue triggers the release of the base from the enzyme[17] (Figure 1.3). For the two cysteine enzyme type, substitution of the second cysteine residue with another amino acid impairs the completion of methylation cycle and results in formation of stable m^5C-RMT-RNA adduct through the covalent connection between the sulfur atom of the first cysteine residue from the enzyme to the C6 position of the target cytosine base within the target RNA molecule.[18, 19]

Based on structural and functional properties, m^5C-RMTs have been subdivided into six families: RsmB/Nol1/NSUN1, RsmF/YebU/NSUN2, RlmI, Ynl022, NSUN6 and DNMT2[5] (Table 1.1). RsmB/Nol1/ NSUN1-family members are conserved, with known substrates (16S rRNA) in E. coli but not in eukaryotes.[5] RsmF/YebU/NSUN2-family enzymes are also conserved, and methylate 16S rRNA in bacteria or tRNA in eukaryotes, though orphan enzymes exist, with unknown substrates.[5] Most tRNAs and several noncoding RNAs have been shown to be the substrates of NSUN2.[6, 20] Studies on RlmI-family members are limited though bacterial RlmI methylates 23S

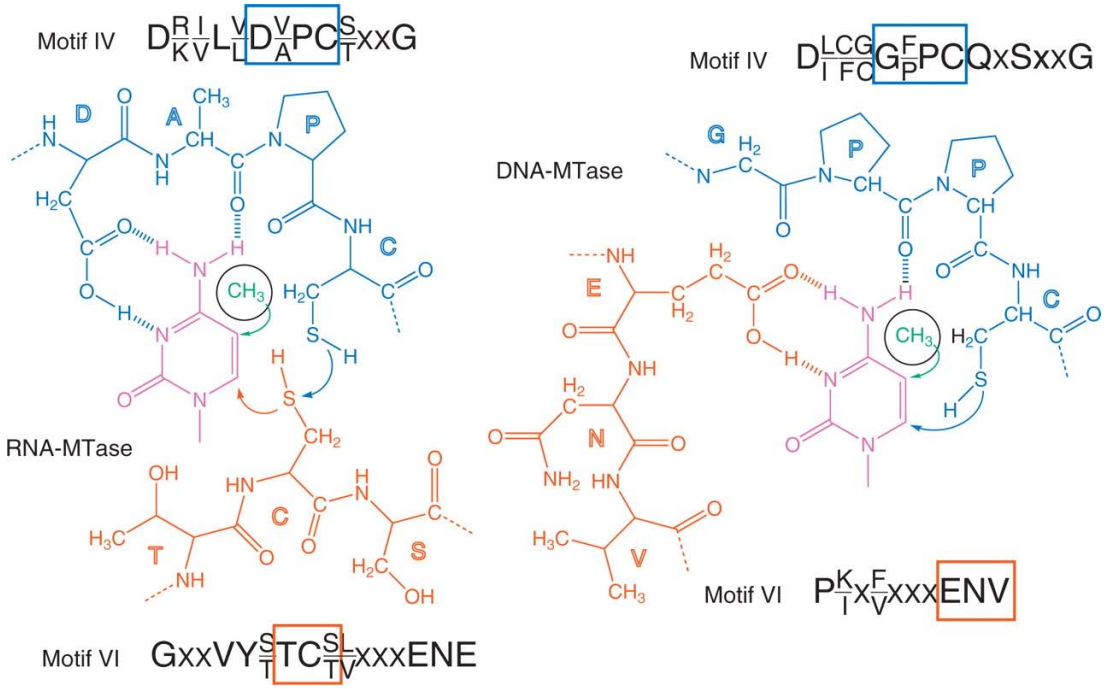


Figure 1.3 | Differences in the mechanism of cytosine methylation by RNA and DNA cytosine methyltransferases.[5] This schematic, from reference,[5] represents the differences between the catalytic mechanism of RNA cytosine methyltransferases (m^5C-RMTs) (left) and DNA cytosine methyltransferases (DNMTs) (right). Most m^5C-RMTs including members of the NSUN family of methyltransferase utilize two cysteines in their catalytic domain. The cysteine residue from motif VI makes the covalent connection and enhances the methylation reaction. The cysteine residue from motif IV then drives the reversal of the covalent connection and release of the enzyme from the base. In contrast, all DNMTs, and also the RNA methyltransferase DNMT2, utilize only one cysteine residue during the entire methylation cycle.

Table 1.1 | Some known and putative m^5C-RMTs.[5] This table, from reference,[5] represents some classified information about different RNA cytosine methyltransferases (m^5C-RMTs), highlighting that substrates of a majority of them are not currently known.

Enzyme family	Enzyme name	Other names	Organism	Life domain	Accession	Identification type	RNA substrate
RsmB family							
RsmB/Nol1	RsmB	Fmu/Fmv	*Escherichia coli*	B	AP_004502	Biochemical	16S rRNA
RsmB/Nol1	P120	NSUN1	*Homo sapiens*	E	P46087	NO	Unknown
RsmB/Nol1	Nop2		*Saccharomyces cerevisiae*	E	YNL061W	NO	Unknown
RsmF/YebU family							
YebU			*Haloferax volcanii*	A	HVO_1594	Bioinformatics	tRNA
YebU			*Pyrococcus abyssi*	A	PAB1947	Biochemical	tRNA
YebU	aTrm4		*Pyrococcus horikoshi*	A	PH1374	Bioinformatics	Unknown
YebU	RsmF	YebU	*Escherichia coli*	B	P76273	Biochemical	16S rRNA
YebU	hTrm4	NSUN2/Misu	*Homo sapiens*	E	NM_017755	Biochemical	pre-tRNA^{Leu}
YebU	FLJ22609	NSUN3	*Homo sapiens*	E	Q9H649	NO	Unknown
YebU	MGC22920	NSUN4	*Homo sapiens*	E	Q96CB9	NO	Unknown
YebU	Trm4	Ncl1	*Saccharomyces cerevisiae*	E	YBL024W	Biochemical	tRNA and pre-tRNA
Dnmt2 family							
DNMT2	Dnmt2		*Drosophila melanogaster*	E	Q9U6H7	Biochemical	tRNA
	Dnmt2	trdmt1	*Drosophila rerio*	E	Q588C1	Biochemical	tRNA
DNMT2	Dnmt2		*Homo sapiens*	E	O14717	Biochemical	tRNA
DNMT2	Dnmt2		*Mus musculus*	E	O55055	Bioinformatics	
DNMT3	pmt1	pmt1	*Saccharomyces pombe*	E	P40999	Bioinformatics	
RlmI family							
COG1092	RlmI	YccW	*Escherichia coli*	B	P75876	Biochemical	23S rRNA
Ynl022 family							
Ynl022c	WBSCR20A	NSUN5A	*Homo sapiens*	E	Q96P11	NO	Unknown
Ynl022c	WBSCR20B	NSUN5B	*Homo sapiens*	E	Q3KNT7	NO	Unknown
Ynl022c	WBSCR20C	NSUN5C	*Homo sapiens*	E	Q63ZY6	NO	Unknown
Ynl022c	WBSCR22		*Homo sapiens*	E	O43709	NO	Unknown
Ynl022c		NSUN7	*Homo sapiens*	E	Q8NE18	NO	Unknown
Ynl022c	Ynl022c		*Saccharomyces cerevisiae*	E	YNL022C	NO	Unknown
NSUN6 family							
NSUN6		NSUN6	*Homo sapiens*	E	Q8TEA1	NO	Unknown

Abbreviations: A: archaea; B: bacteria; E eukaryota; NA not analyzed.

rRNA[5]. Ynl022 family members are solely eukaryotic, and lack defined substrates.[5] NSUN6 is found in vertebrates, and lacks a defined substrate.[5] DNMT2 is now known to function primarily, if not exclusively, as an m^5C-RMT, with three verified tRNA targets: $tRNA^{Asp}$, $tRNA^{Gly}$ and $tRNA^{Val}$.[20, 21]

Functions of m^5C in RNA

The precise locations of many m^5C sites within noncoding RNAs from divergent species are very well conserved during evolution.[5] This is more pronounced in tRNAs partly because they have been studied more. m^5C at C40 of the *S. cerevisiae* tRNAPhe has been shown to be required for proper Mg^{2+} binding triggering the conformational transition of the anticodon loop.[22] m^5C at C34, the anticodon wobble base of leucine tRNAs in *S. cerevisiae,* has been shown to be important for efficient suppressor activity of the leucine-inserting amber suppressor $tRNA3^{Leu}$.[23]

The m^5C sites in ribosomal RNA tend to cluster in close proximity within the ribosomal 3D structure and this is despite the distant locations of m^5C sites within the linear rRNA sequences, indicating the functional importance of m^5Cs.[22] Notably, the antibiotic paromomycin binding site has been mapped to the exact same region in the ribosomes. Paromomycin inhibits the protein synthesis through binding to the ribosomal RNA. Interestingly the yeast strains lacking the corresponding m^5C-RMT enzyme show greater sensitivity to this antibiotic.[24]

Beside these few examples and despite several decades of efforts the exact biological functions of m^5C in RNA remain elusive. However, the importance of m^5C methylation in RNA can become conferred from some severe phenotypic manifestations of the depletion, or misregulation of corresponding m^5C-RMTs in different organisms. For example upregulation of NSUN1 and NSUN2 has been documented in many cancer cell lines or tumor types.[25-27] NSUN2

functions are linked to Myc-induced proliferation of cancer cells,[28] mitotic spindle stability,[29] infertility in male mice, and balancing self-renewal/differentiation of stem cells in the skin.[30] Most importantly, NSUN2 mutations cause an autosomal recessive syndrome in humans conferring intellectual disability and mental retardation.[31-33] Other members of the NSUN family of RNA methyltransferases are also linked to abnormalities such as genetic disorders (NSUN5 variants) and fertility (NSUN7).[5, 34, 35]

Most organisms lacking DNMT2 lack obvious phenotypes,[36] though zebrafish display developmental perturbations.[37] Notably, DNMT2 activities attenuate tRNA cleavage during stress conditions,[21] contribute in RNA processing in stress granules[38] and promote the virus response to positive strand RNA viruses in Drosophila Melanogaster.[39] Moreover, tRNA cytosine methylation by both DNMT2 and NSUN2 homologues promotes tRNA stability and steady-state protein synthesis.[40]

Overall, although efforts in the past and recent discoveries have increased the scope of RNA cytosine methylation, and also revealed a large number of candidate m^5C-RMTs, the exact functions of m^5C sites in RNAs along with the functions and targets of many known and candidate m^5C-RMTs remain elusive.

m^5C RNA methylation dynamics

Studying the dynamic changes of m^5C levels in DNA is indispensable from epigenome research. Similarly, one interesting question in the epitranscriptome field is whether there is such an analogy for RNA modifications like m^5C methyl marks in RNA molecules. Analogous to regulated methylation as well as passive and active DNA demethylation concepts, RNA methylation dynamics could be achieved by any of: RNA turn-over, regulated and targeted methylation, or demethylation. The balance between transcription and RNA decay rates with

methylation rates is a key determinant of RNA methylation dynamics. On the other hand the RNA methylation level at a particular target site and/or differential methylation of target sites in different conditions can become potentially regulated by external factors during stress conditions or immune response, or upon induction by a chemical or biochemical stimulus. Finally, erasing the methyl marks through active enzymatic demethylation is another possibility to achieve the dynamic methylation changes in RNA.

Quantification of tRNA modifications in *Saccharomyces cerevisiae* using a precise mass spectrometric method in a recent work, demonstrated that some modifications, including m^5C, show dynamic changes upon exposure of cells to different toxicants.[41] Specifically m^5C, in parallel with some other modifications such as Cm and $m^2{}_2G$, shows significant increase following hydrogen peroxide exposure, suggesting a possible contribution of dynamic RNA modifications in stress response.[41] Subsequent analysis of substrates of an RNA cytosine methyltransferase; Trm4, provided a model for involvement of dynamic m^5C methylation in stress response.[42] Trm4, the human NSUN2 homologue in yeast, is responsible for methylation of C34, the wobble position, in $tRNA^{LeuCAA}$ anticodon. Oxidative stress induces significant increase in the methylation level at C34 in this tRNA resulting in increased translation of genes with enriched TTG codon notably RPL22A; a ribosomal protein. Hypersensitivity to oxidative stress upon loss of either of Trm4 or RPL22A proteins has been observed indicating the important role of RNA modification-mediated selective protein translation during stress response.[42]

Another example of regulated m^5C RNA methylation is the recent work showing the significant nutritional status-dependent increase in the methylation level of C38 in tRNAAsp in fission yeast. Interestingly, this regulation requires a specific serine/threonine kinase; Sck2, clearly linking the nutrient signaling and RNA modifications pathways. Pmt1, the human DNMT2

homologue in *Schizosaccharomyces pombe*, is responsible for C38 methylation in tRNAs.[43] It is, however, not clear whether the Pmt1 enzyme or its protein partners are the targets of Sck2 or which residue(s) is (are) the targets of phosphorylation, and most importantly what the consequence of C38 methylation during nutrient deficiency is.

Beside regulating the levels and target sites of RNA modifications, erasing the preexisting marks can potentially affect the modification status of RNA molecules. This has been recently very well documented for the case of RNA adenosine methylation (m^6A) by isolating and characterizing the demthylase complexes.[44] Similarly, in explaining the m^5C dynamic changes, although there is no direct evidence for the demethylation scenario, it is a possibility that some m^5C marks on RNA become erased enzymatically. Interestingly hyroxymethyl-cytosine (hm^5C) has been reported in some eukaryotic ribosomal RNAs.[45] Here, m^5C hydroxylation may be considered as an intermediate step in erasing the methyl marks, analogues to Tet protein mediated hydroxylation of m^5Cs in DNA as a prerequisite for the demethylation procedures. Thus, it would be interesting to know whether there is an overlap between the sites of m^5C and hm^5C marks within the RNA molecules and also probe for identification of possible m^5C demethylase complexes by sequence homology surveys.

m^5C profiling methodologies

The first step in RNA modification studies of any kind is to find out what RNA molecules, at which specific location(s), and to what extent are modified. The “RNA modification profiling” term refers to this effort. Different strategies have been used in the past to isolate the modified RNA molecules and define their sites and levels of modifications. Almost all these strategies rely on the differential behavior of unmodified and modified nucleotides during the modification profiling procedures. Most modified nucleotides including m^5C can become determined by

techniques which differentiate the molecules based on their physicochemical properties such as electrophoresis, chromatography and mass spectrometry methods.[46] The exact location of most of the tRNA and rRNA m^5C sites have been determined long ago using these technologies.[5, 45]

The locations of some of these sites as well as their responsible methyltransferases have become confirmed by reconstitution of methyltransferase reactions (MTase assays) in-vitro using purified m^5C-RMTs and RNA substrates followed by chromatography or mass spectrometry analysis.[5]

Finally, a few years ago, based on the differential chemical reactivity of Cs vs m^5Cs with sodium bisulfite mixture, a version of bisulfite sequencing technique suitable for RNA molecules adapted (with some modifications) from DNA bisulfite sequencing methods, proved to be able to precisely report the m^5C sites in RNA molecules.[47] Recently, a similar approach was used to identify the first transcriptome-wide RNA m^5C methylome by performing a high-throughput sequencing of the bisulfite treated RNAs isolated from HeLa cells.[6] This work verified and extended the repertoire of m^5C presence in RNA motivating a more thorough examination of the scope (cell types and developmental contexts) and functions of this modification.

Preview

Here in this dissertation we will discuss the recent improvement of epitranscriptome profiling technologies, made by us and others.

In Chapter 2, we present a modified RNA bisulfite sequencing approach which reports the m^5C sites with high-fidelity and minimum false positive rates. We have applied this technique for exploring the RNA methylome of mouse embryonic fibroblasts (MEFs) and compared the RNA methylome of wt vs. Dnmt2-/- MEFs to find the Dnmt2 RNA targets.

In Chapter 3, we introduce a novel mechanism-based enrichment technique; Aza-IP, for enrichment of the RNA target molecules of all m^5C-RMTs and isolation of the exact target sites within the same experiment. We present the results obtained with Aza-IP for both of human DNMT2 and NSUN2 in HeLa cells.

In Chapter 4, we discuss how similar concepts can be used to study the other RNA modifications and modifiers to provide a high-resolution picture of combinatorial RNA epitranscriptome. We also review all available tools and technologies for epitranscriptome profiling in general.

In Chapter 5, we discuss the on-going projects and future directions of the work with the aim of providing a comprehensive catalogue of the direct targets of all human m^5C-RMTs in tissue culture and selected mouse or zebrafish m^5C-RMTs in-vivo. Our eventual goal for all these experiments is to elucidate the rules and roles of RNA cytosine methylation in living organisms.

References

1. Korlach, J. & Turner, S.W. Going beyond five bases in DNA sequencing. *Curr Opin Struct Biol* **22**, 251-261 (2012).

2. Machnicka, M.A. et al. MODOMICS: a database of RNA modification pathways--2013 update. *Nucleic Acids Res* **41**, D262-267 (2013).

3. Hotchkiss, R.D. The quantitative separation of purines, pyrimidines, and nucleosides by paper chromatography. *J Biol Chem* **175**, 315-332 (1948).

4. Suzuki, M.M. & Bird, A. DNA methylation landscapes: provocative insights from epigenomics. *Nat Rev Genet* **9**, 465-476 (2008).

5. Motorin, Y., Lyko, F. & Helm, M. 5-methylcytosine in RNA: detection, enzymatic formation and biological functions. *Nucleic Acids Res* **38**, 1415-1430 (2010).

6. Squires, J.E. et al. Widespread occurrence of 5-methylcytosine in human coding and non-coding RNA. *Nucleic Acids Res* **40**, 5023-5033 (2012).

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

8. Dubin, D.T. & Stollar, V. Methylation of Sindbis virus "26S" messenger RNA. *Biochem Biophys Res Commun* **66**, 1373-1379 (1975).

9. Dubin, D.T., Stollar, V., Hsuchen, C.C., Timko, K. & Guild, G.M. Sindbis virus messenger RNA: the 5'-termini and methylated residues of 26 and 42 S RNA. *Virology* **77**, 457-470 (1977).

10. Sommer, S. et al. The methylation of adenovirus-specific nuclear and cytoplasmic RNA. *Nucleic Acids Res* **3**, 749-765 (1976).

11. Brule, H., Grosjean, H., Giege, R. & Florentz, C. A pseudoknotted tRNA variant is a substrate for tRNA (cytosine-5)-methyltransferase from Xenopus laevis. *Biochimie* **80**, 977-985 (1998).

12. Santi, D.V., Garrett, C.E. & Barr, P.J. On the mechanism of inhibition of DNA-cytosine methyltransferases by cytosine analogs. *Cell* **33**, 9-10 (1983).

13. Lu, L.W., Chiang, G.H., Medina, D. & Randerath, K. Drug effects on nucleic acid modification. I. A specific effect of 5-azacytidine on mammalian transfer RNA methylation in vivo. *Biochem Biophys Res Commun* **68**, 1094-1101 (1976).

14. Lu, L.J. & Randerath, K. Effects of 5-azacytidine on transfer RNA methyltransferases. *Cancer Res* **39**, 940-949 (1979).

15. Schaefer, M., Hagemann, S., Hanna, K. & Lyko, F. Azacytidine inhibits RNA methylation at DNMT2 target sites in human cancer cell lines. *Cancer Res* **69**, 8127-8132 (2009).

16. Jurkowski, T.P. et al. Human DNMT2 methylates tRNA(Asp) molecules using a DNA methyltransferase-like catalytic mechanism. *RNA* **14**, 1663-1670 (2008).

17. King, M.Y. & Redman, K.L. RNA methyltransferases utilize two cysteine residues in the formation of 5-methylcytosine. *Biochemistry* **41**, 11218-11225 (2002).

18. Redman, K.L. Assembly of protein-RNA complexes using natural RNA and mutant forms of an RNA cytosine methyltransferase. *Biomacromolecules* **7**, 3321-3326 (2006).

19. Sugimoto, Y. et al. Analysis of CLIP and iCLIP methods for nucleotide-resolution studies of protein-RNA interactions. *Genome Biol* **13**, R67 (2012).

20. Khoddami, V. & Cairns, B.R. Identification of direct targets and modified bases of RNA cytosine methyltransferases. *Nat Biotechnol* (2013).

21. Schaefer, M. et al. RNA methylation by Dnmt2 protects transfer RNAs against stress-induced cleavage. *Genes Dev* **24**, 1590-1595 (2010).

22. Chen, Y., Sierzputowska-Gracz, H., Guenther, R., Everett, K. & Agris, P.F. 5-Methylcytidine is required for cooperative binding of Mg2+ and a conformational transition at the anticodon stem-loop of yeast phenylalanine tRNA. *Biochemistry* **32**, 10249-10253 (1993).

23. Strobel, M.C. & Abelson, J. Effect of intron mutations on processing and function of Saccharomyces cerevisiae SUP53 tRNA in vitro and in vivo. *Mol Cell Biol* **6**, 2663-2673 (1986).

24. Vicens, Q. & Westhof, E. Crystal structure of paromomycin docked into the eubacterial ribosomal decoding A site. *Structure* **9**, 647-658 (2001).

25. Fonagy, A. et al. Cell cycle regulated expression of nucleolar antigen P120 in normal and transformed human fibroblasts. *J Cell Physiol* **154**, 16-27 (1993).

26. Frye, M. et al. Genomic gain of 5p15 leads to over-expression of Misu (NSUN2) in breast cancer. *Cancer Lett* **289**, 71-80 (2010).

27. Okamoto, M. et al. Frequent increased gene copy number and high protein expression of tRNA (Cytosine-5-)-Methyltransferase (NSUN2) in human cancers. *DNA Cell Biol* **31**, 660-671 (2012).

28. Frye, M. & Watt, F.M. The RNA methyltransferase Misu (NSun2) mediates Myc-induced proliferation and is upregulated in tumors. *Curr Biol* **16**, 971-981 (2006).

29. Hussain, S. et al. The nucleolar RNA methyltransferase Misu (NSun2) is required for mitotic spindle stability. *J Cell Biol* **186**, 27-40 (2009).

30. Blanco, S. et al. The RNA-methyltransferase Misu (NSun2) poises epidermal stem cells to differentiate. *PLoS Genet* **7**, e1002403 (2011).

31. Abbasi-Moheb, L. et al. Mutations in NSUN2 cause autosomal-recessive intellectual disability. *Am J Hum Genet* **90**, 847-855 (2012).

32. Khan, M.A. et al. Mutation in NSUN2, which encodes an RNA methyltransferase, causes autosomal-recessive intellectual disability. *Am J Hum Genet* **90**, 856-863 (2012).

33. Martinez, F.J. et al. Whole exome sequencing identifies a splicing mutation in NSUN2 as a cause of a Dubowitz-like syndrome. *J Med Genet* (2012).

34. Doll, A. & Grzeschik, K.H. Characterization of two novel genes, WBSCR20 and WBSCR22, deleted in Williams-Beuren syndrome. *Cytogenet Cell Genet* **95**, 20-27 (2001).

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

36. Schaefer, M. & Lyko, F. Solving the Dnmt2 enigma. *Chromosoma* **119**, 35-40 (2010).

37. Rai, K. et al. Dnmt2 functions in the cytoplasm to promote liver, brain, and retina development in zebrafish. *Genes Dev* **21**, 261-266 (2007).

38. Thiagarajan, D., Dev, R.R. & Khosla, S. The DNA methyltranferase Dnmt2 participates in RNA processing during cellular stress. *Epigenetics* **6**, 103-113 (2011).

39. Durdevic, Z. et al. Efficient RNA virus control in Drosophila requires the RNA methyltransferase Dnmt2. *EMBO Rep* **14**, 269-275 (2013).

40. Tuorto, F. et al. RNA cytosine methylation by Dnmt2 and NSun2 promotes tRNA stability and protein synthesis. *Nat Struct Mol Biol* **19**, 900-905 (2012).

41. Chan, C.T. et al. A quantitative systems approach reveals dynamic control of tRNA modifications during cellular stress. *PLoS Genet* **6**, e1001247 (2010).

42. Chan, C.T. et al. Reprogramming of tRNA modifications controls the oxidative stress response by codon-biased translation of proteins. *Nat Commun* **3**, 937 (2012).

43. Becker, M. et al. Pmt1, a Dnmt2 homolog in Schizosaccharomyces pombe, mediates tRNA methylation in response to nutrient signaling. *Nucleic Acids Res* **40**, 11648-11658 (2012).

44. Jia, G., Fu, Y. & He, C. Reversible RNA adenosine methylation in biological regulation. *Trends Genet* **29**, 108-115 (2013).

45. Cantara, W.A. et al. The RNA modification database, RNAMDB: 2011 update. *Nucleic Acids Res* **39**, D195-201 (2011).

46. Kellner, S., Burhenne, J. & Helm, M. Detection of RNA modifications. *RNA Biol* **7**, 237-247 (2010).

47. Schaefer, M., Pollex, T., Hanna, K. & Lyko, F. RNA cytosine methylation analysis by bisulfite sequencing. *Nucleic Acids Res* **37**, e12 (2009).

CHAPTER 2

HIGH RESOLUTION TRANSCRIPTOME-WIDE RNA CYTOSINE METHYLOME OF MOUSE EMBRYONIC FIBROBLASTS

Introduction

Differential behavior of cytosine and 5-methyl-cytosine (m^5C) nucleotides in exposure to sodium bisulfite mixture at acidic pH, has provided a valuable tool for sequencing-based methylation profiling of the genomic DNA.[1] In principle upon bisulfite treatment, all of the unmethylated Cs are deaminated and get converted to U, appearing as T after sequencing, while m^5Cs are refractory to this deamination and remain as Cs after sequencing, an indication of the sites of cytosine methylation. The so called bisulfite sequencing method can define the methylation pattern of individual DNA strands, in single base resolution, and has been successfully scaled up to define the entire eukaryotic genomic DNA methylomes in coupling with high-throughput sequencing.[2-4]

Bisulfite treatment, however, for the second important nucleic acid polymers of the living organisms, the RNA, has been used for more divergent applications in the past. More than four decades ago Shapiro et al. published the first report on specific deamination of cytosine bases in RNA molecules by sodium bisulfite treatment at acidic pH and also demonstrated that bisulfite mediated deamination requires the exposure of cytosines in the single stranded nucleic acid forms.[5] Taking the advantages of these two features, bisulfite treatment was used for defining the secondary structures of bacterial ribosomal RNA, with the fact that bisulfite ions cause C to U conversion of only the exposed cytosines but not the Cs in the C-G hydrogen bounds within the double stranded regions.[6] High-throughput sequencing of the bisulfite treated RNA, has also been used for distinguishing the sense from antisense transcripts, prior to the availability of the directional cDNA sequencing, with the fact that bisulfite conversion changes the sequence context in the way that the sense and antisense strands are no longer completely complementary and therefore distinguishable from each other.[7]

m^5C nucleotides have been detected in many RNA species in organisms from all three kingdoms of life and it is interesting to find their exact positions within the RNA molecules.[8] The first bisulfite treatment based determination of m^5C sites in RNA molecules, backs to the mapping of the methylated cytosine in *Saccharomyces cerevisiae* $tRNA^{His}$ by primer extension over the bisulfite treated RNA.[9] Few years later the RNA bisulfite sequencing technique was established deliberately for mapping of both known and unknown m^5C sites in tRNA and rRNA molecules[10] and successfully been used later on to find the two new tRNA targets of DNA methyltransferase 2 (DNMT2); $tRNA^{Gly}$ and $tRNA^{Val}$ beside its known target $tRNA^{Asp}$.[11] This new RNA bisulfite sequencing technique, however, showed lower C to U conversion rates (about 95% conversion efficiency) when analyzing the larger RNA species, such as 28S rRNA, most likely due to the highly structured nature of the molecules bearing a number of partial double stranded regions, requiring extended treatment time, which itself, resulted in RNA degradation and lower recovery yields.[10] Finally in 2012 the first transcriptome-wide high resolution RNA methylome was reported for HeLa cells. This work reported over 10,000 m^5C sites in tRNAs, mRNAs and other non-coding RNAs (ncRNAs) demonstrating that m^5C is a widespread modification in eukaryotic RNAs. However, since this is the first report and the only existing transcriptome-wide RNA methylome analysis and especially because of the short bisulfite exposure time of 4 hours at 75^oC according to their procedure,[12] it is unclear whether all of the reported m^5C sites are true methylation sites, due to possible incomplete conversion caused by highly structured RNAs and also possible mapping errors (see below). Thus validation of the reported m^5C sites in this work requires reproduction of the results by other groups and /or application of similar protocols in HeLa or other cell types. Here, in order to establish a valid and verified bisulfite sequencing approach we evaluated the existing recipes and examined multiple parameters, testing them at the level of nucleosides, oligonucleotides (methylated and unmethylated, and

linear and structured) and total RNA extracted from the cells, to formulize an effective modified RNA bisulfite sequencing method. Our new protocol utilizes formamide in the bisulfite mix and RNA fragmentation of long RNAs prior to bisulfite treatment to achieve both higher conversion rates and higher recovery yields. This new procedure has proven to produce close to 99% conversion rate with minimal false positive calls, reliably applicable for high-throughput sequencing platforms. Here in this report we present the first high-resolution deep RNA cytosine methylomes of mouse embryonic fibroblasts (MEFs) isolated from wild-type (wt) and Dnmt2 null mice (dnmt2-/-) demonstrating that at normal conditions $tRNA^{Asp}$, $tRNA^{Gly}$ and $tRNA^{Val}$ are the only Dnmt2 targets in MEFs.

Materials and methods

Bisulfite treatment and HPLC analysis of nucleosides

In a 1.5ml eppendorf microtube, 480 nmoles of cytosine or 5-methyl-cytosine nucleosides were mixed with 120 nano-moles of 2'-deoxy-guanosine (as the internal control) and ddH2O was added to the mixture to the final volume of 285µl. In a separate reaction tube, 312µl of the freshly prepared 5M sodium bisulfite (pH 5) were mixed with 3µl of the freshly prepared 100mM hydroquinone and were added to the nucleosides mixture and incubated at 50°C. After 1.5 hrs 100µl of the reaction mixture were directly injected into the column for the HPLC analysis with Polar RP column in 1% TFA buffer. In separate HPLC runs each one of Cytosine (C), 5methyl-Cytosine (m^5C) and deoxy-Guanosine (dG) nucleosides were injected into the same column to draw the reference HPLC peaks. The reference peaks then were overlaid on the peaks obtained from the sodium bisulfite treated samples.

Bisulfite treatment and sequencing of the synthetic RNA oligonucleotides

Sample preparation: Synthetic linear and structured oligonucleotides (methylated and nonmethylated) were used (Table 2.1) to study the efficiency and recovery yield of RNA bisulfite sequencing. The structured RNA oligonucleotides were subjected to refolding by mixing each one of the oligonucleotides, separately, with hybridization buffer, and heating it up to 75^{O}C and letting it to cool down gradually to room temperature. The refolded oligonucleotides were then ethanol precipitated and dissolved in RNase free ddH2O. The quality of the hairpin loop structure formation was then checked on nondenaturing polyacrylamide gels (PAGE) using appropriate size markers.

Denaturation step: For the bisulfite treatment step, for each of the oligonucleotides, in a separate 2ml eppendorf microtube, up to 5µg of the oligonucleotide was mixed with RNase free ddH2O (Ambion) to the final volume of 45µl. For the nonformamide containing reactions 240µl of ddH2O and for the formamide containing reactions 240µl of deionized 100% formamide was added and mixed. The mixtures were then incubated at 95^{O}C for 5min for denaturation. The nonformamide containing reactions were directly exposed to the bisulfite reaction mix (sulfonation step), while the formamide containing reactions were placed on ice for at least 2 min (to avoid causing precipitates) prior to exposure to bisulfite reaction mix.

Table 2.1. Oligonucleotide sequences

Oligo Name	Sequence
60mer linear RNA	5'-GUGUCACAUAGUACCGGAUGUCGACUAAUCGAUAUUGCGCAUCUCGAGUGAAUUCUGAUA-3'
Methylated 60mer linear RNA	5'- GUGUCACAUAGUACMGGAUGUMGACUAAUMGAUAUUGMGCAUCUMGAGUGAAUUCUGAUA-3' *
70mer structured RNA	5'- CUAUCAAAAUUCACUACUUCAAGGUUCCCCGCCUGUCACGCGGGAGACCUUGAUGUAGUGAAUUUUGAUA-3'
Methylated 70mer structured RNA	5'- CUAUCAAAAUUCACUAMUUCAAGGUUMCCCGCCUGUCAMGCGGGAGAMMUUGAUGUAGUGAAUUUUGAUA-3'
Bis-F	5'-CCCATACTCACTATCAAAATTCACT-3'
Bis-R-L	5'-GGTTGGGATGAGGTGTTATATAGTAT-3'
Bis-R-S	5'- GGTTGGGATGAGTTATTAAAATTTAT-3'

* M: methylated cytosine (m^{5}C)

Sulfonation step: After denaturation 312µl of the freshly prepared 5M sodium bisulfite (PH 5) were mixed with 3µl of the freshly prepared 100mM hydroquinone and were added to the nucleosides mixture and incubated at 50°C. For the formamide reaction the tube's cap was tightened by both wrapping parafilm around it and using appropriate plastic clamps. (This was important to keep the tube's cap closed up to the end of the reaction).

Desulfonation step: illustra NAP-10 Columns (GE Healthcare) were used to get rid of the excess amounts of bisulfite ions in the samples. First we equilibrated one column per sample with 15ml of RNase free ddH2O. Then we loaded the 600µl of the reaction mix into the column and after all of it went through the column 400µl of ddH2O was added and allowed to go through the column. Next we added exactly 1ml of ddH20 to the column and collected the eluates of each column in a new 1.5ml eppendorf tube (1ml). Then we split each eluate into two 1.5ml tubes (500µl each) and added 500µl of 2M Tris buffer PH 9.0 (Trizma® Pre-set crystals PH 9.0 – Sigma-Aldrich), and incubated at 37^{O}C for exactly 2 hrs.

Recovery: The bisulfite treated RNA molecules were recovered by ethanol precipitation. First, 110µl of 3M RNase free Sodium Acetate (Ambion) was added to each of the tubes plus 10µl of 15mg/ml GlycoBlue (Ambion) and mixed well. The content of each tube was then split into 3 new 1.5ml tubes (373µl each) (total of 6 1.5ml tubes per bisulfite treated sample), mixed with 933µl absolute ethanol and incubated at -80^{O}C overnight. Next the tubes were spun at 14K rpm at 4^{O}C for 20 minutes, the pellets were washed with 1ml of 70% ice cold ethanol one time and air dried and dissolved in 10µl of RNase free ddH2O. Next the contents of all 6 tubes of the same sample were pooled and stored at -80^{O}C until used.

RT-PCR, cloning and sequencing: The recovered bisulfite treated RNA oligonucleotides were then subjected to reverse transcription reactions (using the forward primer, Bis-F)

followed by PCR (using Bis-F as the forward primer for all oligonucleotides, Bis-R-L and Bis-R-S for the linear and structured oligonucleotides, respectively, as the reverse primer) (Table 2.1). Then the PCR products were cloned individually into TOPO-TA cloning vectors (Invitrogen) and used for transformation of Top-10 Ecoli competent cells (Invitrogen) and plated on the ampicillin containing plates. The next day several colonies for each experiment were picked and inoculated in liquid culture media for plasmid purification. Purified plasmids were then sequenced and analyzed to define the conversion rate and the methylation pattern of the bisulfite treated RNA molecules.

Bisulfite treatment of the RNA isolated from cultured cells

To test the efficacy of the formamide-based bisulfite treatment method on biological RNA samples, RNA from HeLa cells were isolated using Trizol reagent (Invitrogen) and the quality of RNA were tested using Bioanalyzer (Agilent technologies). The isolated RNA was fractionated into small and fragmented total RNA fractions containing all large RNAs (fragmented large RNA fraction). The small RNA fraction, isolated by Mirvana kit (Ambion), and the HeLa cells' total RNA was fragmented to about 60-200bp fragments using RNA fragmentation reagent (Ambion) at 70^{o}C for 15min according to the manufacturer protocol. Small and fragmented large RNA fractions were separately subjected to bisulfite treatment. The recovery yield and RNA integrity were analyzed on Bioanalyzer.

High-throughput RNA bisulfite sequencing of MEFs samples

MEF preparation, high-throughput RNA bisulfite sequencing and analysis were done as previously described[13] and the datasets are publicly available through GEO with the following accession number: GSE44359.

Results

Optimization of the RNA bisulfite sequencing method

In order to find the best conditions to improve the poor C to U conversion rate of the existing methods (especially for the structured large RNAs), and also to increase the recovery rate upon extended treatment time, we performed a series of experimental studies over the bisulfite conversion kinetics using ribo-nucleosides, synthetic oligo-nucleotides and total RNA extracted from cultured cells. We measured the conversion kinetics by bisulfite treatment of the cytidine and 5-methyl-cytidine nucleosides. Our analysis showed that the bisulfite mix efficiently deaminates cytidine but not 5-methyl-cytidine in a short period of time (< 1.5hrs) (Figure 2.1).

We then tested the parameters on synthetic methylated and nonmethylated, linear and structured short RNA oligo-nucleotides and showed that efficient deamination of all cytosines within the RNA polymers in the same bisuflite mix requires longer exposure time of several hours. We noticed that majority of Cs were converted in the first hour (data not shown) but full conversion of all cytosines was only achieved upon exposure period of 12 hours or more. This is likely because of possible base pairing of Cs with neighboring Gs within the same RNA molecules or with Gs in other RNA molecules, decreasing the efficient exposure of each cytosine to bisulfite mix. We also noticed that even in extended exposure times, for RNAs capable of forming strong secondary structures, some cytosines remain unconverted (Figure 2.2).

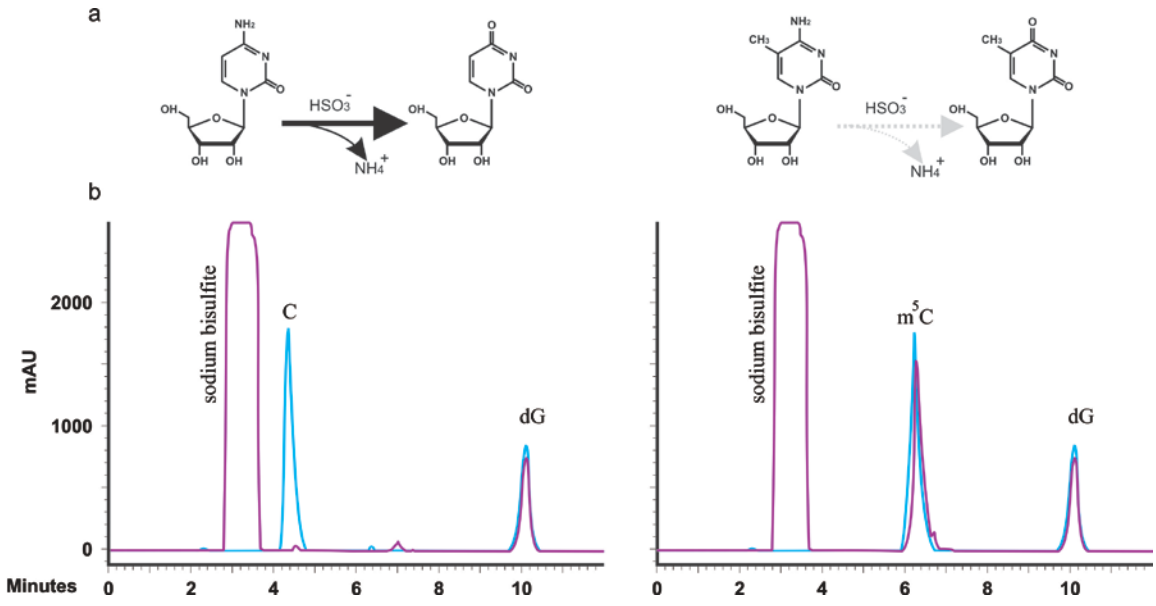
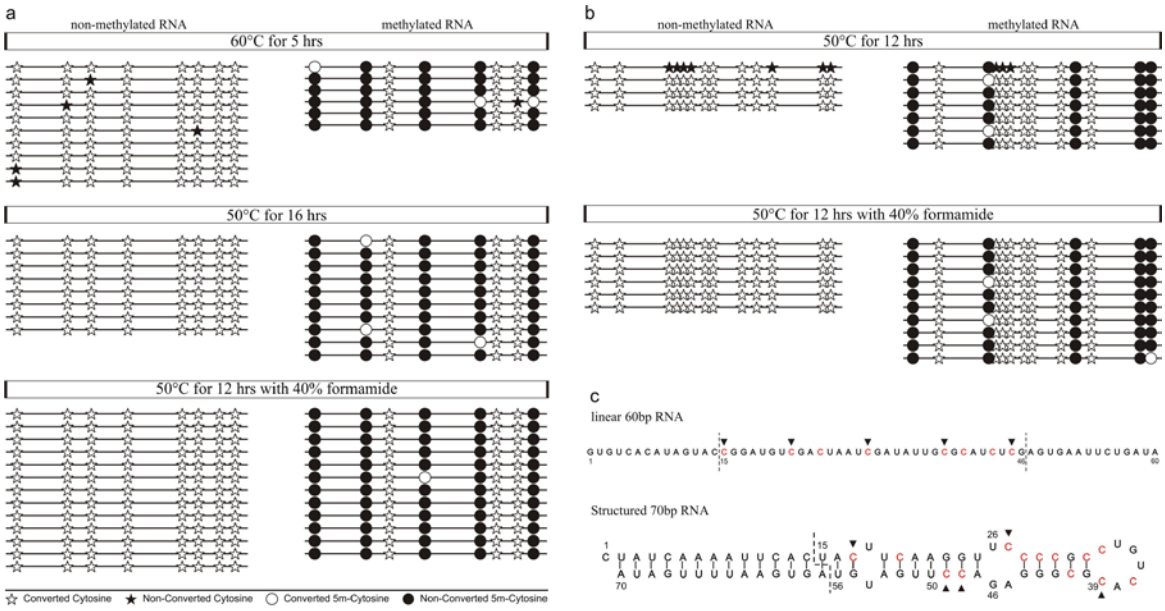


Figure 2.1 | HPLC chromatogram of Cytidine (C) and 5-methyl-cytidine (m^5C) nucleosides before and after sodium bisulfite treatment. a, The schematics of the effect of bisulfite ions on specific deamination of cytidine and 5-methyl-cytidine nucleosides. **b**, HPLC chromatograms showing the kinetics of bisulfite treatment on C and m^5C nucleosides. Blue lines show the non-treated samples and purple lines show the bisulfite treated samples for 1.5 hrs at 50°C. The 2'-deoxy-Guanosine (dG) has been used as an internal control.



d

Type	Linear Oligonucleotide (60nt)			Structured Oligonucleotide (70nt)	
Conditions	no formamide at 60°C (5hrs)	no formamide at 50°C (16hrs)	40% formamide at 50°C (12hrs)	no formamideat 50°C (12hrs)	40% formamide at 50 C (12hrs)
Rate of Cytosine Conversion	92/98 = 93.9 %	94/94 = 100 %	140/140 = 100 %	98/108 = 90.7 %	158/158 = 100 %
Rate of 5m-Cytosine Conversion	3/30 = 10 %	3/50 = 6 %	1/60 = 1.7 %	2/35 = 5.7 %	3/50 = 6 %

Figure 2.2 | Schematic representation of the bisulfite sequencing results of the linear and structured RNA oligonucleotides in different conditions and the effect of formamide. a, Bisulfite sequencing results of the three different conditions tested on linear nonmethylated (left) and methylated (right) synthetic RNA oligonucleotides. b, Bisulfite sequencing results of the two different conditions tested on structured nonmethylated (left) and methylated (right) synthetic RNA oligonucleotides. c, sequence of the linear 60bp and structured 70bp RNA used in this experiment. The regions between the dotted lines show the readable regions after sequencing. (The 5' and 3' sides have been used as the primer binding site for bisulfite specific PCR primer sets for amplification of the bisulfite treated RNAs.) Arrowheads show the place of the methylated cytosine in the methylated oligonucleotides. d, Summary of the bisulfite conversion efficiencies obtained in the different conditions tested in (a) and (b). Note that there is always some background level m^5C to U conversion.

To solve this incomplete conversion rate and also increase the recovery yield we used formamide, a strong denaturing reagent, in the mixture.

In this new procedure we used 100% deionized formamide (final concentration of 84%) to denature the highly structured RNA molecules at elevated temperatures (75-95^{O}C) and also kept the formamide during the bisulfite treatment, at 40% final concentration, to both protecting the RNA molecules from degradation and preventing the denatured secondary structures from reformation, for efficient conversion. Using this recipe we were able to extend the treatment time to 16 hours at 50^{O}C to get the 100% conversion efficiency and about 50-60% recovery rate (Figure 2.2). Next we tested this on small and large RNA fractions isolated from cultured cells. For the small RNA fraction we got comparable conversion efficiency (confirmed by high-throughput sequencing) and recovery rate. However, for the large RNA fraction our initial experiments with the intact total RNA failed to return enough uniform RNA for library preparation (Figure 2.3).

To test whether the fragment size will affect the recovery yield, we fragmented the HeLa cells' total RNA into fragments of about 60-200bp using RNA fragmentation reagent (Ambion), and subjected it to bisulfite treatment. Surprisingly this resulted in uniform RNA population, and recovery yield comparable to the small RNA fraction (Figure 2.3). Interestingly fragmentation of RNA prior to bisulfite treatment may also have some other advantages (see discussion).

High-throughput RNA bisulfite sequencing of MEFs samples

After testing the modified RNA bisulfite sequencing method with synthetic oligonucleotides methylated at defined locations and optimizing the recovery yield by testing

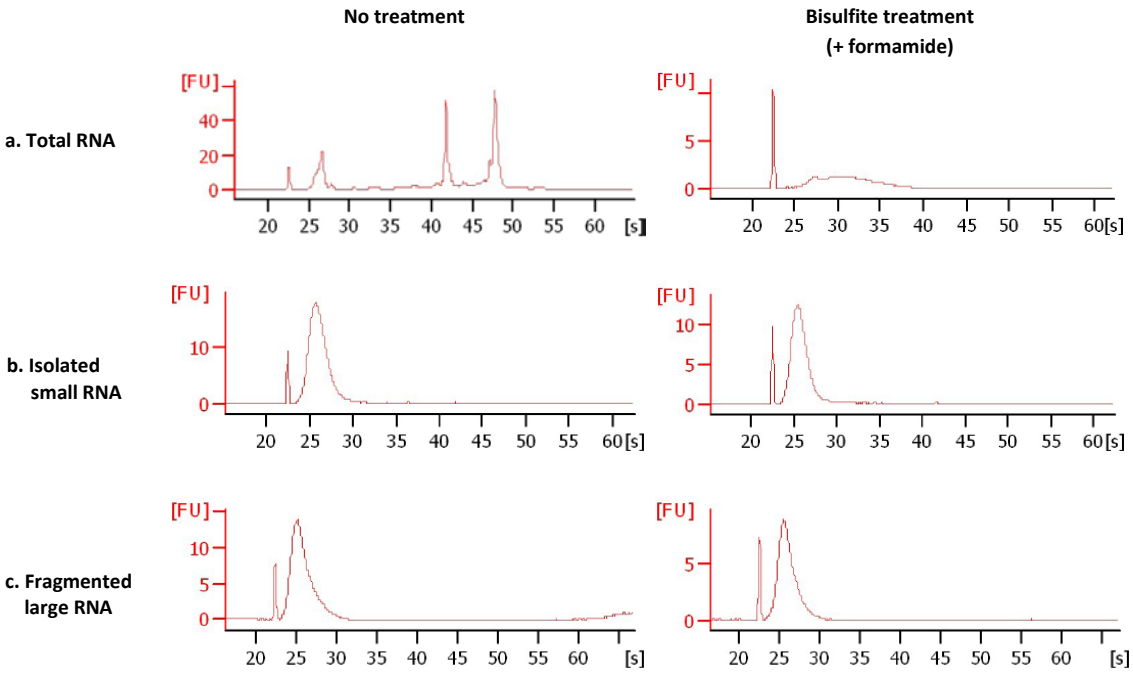


Figure 2.3 | Effect of RNA fragmentation on the recovery yield and uniformity of the fragment sizes after bisulfite treatment. Bioanalyzer results show that bisulfite treatment of the small RNA and fragmented large RNA samples produces uniform population of fragment sizes and better recovery yield, while bisulfite treatment of the intact total RNA results in severe degradation making a mixed population of different fragment sizes with lower recovery yield. The Y axis is the fluorescent unite (FU) and the X axis shows the retention time of each individual fragment in seconds (correlating to its length). This experiment has been done in the presence of formamide.

the parameters on total RNAs extracted from the cells, we applied the technique on RNA samples extracted from mouse embryonic fibroblasts (MEFs) isolated from wild-type (WT) or Dnmt2 null mice. We chose MEFs for their availability and also because that they represent a single cell type. Thus, they are preferred over using a tissue, as this will help in reducing the complexity of the transcriptome and increasing the resolution and coverage of sequencing especially over expected low-copy transcripts.

MEFs were separately isolated from day 13.5 isogenic embryos from either of wild-type (B6129PF2/J) or Dnmt2-/- (B6;129-Trdmt1^{tm1Bes}/J) [14] mice. After passaging for several days, MEFs were harvested and subjected to RNA purification and DNase treatment to remove DNA contaminants. Part of the RNA samples were used for small RNA preparation using MirVana kit (Ambion) and the rest were subjected to ribosomal RNA depletion via Ribominus kit (Invitrogen). The ribosomal RNA deplete fractions were then subjected to chemical fragmentation to yield fragments of 60-200bp sizes. Each of the small and fragmented large RNA fractions from each of WT or Dnmt2-/- MEFs were split into two equal portions: one for direct RNA sequencing and one for RNA bisulfite sequencing. A total of eight samples were then used each separately for library preparation and 101-single end sequencing, each sample on one lane with Illumina's HighSeq-2000. A flowchart representing multiple steps of our high-throughput RNA bisulfite sequencing protocol is provided in Figure 2.4. We sequenced the nonbisulfite treated samples because in our pilot experiments we noticed that after bisulfite sequencing and mapping some small RNAs mapped to some regions of the genome (including coding and repeat regions) even though the surrounding regions in many cases did not show any mapped reads. We considered this as an indication of mapping errors happening due to lower base composition complexity of the RNAs after bisulfite treatment, where the reads 'acquire' a new area of

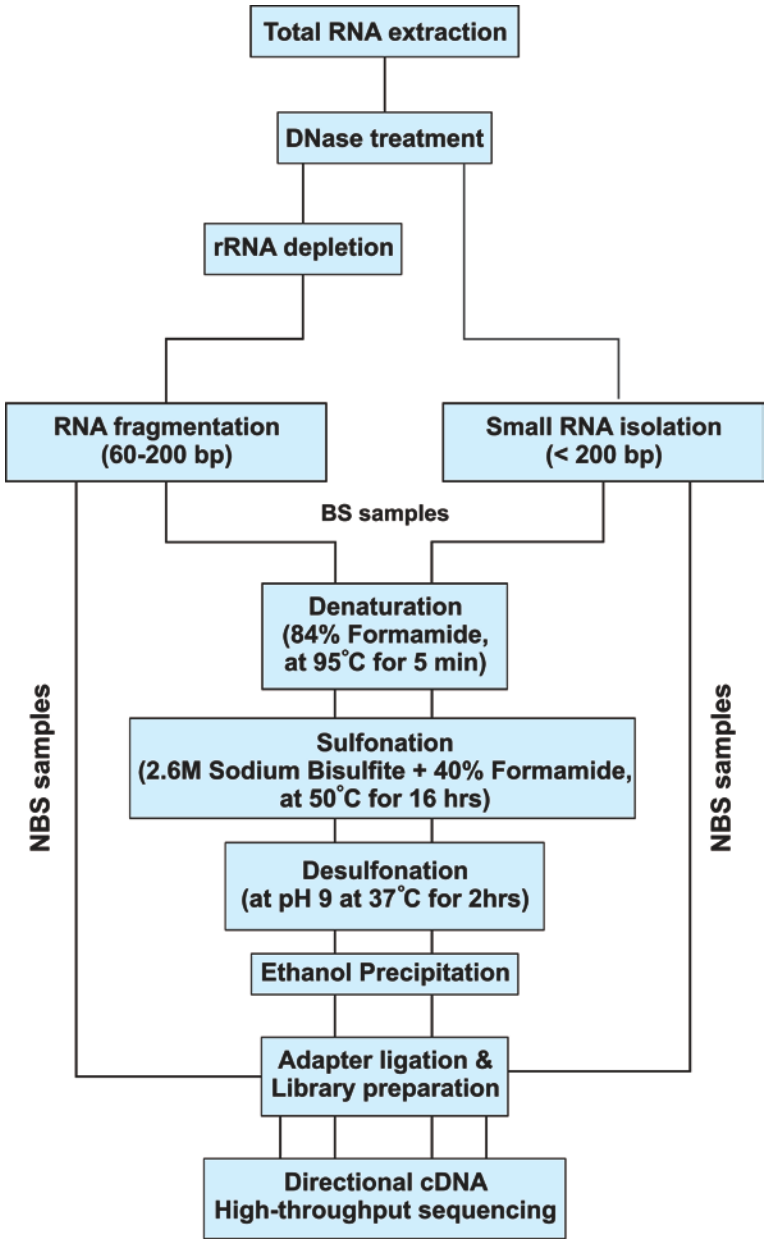


Figure 2.4 | High-throughput RNA bisulfite sequencing flowchart. This flowchart shows the stepwise procedure of high-throughput RNA bisulfite sequencing. See text for details.

mapping due to C>T conversions. Implementing this allowed us to avoid misannotation of the sequencing reads, and introduction of false positives. To remove these artifacts we sequenced the nonbisulfite treated portions of each one of the samples separately and used them as a mapping control at the time of analysis. The assumption is that if a bisulfite treated RNA is mapped to a specific location of the genome, the same location should show some sequencing read coverage in the nonbisulfite treated sample as well. In another word, those reads that are only present in the bisulfite treated but not in the corresponding nonbisulfite treated datasets are most likely due to mapping errors (acquired mapping artifacts) and should not be considered true sequencing reads when calling the methylation sites.

Developing an analysis pipeline for RNA bisulfite sequencing datasets

To facilitate the identification of candidate m^5C sites in the bisulfite treated RNA datasets we developed an analysis pipeline summarized in Figure 2.5. The commercial Novoalign package (http://www.novocraft.com) was used to align the sequencing datasets from each one of the eight MEF RNA samples (Bisulfite (BS) or nonbisulfite (NBS), Small or large RNAs, from wt or dnmt2-/-) to the mouse transcriptome index files. First two separate transcriptome index files, normal (nonbisulfite converted) and bisulfite converted, were made using "Novoindex application" from Novoalign package. To make the Novoindex files, the entire genome sequences (from all chromosomes) from *M. musculus* July 2007 (NCBI37/mm9) genome build was used. In addition to include all known and theoretical splicing events, the "MakeTranscriptome" application from open source USeq package[15] was used to extract, and include, all known and theoretical splice junctions derived from Ensembl transcripts from the same genome build. Sequence alignments were performed using the "Novoalign application"

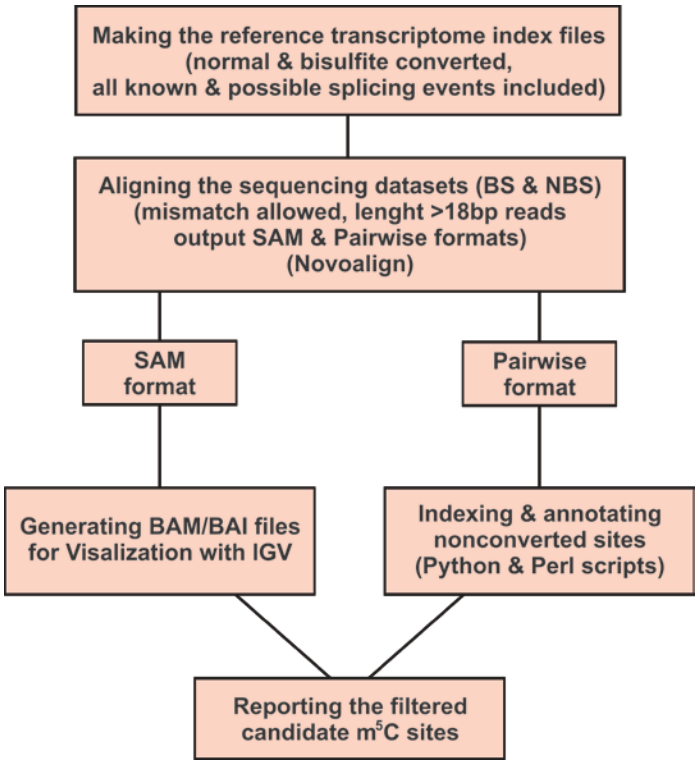


Figure 2.5 | RNA methylation analysis pipeline. This flowchart shows the bioinformatics analysis steps for finding the candidate methylated sites in the transcriptome. See text for details.

from the Novoalign package, with options to allow gaps and mismatches, reporting 18bp or larger inserts, and reporting all of the reads mapped to the repeats. The nonbisulfite treated datasets were aligned to the normal transcriptome indexed file, and alignment output files were generated in the SAM format. The bisulfite treated datasets were aligned to the transcriptome indexed file generated in the bisulfite mode, and alignment output files were generated in both the SAM and Pairwise formats. To visualize the sequencing reads the SAM-formatted alignment files of the bisulfite treated or untreated datasets were processed by the "RNASeq" application of the USeq package to make the BAM/BAI files. The BAM/BAI files were then used to inspect the sequencing reads at base pair resolution using the Integrative Genomics Viewer (IGV).[16] Pairwise-formatted alignment files were processed by custom python scripts to call, annotate and make the tables of candidate methylated cytosines. Custom Perl scripts were used to filter the methylation tables (read coverage $\geq$ 10 and nonconversion rate $\geq$ 20 % in any of the datasets).

To find the Dnmt2 target sites the candidate methylation table was filtered again to keep only sites showing significant methylation in wt but not in dnmt2-/- datasets. The candidate sites were then verified by manual inspection of the individual mapped reads, by filtering out the sites showing coverage only in the bisulfite treated (but not the untreated) datasets (an example is provided in Figure 2.6) due to mapping errors caused by lowered base composition complexity of the reads in the bisulfite treated samples. Further inspections were performed to remove the aggregated candidate sites showing clusters of non-converted Cs in the bisulfite treated datasets. To lower the rate of false positives we removed such sites from the candidate lists. An example of a clustered nonconverted region is presented in Figure 2.7.

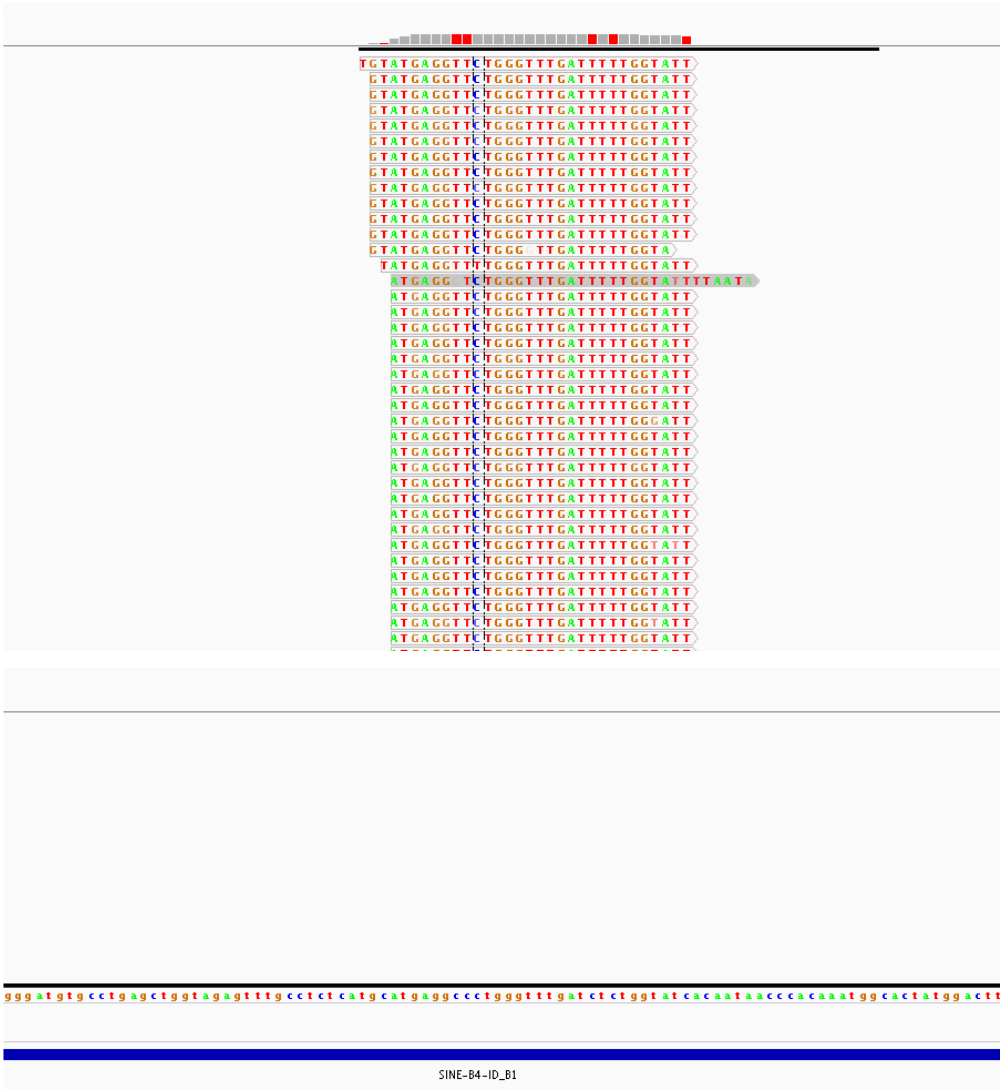


Figure 2.6 | An example of methylated sequenced reads mapped to a SINE repeat with no reads mapped to the same region in the nonbisulfite treated dataset. A random IGV snapshot from a subset of reads mapped to a SINE element. Although in the bisulfite treated datasets (top panel) the mapped reads show significant methylation in a single site, because of the lack of corresponding mapped reads in the nonbisulfite treated datasets (lower panel) this will be considered a mapping error and the corresponding candidate site will be removed from the list of candidate m^5C sites. Only the wt datasets shown as both wt and dnmt2-/- are similar.

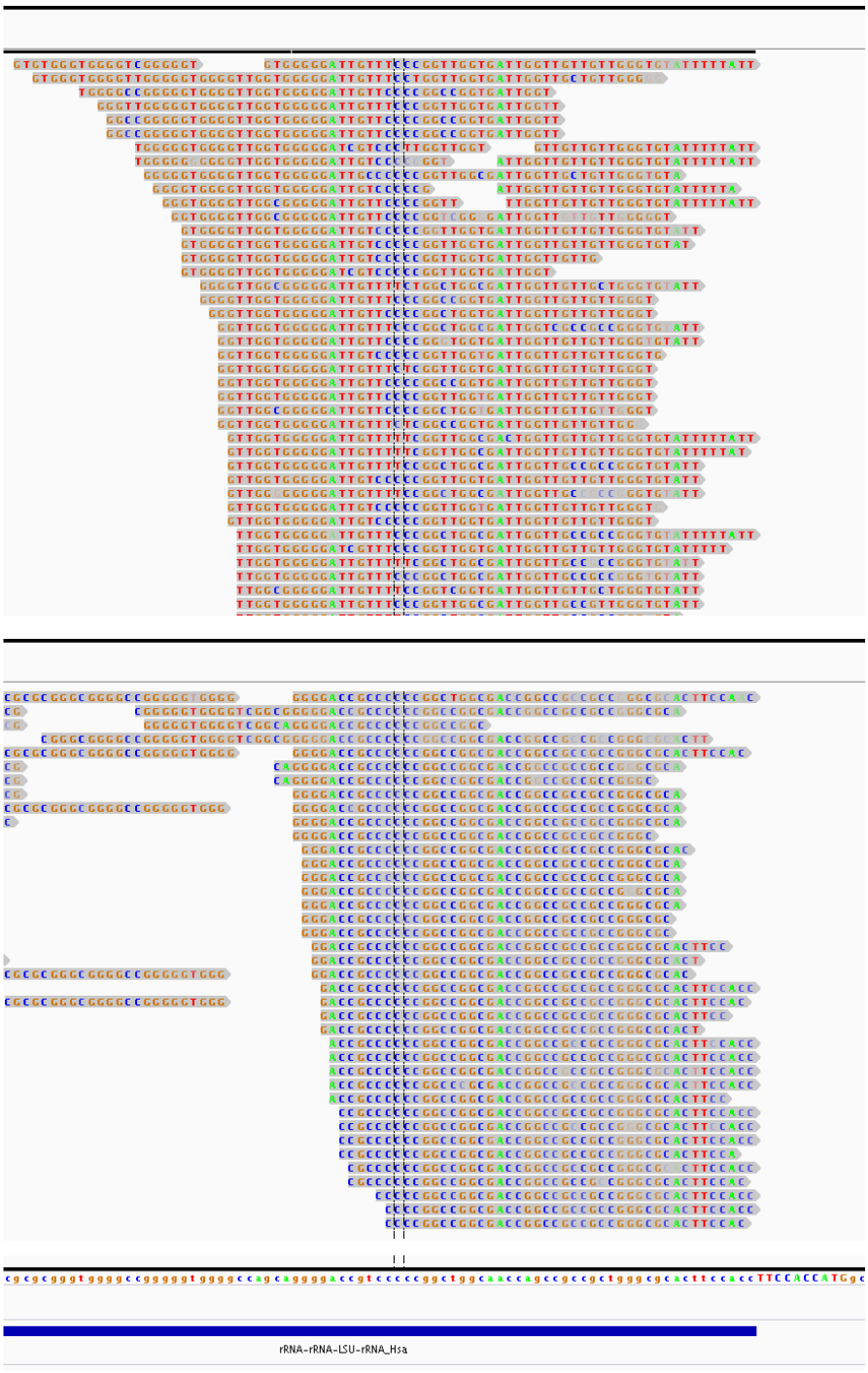


Figure 2.7 | An example of clustered nonconverted cytosines in the reads mapped to an rRNA locus. A random IGV snapshot from a subset of reads mapped to an rRNA. The region shows a high GC content. Several nonconverted cytosines in the bisulfite treated dataset (top), in this high GC rich region of the genome, indicate a possible artifact but not true methylation sites. The lower panel shows the nonbisulfite treated dataset. Only the wt datasets shown as both wt and dnmt2-/- are similar.

The MEFs RNA methylome

Alignment of the sequenced reads from bisulfite treated and nontreated, small and large RNAs from wt and dnmt2-/- samples revealed 90 to 100 million mapped filtered reads per dataset. The small and large RNA datasets from the bisulfite treated samples were combined for methylation analysis. The row methylation analysis of the wt and dnmt2-/- MEFs datasets combined revealed about 99% conversion efficiency with a consistent pattern between the two types of datasets (wt and dnmt2-/-) in all non-Dnmt2 target genes. Figure 2.8 shows the reproducibility of the applied RNA bisulfite sequencing approach by showing the consistency in both of the exact sites and also in their conversion levels for the panel of rRNA genes. Analysis of both wt and dnmt2-/- MEFs datasets, collectively, considering the cutoff of >10 read coverage and >20% nonconversion rate at the sites in either or both of the datasets, revealed a list of 13,520 candidate m^5C sites in the MEFs transcriptome. However, after visual inspection of the mapped sequencing reads in IGV (to filter the table by removing the sites showing sequence coverage only in the BS but not NBS datasets, and also removing the sites showing clustered non-conversions), about 90% of the sites were identified as artifacts, shortening the list of candidate sites to 1305 verified sites (Figure 2.9). It is important to mention that the level of methylation at each particular candidate site was different (i.e. highly, moderately or weakly methylated) (Figures 2.10 and 2.11). As expected most of the candidate sites were mapped to tRNAs as most eukaryotic tRNAs bear one or more m^5C sites in their structures. A lower number of candidate sites were found in rRNAs, snRNAs and scRNAs, and two other ncRNAs; ribonuclease P RNA (Rpph1) and ribonuclease P RNA-like 1 (Rprl1). Surprisingly, protein coding transcripts (mRNAs) showed a high number of candidate sites (total of 388 sites). These sites were mostly concentrated in the coding regions and also in 3'-UTR and fewer sites were found in

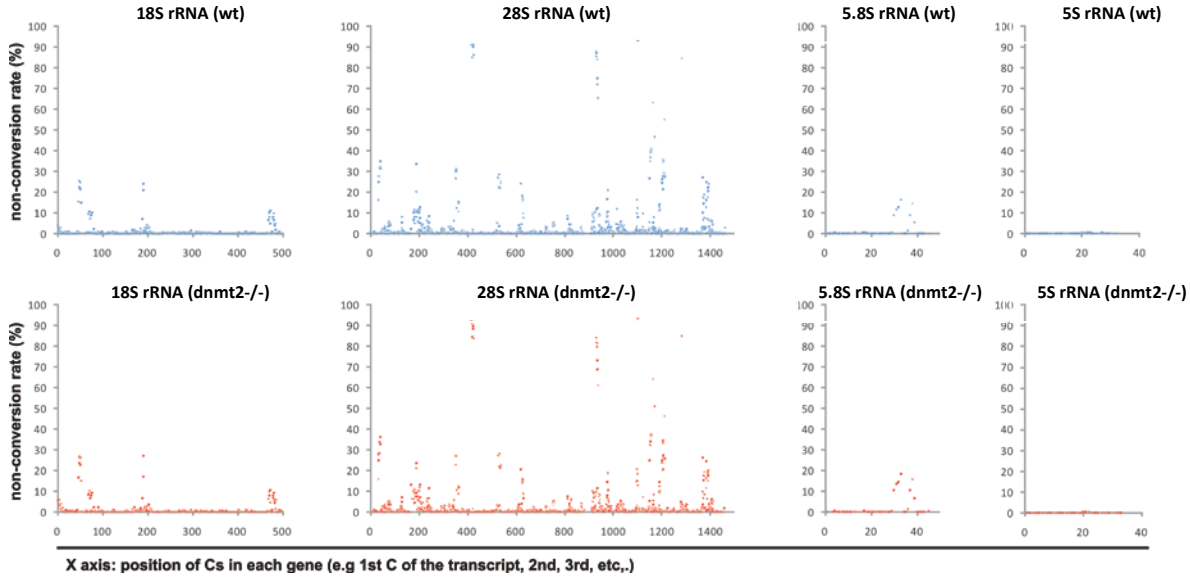


Figure 2.8 | An example of a consistent pattern of nonconverted Cs in the rRNA transcripts. The individual scatter plots of nonconversion levels are presented here showing the pattern of nonconverted Cs and their nonconversion levels in both of the wild type (wt) dataset at top and dnmt2 null (dnmt2-/-) dataset at bottom. The Y axis shows the nonconversion levels (%) and the X axis shows the position of individual Cs in each rRNA transcript. The bisulfite (BS) datasets of each sample (wt and dnmt2-/-) were separately aligned to an index file generated from an artificial transcriptome containing the standard mouse rRNA sequences (18S, 28S, 5.8 and 5S rRNA genes) obtained from gene bank. The level of conversion at each cytosine base in each of the four rRNA genes were then calculated with custom python scripts and plotted in Excel. Comparison of the nonconversion patterns in two datasets clearly shows a consistent trend for both of the exact sites and also the nonconversion levels, indicating the reproducibility of the results with the established high-throughput RNA bisulfite sequencing method. Note the observed sites with some levels of nonconversion are not necessarily true m^5C sites and further filtering steps (explained in the text) are required for defining the most probable candidate m^5C sites.

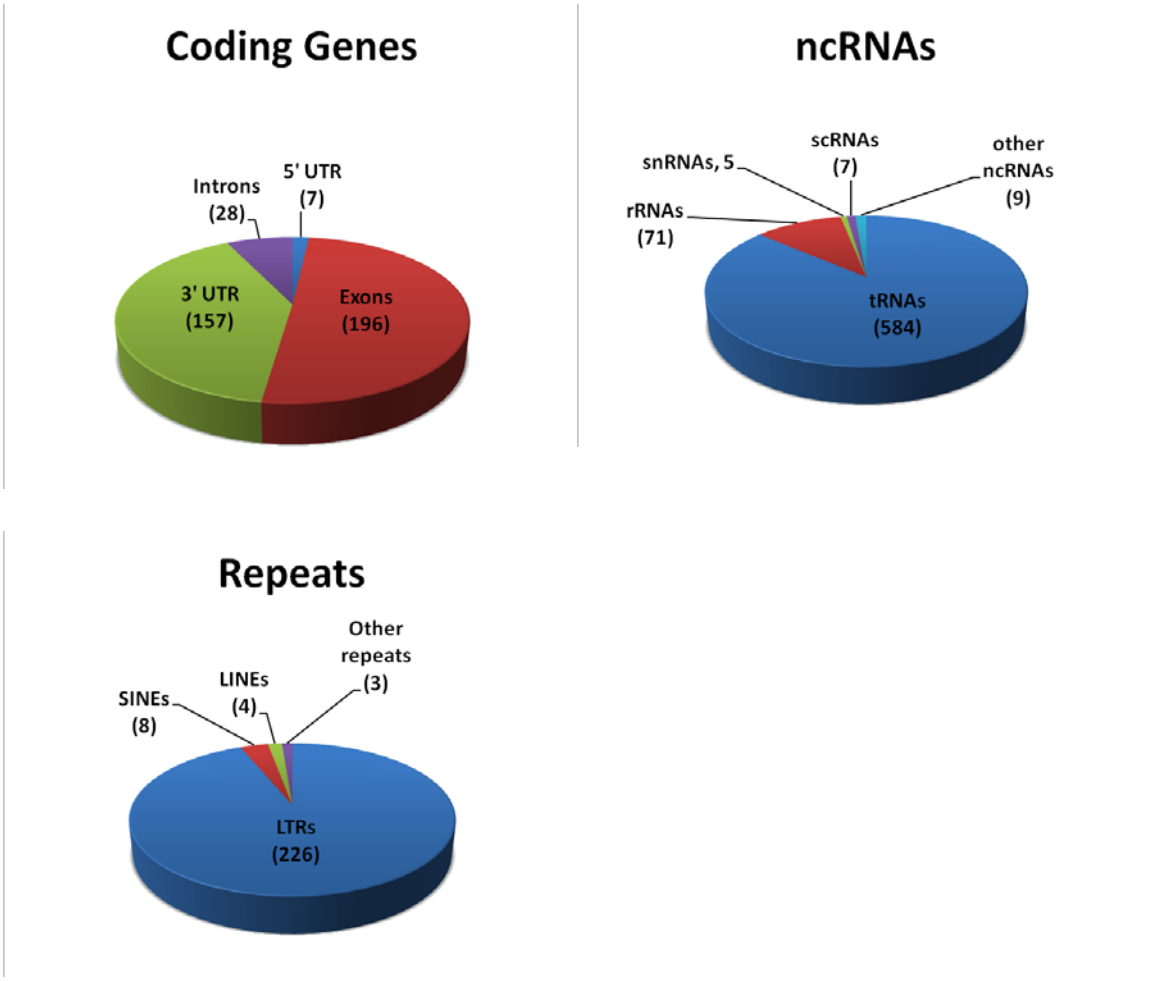


Figure 2.9 | Distribution of candidate m^5C sites in annotated coding and noncoding genes, and repeats. The pie graphs show the distribution of annotated candidate m^5C sites identified in either of the wt or dnmt2-/- datasets under different categories: a, coding genes, b, ncRNAs, and repeats. See text for details and interpretation.

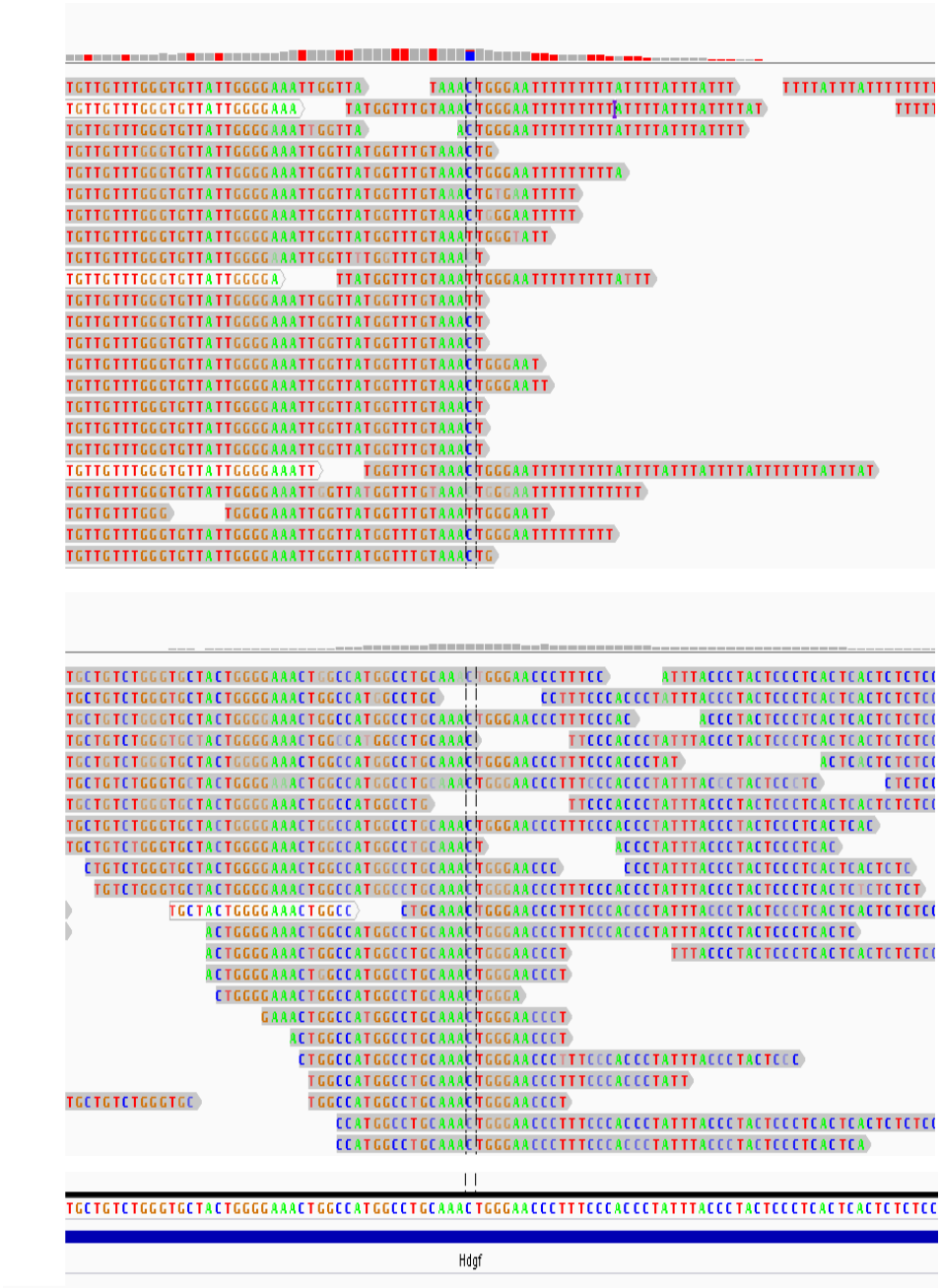


Figure 2.10 | An example of a highly methylated site. A random IGV snapshot from a subset of reads mapped to the 3'-UTR region of hepatoma-derived growth factor (Hdgf) gene. This site shows a high degree of methylation at position chr3: 87718567. The top panel shows the sequencing reads from bisulfite treated RNAs (BS) and the bottom panel shows the sequencing reads from nonbisulfite treated samples (NBS). Only the wt datasets shown as both wt and dnmt2-/- are similar.

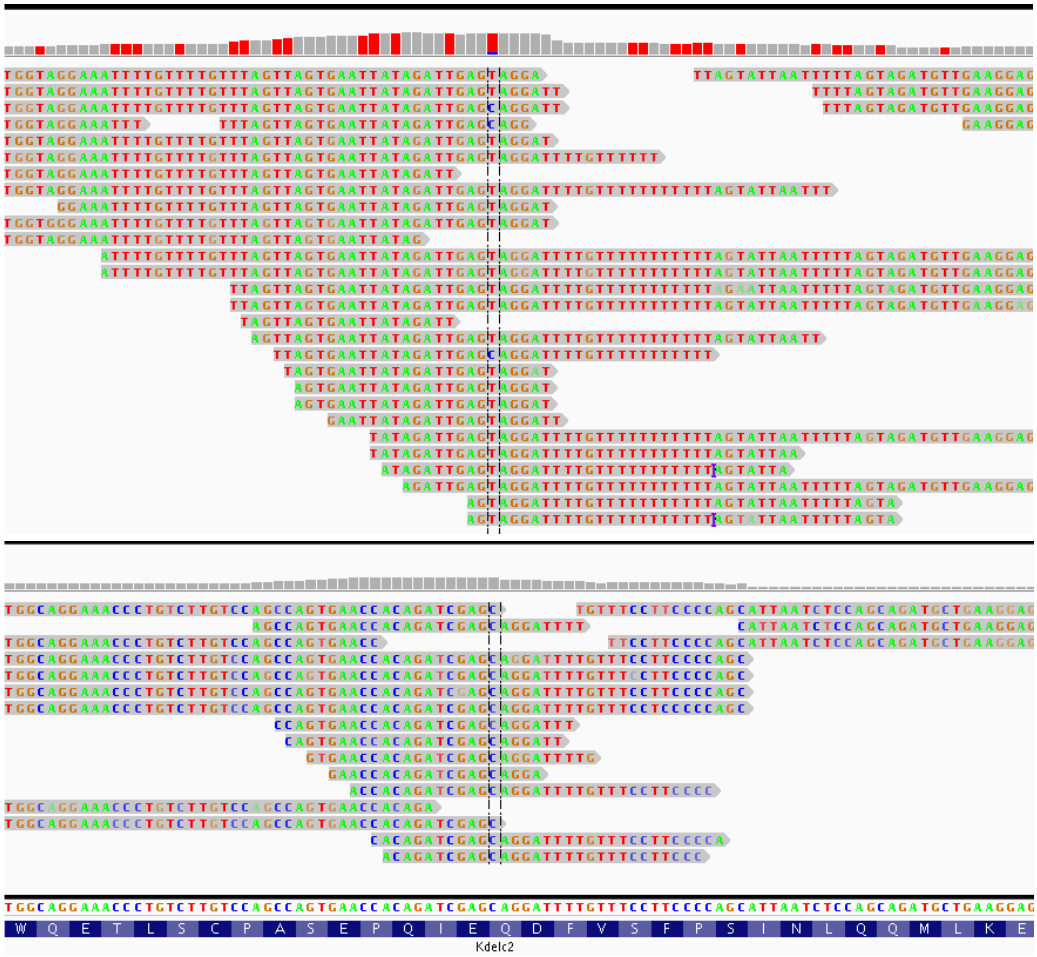


Figure 2.11 | An example of a low/moderately methylated site. A random IGV snapshot from a subset of reads mapped to the coding region of KDEL (Lys-Asp-Glu-Leu) containing 2 protein (Kdelc2) gene. This site shows low to moderate degree of methylation at position chr9:53198635. The top panel shows the sequencing reads from bisulfite treated RNAs (BS) and the bottom panel shows the sequencing reads from nonbisulfite treated samples (NBS). Only the wt datasets shown as both wt and dnmt2-/- are similar.

the reads mapped to the intronic regions or to 5'-UTR of the genes. The methylation report for the candidate m^5C sites within the protein coding genes is provided in Supplementary Table S1 in Appendix A. Most genes show only a single significant candidate m^5C site in their mRNAs; however there were also a few genes with more than one site (Supplementary Table S1 in Appendix A). The gene ontology (GO term) analysis of the genes showing m^5C sites, using the web-based GoMiner application[17] showed enrichment of genes in diverse cellular processes, with "metabolic" and "regulative-related" processes and pathways at the top of the list (Table 2.2). Regarding the sites mapped to the repeat regions quite a high number of sites were identified in LTR retrotransposons, while only few candidate sites were found in LINE, SINE, or other repeat families (Figure 2.9). Notably, almost all of the candidate m^5C sites in LTRs were from short sequences related to the tRNA binding sites of the endogenous retroviruses (ERVs) (an example in Figure 2.12). ERVs, like retroviruses, bear tRNA binding sites in their structures as they rely on tRNA binding for their replication. Because the surrounding regions of tRNA binding sites were devoid of any mapped reads in the datasets (an example in Figure 2.12), it is most likely that the methylated mapped reads are not derived from the LTRs. They are most likely tRNA fragments which can map to the tRNA binding sites of the LTRs. We note that this is not specific to the methylated tRNA fragments as for many of the other annotated LTRs, unmethylated reads were observed, most likely derived from unmethylated tRNAs or the unmethylated portions of the methylated tRNAs.

Dnmt2 targets in MEFs

DNA methyltransferase 2 (DNMT2) which is present in most eukaryotic organisms was first introduced as an enzyme with possible DNA methyltransferase activity.[18] DNMT2 is the most

Table 2.2: Gene Ontology (GO) term analysis of the protein coding genes showing cytosine methylation sites

No.	HYPERLINKED GO CATEGORY	TOTAL GENES	CHANGED GENES	ENRICHMENT	LOG10(p)
1	GO:0009987_cellular_process	12158	156	1.895199	-22.99383
2	GO:0008152_metabolic_process	8152	117	2.119894	-18.287889
3	GO:0044238_primary_metabolic_process	7002	105	2.214928	-17.141621
4	GO:0044237_cellular_metabolic_process	6932	103	2.194679	-16.419523
5	GO:0019538_protein_metabolic_process	2926	58	2.927831	-13.338066
6	GO:0043170_macromolecule_metabolic_process	5745	84	2.159641	-12.29361
7	GO:0048519_negative_regulation_of_biological_process	2292	48	3.093278	-11.761033
8	GO:0044260_cellular_macromolecule_metabolic_process	5139	77	2.213117	-11.621887
9	GO:0048523_negative_regulation_of_cellular_process	2098	45	3.168104	-11.350529
10	GO:0044267_cellular_protein_metabolic_process	2443	49	2.962545	-11.336859
11	GO:0016043_cellular_component_organization	2858	53	2.739088	-11.003136
12	GO:0071840_cellular_component_organization_or_biogenesis	2975	54	2.681014	-10.866587
13	GO:0051179_localization	3440	58	2.490359	-10.433511
14	GO:0006810_transport	2862	49	2.528825	-8.95405
15	GO:0008219_cell_death	1294	31	3.538505	-8.952245
16	GO:0016265_death	1302	31	3.516763	-8.888556
17	GO:0051234_establishment_of_localization	2898	49	2.497411	-8.773695
18	GO:0065007_biological_regulation	7925	94	1.751947	-8.731596
19	GO:0048518_positive_regulation_of_biological_process	2632	46	2.581454	-8.6713
20	GO:0033036_macromolecule_localization	1266	30	3.500096	-8.562557

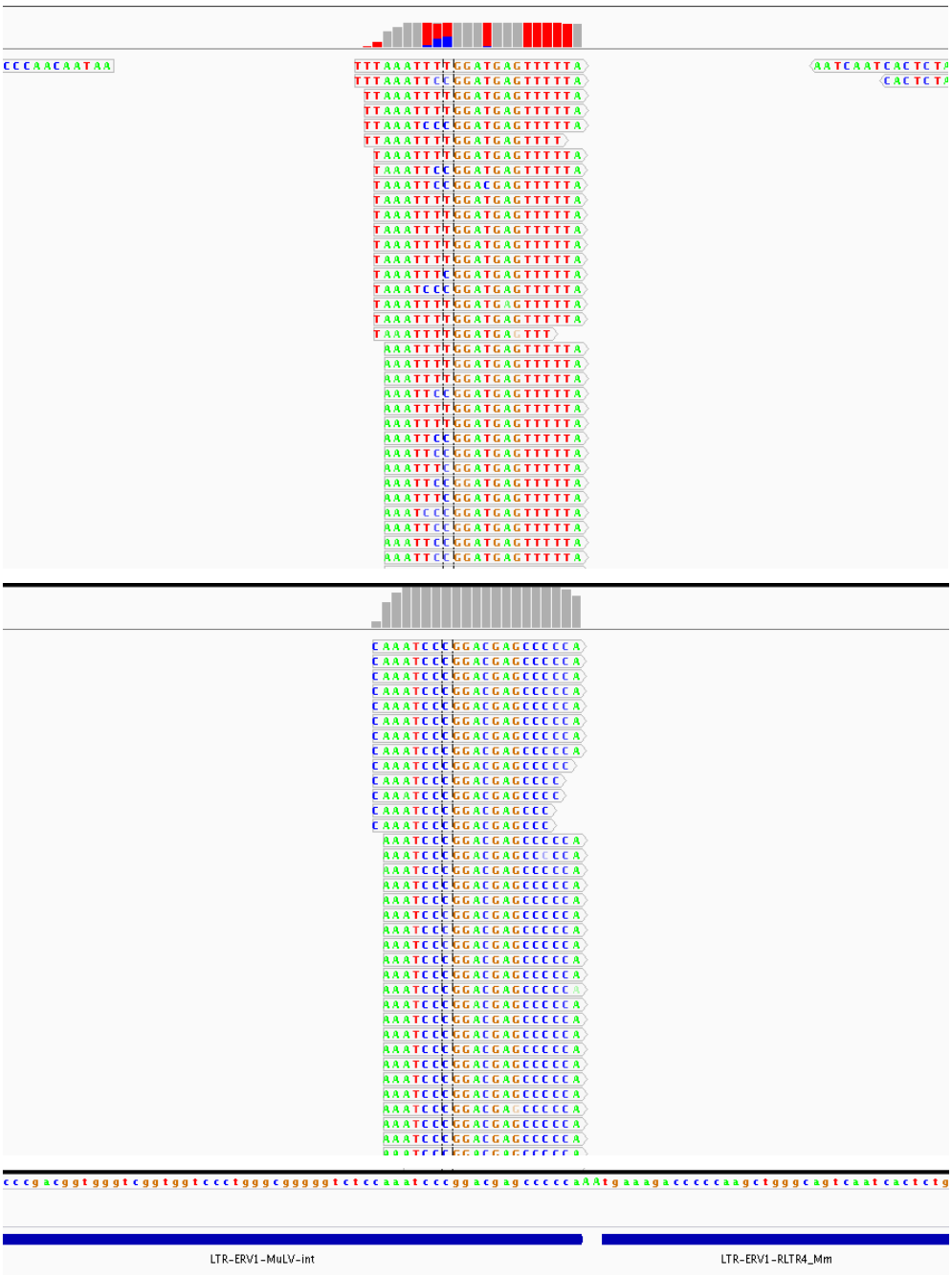


Figure 2.12 | An example of methylated sequenced reads mapped to an LTR. A random IGV snapshot from a subset of reads mapped to an LTR repeat region (LTR-ERV1-MuLV-int). Three sites in the bisulfite treated (BS) panel (top) show methylation. Also in the nonbisulfite treated panel we have sequencing reads mapped to the same region. This indicates that the methylation events are more likely real. However, the sequencing reads are less likely to be driven from the LTR although they can map to it. This is because the surrounding regions do not show any mapped reads and also the reads are mapped to the tRNA binding site of the LTR retrotransposon. Only the wt datasets shown as both wt and dnmt2-/- are similar.

widely conserved DNMT enzyme within eukaryotes, which has all of the required motifs for methyltransferase activity, and its catalytic domain resembles the methyltransferase domain of other active DNMTs, DNMT1 and 3.[18] However, more than a decade of efforts to prove the DNA methyltransferase activity for this enzyme resulted in finding that DNMT2 is actually not a DNA methyltransferase but rather is an RNA methyltransferase instead.[14, 19-21] The first RNA substrate recognized for DNMT2 was $tRNA^{Asp}$ in several organisms and the exact target residue mapped to C38 in the anticodon stem-loop of the tRNA structure.[14, 19, 20] Most organisms lacking DNMT2 lack obvious phenotype,[18] although Dnmt2 morphant zebrafish shows developmental perturbations,[19] however, the high degree of sequence and structural conservation of DNMT2 between divergent organisms prompts more thorough analysis to find its other targets and eventually its functions for the organisms.[18, 22] Since the discovery of RNA methyltransferase activity of DNMT2, only two target sites (C38 in $tRNA^{Gly}$ and $tRNA^{Val}$) were found as other targets of the enzymes.[11] There is currently no transcriptome-wide comparative RNA methylome report comparing the wt and dnmt2-/- null RNA methylomes to find the enzyme's other possible targets in a broader pool of RNA molecules. Here, as the first comprehensive analysis of its type, comparison of the entire wt and Dnmt2-/- MEFs RNA methylomes followed by filtering the false positive m^5C calls revealed that C38 in known DNMT2 tRNA targets ($tRNA^{Asp}$, $tRNA^{Gly}$ and $tRNA^{Val}$) are completely unmethylated in Dnmt2-/- datasets while they show significant methylation at the exact same sites in wt datasets. Beside these we did not detect any other differentially methylated site in the entire transcriptome. This clearly demonstrates that in MEFs, at normal conditions, exclusively C38 sites of these three tRNAs are the DNMT2 targets but no other site in these tRNAs or other coding or noncoding RNAs (Figure 2.13).

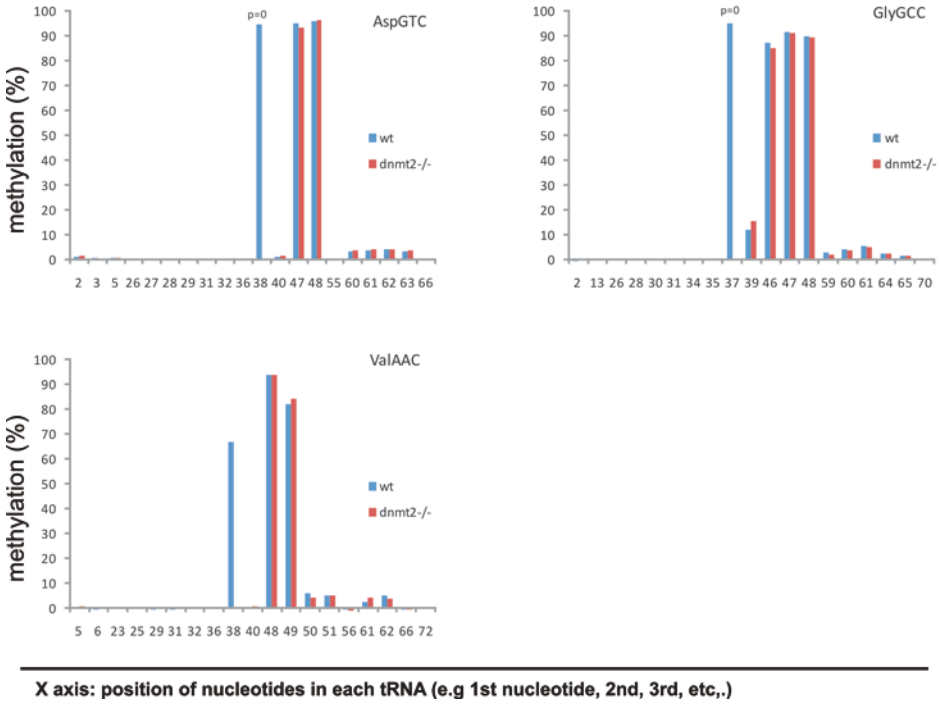


Figure 2.13 | tRNA methylation patterns in known Dnmt2 tRNA targets. The bar graphs show the methylation levels at each cytosine base within the three known Dnmt2 tRNA targets in both of wt (blue) and dnmt2-/- (red) datasets. There are methylation sites in each tRNA which are not the Dnmt2 target sites which show methylated at similar levels in two datasets. The Y axis shows the methylation levels (%) and the X axis shows the position of individual nucleotides (Cs) in each rRNA transcript. The bisulfite (BS) datasets of each sample (wt and dnmt2-/-) were separately aligned to an index file generated from an artificial transcriptome containing only the standard mouse tRNA sequences obtained from genomic tRNA database (http://gtrnadb.ucsc.edu). The methylation levels at each cytosine base in each tRNA were then calculated with custom python scripts and plotted in Excel (here only three tRNAs are shown). Comparison of the methylation patterns in two datasets clearly shows that C38 in $tRNA^{Asp}$, $tRNA^{Val}$ and $tRNA^{Gly}$ are the only known targets of Dnmt2. Note that we did not detect any other Dnmt2 target site in the entire MEFs transcriptome.

Conclusions

In this report we presented a new efficient RNA bisulfite sequencing protocol suitable for high-through transcriptome-wide RNA methylation profiling. To obtain this protocol we systematically tested different parameters at multiple levels including nucleoside, synthetic methylated/unmethyalted, and linear/structured RNA oligonucleotides and eventually total RNA, to formulate the protocol. In this approach, we introduced formamide into the mixture for two main purposes: 1) to denature the local RNA secondary structures in the denaturation and sulfonation steps for efficient conversion, and 2) to protect the RNA from degradation during the procedure to increase the recovery yield.

For efficient methylation profiling in long RNA species we used RNA fragmentation prior to bisulfite treatment. Fragmentation has several advantages. First, RNA fragmentation disrupts the secondary structures within the large RNA molecules, and thus improves the conversion efficiency to great extents. Second, because the fragmentation is a random process, breaking RNA molecules at different locations, it could potentially help to distinguish the true methylation sites from the artifacts produced by extremely strong secondary structures. Third, bisulfite treatment of the fragmented large RNA, after recovery, shows a uniform population of converted RNA molecules suitable for library preparation while the same treatment on non-fragmented total RNA resulted in drastic degradation and a mixed population of different fragment sizes.

To filter out the false positives due to possible mapping errors, as the result of the decreased base composition complexity in the bisulfite treated RNA sequences we sequenced the non-bisulfite treated samples with the logic that a mapping event is only valid if a particular genomic region contains mapped sequenced reads in both of the bisulfite treated and non-treated datasets.

Our results proved that we have formulated an efficient and reproducible RNA bisulfite sequencing approach with about 99% conversion efficiency of the unmethylated cytosines. Due to some extremely strong secondary structures in some RNA molecules we did not expect a complete 100% conversion efficiency. The sites with more probable nonconversion rates are somehow predictable and most of them are expected to reside in the GC rich regions with adjacent palindrome sequences increasing the likelihood of strong secondary structure formation. Such structured hairpin-loops might even be strong enough to resist denaturation in the presence of formamide at high temperatures. The undenatured regions of the RNA molecules will produce incomplete conversion and are the major source of false positives in RNA bisulfite analysis. In order to resolve this, during the analysis, we removed the aggregated nonconverted cytosines appearing in clusters in highly GC reach regions by visual inspection of the sequencing reads. Candidate sites falling in such clustered nonconverted regions were then removed from the initial list. This helps to narrow down the table of candidate m^5C sites to a shorter list with higher confidence level to call the true methylation sites.

It is important to note that the low copy number of most RNA species, as well as the low methylation penetration at particular sites, will decrease the chance of detecting all possible m^5C sites within the transcriptome and always some false negatives are expected. In addition, the stringent filtering parameters we have considered in our analysis, although providing more accurate list of candidate m^5C sites, can cause missing some of the true m^5C sites as well producing some false negatives. To overcome the issues with low copy number RNA species or low methylation penetration the enrichment protocols can be used[13] (see Chapter 3). Also application of emerging single molecule RNA sequencing technologies capable of direct detection of nucleotide modifications[23] (see Chapter 4) might help identify some other true m^5C

sites which may have been filtered as they appear in highly structured regions showing non-converted cytosines in their neighborhood.

Regarding the DNMT2 target identification, we showed that C38 in the stem-loop junction of the anticodon loop of $tRNA^{Asp}$, $tRNA^{Gly}$ and $tRNA^{Val}$ are the only targets of Dnmt2 enzyme in MEFs at normal conditions. This, however, does not rule out the possibility of DNMT2 to methylate other RNA targets, in other cell types or tissues and at other conditions such as in stress or immune response. Especially, Dnmt2 has been shown to be required for efficient control of RNA viruses in *Drosophila Melanogaster.*[24] Limited RNA bisulfite sequencing analysis showed that pmt1, the human DNMT2 homolog in fission yeast, is getting activated upon nutrition deprivation resulting in significant increase at the methylation level of C38 in $tRNA^{Asp}$. Interestingly this process is under the control of the serine/threonine kinase Sck2 arm of the yeast nutrient signaling pathways.[25]

Dnmt2 has been shown to be responsible for retrotransposon silencing in *Drosophila* somatic cells[26] and a recent report indicates that during stem-cell division Dnmt2 is required for chromosome-specific nonrandom segregation of the sister chromatids through a yet to be known mechanism.[27] Since comprehensive DNA methylation analysis demonstrates that DNMT2 has no detectable DNA methyltransferase activity,[21, 28] it is tempting to speculate that the aforementioned observed functions are purely due to the RNA methyltransferase activity of DNMT2. If this is true, two scenarios may be considered: 1) DNMT2 is a pure tRNA methyltransferase and methylated tRNAs or methylated tRNA fragments role in such processes, or 2) DNMT2 can methylate other RNA targets and they are responsible for the observed phenotypes. Although several reports indicate that DNMT2 is a highly specialized tRNA methyltransferase enzyme recognizing a specific cytosine in the CpG context and at the exact stem-loop junction of specific tRNAs in several species, it is possible that it recognizes tRNA-like

structures within other RNA species.[13, 29] We have recently reported that human DNMT2 can methylate a cytosine, resembling the DNMT2 target site, within a tRNA-like structure in KRT18 mRNA at very low levels.[13] It is, however, not clear whether DNMT2 can methylate this or similar sites in other RNAs at biologically relevant levels in special conditions. Overall, a comprehensive RNA methylome profiling in wt and DNMT2 null organisms in different cell types or conditions seems to be required to reveal the other possible RNA targets of this enzyme.

Regarding the MEFs methylome in this report, our transcriptome-wide methylation profiling showed that in MEFs at normal conditions, tRNAs and mRNAs are the most frequent targets of RNA cytosine methylation in comparison to other RNA species. It is important to note that quantification of the bulk m^5C contents in total RNA by chromatography and mass-spectrometry approaches (LC-MS), clearly shows that most of the bulk m^5C contents are coming from tRNAs.[29] This has two reasons: first tRNAs are highly modified as most of the eukaryotic tRNAs, in their short transcripts of about 75pb on average, bear one or more m^5C sites, and second tRNAs are extremely abundant in comparison to most RNA species including mRNAs but not rRNAs. This is evident in organisms lacking the two major tRNA cytosine methyltransferases, DNMT2 and NSUN2, in which the bulk m^5C level in RNA is dropping to close to the background level.[29] Therefore, the m^5C levels in non-tRNA species specifically in mRNAs are considered to make a very small portion of the bulk m^5C content of total RNAs, due to low copy number of the RNAs and scarcity of the sites within the long RNA molecules. However, it is highly interesting to explore the exact functions of such methylated sites.

High-throughput mapping of few RNA modifications such as N6-methyladenosine (m6A) and C to U editing events showed significant enrichment in the 3'-UTR of mRNAs.[30-32] We have also found many m^5C sites in the 3'-UTR of mRNAs. Concentration of modified nucleotides within the untranslated regions can be linked to regulation of processes such as mRNA export,

localization and translation as both 5'- and 3'-UTRs are enriched in regulatory elements. However, unlike previously reported RNA cytosine methylome in HeLa cells claiming that majority of m^5C sites are concentrated within the 3'-UTR of mRNAs,[12] we identified a comparable number of candidate m^5C sites within the coding regions too. It is therefore interesting to speculate possible functions for such sites within the protein coding segments of the mRNAs.

DNMT2 and NSUN2 are believed to be responsible for methylating all of the known sites within tRNA genes.[13] There are at least 8 other RNA cytosine methyltransferases (m^5C-RMTs) in human genome, which are believed to be responsible for methylation of most of non-tRNA sites.[8] Disruption or misregulation of some of these genes are linked to genetic disorders, infertility and cancer[8] and it is therefore interesting to look at their exact direct target molecules/sites as well as their functions in the cells.

References

1. Frommer, M. et al. A genomic sequencing protocol that yields a positive display of 5-methylcytosine residues in individual DNA strands. *Proc Natl Acad Sci U S A* **89**, 1827-1831 (1992).

2. Meissner, A. et al. Genome-scale DNA methylation maps of pluripotent and differentiated cells. *Nature* **454**, 766-770 (2008).

3. Cokus, S.J. et al. Shotgun bisulphite sequencing of the Arabidopsis genome reveals DNA methylation patterning. *Nature* **452**, 215-219 (2008).

4. Lister, R. et al. Human DNA methylomes at base resolution show widespread epigenomic differences. *Nature* **462**, 315-322 (2009).

5. Shapiro, R., Cohen, B.I. & Servis, R.E. Specific deamination of RNA by sodium bisulphite. *Nature* **227**, 1047-1048 (1970).

6. Woese, C.R. et al. Secondary structure model for bacterial 16S ribosomal RNA: phylogenetic, enzymatic and chemical evidence. *Nucleic Acids Res* **8**, 2275-2293 (1980).

7. He, Y., Vogelstein, B., Velculescu, V.E., Papadopoulos, N. & Kinzler, K.W. The antisense transcriptomes of human cells. *Science* **322**, 1855-1857 (2008).

8. Motorin, Y., Lyko, F. & Helm, M. 5-methylcytosine in RNA: detection, enzymatic formation and biological functions. *Nucleic Acids Res* **38**, 1415-1430 (2010).

9. Gu, W., Hurto, R.L., Hopper, A.K., Grayhack, E.J. & Phizicky, E.M. Depletion of Saccharomyces cerevisiae tRNA(His) guanylyltransferase Thg1p leads to uncharged tRNAHis with additional m(5)C. *Mol Cell Biol* **25**, 8191-8201 (2005).

10. Schaefer, M., Pollex, T., Hanna, K. & Lyko, F. RNA cytosine methylation analysis by bisulfite sequencing. *Nucleic Acids Res* **37**, e12 (2009).

11. Schaefer, M. et al. RNA methylation by Dnmt2 protects transfer RNAs against stress-induced cleavage. *Genes Dev* **24**, 1590-1595 (2010).

12. Squires, J.E. et al. Widespread occurrence of 5-methylcytosine in human coding and non-coding RNA. *Nucleic Acids Res* **40**, 5023-5033 (2012).

13. Khoddami, V. & Cairns, B.R. Identification of direct targets and modified bases of RNA cytosine methyltransferases. *Nat Biotechnol* (2013).

14. Goll, M.G. et al. Methylation of tRNAAsp by the DNA methyltransferase homolog Dnmt2. *Science* **311**, 395-398 (2006).

15. Nix, D.A., Courdy, S.J. & Boucher, K.M. Empirical methods for controlling false positives and estimating confidence in ChIP-Seq peaks. *BMC Bioinformatics* **9**, 523 (2008).

16. Thorvaldsdottir, H., Robinson, J.T. & Mesirov, J.P. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Brief Bioinform* **14**, 178-192 (2013).

17. Zeeberg, B.R. et al. GoMiner: a resource for biological interpretation of genomic and proteomic data. *Genome Biol* **4**, R28 (2003).

18. Schaefer, M. & Lyko, F. Solving the Dnmt2 enigma. *Chromosoma* **119**, 35-40 (2010).

19. Rai, K. et al. Dnmt2 functions in the cytoplasm to promote liver, brain, and retina development in zebrafish. *Genes Dev* **21**, 261-266 (2007).

20. Jurkowski, T.P. et al. Human DNMT2 methylates tRNA(Asp) molecules using a DNA methyltransferase-like catalytic mechanism. *RNA* **14**, 1663-1670 (2008).

21. Raddatz, G. et al. Dnmt2-dependent methylomes lack defined DNA methylation patterns. *Proc Natl Acad Sci U S A* (2013).

22. Dong, A. et al. Structure of human DNMT2, an enigmatic DNA methyltransferase homolog that displays denaturant-resistant binding to DNA. *Nucleic Acids Res* **29**, 439-448 (2001).

23. Korlach, J. & Turner, S.W. Going beyond five bases in DNA sequencing. *Curr Opin Struct Biol* **22**, 251-261 (2012).

24. Durdevic, Z. et al. Efficient RNA virus control in Drosophila requires the RNA methyltransferase Dnmt2. *EMBO Rep* **14**, 269-275 (2013).

25. Becker, M. et al. Pmt1, a Dnmt2 homolog in Schizosaccharomyces pombe, mediates tRNA methylation in response to nutrient signaling. *Nucleic Acids Res* **40**, 11648-11658 (2012).

26. Phalke, S. et al. Retrotransposon silencing and telomere integrity in somatic cells of Drosophila depends on the cytosine-5 methyltransferase DNMT2. *Nat Genet* **41**, 696-702 (2009).

27. Yadlapalli, S. & Yamashita, Y.M. Chromosome-specific nonrandom sister chromatid segregation during stem-cell division. *Nature* (2013).

28. Schaefer, M. & Lyko, F. Lack of evidence for DNA methylation of Invader4 retroelements in Drosophila and implications for Dnmt2-mediated epigenetic regulation. *Nat Genet* **42**, 920-921; author reply 921 (2010).

29. Tuorto, F. et al. RNA cytosine methylation by Dnmt2 and NSun2 promotes tRNA stability and protein synthesis. *Nat Struct Mol Biol* **19**, 900-905 (2012).

30. Meyer, K.D. et al. Comprehensive analysis of mRNA methylation reveals enrichment in 3' UTRs and near stop codons. *Cell* **149**, 1635-1646 (2012).

31. Rosenberg, B.R., Hamilton, C.E., Mwangi, M.M., Dewell, S. & Papavasiliou, F.N. Transcriptome-wide sequencing reveals numerous APOBEC1 mRNA-editing targets in transcript 3' UTRs. *Nat Struct Mol Biol* **18**, 230-236 (2011).

32. Dominissini, D. et al. Topology of the human and mouse m6A RNA methylomes revealed by m6A-seq. *Nature* **485**, 201-206 (2012).

CHAPTER 3

IDENTIFICATION OF DIRECT TARGETS AND MODIFIED BASES OF RNA CYTOSINE METHYLTRANSFERASES

LETTERS

nature biotechnology

Identification of direct targets and modified bases of RNA cytosine methyltransferases

Vahid Khoddami & Bradley R Cairns

The extent and biological impact of RNA cytosine methylation are poorly understood, in part owing to limitations of current techniques for determining the targets of RNA methyltransferases. Here we describe 5-azacytidine–mediated RNA immunoprecipitation (Aza-IP), a technique that exploits the covalent bond formed between an RNA methyltransferase and the cytidine analog 5-azacytidine to recover RNA targets by immunoprecipitation. Targets are subsequently identified by high-throughput sequencing. When applied in a human cell line to the RNA methyltransferases DNMT2 and NSUN2, Aza-IP enabled >200-fold enrichment of tRNAs that are known targets of the enzymes. In addition, it revealed many tRNA and noncoding RNA targets not previously associated with NSUN2. Notably, we observed a high frequency of C→G transversions at the cytosine residues targeted by both enzymes, allowing identification of the specific methylated cytosine(s) in target RNAs. Given the mechanistic similarity of RNA cytosine methyltransferases, Aza-IP may be generally applicable for target identification.

Although cytosine methylation is most commonly studied in DNA, it is also found in RNA[1]. As with DNA, RNA cytosine methylation occurs at the C5 position (m^5C). RNA methylation has been detected in both prokaryotic and eukaryotic noncoding RNAs (ncRNAs) such as tRNA and rRNA[1]. Recent high-throughput RNA methylation profiling by bisulfite sequencing in HeLa cells verified and extended the repertoire of m^5C modifications in RNA[2], motivating a more thorough examination of the scope (cell types and developmental contexts) and functions of RNA methylation.

The m^5C-RNA methyltransferases have been subdivided into six families based on structural and functional properties: RsmB/Nol1/NSUN1, RsmF/YebU/NSUN2, RlmI, Ynl022, NSUN6 and DNMT2 (ref. 1). Only DNMT2-family enzymes are of the single-cysteine type; similar to DNA methyltransferases, they utilize a single cysteine in their catalytic pocket[3], whereas the other m^5C-RNA methyltransferase family enzymes utilize two cysteines[4]. Here we focus on DNMT2 and NSUN2, as they represent one member of each family that is either highly studied (DNMT2) or highly relevant to a specific disease (NSUN2).

DNMT2 functions primarily, if not exclusively, as an m^5C-RNA methyltransferase, with three verified tRNA targets: $tRNA^{Asp}$, $tRNA^{Gly}$ and $tRNA^{Val}$ (refs. 3,5–7). In most organisms, lack of DNMT2 is not phenotypically evident[8], although DNMT2-deficient zebrafish display developmental perturbations[6]. Notably, DNMT2 activity attenuates tRNA cleavage during stress conditions, and promotes response to RNA viruses in *Drosophila*[7,9]. NSUN2 also methylates cytosines in tRNAs as well as in the ncRNA subunit of RNase P and possibly mRNA substrates[2,10,11], but the links between particular NSUN2 targets and cellular functions are unknown. NSUN2 has been associated with Myc-induced proliferation of cancer cells[12], mitotic spindle stability[13], infertility in male mice, and the balance of self-renewal and differentiation in skin stem cells[14]. In humans NSUN2 mutations cause an autosomal recessive syndrome characterized by intellectual disability and mental retardation[15–17]. Furthermore, tRNA cytosine methylation by both Dnmt2 and Nsun2 in mice increases tRNA stability and steady-state protein synthesis[10].

In principle, RNA targets of m^5C-RNA methyltransferases could be identified by deep RNA bisulfite sequencing of cell lines or tissues in which a particular m^5C-RNA methyltransferase has been knocked down or knocked out[2]. However, this approach is labor intensive and requires effective enzyme knockout methods. In addition, in cases in which other enzymes are redundant with the m^5C-RNA methyltransferase under study, targets of interest may be missed. Although such an approach could identify candidate targets of the m^5C-RNA methyltransferase, it could not distinguish between direct and indirect targets. Lastly, this approach would require extremely deep sequencing to reveal modifications on RNAs of low abundance or low methylation penetrance. To circumvent these and other issues, we developed Aza-IP, a technique that enriches the direct RNA targets of specific m^5C-RNA methyltransferases and identifies the precise cytosine(s) targeted by the enzyme.

Like m^5C-DNA methyltransferases all m^5C-RNA methyltransferases tested to date form a covalent enzyme-substrate intermediate with their target[1]. Specifically, the sulfur atom of a cysteine residue in the m^5C-RNA methyltransferase catalytic domain covalently bonds to the C6 position of the base in the target RNA. Covalent linkage precedes methylation, which occurs by enamine methylation of the C5 position of the target cytosine using the methyl donor *S*-adenosyl methionine (SAM). Free enzyme is regenerated by subsequent beta elimination[1] (**Fig. 1a**).

This catalytic mechanism is disrupted by the suicide inhibitors 5-azacytidine (5-aza-C) and 5-aza-2′-deoxycytidine (5-aza-dC)[18]. These cytidine analogs are randomly incorporated by RNA and

Howard Hughes Medical Institute, Department of Oncological Sciences, Huntsman Cancer Institute, University of Utah School of Medicine, Salt Lake City, Utah, USA. Correspondence should be addressed to B.R.C. (brad.cairns@hci.utah.edu).

Received 1 January; accepted 2 April; published online 21 April 2013; doi:10.1038/nbt.2566

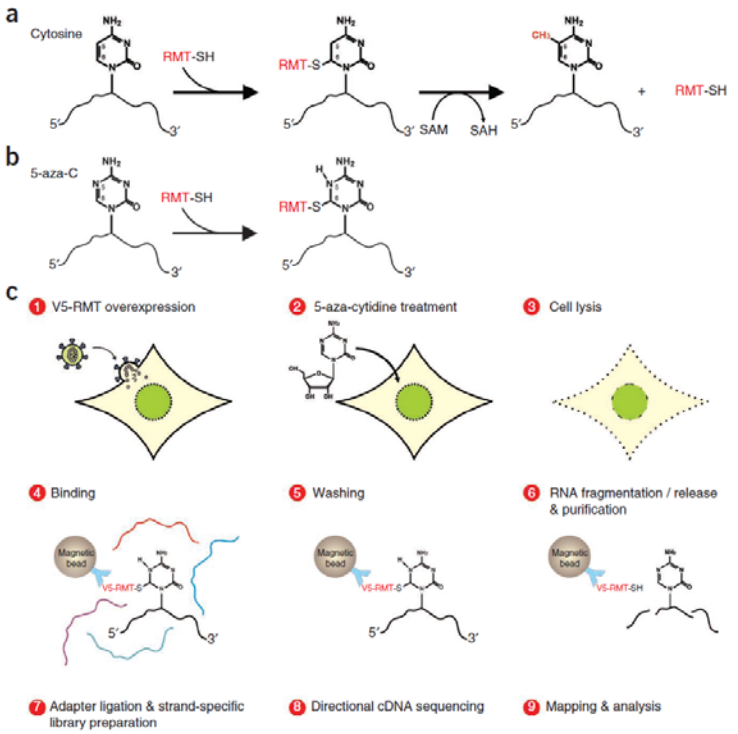


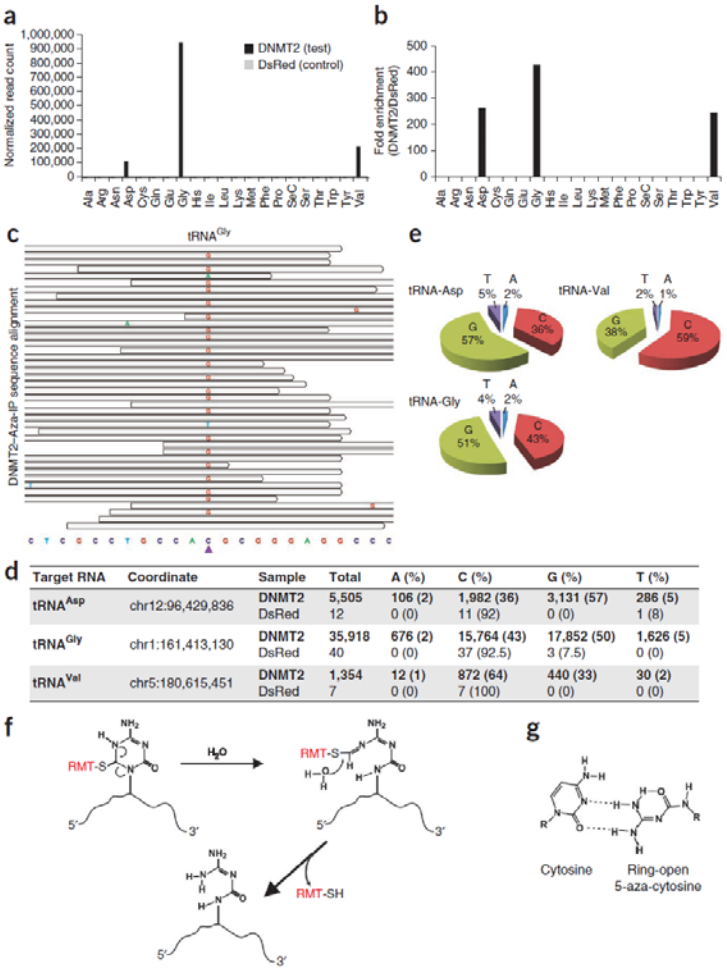
Figure 1 RNA cytosine methylation mechanism and Aza-IP experimental design. (**a**) m^5C-RNA methyltransferases (RMT) catalyzing methylation of C5 of cytosine. First, the enzyme forms a covalent thioester bond, connecting the cysteine residue of its catalytic domain to the C6 position of the target cytosine, forming an RNA methyltransferase–RNA adduct. Next, the RNA methyltransferase transfers a methyl group from cofactor SAM to the C5 of the target cytosine. The enzyme is then released from the adduct by β-elimination. Methylated RNA and *S*-adenosyl-L-homocysteine (SAH) are the product and by-product of this reaction, respectively. (**b**) 5-azacytidine (5-aza-C) is a suicide inhibitor that traps the enzyme by forming a stable RNA methyltransferase-RNA adduct. (**c**) Aza-IP technique (see text).

DNA polymerases into nascent RNA or DNA molecules, respectively. Owing to nitrogen substitution at C5, RNA and DNA methyltransferase enzymes remain covalently bound to the target RNA or DNA molecule (**Fig. 1b**), thereby depleting cells of the endogenous enzymes and resulting in hypomethylation of RNA and DNA[18–21]. Thus, in the presence of 5-aza-C, even overexpression of an m^5C-RNA methyltransferase should result in only a small amount of short-lived free active enzyme, greatly reducing concern that enzyme overexpression will result in the methylation of nonphysiological, irrelevant targets. Given their effectiveness in enzyme depletion, 5-aza-C and 5-aza-dC are currently used for a variety of experimental and clinical applications, including linkage of a DNMT to a known target DNA *in vivo*[22], visualization and monitoring of DNMTs in living cells[23], and inducing DNA hypomethylation of tumor cells in patients[24].

Aza-IP involves nine steps: (i) expression of an epitope-tagged m^5C-RNA methyltransferase derivative in cells (or use of an antibody capable of immunoprecipitating the endogenous RNA-bound enzyme); (ii) cell growth in the presence of 5-aza-C, which is incorporated at low or moderate levels into nascent RNA; (iii) cell lysis; (iv) immunoprecipitation of the m^5C-RNA methyltransferase of interest, a portion of which is covalently attached to target RNAs bearing 5-aza-C; (v) stringent washing to remove RNA contaminants; (vi) RNA fragmentation, release and purification; (vii) ligation of adaptor oligos to the RNA, and creation of a cDNA library in a manner that enables strand-specific assignments; (viii) cDNA sequencing (50-bp single end); (ix) mapping and analysis of sequence reads to define RNA targets and site of cross-linking and/or catalysis (**Fig. 1c** and Online Methods).

We chose HeLa cells because of their favorable growth properties and ease of infection by lentiviruses. We overexpressed V5-tagged (epitope derived from simian virus 5) human DNMT2 (test) or V5-tagged DsRed (control) proteins using a lentiviral expression system and the relatively strong cytomegalovirus (CMV) immediate early promoter. We grew HeLa cells in 3 µM 5-aza-C for a time (12 h) empirically determined as sufficient for incorporation of 5-aza-C into nascent RNA and efficient RNA target detection (3–5 µM 5-aza-C facilitates 5-aza-C incorporation without detectable toxicity, and 6- to 24-h time ranges should be tested). Cells were lysed in a stringent denaturing buffer supplemented with an RNase inhibitor, briefly sonicated and cleared to yield a cell lysate. The lysate was incubated with magnetic beads coated with anti-V5, and the beads were stringently washed, taking advantage of the covalent association between DNMT2 and target RNA. We then fragmented the enriched RNA molecules while they were still bound to the beads, generating RNA fragments of appropriate size (60–200 bp) for the construction of sequencing libraries. These RNA fragments were then isolated, ethanol precipitated and extracted. RNA samples were then used for directional (strand-specific) cDNA library preparation and, after application of quality control procedures, the cDNA library was subjected to high-throughput sequencing using Illumina 50-bp single-end sequencing (Online Methods).

We used the USeq analysis package and Biotoolbox to identify RNAs showing statistically significant enrichment in DNMT2 immunoprecipitates (Online Methods). The known DNMT2 targets ($tRNA^{Asp}$, $tRNA^{Gly}$ and $tRNA^{Val}$)[7] were detected at background levels in the control V5-DsRed immunoprecipitates, but were markedly enriched 271-, 431- and 255-fold, respectively, compared to V5-DsRed control in the V5-DNMT2 immunoprecipitates (**Fig. 2a**,**b** and **Supplementary Data set 1**). It appeared that DNMT2 was indeed released from the target 5-aza-C base during the fragmentation step, as sequence reads covering the known target cytosine (C38) were plentiful and not depleted relative to reads covering the flanking sequences. Beyond these three tRNAs, we found no comparable enrichment of other ncRNAs (rRNAs, small nuclear RNAs, small nucleolar RNAs, small cytoplasmic RNAs, microRNAs) or mRNAs, with two preliminary exceptions; the KRT18 mRNA and the KRT18 pseudogene mRNA displayed moderate enrichment and a C→G transversion (a purine replacement of a pyrimidine, explained below) (**Supplementary Data set 2** and **Supplementary Fig. 1**). Although we did not rigorously



Target RNA	Coordinate	Sample	Total	A (%)	C (%)	G (%)	T (%)
tRNA^{Asp}	chr12:96,429,836	**DNMT2**	**5,505**	**106 (2)**	**1,982 (36)**	**3,131 (57)**	**286 (5)**
		DsRed	12	0 (0)	11 (92)	0 (0)	1 (8)
tRNA^{Gly}	chr1:161,413,130	**DNMT2**	**35,918**	**676 (2)**	**15,764 (43)**	**17,852 (50)**	**1,626 (5)**
		DsRed	40	0 (0)	37 (92.5)	3 (7.5)	0 (0)
tRNA^{Val}	chr5:180,615,451	**DNMT2**	**1,354**	**12 (1)**	**872 (64)**	**440 (33)**	**30 (2)**
		DsRed	7	0 (0)	7 (100)	0 (0)	0 (0)

Figure 2 Aza-IP analysis of DNMT2 RNA targets. (**a**) Normalized reads mapping to each tRNA in the V5-DNMT2 (test) and V5-DsRed (control) data sets (one replicate of each shown). Each tRNA is designated by a three-letter amino acid abbreviation. (**b**) Fold enrichment was calculated from the data shown in **a** by dividing the normalized RPKM values for each tRNA type in the V5-DNMT2 data set by the values in the V5-DsRed data set. (**c**) A representative snapshot from the Integrative Genomics Viewer (IGV, Broad Inst.) browser depicting a subset of the sequencing reads mapped to a tRNA^{Gly} locus (chr1:161,413,119–161,413,141, human genome version 19 (hg19) bottom) at base pair resolution. The gray bars span the start and stop of individual sequencing reads mapped to the locus. The mismatched nucleotides are shown with colored letters and the matched nucleotides are hidden (gray). The purple arrowhead points to the tRNA^{Gly} C38 nucleotide (chr1:161,413,130). (**d**) Summary of the base distribution at the known DNMT2 target sites in tRNA^{Asp}, tRNA^{Gly} and tRNA^{Val}. The coordinate indicates the genomic location of the target cytosine in the human genome and the raw numbers are reported for both the V5-DNMT2 and V5-DsRed Aza-IP data sets. (**e**) Base distributions at the target nucleotide in the mapped reads. The numbers for the tRNAs are averaged over all annotated tRNA loci of the same type in the human genome showing coverage over the target nucleotide (C38). (**f**) RNA methyltransferase–induced ring opening and RNA methyltransferase–RNA dissociation model, as proposed for mammalian DNA cytosine methyltransferases[25] and adapted here for RNA methyltransferases. RNA methyltransferase covalent linkage to the C6 position of 5-aza-C induces the rearrangement and ring opening and results in dissociation of the RNA methyltransferase from the target RNA molecule. (**g**) Base-pairing behavior of ring-open 5-aza-C. The ring-open 5-aza-C prefers to pair with cytosine and is therefore read as guanosine after RT-PCR and sequencing.

validate these mRNA enrichments *in vivo*, we instead used KRT18 RNA fragments and tRNAs to reveal that DNMT2 activity requires a CpG within a tRNA-type stem-loop junction (**Supplementary Result 1** and **Supplementary Figs. 2**–7). Together, these findings indicate that Aza-IP highly enriched only the three known tRNA targets of DNMT2.

Notably, a majority of the reads that mapped to the three known target tRNAs contained a single-nucleotide polymorphism (SNP) present solely at the known DNMT2 target cytosine (C38) (**Fig. 2c–e** and **Supplementary Data set 1**). Typically, the base change involved transversion to guanosine; however, adenosine and thymine were also observed at levels well above the estimated error rate of sequencing (~0.5%). Transversion was observed at the target cytosine but almost never (<1%) at other cytosines in all three DNMT2 target tRNAs. Over 200 m^{5}C sites have been identified in different human tRNAs and their isoacceptors or isodecoders[2] (tRNAs that accept the same amino acid but have different anticodons, or have identical anticodons but different body sequences, respectively). However, only C38 and not the other sites bearing m^{5}C in these three tRNAs exhibited transversion in our data sets. Thus, Aza-IP does not simply cause mutations at cytosines, or at m^{5}C; rather, transversion is observed specifically at DNMT2 target sites in the RNA enriched by our technique. These are true transversions, not SNPs, because standard high-throughput RNA sequencing profiles from HeLa cells revealed cytosine at C38 in >99% of the reads (data not shown). Notably, previous work with DNMTs revealed C→G transversion occurring within DNA following the growth of cells in 5-aza-C; the authors proposed that covalent attachment of the DNMT to the target 5-aza-C enables the opening of the 5-aza-C ring and its subsequent pairing with cytosine at replication[25] (**Fig. 2f,g**).

To validate the DNMT2 targets identified by Aza-IP, we performed transcriptome-wide RNA bisulfite sequencing in wild-type (WT) and Dnmt2-knockout mouse embryonic fibroblasts (MEFs); we used MEFs because knockdown of DNMT2 in HeLa could not be validated with current antibodies. We adapted existing methods and criteria (coverage >10 reads, methylation level >20%)[2], defined the significantly methylated cytosines, and compared the WT and Dnmt2-knockout data sets (**Supplementary Data set 3**). In Dnmt2-knockout but not WT MEFs, we observed a total loss of RNA methylation at the three known DNMT2 tRNA targets identified by Aza-IP;

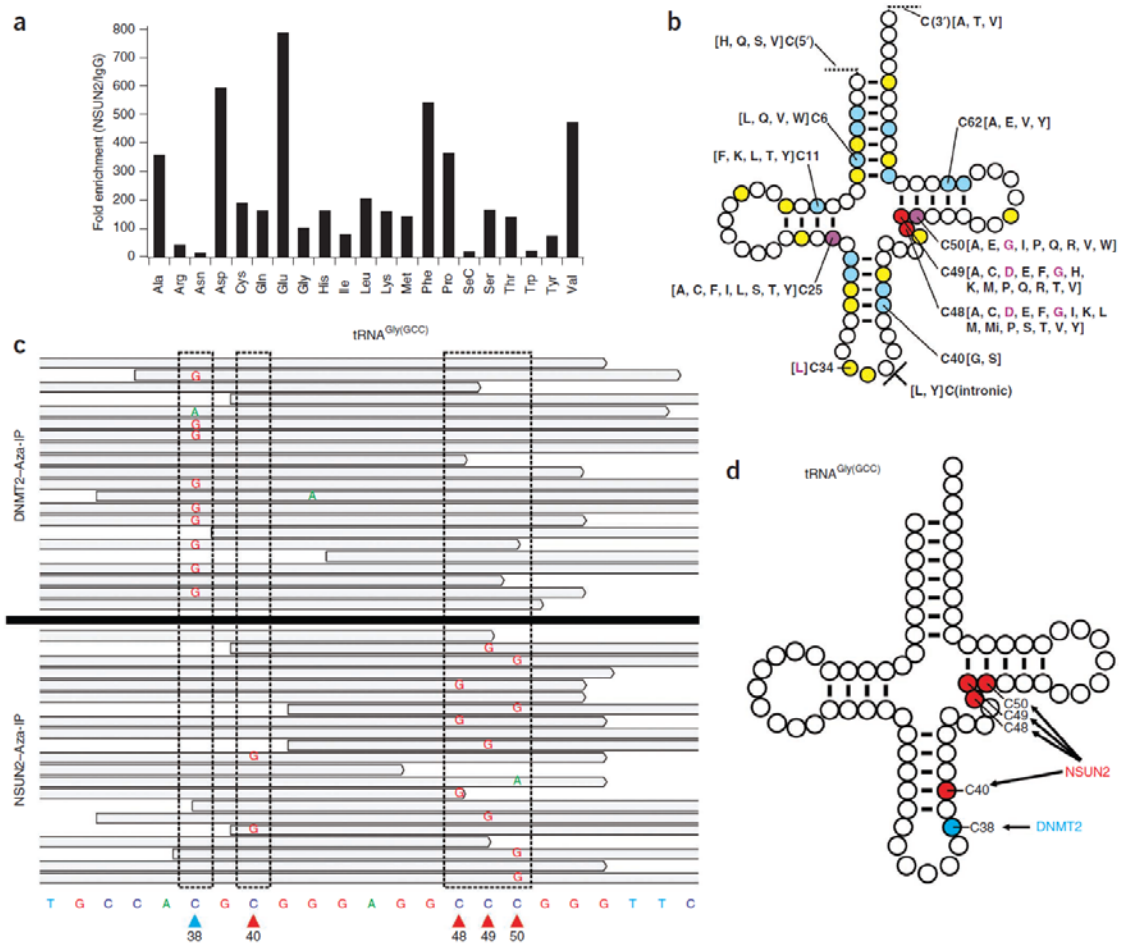


Figure 3 Aza-IP analysis of NSUN2 RNA targets. (**a**) Graph depicts fold enrichment of human tRNAs in V5-NSUN2 immunoprecipitates. Fold enrichment was calculated by dividing the normalized RPKM values for each tRNA type in the V5-NSUN2 replicate data sets (combined) by the values in the IgG control data set. Each tRNA is designated by a three-letter amino acid abbreviation. (**b**) A 'standardized' tRNA summarizing the human NSUN2 target cytosines revealed by Aza-IP in HeLa cells. Cytosines are color-coded based on the number of tRNA types that we found to be NSUN2 target sites: yellow, one tRNA type; blue, 2–5 tRNA types; purple, 6–9 tRNA types; and red >10 tRNA types. For each position, the individual tRNAs are designated by their single letter amino acid abbreviation, grouped in square brackets; purple letters refer to previously known NSUN2 target sites in the designated tRNA types (in human)[2,11,26]. For clarity of presentation, only selected positions are depicted (see **Supplementary Data set 5** for all tRNAs, their isoacceptors/isodecoders, and target positions). (**c**) Integrative Genomics Viewer (IGV) browser snapshots of a random subset of the sequencing reads mapped to a $tRNA^{Gly(GCC)}$ locus (chr17:8,029,095–8,029,117) from the separate DNMT2 (top) or NSUN2 Aza-IP (bottom) data sets. Stippled boxes show locations that meet target criteria with either enzyme. Arrowheads at bottom depict the sole DNMT2 target site (blue) or the four NSUN2 target sites (red). (**d**) A standardized $tRNA^{Gly(GCC)}$ with all five known m^5C bases depicted, all of which (and no other resident cytosine) were specifically and selectively identified by Aza-IP of DNMT2 or NSUN2.

other sites of RNA methylation were not affected. However, the site of methylation observed in human KRT18 mRNA is not conserved in the mouse Krt18 mRNA.

We then applied Aza-IP to identify RNA targets of the 'two cysteine' human enzyme NSUN2. Previous studies identified three direct tRNA targets of human NSUN2 ($tRNA^{Leu}$ (C34), $tRNA^{Asp}$ (C48, 49) and $tRNA^{Gly}$ (C48, 49, 50))[2,11,26], four direct tRNA targets of mouse Nsun2 ($tRNA^{Leu}$ (C34); $tRNA^{Asp}$ (C48, 49), $tRNA^{Gly}$ (C40, 48, 49, 50) and $tRNA^{Val}$ (C48, 49))[10], and seven additional target cytosines in the tRNA targets for yeast Trm4/Nsun2 (refs. 27,28) (**Fig. 3** and **Supplementary Fig. 8**). We expressed V5-NSUN2 (with modifications; see Online Methods) in HeLa cells, and performed two anti-V5 biological replicates (Rep1 and Rep2), and one IgG-only (control) replicate, of Aza-IP. Rep1, Rep2 and IgG-only experiments yielded 55,180,207, 56,051,845 and 57,133,775 mapped filtered reads, respectively.

We applied three criteria for identifying candidate NSUN2 RNA targets: reads per kilobase per million mapped reads (RPKM > 3); enrichment (replicates RPKM/control RPKM over threefold and false-discovery rate (FDR) < 0.01); and transversion frequency (>4%, and $P < 0.01$). Our mapping and analyses enabled attribution of reads to

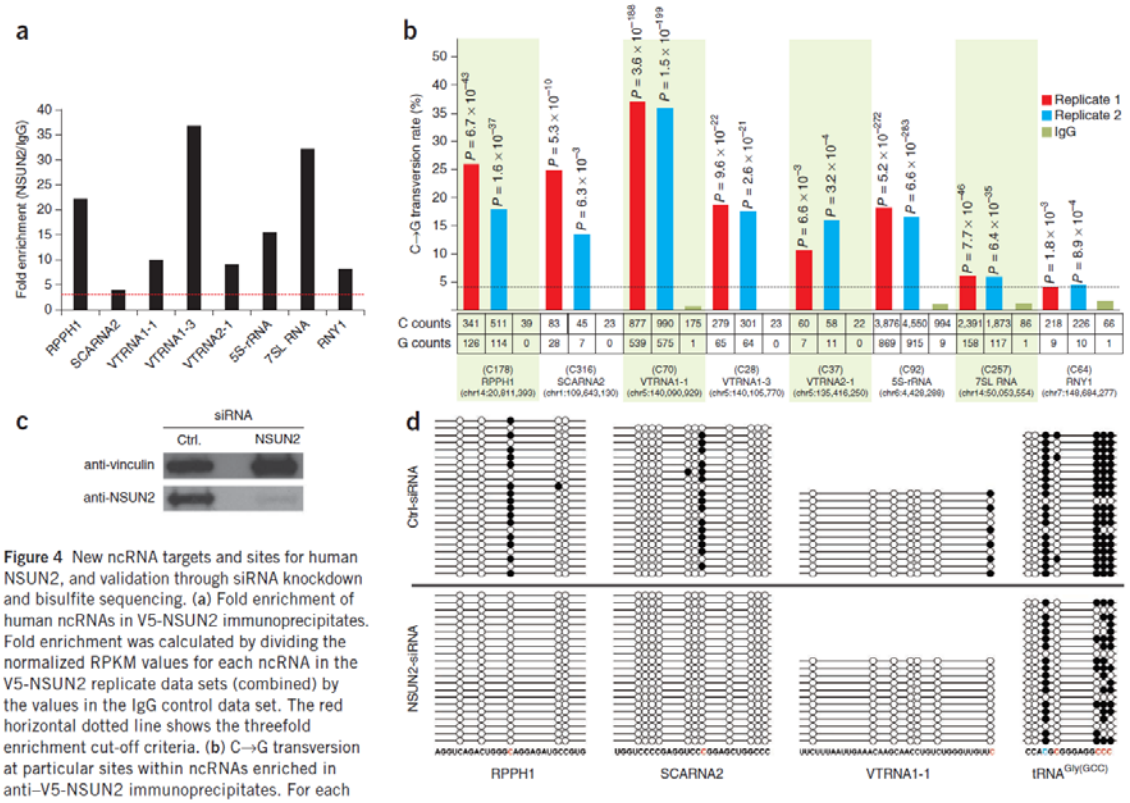


Figure 4 New ncRNA targets and sites for human NSUN2, and validation through siRNA knockdown and bisulfite sequencing. (**a**) Fold enrichment of human ncRNAs in V5-NSUN2 immunoprecipitates. Fold enrichment was calculated by dividing the normalized RPKM values for each ncRNA in the V5-NSUN2 replicate data sets (combined) by the values in the IgG control data set. The red horizontal dotted line shows the threefold enrichment cut-off criteria. (**b**) C→G transversion at particular sites within ncRNAs enriched in anti–V5-NSUN2 immunoprecipitates. For each of the eight enriched ncRNAs, we depict the single target site with the highest C→G transversion (statistics for other candidate sites in **Supplementary Data set 5**). The bar graph shows the C→G transversion occurrence (% of total) in each V5-NSUN2 Aza-IP replicate and IgG-only control; the horizontal dotted line shows the 4% transversion cut-off. *P*-values calculated by VarScan are indicated above columns. Provided underneath are the total number of sequenced C and G nucleotides (counts), the base position of the target cytosine within the ncRNA, and its encoded location in the genome (hg19). (**c**) Extracts (60 µg of protein was loaded on the gel) of HeLa cells expressing NSUN2 or nonspecific (control) siRNA pools were probed with anti-hNSUN2 or anti-vinculin (control) (**Supplementary Fig. 10**). (**d**) Total RNA was extracted from HeLa cells treated with NSUN2 or control siRNAs; RNA was subjected to bisulfite treatment, PCR amplification, cloning and sequencing of several clones per sample (Online Methods). Candidate target sites of NSUN2 (in red; C178 in RPPH1, C316 in SCARNA2, C70 in VTRNA1-1, and C40, C48, C49 and C50 in $tRNA^{Gly(GCC)}$) and DNMT2 (in blue; C38 in $tRNA^{Gly(GCC)}$) were analyzed.

particular isoacceptors and/or isodecoders (**Supplementary Data set 4**), and we provide in **Supplementary Data set 5** the precise human genome coordinate (hg19) of the methylated cytosine and *P* values for transversion. Almost all tRNAs were enriched in both replicates of anti-V5-NSUN2 immunoprecipitates ($tRNA^{Asn}$ and $tRNA^{SeC}$ excepted), with many enriched over 100-fold (**Fig. 3a**). The diversity of candidate NSUN2 RNA targets and candidate target sites required a rigorous statistical approach. Therefore, we used VarScan, a package for analyzing sequence variants in parallel sequencing data, to define locations where C→G transversions were both frequent (≥4%) and highly significant ($P < 0.01$) in both anti-V5-NSUN2 replicates, but not in the IgG-only control. We also applied VarScan to RNA-seq data sets of HeLa cells, to filter out SNPs. Transversion was clear and significant (**Fig. 3b** and **Supplementary Data set 5**) at the known human NSUN2 tRNA targets ($tRNA^{Leu}$ (C34), $tRNA^{Asp}$ (C48, 49) and $tRNA^{Gly}$ (C48, 49, 50))[2,11,26], as well as at C48, C49 and C50 within most tRNAs ($tRNA^{Asn}$ and $tRNA^{SeC}$ excepted), greatly expanding the known repertoire. Importantly, although the target tRNAs of DNMT2 were robustly enriched within anti-V5-NSUN2 immunoprecipitates, we did not observe C→G transversion at the DNMT2 target site (C38) within the anti-V5-NSUN2 immunoprecipitates, highlighting the specificity of both the enzyme and the Aza-IP technique (**Fig. 3c,d**). Furthermore, we identified a large number of additional candidate NSUN2 target sites within particular tRNAs; these sites were not previously described in any organism ($P < 0.01$) (**Fig. 3b** and **Supplementary Data set 5**). Moreover, we observed significant transversion sites within introns, and also upstream and downstream, of particular preprocessed tRNAs ($P < 0.01$) (**Fig. 3b** and **Supplementary Data set 5**).

Notably, within a particular RNA, the extent of C→G transversion was proportionally lower when there were multiple target sites. For example, $tRNA^{Gly(GCC)}$ bears one DNMT2 target site (C38) and four NSUN2 target sites (C40, 48, 49, 50) (**Fig. 3d**); here, the transversion frequency at C38 with DNMT2 was comparable to the sum at the four NSUN2 target sites (**Fig. 3c** and **Supplementary Fig. 9**), consistent with covalent linkage of the m^5C-RNA methyltransferase to only one target site in any individual isolated tRNA.

Human NSUN2 has been reported to methylate cytosines in four other RNAs: ribosomal RNA[12], the RNA subunit of RNaseP (RPPH1)

and two mRNAs (CINP and NAPRT1)[2]. Our Aza-IP analyses yielded eight candidate NSUN2 target ncRNAs: 5S rRNA, RPPH1, vault RNAs (VTRNA1-1, VTRNA1-3 and VTRNA2-1), the small cajal body–specific RNA 2 (SCARNA2, a C/D box snoRNA), a Y RNA (RNY1) and the signal recognition particle RNA (7SL RNA) (**Fig. 4a**). All of these RNAs contain one very clear and significant target site (**Fig. 4b**), and most contain one or more additional sites that also pass all three thresholds in both anti-V5-NSUN2 replicates but not the IgG-only control (**Supplementary Data set 5**); these findings suggest that, as with tRNAs, most ncRNA targets have more than one NSUN2 target cytosine. Notably, Aza-IP analysis did not detect enrichment of CINP or NAPRT1 mRNAs in anti-V5 NSUN2 immunoprecipitates.

To validate the NSUN2 target RNAs identified by Aza-IP, we used short interfering RNA (siRNA) to knock down NSUN2 expression (**Fig. 4c**) and then tested the location and extent of methylation on selected candidate target RNAs by conventional RNA bisulfite sequencing (**Fig. 4d**). Here, we tested RNAs at the top (tRNA), middle (RPPH1) and bottom (SCARNA2 and VTRNA1-1) of our enrichment results. In control siRNA-treated HeLa cells, all tested RNA candidates displayed a methylation (cytosine retention) at the precise site(s) predicted by the Aza-IP transversion, but not at other flanking cytosines. We noted a marked diminishment or elimination of methylation at these sites in NSUN2 siRNA-expressing HeLa cells, validating the involvement of NSUN2 in target methylation (**Fig. 4d**). For several reasons, we suggest that diminishment rather than elimination in $tRNA^{Gly}$ methylation is predicted. Our protocol analyzes methylation at day 6, whereas NSUN2 protein knockdown via siRNA requires several days to reach >90% reduction; also tRNAs are long lived (~30–60 h), so those tRNAs made 2 days prior will still be largely present and methylated. Also, tRNAs, given their exceptional enrichment in Aza-IP, may be preferred by NSUN2 over the ncRNAs and may compete more effectively for the remaining NSUN2 pool. Regardless, these siRNA experiments validate the conclusions of the Aza-IP analysis.

The NSUN2 candidate ncRNA targets identified here include RNAs with central functions in the processing, folding and modification of other ncRNAs (RPPH1, Y RNA, SCARNA2), RNAs important for protein synthesis and trafficking (5S rRNA and 7SL RNA) and RNAs involved in multidrug resistance and other processes (Vault RNAs)[29,30]. Notably, all of the NSUN2 targets revealed by Aza-IP are either transcribed by RNA Pol III[29] in the nucleolus (SCARNA2 excepted[30]), or function in the nucleolus (SCARNA2)[30], where NSUN2 is known to reside[12,31]. The biological functions of RNA methylation at these sites remain to be explored, but they could affect RNA structure, association with ribonucleoprotein complexes, or complex activity. Taken together, our work greatly increases the set of target site candidates for NSUN2 that should be considered as possible contributors to enzyme function, including in pathologies related to cancer, stem cells and intellectual disability[12,14–17].

Comparisons of our NSUN2 data sets to the recent RNA bisulfite sequencing (RBS-seq, with ABI sequencing platform) data set from HeLa cells[2] reveals high overlap, with the minor deviations likely resulting from the high filtering thresholds used in the RBS-seq method. Going forward, we envision Aza-IP and RBS-seq as complementary approaches. However, as Aza-IP enriches target RNAs (revealing low-copy RNAs), identifies only direct targets, and reveals the precise methylation sites in these targets (even in situations of m^5C-RNA methyltransferase redundancy and/or low methylation penetrance), we intend to apply Aza-IP to discover new m^5C-RNA methyltransferase targets and to validate these using focused RBS-seq or other approaches. Finally, additional mechanism-based 'adduct-IP' trapping techniques (using other nucleotide analogs[32]) may help identify the targets of other RNA-modifying enzymes.

METHODS

Methods and any associated references are available in the online version of the paper.

Accession codes. GEO: GSE38957 and GSE44359.

Note: Supplementary information is available in the online version of the paper.

ACKNOWLEDGMENTS

We thank C. Clapier (hDNMT2 protein expression and purification), K. Rai (MTase assay set-up), V. Planelles, University of Utah (gift of the lentiviral expression construct), C. Maximiliano Rêgo Monteiro Filho and S. Dehghanizadeh (lentiviral protein expression, assays and IP experiments), J. Xu and A. Yerra (help on data not shown), and X. Cheng, Emory University (gift of pQE9 plasmid). We thank B. Dalley and N. Moss (library preparation and sequencing), T. Parnell, D. Nix, B. Milash, Y. Sun and K. Boucher (help and advice on analysis) and the Center for High Performance Computing, especially W.R. Cardoen. We thank C.J. Burrows for advice on reaction mechanisms, and D.A. Jones, C.J. Burrows and D.R. Davis for many helpful comments. This work was supported by the Howard Hughes Medical Institute, the Samuel Waxman Foundation and the National Cancer Institute CA24014 (for core facilities).

AUTHOR CONTRIBUTIONS

V.K. contributed to experimental design and approaches, performed all experiments and analyses, and helped write the paper; B.R.C. contributed to experimental design and approaches, data interpretation and wrote (with V.K.) the manuscript.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

1. Motorin, Y., Lyko, F. & Helm, M. 5-methylcytosine in RNA: detection, enzymatic formation and biological functions. *Nucleic Acids Res.* **38**, 1415–1430 (2010).
2. Squires, J.E. *et al.* Widespread occurrence of 5-methylcytosine in human coding and non-coding RNA. *Nucleic Acids Res.* **40**, 5023–5033 (2012).
3. Jurkowski, T.P. *et al.* Human DNMT2 methylates tRNA(Asp) molecules using a DNA methyltransferase-like catalytic mechanism. *RNA* **14**, 1663–1670 (2008).
4. King, M.Y. & Redman, K.L. RNA methyltransferases utilize two cysteine residues in the formation of 5-methylcytosine. *Biochemistry* **41**, 11218–11225 (2002).
5. Goll, M.G. *et al.* Methylation of tRNAAsp by the DNA methyltransferase homolog Dnmt2. *Science* **311**, 395–398 (2006).
6. Rai, K. *et al.* Dnmt2 functions in the cytoplasm to promote liver, brain, and retina development in zebrafish. *Genes Dev.* **21**, 261–266 (2007).
7. Schaefer, M. *et al.* RNA methylation by Dnmt2 protects transfer RNAs against stress-induced cleavage. *Genes Dev.* **24**, 1590–1595 (2010).
8. Schaefer, M. & Lyko, F. Solving the Dnmt2 enigma. *Chromosoma* **119**, 35–40 (2010).
9. Durdevic, Z. *et al.* Efficient RNA virus control in *Drosophila* requires the RNA methyltransferase Dnmt2. *EMBO Rep.* **14**, 269–275 (2013).
10. Tuorto, F. *et al.* RNA cytosine methylation by Dnmt2 and NSun2 promotes tRNA stability and protein synthesis. *Nat. Struct. Mol. Biol.* **19**, 900–905 (2012).
11. Brzezicha, B. *et al.* Identification of human tRNA:m5C methyltransferase catalysing intron-dependent m5C formation in the first position of the anticodon of the pre-tRNA Leu (CAA). *Nucleic Acids Res.* **34**, 6034–6043 (2006).
12. Frye, M. & Watt, F.M. The RNA methyltransferase Misu (NSun2) mediates Myc-induced proliferation and is upregulated in tumors. *Curr. Biol.* **16**, 971–981 (2006).
13. Hussain, S. *et al.* The nucleolar RNA methyltransferase Misu (NSun2) is required for mitotic spindle stability. *J. Cell Biol.* **186**, 27–40 (2009).
14. Blanco, S. *et al.* The RNA-methyltransferase Misu (NSun2) poises epidermal stem cells to differentiate. *PLoS Genet.* **7**, e1002403 (2011).
15. Abbasi-Moheb, L. *et al.* Mutations in NSUN2 cause autosomal- recessive intellectual disability. *Am. J. Hum. Genet.* **90**, 847–855 (2012).
16. Khan, M.A. *et al.* Mutation in NSUN2, which encodes an RNA methyltransferase, causes autosomal-recessive intellectual disability. *Am. J. Hum. Genet.* **90**, 856–863 (2012).
17. Martinez, F.J. *et al.* Whole exome sequencing identifies a splicing mutation in NSUN2 as a cause of a Dubowitz-like syndrome. *J. Med. Genet.* **49**, 380–385 (2012).

18. Santi, D.V., Garrett, C.E. & Barr, P.J. On the mechanism of inhibition of DNA-cytosine methyltransferases by cytosine analogs. *Cell* **33**, 9–10 (1983).
19. Lu, L.W., Chiang, G.H., Medina, D. & Randerath, K. Drug effects on nucleic acid modification. I. A specific effect of 5-azacytidine on mammalian transfer RNA methylation *in vivo*. *Biochem. Biophys. Res. Commun.* **68**, 1094–1101 (1976).
20. Lu, L.J. & Randerath, K. Effects of 5-azacytidine on transfer RNA methyltransferases. *Cancer Res.* **39**, 940–949 (1979).
21. Schaefer, M., Hagemann, S., Hanna, K. & Lyko, F. Azacytidine inhibits RNA methylation at DNMT2 target sites in human cancer cell lines. *Cancer Res.* **69**, 8127–8132 (2009).
22. Liu, K., Wang, Y.F., Cantemir, C. & Muller, M.T. Endogenous assays of DNA methyltransferases: evidence for differential activities of DNMT1, DNMT2, and DNMT3 in mammalian cells *in vivo*. *Mol. Cell Biol.* **23**, 2709–2719 (2003).
23. Schermelleh, L. *et al.* Trapped in action: direct visualization of DNA methyltransferase activity in living cells. *Nat. Methods* **2**, 751–756 (2005).
24. Yang, X., Lay, F., Han, H. & Jones, P.A. Targeting DNA methylation for epigenetic therapy. *Trends Pharmacol. Sci.* **31**, 536–546 (2010).
25. Jackson-Grusby, L., Laird, P.W., Magge, S.N., Moeller, B.J. & Jaenisch, R. Mutagenicity of 5-aza-2′-deoxycytidine is mediated by the mammalian DNA methyltransferase. *Proc. Natl. Acad. Sci. USA* **94**, 4681–4685 (1997).
26. Auxilien, S., Guerineau, V., Szweykowska-Kulinska, Z. & Golinelli-Pimpaneau, B. The human tRNA m^5C methyltransferase Misu is multisite-specific. *RNA Biol.* **9**, 1331–1338 (2012).
27. Becker, M. *et al.* Pmt1, a Dnmt2 homolog in *Schizosaccharomyces pombe*, mediates tRNA methylation in response to nutrient signaling. *Nucleic Acids Res.* **40**, 11648–11658 (2012).
28. Motorin, Y. & Grosjean, H. Multisite-specific tRNA:m^5C-methyltransferase (Trm4) in yeast *Saccharomyces cerevisiae*: identification of the gene and substrate specificity of the enzyme. *RNA* **5**, 1105–1118 (1999).
29. Hu, S., Wu, J., Chen, L. & Shan, G. Signals from noncoding RNAs: unconventional roles for conventional pol III transcripts. *Int. J. Biochem. Cell Biol.* **44**, 1847–1851 (2012).
30. Gerard, M.A. *et al.* The scaRNA2 is produced by an independent transcription unit and its processing is directed by the encoding region. *Nucleic Acids Res.* **38**, 370–381 (2010).
31. Sakita-Suto, S. *et al.* Aurora-B regulates RNA methyltransferase NSUN2. *Mol. Biol. Cell* **18**, 1107–1117 (2007).
32. Guelorget, A. & Golinelli-Pimpaneau, B. Mechanism-based strategies for trapping and crystallizing complexes of RNA-modifying enzymes. *Structure* **19**, 282–291 (2011).

ONLINE METHODS

DNMT2/DsRed Aza-IP experimental design and data analysis. *Expression vector construction.* Total RNA was extracted from HeLa cells using Trizol reagent (Invitrogen) and first-strand cDNA synthesis was performed with SuperScript III First-Strand Synthesis System (Invitrogen), using the Oligo(dT) (Invitrogen), according to the manufacturer's protocol. A DNMT2 clone (or DsRed) bearing a V5 tag was obtained by PCR from the HeLa cDNA (for hDNMT2), or from a DsRed plasmid (Clontech) as templates, using specific primer sets in a two-step PCR format. The primer sets (**Supplementary Table 2**) replaced the first (ATG) codon with an XhoI restriction site, and inserted the Kozak consensus sequence containing the start codon (CACCATGG), and the sequence corresponding to the V5 tag (GGTAAG CCTATCCCTAACCCTCTCCTCGGTCTCGATTCTACG) at the 5′ end of the amplicons. The primers also placed a BamHI restriction site at the 3′ end of these amplicons, right after the stop codon. Validated PCR products were double digested with XhoI and BamHI enzymes and were cloned into the multiple cloning sites of the pPR-lentiviral plasmid (a gift from V. Planelles, Professor of Pathology at University of Utah).

Virus production and titration. The expression vectors (hDNMT2 and DsRed) were used to produce viral particles via transfection. The expression (transfer) plasmids were mixed with packaging and VSVG envelope plasmids (gifts from V. Planelles) and transfected into the HEK-293-FT cells (Invitrogen) using polyethyleneimine (Polysciences). The viral particles were then harvested, concentrated by ultracentrifugation and titrated on HeLa cells using the EGFP marker (which exists on pPR-lentiviral plasmid backbone) by flow-cytometry. Expression of the V5-tagged transgenes was checked by western blotting of the protein extracted from the infected HeLa cells using anti-V5 antibody (Invitrogen). The concentrated viral particles were stored at −80 °C until used.

Lentiviral infection. A day before infection, 20 100-mm plates per experiment (e.g., hDNMT2 and DsRed) were each seeded with 2 million HeLa cells. 24 h later the proper amount of the titered virus, calculated to give ~100% infection rate, were mixed with 6 ml of DMEM media (Invitrogen) supplemented with 10% FBS and 4 μg/ml (final concentration) Polybrene (Millipore) and added to each plate. 12 h post infection, the media was removed and cells from all 20 plates were washed with 1× PBS and trypsinized using TrypLE Express (Invitrogen), and then were pooled and dispensed into fifteen 150 mm plates.

5-Azacytidine treatment. After 6 h the media of each plate was replaced with freshly prepared 5-Azacytidine containing DMEM media (final concentration of 3 μM). To prepare, 5-Azacytidine powder (Sigma) was dissolved in DMEM media to the concentration of 300 μM (100×), filtered and used to make the 1× (final concentration of 3 μM 5-Azacytidine) DMEM media supplemented with 10% FBS.

Preparing the pre-clearing beads. For each experiment, 750 μl of Dynabeads Pan Mouse IgG (Invitrogen) were washed in 1 ml of RIPA buffer (50 mM Tris PH 7.5, 1% Nonidet P-40 (NP-40), 0.5% sodium deoxycholate, 0.1% SDS, 1mM EDTA, 150 mM NaCl + Protease Inhibitor cocktail) supplemented with 5 mg/ml protease free bovine serum albumin (BSA) (Sigma) three times, 2 min each, and re-suspended in 1.5 ml of RIPA buffer + BSA + 30 μl of RNasIN (Promega).

Preparing the antibody coated beads. For each experiment, 1.5 ml of the Dynabeads Pan Mouse IgG were split into two 1.5 ml tubes (750 μl each) and washed with 1 ml of RIPA buffer + BSA three times, 2 min each. Next the beads of each tube were re-suspended in 1.5 ml of RIPA buffer + BSA + 45 μg of the mouse anti-V5 tag antibody (Invitrogen) and incubated at room temperature, rotating for 2 h. The beads of each tube were then washed with 1 ml of RIPA buffer + BSA three times, 2 min each, and were re-suspended in 1.2 ml of RIPA buffer + BSA + 25 μl of RNasIN.

Cell lysis and solubilization. 12 h after 5-Azacytidine treatment, cells were washed with 1× PBS and trypsinized and quenched with complete DMEM media. Then the contents of each group of three 150 mm plates were pooled in a single 15 ml conical tube, spun at 2,000 r.p.m. (~650*g*) at 4 °C for 5 min. Next, 2.5 ml of 1× PBS were added to the cells of each of the 15 ml conical tubes (5 conical tubes per experiment). Then the cells were re-suspended and all of them were pooled in a single 15 ml conical tube, spun at 2,000 r.p.m. at 4 °C for 5 min and the supernatant was discarded. Cells were then lysed in RIPA buffer (total volume of 6 ml) supplemented with 30 μl of RNaseIN, by pipetting up and down several times to get a homogenous lysate. The 6 ml cell lysate was then subjected to sonication, using Misonix Sonicator XL2020, 6 times for 30 s (at setting #4 (0.9 ON time/0.1 OFF time)) followed by 30 s on ice at each round. After sonication another 6 ml of the RIPA buffer was added to the lysate (total of 12 ml) and the diluted lysate was dispensed into twelve 1.5 ml Eppendorf tubes, 1ml each and spun at 14,000 r.p.m. (~21,000*g*) at 4 °C for 10 min. The cleared lysates were then transferred to clean 1.5 ml tubes and kept on ice.

Immunoprecipitation. For each 1 ml of the lysate (in one 1.5 ml tube), 125 μl of the pre-clearing beads were added and incubated at room temperature, rotating. After 2 h, using a magnetic stand, the supernatant of each tube was removed and transferred to a new 1.5 ml tube. Next, 200 μl of the re-suspended antibody coated beads were added to each of the tubes and the mixtures were incubated at room temperature, rotating. After 4 h the supernatant of each tube was removed and discarded and the beads were washed with 850 μl of RIPA + BSA + RNaseIN, 3 times, 5 min each at room temperature. Then the beads were re-suspended in 500 μl of RIPA + BSA + RNaseIN and were transferred to a 0.5 ml clean Eppendorf tube. The supernatant of each tube was then removed and discarded.

RNA fragmentation. For each of the 12 small (0.5 ml) Eppendorf tubes, 7.5 μl of the RNA fragmentation reagent (Ambion) was mixed with 67.5 μl of RNase free ddH_2O (Ambion) in one tube (total of 75 μl per tube) and in another tube 7.5 μl of the fragmentation stop solution (Ambion) was mixed with 7.5 μl of RNase free ddH_2O (total of 15 μl per tube). For the fragmentation, 75 μl of the diluted RNA fragmentation reagent was added to beads of each tube, followed by incubating the tubes at 94 °C for exactly 5 min in a thermocycler, chilled on ice for 1 min, and then the fragmentation was stopped by adding 15 μl of the diluted stop solution.

Ethanol precipitation and RNA extraction. After fragmentation, the supernatant of all small tubes was removed and pooled in a clean 1.5 ml tube (total of 1080 μl). Next 40 μl of the 15 mg/ml GlycoBlue (Ambion) and 104 μl of 3 M Sodium Acetate (Ambion) were mixed with the fragmented RNA and the mixture was split into three 1.5 ml tubes (346 μl each). Then 865 μl absolute ethanol was added to each tube and the tubes were incubated at −80 °C overnight, and spun at maximum speed at 4 °C for 20 min. The pellets were then washed once with 1 ml of 70% ice cold ethanol, air dried and dissolved in total of 40 μl RNase free ddH_2O (for all three of the tubes per experiment) and pooled in a single tube. Next 1 ml of Trizol reagent was added and the RNA was extracted according to the manufacturer's protocol.

Library preparation and high-throughput sequencing. Illumina's directional mRNA-Seq sample preparation protocol was used to prepare the libraries, by ligating the adapters to the fragmented RNAs followed by reverse transcription, PCR and sample clean-up using AMPure beads (Beckman Coulter Genomics). The libraries were then subjected to 50-cycle single-end high-throughput sequencing using Illumina's HiSeq 2000 sequencing system.

Computational analytical methods. Sequenced reads were aligned to the *H. sapiens* Feb 2009 genome build plus all known and theoretical splice junctions derived from Ensembl transcripts (see "MakeTranscriptome" application from open source USeq package[33]). Sequence alignments was performed using the commercial Novoalign package (http://www.novocraft.com) with options to allow gaps and mismatches, reporting 18 bp or larger inserts and reporting all of the reads mapped to the repeats and generating SAM-formatted alignment files. To assess the enriched regions and also making the BAM/BAI files for visualization the "RNASeq" application of the USeq package was used. The Integrative Genomics Viewer (IGV)[34] was used to visualize the alignment files, inspect the mapped reads at base pair resolution and obtain the base distribution reports at desired locations (e.g., C38 position of DNMT2 target tRNAs). The data sets from mRNA sequencing of the input samples were processed via the same pipeline. For calculating the total number of reads mapped to tRNA loci in Aza-IP data sets, applications from the open source Biotoolbox (http://code.google.com/p/biotoolbox) were used. First, the total number of reads mapped to individual annotated tRNAs was calculated and then the numbers for each tRNA types were summed manually to calculate the total number of reads belonging to a particular tRNA type. The total numbers were then normalized to the total number of mapped reads for each dataset. The fold enrichment values for all tRNAs were calculated by dividing the RPKM

values obtained from DNMT2 data set by the RPKM values obtained from DsRed Dataset. Data sets are publicly available through GEO with the following accession number: GSE38957.

MEF preparation, high-throughput RNA bisulfite sequencing and analysis. *MEF isolation and culture.* Wild-type (B6129PF2/J) and Dnmt2$^{-/-}$ (B6;129-Trdmt1^{tm1Bes}/J)[5] mice were obtained from Jackson Laboratory. Mouse Embryonic Fibroblasts (MEFs) were harvested from 13.5 isogenic wt or Dnmt2$^{-/-}$ embryos. Single embryos were dispersed and trypsinized in 10 cm dishes, and were maintained in DMEM (Invitrogen) supplemented with 10% FBS (GIBCO) and passaged several times as reported previously[35].

RNA isolation and fractionation. Harvested MEFs were subjected to total RNA isolation using Trizol reagent (Invitrogen) followed by DNase treatment using TURBO DNA-free Kit (Ambion). The small RNA fraction was separated using mirVana kit (Ambion). The large RNA fraction was prepared by ribosomal RNA depletion using RiboMinus Transcriptome Isolation Kit (Invitrogen) followed by RNA fragmentation using RNA Fragmentation Reagent (Ambion). Small and ribominus fragmented large RNA fractions were separately concentrated via ethanol precipitation. Each one of the fractions (small or large RNA) from each one of the samples (wt or Dnmt2$^{-/-}$) were split into two equal portions: one for direct RNA sequencing and one for RNA bisulfite sequencing.

Bisulfite treatment. Small and large RNA fractions of each sample (wt or Dnmt2$^{-/-}$) were separately subjected to bisulfite treatment. First 5 µg of RNA was dissolved in 45 µl of RNase free ddH_2O and added to 240 µl of de-ionized formamide, mixed well and incubated at 95 °C for 5 min then placed on ice for 2 min. Next 3 µl of the 100 mM hydroquinone and 312 µl of 5 M sodium bisulfite (pH 5) were added and the mixture was incubated at 50 °C, rotating. After 16 h, the mixture was cleaned up using Illustra NAP-10 Columns (GE Healthcare Life Sciences) and de-sulfonation was performed for 2 h at 37 °C in 1 M Tris buffer pH 9.0, followed by ethanol precipitation.

Library preparation and high-throughput sequencing. The bisulfite treated and untreated small and large RNA fractions of each sample (wt or Dnmt2$^{-/-}$) (total of 8 samples) were subjected to high-throughput sequencing. Illumina's directional mRNA-Seq sample preparation protocol was used to prepare the libraries, by ligating the adapters to the RNAs followed by reverse transcription, PCR and sample clean-up using AMPure beads (Beckman Coulter Genomics). The libraries were then subjected to 101-cycle single-end high-throughput sequencing using Illumina's HiSeq 2000 sequencing system.

RNA methylome analysis. Sequenced reads were aligned to the *M. musculus* July 2007 (NCBI37/mm9) genome build plus all known and theoretical splice junctions derived from Ensembl transcripts (see "MakeTranscriptome" application from open source USeq package[33]). Sequence alignments were done using the commercial Novoalign package (http://www.novocraft.com) with options to allow gaps and mismatches, reporting 18 bp or larger inserts and reporting all of the reads mapped to the repeats and generating SAM-formatted alignment files. For alignment of the bisulfite-treated data sets a separate transcriptome index file was generated in bisulfite mode using Novoalign package (Novoindex function) and the alignment output files were generated in both of the SAM and Pairwise formats. SAM-formatted alignment files of the bisulfite treated or untreated data sets were processed by the "RNASeq" application of the USeq package to make the BAM/BAI files, for visualization of the mapped reads using the Integrative Genomics Viewer (IGV)[34]. Pairwise-formatted alignment files were processed by custom python scripts (available on request) to call, annotate and define the candidate differentially-methylated cytosine residues in wt vs Dnmt2$^{-/-}$ data sets. The candidate sites were then verified by manual inspection of the individual mapped reads, by filtering out the sites showing coverage only in the bisulfite-treated (but not the untreated) data sets due to mapping errors caused by lowered base composition complexity of the reads in the bisulfite treated samples. Data sets are publicly available through GEO with the following accession number: GSE44359.

hDNMT2 methyltransferase assay (MTase assay). *hDNMT2 protein expression and purification.* Cloned His-tagged-hDNMT2 in pQE9 plasmid was obtained as a gift from X. Cheng (Emory University) and the protein was expressed in *E. coli* and purified as described previously[36].

Substrate RNA preparation for in-vitro MTase assay. tRNA^{Asp} was made synthetically, or *in vitro* transcribed using T7-RNA polymerase. The synthetically-made tRNA^{Asp} has the endogenous sequence while the tRNA obtained from *in vitro* transcription has the first base changed (U1G); a requirement to be transcribed by T7-RNA polymerase, and the compensatory A71C mutation to maintain the base-pairing, as described previously[5]. The sequences of the wild type and mutant tRNAs and the tRNA-like structure of the KRT18 mRNA have been presented in the (**Supplementary Table 3**). The T7-promoter was attached to the 5′ of each of these sequences by PCR using specific primer sets on ssDNA templates (**Supplementary Table 4**). The PCR products were then subjected to *in vitro* transcription using T7-RNA polymerase (Ambion), followed by DNase treatment (Ambion) and ethanol precipitation of the RNA.

MTase assay. 5 µg of the RNA substrate were mixed with 400 ng of the purified hDNMT2 in the 1× DNMT2-MTase buffer (100 mM Tris-HCl pH 7.5, 5% glycerol, 5 mM $MgCl_2$, 100 mM NaCl, 1 mM DTT and 5 U RNaseIN) supplemented with 1 µCi of *S*-[Methyl-3H]-Adenosyl-L-methionine (PerkinElmer) and incubated at 37 °C for 5 h. After removing the unincorporated SAM using Micro Bio-Spin Columns (BIO-RAD), the RNA was ethanol precipitated and the tritium-labeled methyl group incorporation was evaluated by scintillation counting, and reported as counts per minute (CPM).

Bisulfite sequencing of the in-vitro methylated tRNAs. 5 µg of the purified *in vitro* transcribed and methylated tRNA^{Asp} or tRNA^{Gly} were dissolved in 45 µl of RNase free ddH_2O and added to 240 µl of de-ionized formamide, mixed well and incubated at 95 °C for 5 min then placed on ice for 2 min. Next 3 µl of the 100 mM hydroquinone and 312 µl of 5 M sodium bisulfite (pH 5) were added and the mixture was incubated at 50 °C, rotating. After 16 h, the mixture was cleaned up using Illustra NAP-10 Columns (GE Healthcare Life Sciences) and de-sulfonation was performed for 2 h at 37 °C in 1 M Tris buffer pH 9.0, followed by ethanol precipitation. The first-strand DNA synthesis and PCR amplification was performed using specific primer sets (**Supplementary Table 5**) designed for the bisulfite-converted tRNAs. The PCR product was cloned using TOPO TA Cloning Kit (Invitrogen) and the resulting plasmids purified from several clones were sequenced. The sequences were aligned with the corresponding tRNA sequences and the maintained cytosines (representing methylated cytosines) are depicted in the figures as closed circles.

***In silico* RNA folding analysis of the KRT18 tRNA-like structure.** Sequences representing RNA 'fragments' (55 or 75 bases) from KRT18 were centered on the candidate target cytosine, which were flanked on either side by the endogenous 27 or 37 bases, respectively. They were then analyzed by the Mfold web server[37] using the RNA Folding Form (version 2.3 energies). The default parameters were used except for the percent suboptimality number which was set to 10 to allow reporting of suboptimal structures. The structure of the 75 bp fragment was then manually adjusted (outside the critical anticodon-resembling stem-loop) for comparison with canonical tRNA structures (**Supplementary Fig. 6**).

NSUN2/IgG Aza-IP experimental design and data analysis. *Expression vector construction, and virus production and titration.* Total RNA was extracted from HeLa cells using Trizol (Invitrogen) and first-strand cDNA synthesis was performed with SuperScript III First-Strand Synthesis System (Invitrogen), using Oligo(dT) (Invitrogen), according to the manufacturer's protocol. An NSUN2 clone bearing a V5 tag was obtained by PCR using specific primer sets and HeLa cDNA as template in a two-step PCR format. The primer sets (**Supplementary Table 2**) replaced the first (ATG) codon with an AgeI restriction site, and inserted the Kozak consensus sequence containing the start codon (CACCATGG), and the sequence corresponding to the V5 tag (GGTAAGCCTATCCCTAACCCTCTCCTCGGTCTCGATTCTACG) at the 5′ end of the amplicon. The primers also placed an NheI restriction site at the 3′ end of the amplicon, right after the stop codon. Validated PCR products were double-digested with AgeI and NheI, and cloned into an AgeI/NheI-modified pPR-lentiviral plasmid (a gift from V. Planelles, Utah). Virus production used standard packaging and VSVG envelope plasmids, HEK-293-FT cells (Invitrogen) transfection with polyethyleneimine (Polysciences), harvesting, and titration on HeLa cells (using the EGFP marker) by flow-cytometry.

doi:10.1038/nbt.2566

Expression of V5-tagged NSUN2 was confirmed by western blot and the concentrated viral particles were stored at −80 °C.

Lentiviral infection. Forty 100 mm plates (total, for replicates and one control) were each seeded with two million HeLa cells, followed by infection (24 h later) with concentrated virus mixed with DMEM (3 ml, each plate, Invitrogen) supplemented with 10% FBS and 4 µg/ml (final concentration) Polybrene (Millipore). 18 h later, cells were washed with 1× PBS, trypsinized using TrypLE Express (Invitrogen), pooled and dispensed into 60 150 mm plates.

5-Azacytidine treatment. After 14 h growth, media was replaced with DMEM media containing freshly-prepared 5-Azacytidine (Sigma, 5 µM final), incubated for 12 h, followed by a second media exchange again with freshly-prepared 5-Azacytidine (5 µM final), and incubation for another 12 h, followed by harvesting (see below).

Preparing the pre-clearing beads. For each replicate (1 & 2) and IgG control experiments, 750 µl of Dynabeads Pan Mouse IgG (Invitrogen) were washed in 1 ml of diluted modified RIPA buffer (50 mM Tris PH 7.5, 1% Nonidet P-40 (NP-40), 0.1% sodium deoxycholate, 0.025% SDS, 1mM EDTA, 150 mM NaCl + Protease Inhibitor (PI) cocktail) supplemented with 5 mg/ml protease free bovine serum albumin (BSA) (Sigma) three times (for 2 min each), and re-suspended in 1.5 ml of diluted modified RIPA buffer + BSA + 20 µl of RNasIN (Promega).

Preparing the antibody-coated beads. For each experiment, 1.5 ml of the Dynabeads Pan Mouse IgG were split into two 1.5 ml tubes (750 µl each) and washed three times (for 2 min each) with 1 ml of diluted modified RIPA buffer + BSA. Next the beads of each tube were re-suspended in 1.5 ml of diluted modified RIPA buffer + BSA + 45 µg of the Invitrogen's mouse anti-V5 antibody (for replicate 1 & 2) or IgG (for control) and incubated at room temperature, rotating for 2 h. The beads of each tube were then washed three times (2 min each) with 1 ml of diluted modified RIPA buffer + BSA, and then re-suspended in 1.2 ml of diluted modified RIPA buffer + BSA + 15 µl of RNasIN.

Cell lysis and solubilization. After 24 h of growth in 5-Azacytidine, cells from each plate were washed with 1× PBS and trypsinized with 5 ml of TrypLE Express, and quenched with 5 ml of complete DMEM media. The contents of all 60 150 mm plates were pooled, spun at 2,000 r.p.m. at 4 °C for 10 min, washed with 15 ml of 1× PBS, spun again at 2,000 r.p.m. at 4 °C for 5 min, and pellets were flash frozen in liquid nitrogen and stored at −80 °C. Cells were thawed in 6 ml of modified RIPA buffer (50mM Tris PH 7.5, 1% Nonidet P-40 (NP-40), 0.2% sodium deoxycholate, 0.05% SDS, 1 mM EDTA, 300 mM NaCl + PI cocktail) supplemented with 30 µl of RNaseIN, and pipetted to homogeneity. The cell lysates (6 ml) were sonicated using a Misonix Sonicator XL2020, 6 pulses of 30 s (setting 4 (0.9 ON time/0.1 OFF time)) interposed with 30 s on ice. After sonication the lysates were pooled, mixed with 18 ml of dilution buffer (50mM Tris buffer (pH 7.5), 1% Nonidet P-40 (NP-40)). The twofold diluted lysate was dispensed into 36 1.5 ml Eppendorf tubes (1ml each), spun at 14,000 r.p.m. at 4 °C for 10 min, transferred to clean 1.5 ml tubes and kept on ice.

Immunoprecipitation. The 36 tubes were split into 3 groups (12 tubes each), the first and second groups were considered replicates (1 & 2), and separately mixed with anti-V5 coated beads. The third group was considered 'control', and mixed with IgG-coated beads. For each 1 ml of the lysate, 125 µl of the pre-clearing beads were added, incubated at RT with rotation for 2 h. Using a magnetic stand, the supernatant was transferred to a new 1.5 ml tube, and 200 µl of V5-coated beads (or IgG) were added and rotated for 4 h at RT. Beads were collected by the magnet, washed with 850 µl of modified RIPA + PI (3 times, 5 min each at RT), re-suspended in 500 µl of modified RIPA+ PI and transferred to a clean 0.5 ml Eppendorf tube. The beads (in 12 separate tubes) were collected and the supernatants discarded.

RNA fragmentation. For each sample (0.5 ml), 7.5 µl of the RNA fragmentation reagent (Ambion) was mixed with 67.5 µl of RNase free ddH_2O (Ambion) in one tube (total of 75 µl per tube), and in another tube 7.5 µl of the fragmentation stop solution (Ambion) was mixed with 7.5 µl of RNase free ddH_2O (total of 15 µl per tube). For the fragmentation, 75 µl of the diluted RNA fragmentation reagent was added to the sample, incubated at 94 °C for exactly 5 min in a thermocycler, chilled on ice for 1 min, and terminated by adding 15 µl of the diluted stop solution.

Ethanol precipitation and RNA extraction. After fragmentation, the supernatants derived from each replicate were collected and pooled in a clean 1.5 ml tube (total of 1080 µl). Next, 40 µl of the 15 mg/ml GlycoBlue (Ambion) and 104 µl of 3 M Sodium Acetate (Ambion) were mixed with the fragmented RNA and the mixture was split into three 1.5 ml tubes (346 µl each). Then, 865 µl absolute ethanol was added to each tube and the tubes were incubated at −80 °C overnight, and spun at maximum speed at 4 °C for 20 min. The pellets were washed once with 1 ml of 70% ice cold ethanol, air dried and dissolved in total of 40 µl RNase free ddH_2O (for all three of the tubes per experiment) and pooled in a single tube. Next 1 ml of Trizol reagent was added and the RNA was extracted according to the manufacturer's protocol.

Library preparation and high-throughput sequencing. Libraries involved Illumina's directional mRNA-Seq protocol involving the ligation of adapters to fragmented RNAs followed by reverse transcription, PCR and sample clean-up using AMPure beads (Beckman Coulter Genomics). Libraries were subjected to 50-cycle single-end high-throughput sequencing using Illumina's HiSeq 2000 system.

Computational analytical methods. Sequenced reads were aligned to the H. sapiens Feb 2009 genome build plus all known and theoretical splice junctions derived from Ensemble transcripts (see "MakeTranscriptome" application from open source USeq package[33]). Sequence alignments was performed using the commercial Novoalign package (http://www.novocraft.com) with options to allow gaps and mismatches, reporting 18 bp or larger inserts and reporting all of the reads mapped to the repeats and generating SAM-formatted alignment files. The "RNASeq" application within USeq was used to define enriched regions and obtain the RPKM (Reads Per Kilobase per Million mapped reads) values, and to make BAM/BAI files for visualization. The Integrative Genomics Viewer (IGV)[34] was used to visualize the alignment files and inspect the mapped reads at base pair resolution. Next, a combination of SAMtools (mpileup function)[38] and VarScan (mpileup2cns function)[39] were used to identify the cytosines showing the significant C>G transversion signatures (FDR<0.01 & transversion frequency > 4%) within the RNAs identified as significantly enriched by RNASeq application (FDR<0.01, Fold enrichment >3 & RPKM>3). For NSUN2 tRNA target sites, to convert the genomic coordinates of the candidate sites to the tRNA nucleotide numbers according to the standard tRNA numbering system[40], the candidate sites within the sequenced reads (visualized in IGV) were compared to the tRNA alignments from Genomic tRNA Database[41] for *Homo sapiens* genome (hg19 - NCBI Build 37.1 Feb 2009) and the corresponding residue numbers were deduced and reported according to the canonical tRNA cloverleaf secondary structure and numbering system[40]. To filter for possible C>G SNPs in the HeLa transcriptome, available RNA Seq data sets were used for comparison. Datasets are publicly available through GEO with the following accession number: GSE38957.

RNAi-mediated hNSUN2 knockdown. For each sample, 3×10^5 HeLa cells were seeded in a single well of a 6-well plate. The next day, cells were transfected with 60 pmol of Dharmacon's siGENOME Human NSUN2 siRNA – SMARTpool (M-018217-01-0005) or siGENOME Non-Targeting siRNA Pool #1 (D-001206-13-05) using Lipofectamine RNAiMAX transfection reagent (Invitrogen). After 72 h, cells of each group were passaged and transfected with 120 pmol of the same siRNA pools. The cells were harvested 72 h after the second transfection and subjected to RNA and protein extraction. Protein extracts were evaluated by western blotting for knockdown efficiency by immuno-blotting with hNSUN2 polyclonal (Proteintech-20854-1-AP) or hVinculin monoclonal (Sigma-V9131) antibodies.

Conventional RNA bisulfite sequencing. Purified RNA from hNSUN2 or control siRNA knockout HeLa cells were subjected to DNase treatment and fragmentation. For bisulfite treatment, 5 µg of the fragmented RNA was dissolved in 45 µl of RNase-free ddH2O and added to 240 µl of de-ionized formamide, mixed well and incubated at 95 °C for 5 min then placed on ice for 2 min. Next, 3 µl of the 100 mM hydroquinone and 312 µl of 5 M sodium bisulfite (pH 5) were added and the mixture was incubated at 50 °C, rotating. After 16 h, the mixture was cleaned up using Illustra NAP-10 Columns (GE Healthcare Life Sciences) and de-sulfonation was performed for 2 h at 37 °C in 1 M Tris buffer pH 9.0, followed by ethanol precipitation. The first-strand DNA synthesis and PCR amplification was performed using specific primer sets (**Supplementary Table 6**) designed for the bisulfite-converted tRNAs. The PCR product was

cloned using TOPO TA Cloning Kit (Invitrogen) and the resulting plasmids purified from several clones were sequenced. The sequences were aligned with the corresponding tRNA sequences and the maintained cytosines (representing methylated cytosines) are depicted in the figures as closed circles.

33. Nix, D.A., Courdy, S.J. & Boucher, K.M. Empirical methods for controlling false positives and estimating confidence in ChIP-Seq peaks. *BMC Bioinformatics* **9**, 523 (2008).
34. Robinson, J.T. *et al.* Integrative genomics viewer. *Nat. Biotechnol.* **29**, 24–26 (2011).
35. Todaro, G.J. & Green, H. Quantitative studies of the growth of mouse embryo cells in culture and their development into established lines. *J. Cell Biol.* **17**, 299–313 (1963).
36. Dong, A. *et al.* Structure of human DNMT2, an enigmatic DNA methyltransferase homolog that displays denaturant-resistant binding to DNA. *Nucleic Acids Res.* **29**, 439–448 (2001).
37. Zuker, M. Mfold web server for nucleic acid folding and hybridization prediction. *Nucleic Acids Res.* **31**, 3406–3415 (2003).
38. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
39. Koboldt, D.C. *et al.* VarScan: variant detection in massively parallel sequencing of individual and pooled samples. *Bioinformatics* **25**, 2283–2285 (2009).
40. Laslett, D. & Canback, B. ARAGORN, a program to detect tRNA genes and tmRNA genes in nucleotide sequences. *Nucleic Acids Res.* **32**, 11–16 (2004).
41. Chan, P.P. & Lowe, T.M. GtRNAdb: a database of transfer RNA genes detected in genomic sequence. *Nucleic Acids Res.* **37**, D93–D97 (2009).

doi:10.1038/nbt.2566

CHAPTER 4

DISCOVERING THE EPITRANSCRIPTOME: POTENTIALS, CHALLENGES AND FUTURE DIRECTIONS

Introduction

Natural nucleic acid polymers are made of only four standard nucleotides: A, C, G and T/U in DNA/RNA. However, after replication and transcription some of the incorporated standard nucleotides, at specific positions, undergo chemical changes to produce "modified nucleotides," in both DNA and RNA.

The first modified nucleotide, m^5C, was identified in 1948 in genomic DNA.[1] Since then, however, only a few other modified nucleotides have been characterized in DNA, which can be categorized under two groups. The first group is composed of those wanted events, which are made enzymatically such as hm^5C, f^5C, ca^5C, m^6A and Base J.[2] These are generated on purpose and have substantial known, or to be known, functions. Members of the second group are those unwanted events, which are produced as the result of DNA damage upon chemical or UV exposure, such as 8-oxo-A, Xanthine, Inosine etc., known as DNA lesions[2]. Cells have adopted elaborated pathways to actively and selectively recognize and replace these with normal nucleotides to prevent interruptions and errors during replication. Overall, it is evident that accommodation of natural modified nucleotides in genomic DNA has been restricted to very few instances during evolution, more likely to avoid compromising the processivity and fidelity of replication and transcription processes. This is because that some modifications can pause or slow down DNA/RNA polymerases, on DNA, and some others can affect the standard A:T, C:G base paring rules and therefore compromise the maintenance and flow of genetics information.

RNA, in contrast, does not serve as the genetic repertoire of living organisms, except in a few viral examples, and once is transcribed it will no longer be used as polymerization template, except in a few examples of RNAs involved in telomere maintenance or propagation of infective retroviruses and retrotransposons. Thus there has been no evolutionary pressure to prevent the standard nucleotides of RNA molecules from being evolved into new modified entities. In

addition unlike the single function of DNA, the genetic reservoir, RNAs have become specialized for more diverse responsibilities in living organisms, which partly explains their greater tendency to accommodate much more modified nucleotides, something over 100 distinct structures, characterized so far.

RNA modification can be defined as any post-transcriptional alteration of the structure of standard nucleotides within the RNA polymers once they are being transcribed. According to this definition from simple deamination and editing of C to U bases in mRNAs to the formation of the hypermodified wybutosine (yW) in the anticodon of eukaryotic phenylalanine tRNAs are considered RNA modifications. Generation of almost all modified nucleotides (except for few that are produced upon RNA damage) is enzymatic requiring one or more specialized enzymes depending on the reactions. In addition the modification processes often take place within complexes composed of one or more accessory proteins and/or RNAs, each with specific roles such as providing the scaffold for the modification complex or guiding the modifier enzymes to specific RNA and/or specific nucleotide targets.

In principle, all accessible portions of a given nucleotide, base, sugar and phosphate, can become targeted by RNA modifier enzymes (Figure 4.1). Although sugar and phosphate modification types are limited (Figure 4.1a,b), the base modification is highly diverse, which is achieved through one of the four reaction types: functional group substitution, or addition, and base exchange, or isomerization (Figure 4.1c). All these modifications together shape the cellular RNA “epitranscriptome.”

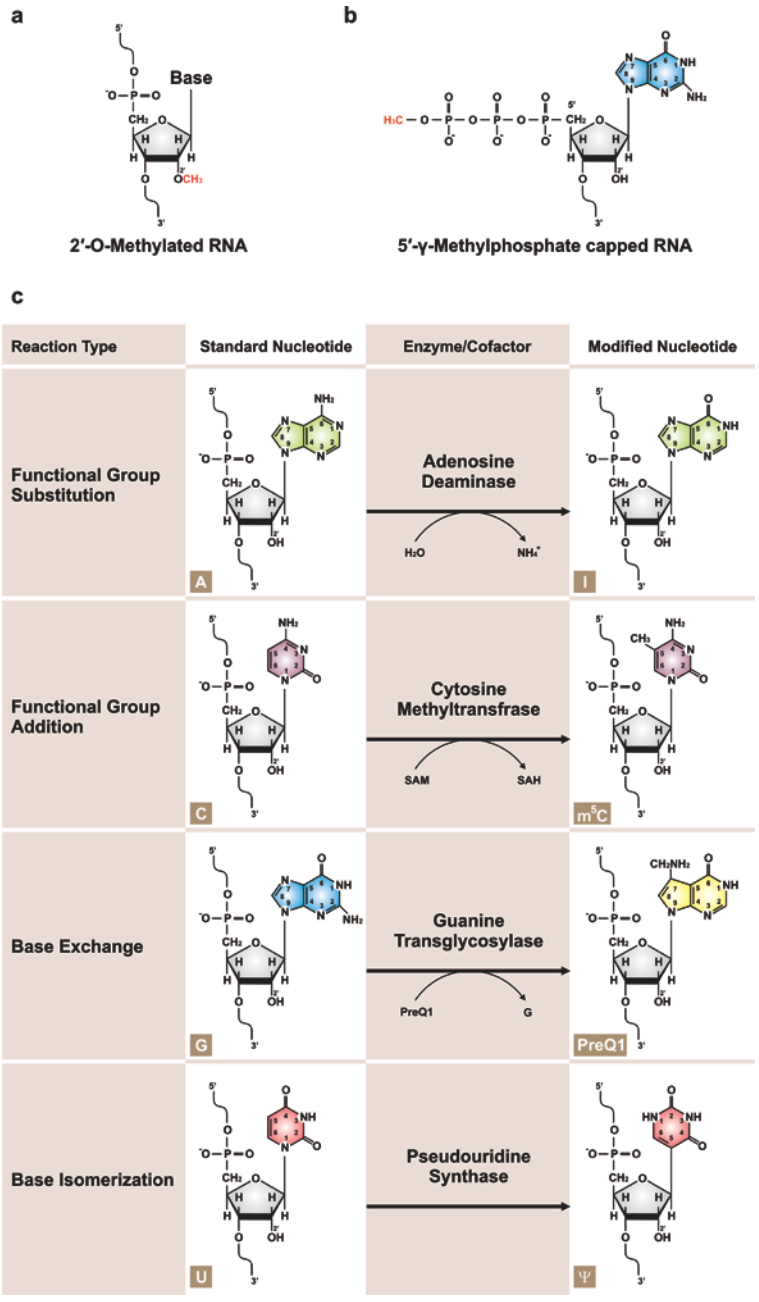


Figure 4.1 | Classification of RNA nucleotide modifications. a, Sugar modifications. There are few types of sugar modifications. Here a single 2′-O-methylated nucleotide (Nm) as part of an RNA molecule is shown with a methyl group at the 2′ position of the sugar. The "Base" can be any of A, C, G or U. **b**, Phosphate modifications. There are few types of phosphate modifications. Here a 5′-γ-methylphosphate capped RNA is shown. **c**, Base modifications. There are over 100 distinct enzymatically made base modifications in RNA. Most base modifications are made through either functional group substitution or addition. Limited modifications are generated through base exchange or isomerization. One example for each reaction type is given.

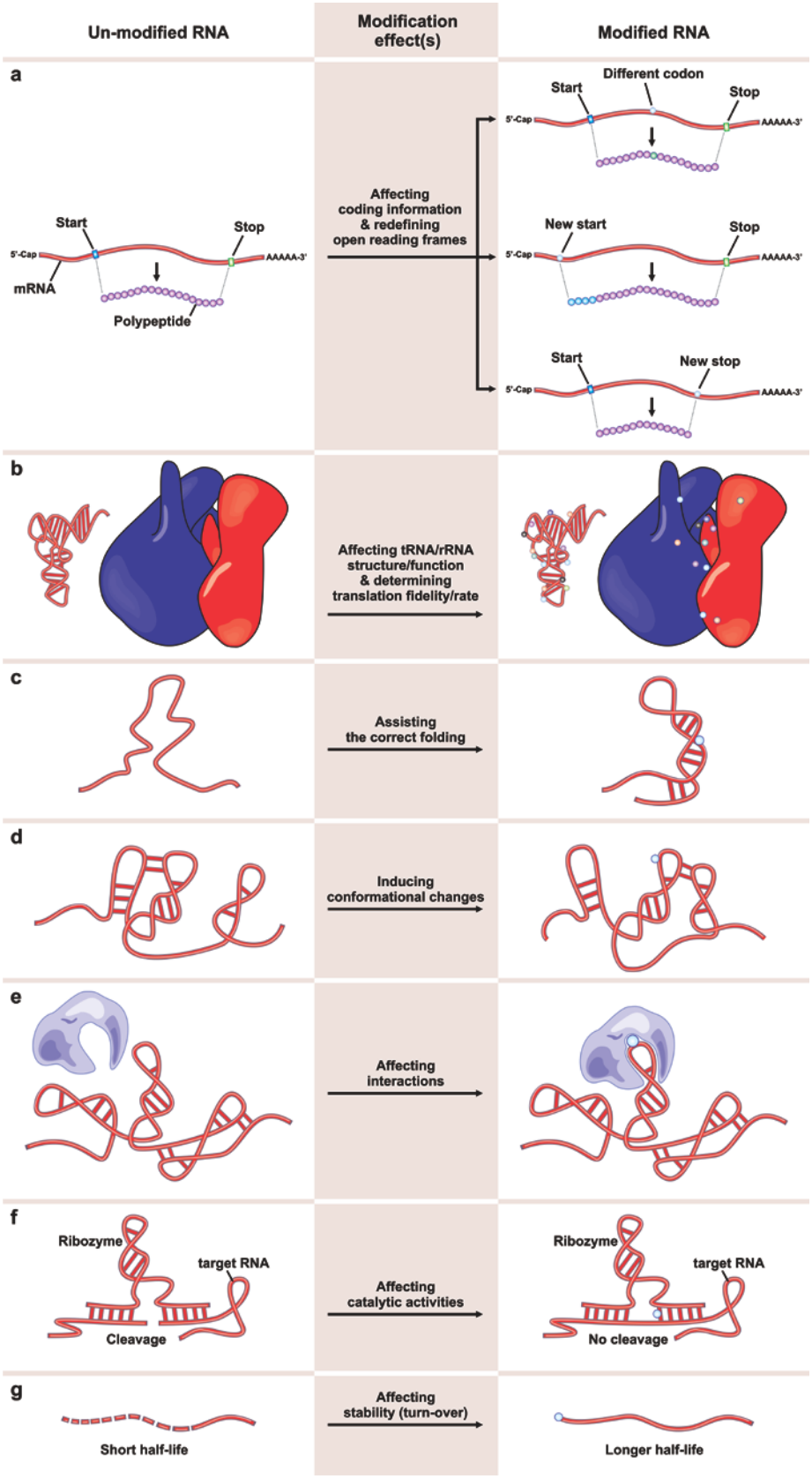
Epitranscriptome landscape

Modified nucleotides are present, in different frequencies, in both coding and non-coding RNA molecules of different types. Modified nucleotides function in regulating the mRNA processing and translation, or affect their stability and turnover rates. They can also provide additional letters to RNA alphabet or redefine the open reading frames (ORFs) increasing the coding capacity of mRNAs. In noncoding RNAs (ncRNAs) modifications provide new structural and functional features, expanding the flexibility and operative capabilities of diverse ncRNAs serving as transfer, catalytic, structural, regulatory or guide RNAs (Figure 4.2). Here, to emphasize on the importance of RNA modification research, we briefly introduce some selected modified nucleotides that have been studied more in the past.

N7-methylguanosine (m^7G) is the most studied modified nucleotide in mRNAs as it is the main component of the eukaryotic mRNA 5'-cap structure.[3] 5'-cap is made by addition of a non-templated guanosine to the 5' end of mRNAs through formation of an unusual 5' to 5' triphosphate linkage followed by enzymatic methylation of G at position N7. It functions in regulation of mRNA stability, splicing, transport, and translation.[4] 5'-cap structure has also been found at 5' end of some noncoding RNAs including a subset of small RNAs, in which its function is yet to be explored.[5] Internal m^7G is also widely present in both prokaryotic and eukaryotic tRNAs.[6]

N6-methyladenosine (m^6A) is probably the most abundant internal modification in eukaryotic mRNAs from yeast to human, and is also present in their tRNAs, rRNAs and snRNAs, as well as in bacterial and archeal tRNAs and rRNAs and some viral RNAs.[7] In Saccharomyces cerevisiae, meiotic mRNAs show much higher m^6A levels as compared to mitotic mRNAs, and m6A content of polyA RNA fraction shows remarkable increase upon induction of sporulation. Several lines of evidence indicate that N6 adenosine methylation function in stress response,

Figure 4.2 | Functional classification of RNA modifications. a, Modifications affecting the coding information of mRNAs. Some post-transcriptional modifications within the open reading frames (ORFs) can change the way the triplets are decoded in the ribosomes introducing a different amino acid and changing the protein sequence. Nucleotide modifications can also change the definition of ORFs to produce shorter or longer protein chains through introduction of new or removal of the existing start/stop codons. **b**, Modifications affecting the structure or functions of decoding (translation) machinery. tRNAs and rRNAs are the most highly modified RNA species and the modifications are required for their proper folding and activity. Both the rate and fidelity of translation can be changed by changes in tRNA or rRNA modifications. **c,** Modification assisted folding. Some RNA molecules require nucleotide modifications to obtain their proper folding. **d**, Modification induced conformational changes. Nucleotide modifications might help switching the conformation of a folded RNA from one state to the other. **e**. Modifications affecting the interactions. Nucleotide modifications can provide new structural properties for RNAs to absorb or repel specific proteins binders. The binders might prefer binding to either of the modified or unmodified RNAs. Here protein-RNA interaction is shown, but in general nucleotide modifications might affect other interactions as well as interaction of RNA with DNA, RNA, small molecules, ions, etc. **f.** Modifications affecting the catalytic activity of the ribozymes. It is anticipated that nucleotide modifications in ribozymes affect their substrate specificity or catalytic activity for specific RNA substrates. **g**. Modifications affecting the RNA stability (turnover rates). Modifications can prevent degradation of RNAs. Examples are mRNA 5'-capping and changing the miRNA binding sites of mRNAs through RNA editing. There may also be modifications marking the RNAs for more rapid degradation and turnover.


Un-modified RNA
Modification effect(s)
Modified RNA
a
Start
Stop
5'-Cap
AAAAA-3'
mRNA
Polypeptide
Affecting coding information & redefining open reading frames
Different codon
New start
New stop
b
Affecting tRNA/rRNA structure/function & determining translation fidelity/rate
c
Assisting the correct folding
d
Inducing conformational changes
e
Affecting interactions
f
Ribozyme
target RNA
Cleavage
Affecting catalytic activities
No cleavage
g
Affecting stability (turn-over)
Short half-life
Longer half-life

cell fate control and initiation of meiosis in budding yeast.[8, 9] m^6A contents in Arabidopsis thaliana is tissue specific with higher levels in flower buds in comparison to leaves and roots.[10] N6-adenosine methyltransferase gene (MTA) is required for normal growth and organ definition as well as prompt and proper responses to environmental stresses and stimuli in plants.[11] This modification is also essential for development and gametogenesis of drosophila melanogaster and complete deletion of the methyltransferase gene (Dm ime4) is lethal for the organism.[12] m6A was known for years to be present in mammalian mRNAs, with only one mapped m^6A in bovine prolactin (bPRL) mRNA.[13] Recently two transcriptome-wide profiling of this methyl mark in human and mouse cells revealed a widespread occurrence of this modification in both coding and noncoding RNAs (more than 12,000 sites). Many of these sites were shown to be evolutionary conserved between human and mouse, with more incidents around the stop codons and within the long internal exons.[14, 15]

Conserved locations in tRNAs and rRNAs contain 5-methylcytosine (m^5C) and a recent study shows the widespread occurrence of cytosine methylation in other noncoding RNAs and also in many mRNAs.[16, 17] Although the exact function of cytosine methylation is still unknown, gene disruption or miss-regulation of some known cytosine methyltransferases can result in variety of different abnormalities such as developmental defects, mental retardation, infertility, growth problems and cancer.[16, 18-26] In addition, recent works by several groups suggest that the main functions of cytosine methylation could be more involved during the stress conditions and/or in cellular response to pathological conditions such as in defense against the infective RNA viruses.[27-29]

Inosine (I) is another widespread modified nucleotide produced upon deamination of adenosine by adenosine deaminase enzymes, in a process known as RNA editing. Because I prefers to base pair with C (instead of U), A to I editing changes the sequence information of

mRNAs, post-transcriptionally. Editing within the open reading frame (ORF) can then change the transcribed codes and produce new protein isoforms and contribute in functional diversification of the proteome. Indeed possible edition of the adenosine residues may result in generation or removal of the start (AUG) or stop (UGA, UAG, UAA) codons, redefining the ORFs in the edited mRNAs. ORF extension as the result of changing non-sense to sense codon upon adenosine deamination has been reported for viral RNAs.[30] A recent transcriptome-wide profiling of the edited sites in human, however, shows that in mRNAs, A to I editing events are more frequent in 3' UTRs and introns and moderately abundant in 5' UTRs. Although the exact contribution of majority of these events are not known, some of the edited sites in 3'UTRs were mapped to the miRNA binding sites suggesting the regulatory roles of this process. A number of edited sites in small and long noncoding RNAs have also been reported including some that happen within miRNA precursors affecting their biogenesis.

The C to U transition through post-transcriptional deamination of cytosines is another way of changing the coding information of mRNAs. This is very well documented in the case of APOBEC1 mediated mRNA editing, generating new in-frame stop codons in apolipoprotein B and neurofibromin-1 mRNAs to produce shorter forms of the proteins in special cases or tissues.[31-34] Numerous APOBEC1 dependent C to U editing events have also been recently found in 3'-UTR of many mRNAs although their exact function remains to be addressed.[35]

Pseudouridine (Ψ), originally designated the "fifth nucleotide,"[36] is the first of the nucleotide modifications detected in RNA and still the most abundant of them. In human about 100 Ψ residues have been mapped to the small and large subunits of cytoplasmic and mitochondrial rRNAs.[37] Some pseudouridylation sites are specifically targeted (guided) by H/ACA box snoRNAs[38] and miss-regulation of snoRNPs, targeting U to Ψ isomerization at ribosomal peptidyl-transferase center, alters the structure and activity of ribosomes.[39] Ψ in tRNAs roles in

structural stabilization of helices, mediated by hydration of its major-groove imine group that generally makes the Ψ-containing RNAs a little stiffer.[40, 41] The complementary region of U2 snRNA, which makes a short seven base pair helix with a consensus sequence in the intron of precursor-mRNAs, has a conserved Ψ residue in nearly opposite of the branch site. This Ψ induces the branch site adenosine residue of the intron to bulge out of the duplex, placing its nucleophilic 2'OH in an appropriate position to start the first trans-esterification step of splicing.[42] Although it is not clear whether mRNAs are naturally subjected to pseudouridylation, Karijolich and Yu showed that during mRNA translation ribosomes do not terminate the polymerization at Ψ-containing stop codons, producing longer poly-peptides, both in-vitro and in-vivo. In order to produce Ψ-containing stop codons in-vivo, they engineered specific H/ACA box snoRNAs to target the isomerization of the first nucleotide (U) of the stop codons into Ψ in mRNA sequences.[43] This raises the possibility that mRNAs and other noncoding RNAs can be the natural targets of pseudouridylation due to the diverse pool of expressed H/ACA box snoRNAs, most of them uncharacterized.[44] It also indicates the possible therapeutic application of custom-designed H/ACA box snoRNAs to remove unwanted stop codons in mRNAs generated as the result of point mutations in some genetic disorders.[45]

N1-methyladenosine (m^1A) is most studied in tRNAs as it is a highly conserved modification at position 58 (m^1A58) of almost all eukaryotic tRNAs with few exceptions.[46] Methylation of A58 provides a positive charge, which is believed to stabilize the tertiary structure of certain tRNAs, and enzymes involved in its synthesis are essential for organismal viability.[47, 48] In addition, methylation at N1 position (m^1A), as it affects the Watson-Crick interface, unlike methylation at N6 position (m^6A), affects the normal base pairing capability of the base thus can immediately influence the folding of RNA. This is evident in observation showing that N1-methylation of A9 residue (m^1A9) in human mitochondrial $tRNA^{Lys}$ shifts the

conformational equilibrium of unmatured tRNA precursor toward the functional cloverleaf structure.[49] This is an example of modification-assisted/-induced RNA-folding. m^1A has been found in ribosomal RNAs and it would be interesting to see if it naturally appears in mRNAs and other noncoding RNAs as well.

4-thio-uridine (s^4U) has been identified in prokaryotes. NuvA and NuvC are the two enzymes responsible for modifying the U8 in some E. coli tRNAs into s^4U.[50] Although the exact functions of s^4U are not known failure to make the s^4U results in higher susceptibility of bacteria to UV. Interestingly the absorbance spectrum of s^4U extends into the near-UV range and UV exposure induces the cross-linking of s^4U8 to its nearby C13. This in-vivo cross-linking somehow induces the accumulation of nucleotide ppGpp, which itself is a key regulator of stress response in bacteria.[51] It is interesting to find the other possible sites of s^4U in tRNAs and other ncRNAs or even mRNAs.

The ribose of the immediate first two nucleotides of the capped mRNAs is often subjected to methylation making 2'-O-methylated modified nucleotides (Nm). This modification is believed to function in cellular immunity by providing a signature to distinguish self-mRNAs from invading non-self RNAs, recognized by specialized RNA sensors.[52, 53] The C/D box snoRNAs guide the 2'-O-methylation of internal nucleotides in essential regions of tRNAs, rRNAs and snRNAs, important for their proper functions.[38] Internal mRNA methylation, by a viral encoded 2'-O methyltransferase enzyme, has also been shown to inversely correlate with RNA translation rate. This methyltransferase enzyme may function to methylate the viral genomic RNA as a protective strategy against host immune response, and/or modulate the host physiology through methylation of host RNAs.[54] In addition, terminal 2'-O-methylation of endo-siRNAs, piRNAs and plant miRNAs seems to protect their 3'-end or mark them to be recognized by specific proteins for downstream functions.[55, 56]

The phosphate group can also be the target of enzymatic post-transcriptional modifications. While most RNA polymerase II (PolII) transcripts contain the famous m7G containing 5'-cap structure, some of the PolIII transcripts such as U6, 7SK and B2 small RNAs represent a special 5'-Y-methylphosphate cap structure required for stability of these transcripts.[57] Enzymatic methylation of α-phosphate at 5' end of miRNAs has also been shown to role in regulation of miRNA processing.[58]

Epitranscriptome dynamics

Appearance of some modifications in RNA shows cell type-, tissue- or developmental stage-specific differences. One of the best examples is the C>U editing in apolipoprotein B mRNA exclusively in the small intestine to make a specific isoform, ApoB48, which is shorter than the ApoB100 produced in the liver.[31, 32] Other examples are the marked differences between the m6A levels in mRNAs from meiotic and mitotic cells and during sporulation of budding yeast, or between different plant tissues.[8-10]

Beside these and some other examples, however, during the lifetime of a single cell, many RNA modifications show appearance and disappearance cycles upon stimulation, providing dynamics to the cellular epitranscriptome. For example a precise quantitation of modified nucleotides in tRNAs revealed that some modifications such as Cm, m^5C and $m^2{}_2G$ show increased levels upon hydrogen peroxide exposure.[59] Another study shows that oxidative stress-induced increase in the methylation of C34 (by Trm4) in $tRNA^{Leu(CAA)}$ in yeast directly correlates with the translation level of proteins enriched in TTG codons notably the ribosomal protein RPL22A. This could become achieved through increased stability of the anticodon loop of $tRNA^{Leu(CAA)}$ upon methylation of C34. Interestingly loss of either Trm4 or RPL22A makes the

cells more susceptible to hydrogen peroxide.[27] A remarkable increase in the methylation level of C38 in $tRNA^{Asp}$ in yeast in response to nutrient deprivation has also been reported. pmt1, the homolog of human DNMT2, is the responsible methyltransferase for methylating the C38 position.[28] For many years, pmt1 was considered an inactive methyltransferase enzyme because of lack of known substrates and also due to the insertion of a serine amino acid within its catalytic site as compared to DNMT2.[60] This targeted survey, however, demonstrated that not only pmt1 is an active RNA cytosine methyltransferase but its activity is also under the control of serine/threonine kinase Sck2 arm of the yeast nutrient signaling pathways.[28]

Another example of translation of signals to RNA modifications is appearance of new pseudouridylation sites in U2 snRNA upon external stimuli in yeast. U2 in budding yeast has only three Ψ residues: Ψ35, Ψ42 and Ψ44 in normal conditions; however, upon heat shock Ψ56 and upon nutrient deprivation Ψ56 and Ψ93 appear in the molecule. This indicates that pseudouridylation can be regulatory at specific locations, differentially inducible by different environmental stimuli. The inducible Ψ93 has been shown to function during pre-mRNA splicing.[61] In addition the dynamic changes of RNA modifications has been demonstrated in subsets of m^6A marks in eukaryotic mRNAs showing increase or decrease through stimulus-dependent regulations.[14, 15]

Increase in the level of modification at a specific site in RNA can become achieved through expression or up-regulation of the modifier enzymes. Appearance of new sites can be achieved through activation of the existing enzymes, or activation, or changes in the combination, of RNA and/or protein components of the modifier complexes to guide for modifying new RNAs and/or sites. Conversely, modifications may passively disappear upon depletion or inactivation of the modifier enzymes/complexes or through increased RNA turnover rates to make more unmodified RNAs, or the combination of both. The most intriguing

finding in the field of dynamic RNA modifications in the past few years, however, is the discovery of FTO, a demethylase enzyme that can actively remove the methyl marks from m6A in RNA molecules. This is analogous to active demethylation of methyl-cytosines in DNA[62]. FTO shows strong genetic association with body weight and obesity and its loss of function results in problems such as postnatal growth defects as well as some cardiac abnormalities and brain malformations, indicating its important functions[63].

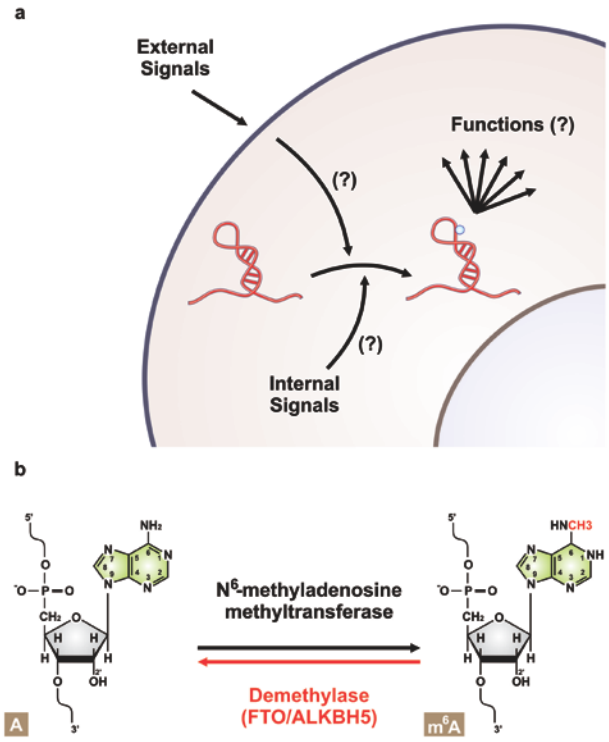


Figure 4. 3 | The dynamic epitranscriptome. a, Translation of external and internal signals into RNA modifications. External regulative or stress conditions, and internal signals such as changes during the cell cycle events, can drive the signaling pathways resulting in induction of RNA nucleotide modifications. The increase (or decrease) in the level of existing modifications as well as appearance of new modification sites within RNA molecules has been observed. The extent of modification changes, as well as the underlying signaling pathways and key regulators, and most importantly the outcomes (functions) of such RNA modifications remain to be explored. **b**, Active erasure of nucleotide modifications. For some RNA nucleotide modifications decrease in the level of modifications might be due to active enzymatic reversal of the modifications. This has been proven by discovery m^6A demthylase enzymes (FTO and AlKBH5). These enzymes can actively change the m^6A to A within an RNA molecule by removing the methyl mark from the N6-methyladensoine.

Epitranscriptome profiling

Despite decades of extensive research on profiling of modified nucleotides in RNA and their functional analysis, and enzymology of the RNA modifiers, still a comprehensive picture of the cellular epitranscriptome is lacking. Recently this field has attracted more attention mainly for two reasons: First, due to the recent technical improvements providing the possibility of exploring RNA modifications in more details and accuracy, and at larger scales, higher speed and lower prices; second, due to the discovery of a number of new small and large noncoding RNAs in different organisms leading to proposing and proving novel functions for RNA molecules.

The ultimate goal of studying RNA modifications is to fully understand the contribution and function of each individual modification, and the interplay between different modifications. However, the first step is to address how these over 100 modifications are distributed in the sequence of coding and noncoding RNA molecules or what the entire epitranscriptome looks like. Here, from the technical standpoint, we discuss “epitranscriptome profiling” by providing a brief overview of the classical methodologies used so far, elaborating the recent advancements, as well as proposing potential new directions.

Classical methodologies

Different strategies have been used in the past decades to detect and locate specific modifications in RNA molecules (Table 4.1). In principle, a majority of these methods belong to one of the two major groups. Group one encompasses those techniques that are based on the unique physicochemical properties of specific modifications. These include thin layer chromatography (TLC), gas chromatography (GC), liquid chromatography (LC), and its coupling

Table 4.1. Classical methodologies for RNA modification profiling

Approach	Principle	Applications (examples)
Physicochemical property-based		
Capillary electrophoresis	Hydrolyzed RNAs are subjected to separation of nucleotides/nucleosides, based on differential electromobility of charged particles (due to their specific mass-to-charge ratio) in an electrical field, followed by UV detection.	Rarely used in the past for separation and detection of different nucleosides and nucleotides derived from RNA samples.
Chromatography	Hydrolyzed RNAs are subjected to separation of nucleotides/nucleosides, based on their differential solubility in different solvents resolved by their retention characteristics in chromatography approaches such as TLC, LC or GC, followed by UV detection.	Widely used in the past and still in use for separation and detection of different nucleosides and nucleotides derived from RNA samples.
Mass Spectrometry	Oligonucleotides/nucleotides/nucleosides separated by chromatography approaches are specifically subjected to ionization followed by mass analyzer transfer and detection by LC-MS or LC-MS/MS (tandem-) mass spectrometry.	Widely used in the past and still in use for characterization of nucleosides and nucleotides as well as defining the position of modified nucleotides in short stretches of RNA oligonucleotides.
Reverse transcription (RT)-based		
RT on intact RNA	a) cDNA sequencing determines the location of those modified nucleotides with naturally altered base pairing behavior. b) Gel electrophoresis can determine the location of those modifications naturally slowing down or stopping the RT enzyme during first strand cDNA synthesis producing shorter cDNA fragments which can be resolved on the gel.	a) Widely used in the past and still in use for determination of the RNA editing sites (i.e. A>I or C>U changes). b) Example: Determination of the location of 2′-O-methylated nucleotides (Nm) in RNA templates using low dNTP concentrations during cDNA synthesis.
RT on pretreated RNA	a) Chemical pretreatment followed by cDNA sequencing can determine the location of those modifications differentially react with specific chemicals that change the sequence information of the RNA molecules. b) Chemical pretreatment can specifically target and change the structure of some modified nucleotides enabling them to slow down or stop the RT enzyme during first strand cDNA synthesis. c) Some reagents can cleavage the RNA specifically at special modified nucleotides. After RT reaction, the shorter cDNA fragments will be resolved on the gel.	a) Example: Cytosine RNA methylation (m^5C) profiling by RNA bisulfite sequencing approaches. b) Example: Determination of pseudouridine sites by CMCT treatment of RNA which exclusively makes the pseudouridines bulky arresting the RT enzyme during cDNA synthesis. c) Example: Determination of the location of 2′-O-methylated nucleotides (Nm) in RNA templates by partial alkaline hydrolysis resulting in cleavage of RNA at random locations but never at Nm site as 2′- O-methylation protect the nucleotides from nucleophilic attacks.
Others		
Ligation-based	For some modifications, if the sequence context of the RNA and location of the modified nucleotide is known, two juxtaposed DNA oligonucleotides can be designed to anneal to the RNA template, one sitting exactly on top a modified nucleotide. The differential ligation affinity of T4 DNA ligase to join the oligonucleotides annealed to unmodified or modified transcripts will then show the modification status of the pool of RNAs for that specific RNA/modification.	Very limited application as an analytical method for evaluation of the level of defined modifications (i.e. 2′-O-methylated nucleotides (Nm)) at defined positions within the RNA templates.
Hybridization-based	Those modifications resulting in differential hybridization affinity of modified RNAs are analyzed by microarray platforms harboring the complementary probes.	Has been applied for detection of dihydrouridine, m^1G, $m^2{}_2G$, m^1A and $m^6{}_2A$ in RNA molecules on array platforms.

with mass spectrometry (LC-MS), and capillary electrophoresis (CE) to separate and identify the modified nucleotides based on their physicochemical properties.[64] Although these techniques are best for characterization and quantification of modified nucleotides and have provided us with a massive backup of knowledge about the diversity and levels of modified nucleotides in RNA, except in limited cases they fail to show the exact location of the modifications within the RNA molecules. Members of the second group of technologies, in contrast, can show the exact location of the modifications as they rely on the effect of modified nucleotides on fidelity, speed and processivity of reverse transcriptase (RT) enzymes during first strand cDNA synthesis.[65] It is also worth mentioning that the differential chemical reactivity of many of modified nucleotides has been recruited efficiently in order to expand the applications of these two methodologies for studying more diverse modifications and/or to increase the resolution of the techniques.[66] In addition to the two major methodologies mentioned above a few other strategies have also been devised and used recently with slightly different principles. These include a microarray-based technique, which is based on the differential hybridization affinity of modified RNA molecules to complementary probes on an array platform,[67] and a ligation-based technique, which relies on the differential ligation affinity of T4 DNA ligase to join two juxtaposed DNA oligonucleotides annealed to an RNA template, one sitting exactly on top of a modified nucleotide.[68, 69] These two strategies, though used successfully to detect or verify the modified nucleotides in real biological samples, are yet to be fully implemented for broader applications due to some technical limitations.

All these technologies together have helped the scientists to discover more than 100 distinct modified nucleotides in RNA, and determine the distribution of some of them in different transcripts. However, the greater extent of knowledge about tRNA and rRNA modifications is not comparable to what we know about modification of other RNA species,

while the presence of many modifications is anticipated in most RNAs. This is probably the result of at least three different factors. First, the tRNAs and rRNAs are highly modified transcripts, a requirement for effective regulation of the fidelity and rate of translation both in normal and stress conditions. Especially tRNAs with their short length of about 75 bp, on average bear about 13 modified nucleotides in each transcript. This greater level of modification has made them the first obvious targets of RNA modification studies. Second, tRNAs and rRNAs are the most highly abundant transcripts in every cell type. This higher level of transcripts has made them the first available targets of RNA modification studies. Conversely, the higher transcript level of the highly modified tRNAs and rRNAs makes it extremely difficult to purify any of the other RNA species without contamination with these two, even with today's protocols. The impurities in the RNA preps increase the false discovery rates and decrease the resolution of the surveys. Finally, the detection limit of the classical epitranscriptome profiling approaches as well as their low-throughput format has made them most suitable for epitranscriptome profiling of the highly abundant and modified RNA species. Therefore, characterization of low copy modified RNA species and/or RNAs with low modification penetrance has been hampered by lack of powerful sensitive technologies.

High-throughput sequencing approaches

Concomitant with revolutionary impact of high-throughput sequencing on large-scale and comprehensive genome studies,[70] sequencing reverse transcribed RNA molecules (cDNAs), using the same platforms, known as RNAseq, indeed revolutionized the area of transcriptome studies.[71] RNA-seq is currently used for transcriptome-wide expression profiling and has clearly proven its superiority over microarray-based techniques. The still increasing power and

decreasing cost of the technique in conjunction with the possibility of barcoding, which allows sequencing of several samples in a single lane, has made this technique the method of choice for almost all transcriptome-wide studies. This technology has recently been used for epitranscriptome profiling of few modified residues, notably m^5C, m^6A and I revealing some new biological insights for them. In principle the high-throughput sequencing approaches we discuss here are the extension of classical RT-based approaches with much higher sensitivity, and sometimes with some changes enabling their application as more of discovery rather than analytical techniques. Detection of many modifications using such approaches does not require much prior knowledge about the extent and distribution of the modified nucleotides within the transcripts. Here, we introduce different strategies for epitranscriptome profiling using the high-throughput sequencing approaches (see Figure 4.4).

Direct detection by high-throughput sequencing

A majority of modified nucleotides behave the same as standard nucleotides during reverse transcription and cDNA synthesis and are therefore called “RT-silent.” However, some modifications change the standard A:T, C:G base pairing rules and some others may decrease the RT rates or even completely block the progression of cDNA synthesis. Such modifications can introduce single nucleotide variations (SNVs) or produce truncated cDNAs indicating the sites of the modified nucleotides upon analyzing the sequencing reads at base pair resolution. Detection of the frequent sites of modifications regardless of the modification type could be possible through analyzing the available RNA-Seq datasets generated for other purposes. In addition RNA samples can be sequenced and analyzed, considering proper controls, with the intention of identifying the sites of specific modifications.

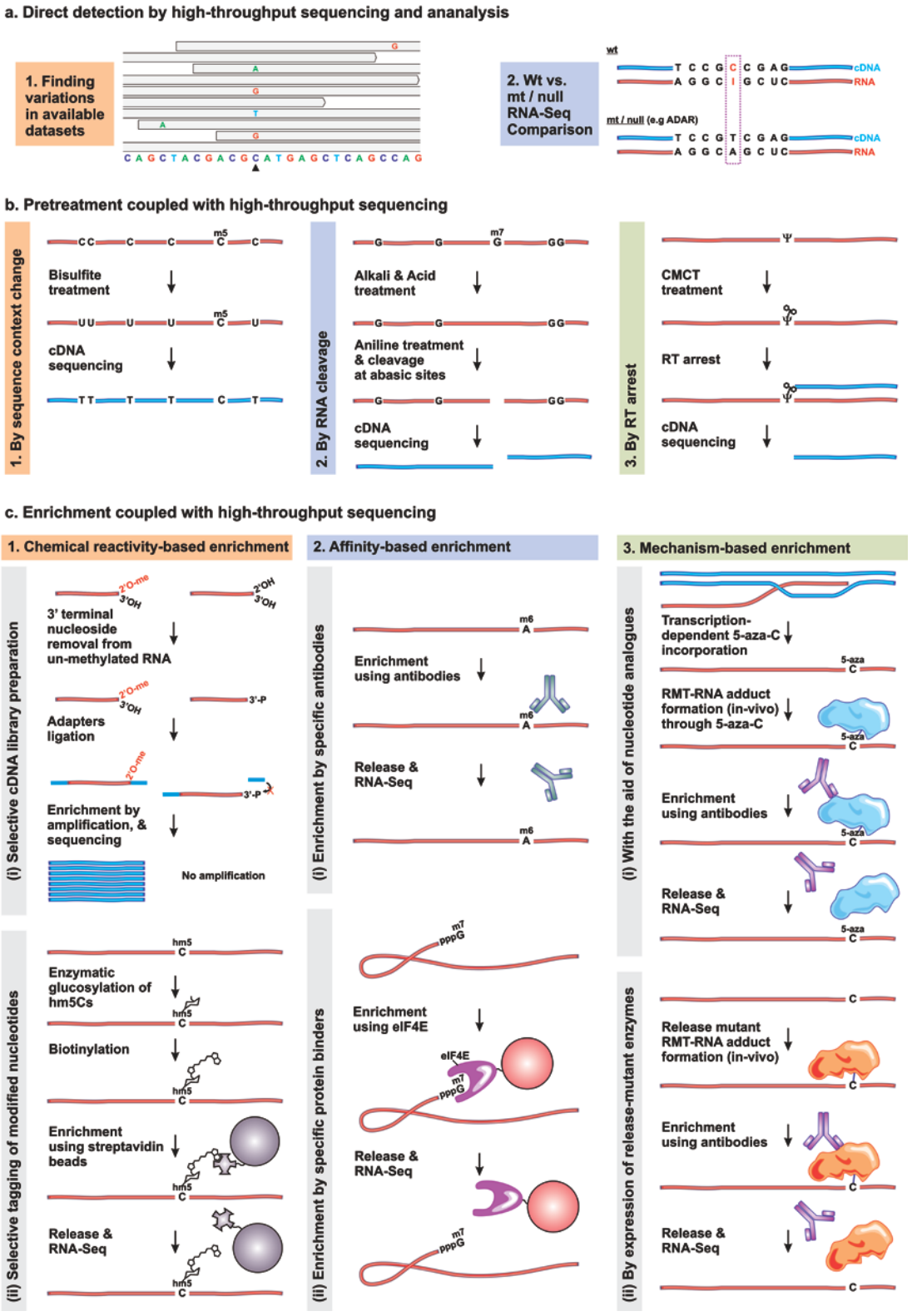


Figure 4.4 | High-throughput epitranscriptome profiling approaches. See text for details.

Over the past few years massive RNA-Seq datasets covering the entire transcriptome of many organisms have become publically available. Limited bioinformatics analysis of such datasets has been specifically used to find the locations of frequent sequencing errors as possible sites of modified nucleotides. These works revealed a number of known and new modification sites within small RNAs such as tRNAs and miRNAs, demonstrating that not all of the so called "sequencing errors" are true artifacts but rather a subset of them could be simply real sites of post-transcriptional RNA modifications.[72, 73] This may explain, to some extent, the controversies over a work reporting widespread differences between the sequences of RNA and DNA in the human transcriptome. The report claimed the identification of more than 10,000 sites in human mRNAs.[74] Several groups later on commented on the work and suggested that majority of these sites are most likely artifacts due to misinterpretation as the result of analysis errors such as neglecting the genetic variations as well as sequencing or mapping errors.[75-77] However, neither of the report itself nor the comments specifically considered the post-transcriptional nucleotide modification as one potential cause of at least a subset of those RNA/DNA differences, while this is practically anticipated. Overall, bioinformatics analysis of the available datasets can help to identify the frequent and reproducible sites of RNA/DNA discrepancies as the candidate sites of RNA modifications become experimentally verified.

For a few defined RNA modifications, capable of changing the standard base pairing rules, comparison of transcriptome with the source genome has been efficiently used to readily identify the sites of modifications. The A>I and C>U editing are the well established examples of the application of this approach and there are a number of reports for detection of such editing sites using both classical approaches in the past and high-throughput sequencing approaches in recent years. To accurately identify the sites of such modifications RNA extracted from the cells lacking the responsible enzyme could be used as a definitive control. It would be interesting to

experimentally survey the base pairing behavior of all other modified nucleotides to find those that can become simply identified by row sequencing of the RNA. For example, s4U shows a greater tendency to base pair with G than A inducing some levels of U>C transition upon reverse-transcription and sequencing, thus potentially becoming readily identified, although it has not been explored this way yet.[78] In addition, for those modifications that cause the delay or pause during reverse transcription, mapping might be possible through data analysis by looking for the sites of cDNA truncation.

Overall, direct detection of the sites of modifications within RNA transcripts via analysis of the high-throughput sequencing datasets is a quick way of epitranscriptome profiling for limited types of modifications. However, this approach, like other approaches discussed in this report, is prone to both false positives and negatives. The false positives may come from sequencing and mapping errors or genetic variations. The false negatives may be due to multiple sources such as the higher turn-over rates of modified RNAs, the transcripts' low copy number, the low penetration of modification at a given site, the incomplete effect of some modifications on the base change or RT arrest, the dynamic changes of modifications (erasure and writing cycles), possible transcript destabilizing modifications, and placement of the RT pausing or blocking modifications at the very 3' end of the transcript preventing the primer extension and detection of the modified RNA. All these possibilities should be considered in any discovery or analytical approaches to avoid errors in mapping of the modified nucleotides.

Pretreatment coupled with high-throughput sequencing

Pretreatment of RNA with chemical or enzymatic mixtures prior to cDNA synthesis followed by sequencing can expand the applications of RNA-Seq for modification detection.[65, 66]

The principle is the differential reactivity of the specific modification under study with a chemical compound or in a special enzymatic reaction. The goal could be either of (1) changing the base pairing properties of the nucleotides, (2) cleaving the RNA at or next to the modified nucleotides, or (3) converting the modified nucleotides into bulky structures to block the RT enzymes during cDNA synthesis. Single nucleotide variations can be deduced directly by inspecting the mapped sequencing reads at base pair resolution, and sites of cleavage or RT arrest can be found by observation of frequent split or truncation of the reads. For either of these experiments RNA samples extracted from knock-down or –out cells lacking the corresponding RNA modifier could be used in parallel to filter for the possible false positives.

Detection of m^5C in RNA by bisulfite sequencing is the most well known example of pretreatment coupled with sequencing which has been applied in both of small and large (high-throughput) scales.[17, 79] In this procedure all the Cs but not m^5Cs in RNA get converted to Us which after sequencing shows the sites of Cs as Ts and sites of m^5Cs as maintained Cs.

Recently chemical oxidation of 5-hydroxymethylcytosine (hm^5C) to 5-formylcytosine (f^5C) prior to bisulfite sequencing has been applied to identify the locations of hm^5C sites in DNA. f^5C unlike m^5C and hm^5C undergoes deamination and therefore gets converted to U (T after sequencing) marking the precise location of hm^5C in the genome.[80] In another report enzymatic oxidation of hm^5C to 5-carboxylcytosine (ca^5C) by TET proteins was used for a similar purpose as again bisulfite treatment efficiently deaminates the ca^5Cs.[81] As hm^5C has been detected in RNA similar strategies can be used to differentiate it from m^5C and find the exact locations of hydroxymethylated cytosines in bisulfite treated RNA.[82]

The other example of similar strategies is the direct detection of A>I editing sites by cyanoethylation of inosine. While knockdown of adenosine deaminase enzymes as the standard control for detection of true editing sites can introduce some unwanted transcriptome-wide

changes, cyanoethylation of inosine on the wild-type RNA samples can show the editing sites by blocking the primer extension over the cyanoethylated inosines resulting in loss of edited RNAs during library preparation and manifested by lack of A>G changes in those sites exclusively in the treated but not untreated samples.[83]

Some reagents are known to selectively depurinate specific modified nucleotides.[64] If the integrity of RNA remains intact during the procedure the depurinated sites can become possibly identified by cDNA sequencing indicating the sites of sequencing errors (modifications sites).

Detection of specific modifications through induction of selective RNA cleavage or RT arrest, upon special chemical pretreatment, has been used for detection of m^7G or Ψ, respectively, by small scale Sanger sequencing. The RNA cleavage produces shorter fragments that can be used separately for library preparation, and the RT arrest produces truncated cDNAs.[65, 66] Thus, the outputs of both of these procedures can be directly sequenced along with proper controls using high-throughput sequencing approaches to find the locations of RNA cleavage or RT arrest as the possible sites of RNA modifications.

Overall, a comprehensive thorough survey seems to be required to explore the differential reactivity of all modified nucleotides, with different chemical reagents or within specific enzymatic reactions, in order to develop specific pretreatment procedures facilitating the detection and mapping of modified nucleotides within the RNA molecules.

Enrichment coupled with high-throughput sequencing

Beside the highly abundant tRNAs and rRNAs, other ncRNA species and mRNAs are expressed at moderate to very low levels. Some RNA species are so scarce that their abundance

is considered generally less than one transcript per cell, meaning that in several cells only one copy may be found[44]. In addition the penetrance of modification at a particular site as well as the turnover rate or dynamic manifestation of a given modification, as the result of cyclic writing/erasing processes, can bring the level of modified RNA species below the detectable threshold. Finally the combination of these two scarcities (low-copy RNAs and low-penetration modifications) will make it even worse to capture and study such modified RNAs. The less abundant RNA species including the very rare ones can be considered equally, if not more, interesting in comparison to their abundant tRNA and rRNA counterparts. Also the low-level modification events may contribute in marking a subset of transcripts for specialized functions and thus it could be interesting to isolate and characterize them. All these indicate the importance of development and application of enrichment techniques for detection of rare modification events. Enrichment can be achieved through depletion of the abundant tRNAs and rRNAs from the total RNAs. However, even with current depletion protocols a clean RNA preparation is hard to obtain. Alternatively the modified RNAs themselves can be the subjects of enrichment. Here we introduce, and give examples for, different applied or novel strategies of enriching the modified RNAs, including techniques based on (A) differential chemical reactivity of modified RNAs, (B) affinity of antibodies or specific binding proteins for modifications, or (C) reaction mechanisms of the corresponding RNA modifier enzymes.

Differential chemical reactivity-based enrichment

Each modified nucleotide is a distinct structure with specific chemical properties. Thus it may be possible to use special reagents to specifically enrich the modified RNAs. piRNAs and miRNAs in some species contain 3’ terminal 2’-O-methylated nucleotides. Taking advantage of

the methyl group at the 2’ position of the terminal sugar, to enrich for the methylated piRNAs, researchers have developed a smart protocol to make cDNA selectively from the methylated RNAs. In this protocol treatment of small RNAs with NaIO4 followed by β-elimination results in the removal of the 3’ terminal nucleoside leaving a 3’ phosphate. The terminally 2’-O-methylated nucleotides are refractory to this reaction and remain intact. During library preparation for sequencing, the 3’ phosphate at the end of small RNAs blocks the 3’ adapter ligation resulting in the loss of corresponding cDNAs in the library and enrichment of cDNAs made exclusively from the methylated RNAs.[84] With a similar concept the m^7G capped RNAs can be enriched by treating the total RNA with calf intestinal alkaline phosphatase (CIP) to remove the phosphate group from all un-capped RNAs and leave a 5’-OH followed by tobacco acid pyrophosphatase (TAP) treatment to break the cap structure and leave the capped RNAs with a clonable 5’ monophosphate. The 5’ adapter can then be ligated exclusively to the 5’ monophosphate RNAs to reveal the profile of capped RNAs in the cells.[5]

Differential chemical reactivity has also been used for enrichment of DNA fragments harboring hm5C through enzymatic glycosylation of hm^5Cs followed by biotinylation, capture by streptavidin beads, release and then sequencing. The exact same protocol might be useful to detect hm^5Cs in RNAs, and similar strategies can be developed to capture the RNA molecules containing other modified nucleotides.[85]

Affinity-based enrichment

Most modified nucleotides are distinct enough and immunogenic to raise specific antibodies against them and some may have natural protein binders recognizing them in the cells. Anti-m^6A antibody has recently been used to explore the transcriptome wide distribution

of this modification in RNAs.[14, 15] Anti-m^7G cap structure antibodies have also been used for enrichment of capped RNAs.[5] Alternatively, the cap binding protein eIF4E can be used to specifically enrich the m^7G capped RNAs.[86]

Additionally RNA targets of modifier enzymes can be captured through UV or chemical reagent enhanced cross-linking of the enzymes to their RNA substrates followed by enrichment of the complex via enzyme specific antibodies. The reverse cross-linked RNAs can then become recovered for cDNA library preparation and high-throughput sequencing to identify the target RNAs. Development of antibodies against the modifications or the modifier enzymes or discovery and application of natural protein binders seems to be a more available strategy for a majority of RNA modifications/modifiers.

Mechanism-based enrichment

As explained above capturing the substrates of RNA modifying enzymes is one way of studying the modified RNAs, which can be achieved through cross-linking of the target RNAs to the enzymes via UV irradiation or use of different chemical cross-linkers. This approach is an "interaction-based" enrichment strategy that may suffer from both false positives and negatives. A given modifier enzyme in order to reach its target molecule may encounter some nontarget RNAs and once finding the right target may inspect many nucleotides to find and modify the exact target nucleotide. An affinity-based technique potentially not only enriches the true substrates but may also enrich some of the nonsubstrate interacting bound RNAs. On the other hand the on/off rate of most enzymatic reactions is very fast decreasing the chance of capturing the true RNA substrate-enzyme complexes, which can be the source of false negatives. It is therefore useful to apply "reaction (mechanism)-based" enrichment approaches

to capture the RNA substrate-enzyme complexes only when the real modification reaction is involved. This could potentially generate higher resolution and cleaner background.

Recently we have reported a mechanism-based enrichment method called Aza-IP to capture the direct in-vivo targets of RNA cytosine methyltransferases (m^5C-RMTs).[87] Aza-IP takes the advantage of the reaction mechanism of cytosine methyltransferases involving a covalent connection during the methylation process. Random transcription-dependent incorporation of nucleotide analogues such as 5-aza-cytidine (5-aza-C) at the exact target site of the m^5C-RMTs within a given target RNA stops the methylation cycle in the middle and results in formation of a stable RNA-RMT adduct intermediate, connected through the covalent linkage between the active site of the enzyme and 5-aza-C. This RNA-RMT adduct can be enriched via application of specific antibodies against the enzyme. The covalent linkage is stable enough for efficient immunoprecipitation, but is also reversible by application of heat and/or chemical reagents to release the RNA for cDNA synthesis and sequencing. Surprisingly, the release step induced a C>G transversion signature at the exact target sites of known m^5C-RMT target sites within the RNA molecules enabling the direct detection of the precise target sites within the target RNA substrates. This approach helped to identify all known tRNA targets of human DNMT2 and NSUN2 and also revealed many novel NSUN2 target sites within tRNAs and several ncRNAs. Similar strategies under a general term of "Adduct-IP" have also been suggested for identification of the direct targets of some other RNA modifiers.[87] In principle if a covalent linkage is involved in a given modification reaction and upon availability of appropriate nucleotide analogues Adduct-IP can be applied to find the targets. We propose the application of Adduct-IP for target identification of at least some other RNA modifiers such as pseudouridine synthases and uracil methyltransferases. The 5-flurouracil has been shown to block these enzymes and produce RNA-enzyme adducts through a covalent connection.[88]

Adduct-IP may also be done without using the nucleotide analogues in special cases. For example the NSUN family of methyltransferases utilizes two cysteine residues for the catalysis, one for forming the covalent connection with the target cytosine base and the other for releasing the RNA after the completion of methylation.[16, 89] Experiments have shown that mutation of the second cysteine to another amino acid will interfere with the release step and results in the formation of an RNA-RMT adduct.[90, 91] This RNA-RMT adduct can become efficiently enriched via immunoprecipitation and the RNA can then be released and sequenced. Interestingly, because the target cytosines of the released RNAs are methylated, bisulfite sequencing of the enriched RNAs might be helpful for identification of the RMT target sites. Alternatively, the target site can be defined by degrading the RMT enzyme of the adduct by proteinase K treatment leaving few amino acids attached to the RNA at the exact target site. Upon cDNA synthesis the protein left-over will block the RT enzyme revealing the exact target sites.[91] Overall such release mutant RNA modifiers can be over-expressed in the cells for isolation of the direct targets of some modifier enzymes. Even for those modifier enzymes in which the covalent intermediate is not involved or the release mutant cannot be efficiently made any mutation that decreases the off rate, but does not affect the target specificity of the enzyme, can help to establish the Adduct-IP procedure. In such cases chemical cross-linking can also come into play to boost the stability of the RNA-RMT adduct during immunoprecipitation. Overall, it would be interesting to explore the catalytic mechanism of all RNA modifier enzymes in order to construct suicide nucleotide analogue inhibitors, or to find the appropriate mutations that may decrease the off rate of the enzymes suitable for application of the Adduct-IP type of approaches for their target profiling.

Modification detection by advanced sequencing technologies

Current high-throughput sequencing approaches rely on utilizing the DNA/cDNA libraries made from small (fragmented) pieces of nucleic acid polymers, which have been copied many times from the original templates. The smaller sizes often cause problems in accurate mapping of the sequenced reads, and amplification cycles quite often introduce biases and also unwanted single nucleotide errors. Especially, for the case of modification profiling in both the DNA and RNA current approaches and strategies allow detection of only a single type of modification at a time, and when looking for multiple modifications in the same molecules multiple sequencing experiments should be performed. Finally there are still no reliable approaches or tricks for differentiation and detection of many modifications with current sequencing technologies.

To circumvent the current size limit of the sequencable molecules and to bypass the amplification cycles during library preparations, new single molecule sequencing technologies are emerging (Figure 4.5). Notably, these technologies can readily differentiate the modified from standard nucleotides, and in principle are able to detect multiple modifications at a time, in both DNA and RNA.[2] In addition, by such technologies, prior cDNA synthesis is no longer required as RNA can become directly sequenced. Here we introduce three different platforms; each utilizes a substantially unique principle for single-molecule real-time sequencing. Although these technologies have not yet been fully implemented for RNA sequencing some successful reports of RNA sequencing as well as efficient nucleotide modification profiling in DNA indicate their potentials to be applied more widely for both RNA sequencing and modification profiling. Here we briefly discuss these new technologies:

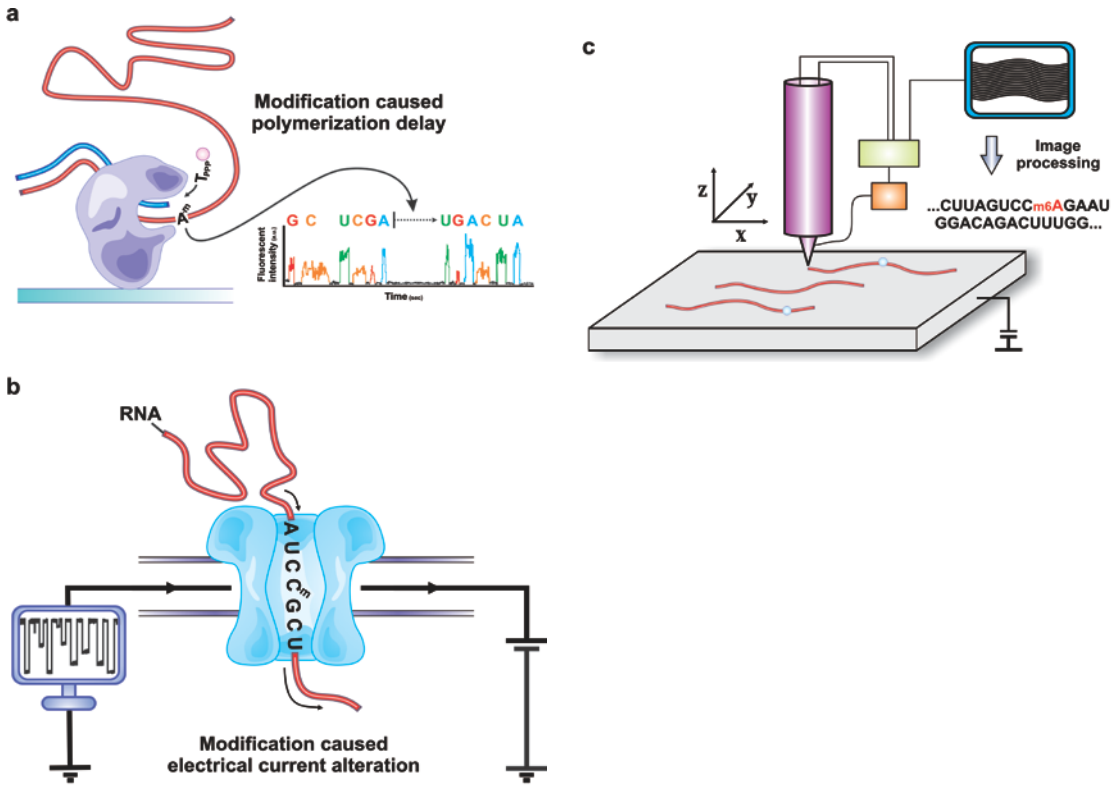


Figure 4.5 | Advanced single-molecule RNA sequencing approaches. a, Single-molecule, real-time (SMRT®) RNA sequencing. Schematic shows a reverse transcriptase enzyme fixed on a surface at the bottom of a nanochamber. The RNA molecule (red) is used as the template for real-time sequencing through cDNA synthesis. Special phospholinked fluorescent standard nucleotides (A, C, G and T), labeled with different fluorescent molecules at their terminal phosphates, are present in the mixture. The cDNA synthesis is catalyzed continuously without interruption. Each time a nucleotide is added to the growing cDNA chain a distinct fluorescent pulse is generated and recorded. The pulse width (PW; the nucleotide retention time in the polymerase active site) and the interpulse duration (IPD; the time interval between the nucleotide-bond states) generated as the result of addition of the complementary nucleotide in front of the modified nucleotide (within the RNA) will be distinct enough to make a unique signature for detection of individual modifications. **b**, Nanopore RNA sequencing. The RNA molecule can become driven through a nanopore allowing the passage of individual nucleotides through the channel one-by-one. Each individual nucleotide (standard or modified) will specifically change the flow of ions in the channel resulting in generation of a unique change in the electrical current to be used as a unique identifier of that standard or modified nucleotide. **c**, Microscopy-based RNA sequencing. RNA molecules can become deposited individually on special surfaces, like single-layer graphene, for taking their image by transmission electron microscopy (TEM) or scanning tunneling microscopy (STM). The image processing can reveal the sequence of the RNA molecules and possibly find the location and nature of the modified nucleotides. A schematic example using the STM technology is given here.

1. Single-molecule, real-time (SMRT®) DNA sequencing uses a DNA polymerase fixed at a bottom of a nano-chamber, and a single stranded DNA template for real-time DNA polymerization and sequencing. In this procedure special phospholinked fluorescent standard nucleotides, labeled with different fluorescent molecules at their terminal phosphates, are sequentially added to the growing DNA strand without interruption for washing. Upon addition of each nucleotide, a specific fluorescent pulse is generated and recorded showing the type of the added nucleotide. During this continuous polymerization, when facing a modification in the template, the pulse width (PW; the nucleotide retention time in the polymerase active site) and the interpulse duration (IPD; the time interval between the nucleotide-bond states) will be distinct enough to make a unique signature for each individual modification.[92] SMRT® DNA sequencing has been successfully applied for simultaneous detection of multiple natural DNA modifications such as m^5C, hm^5C, m^6A and m^4C as well as characterization of many DNA damage induced modifications such as 8-oxo-G, 8-oxo-A, O6-methyl-G, etc.[2, 93-96] It has also recently been applied for direct m6A profiling in RNA through replacement of the DNA polymerase with reverse transcriptase in the platform.[97]

2. Nanopore single-molecule DNA sequencing is another promising sequencing approach potentially capable of detecting modifications in DNA and RNA.[2] In contrast to all previously applied sequencing approaches this approach does not require any polymerization reactions. Instead the nucleotides of a single-stranded DNA molecule are driven through a nanoscale pore one-by-one and the changes in the electrical current, caused by passage of individual nucleotides and alteration of ion current within the channel, are recorded to determine the nucleotide type.[98] Theoretically any distinct nucleotide structure including the modified ones should produce a unique signal enabling their detection, and in practice nanopore sequencing has been successfully applied for m^5C and hm^5C profiling in DNA.[99, 100]

Development of nanopore technology for direct protein sequencing[101] as well as for direct RNA sequencing and RNA modification mapping[102] is under progress.

3. Microscopy-based imaging has also been added to the list of single-molecule DNA sequencing technologies.[2] In this approach pictures obtained by transmission electron microscopy (TEM) or scanning tunneling microscopy (STM) from ordered arrays of DNA molecules, deposited individually on special surfaces like single-layer graphene, are analyzed to detect the order of nucleotides in each DNA molecule.[2, 103, 104] Although this strategy still requires full implementation it has been applied, with some modifications, for detection of m^5C in DNA.[105] RNA sequencing and RNA modification profiling with this technology might become available in future.

Outlook

Post-transcriptional RNA modifications, in parallel with postreplicational DNA and post-translational protein modifications provide extra layers of information complexity for the flow of genetic information and are fundamental for survival, adaptation and evolution of living organisms from all three kingdoms of life.

All modified nucleotides, mentioned in this review, along with tens of others that are not discussed, shape the complex RNA “epitranscriptome” of the cells. Many of these modifications are essential for life and loss or gain of function of the responsible modifier enzymes or their accessory components can result in cancer, infertility, metabolic and immune diseases, as well as genetic disorders such as neurological or cognition abnormalities. Understanding the distribution and function of individual modified nucleotides, and their possible interplay and interactions with each other, as well as characterization of the modifier

enzymes and complexes, will provide a valuable level of knowledge for better understanding of life. Elucidating the epitranscriptome differences in different cell types or tissues as well as changes during development and upon stimulation or during stress conditions or immune response can also provide additional layers of knowledge about the impact of dynamic RNA modifications for the cells.

For several decades, due to technical limitations our knowledge about the scope and functions of RNA modifications was restricted to the highly abundant and highly modified tRNAs and rRNAs. Pretty recently, however, results obtained by application of high-throughput sequencing technologies along with discovery of diverse families of novel ncRNAs attracted more attentions to the field, with the aim of discovering possible nucleotide modifications affecting the structure or function of all or a subset of transcripts. Targeted projects by several research groups revealed many new sites of modifications of different types in both of coding and non-coding RNAs proving that RNA modifications are more widespread than it was initially thought to be. Currently the epitranscriptome field is open for both of evidence or hypothesis driven, and small- or large-scale (transcriptome-wide) studies to explore new biological territories and answer to a number of long lasting questions in the field, and several live projects are currently going on in different laboratories worldwide.

To enhance the growth of knowledge in the field, systematic comparison of massive available row RNA sequencing datasets deposited in public databases with the intention of finding the sites of frequent, consistent and/or conserved DNA/RNA discrepancies can help to prepare a list of candidate modified sites within the transcriptome for further verifications. Examining the base pairing behavior of all known modified nucleotides will help to design experiments for their direct mapping within the RNA by parallel sequencing of the RNA samples extracted from wild-type or RNA modifier null cells/organisms.

It is also interesting to expand our knowledge about the differential reactivity of all modified nucleotides, with different chemical reagents or within specific enzymatic reactions for developing specific pretreatment procedures facilitating their detection/enrichment and mapping.

Development and use of antibodies against either of modifications, themselves, or their corresponding modifier enzymes can help for enrichment and more focused analysis of the modified RNAs. Similarly isolation and characterization of possible proteins capable of binding specifically to the modified nucleotides would be helpful. Elucidation of reaction mechanisms of different RNA modifiers at molecular level can help to establish reaction-based enrichment methods for more reliable detection and mapping of the direct substrate RNAs. Finally, application of the emerging real-time single-molecule sequencing technologies can help to increase the spectrum as well as resolution of the epitranscriptome mapping.

Altogether, it is anticipated that in the near future, these diverse types of classical, current and future technologies will help improve our understanding of the cellular epitranscriptome, its logic, dynamics, diverse functions, and overall impact on living organisms.

References

1. Hotchkiss, R.D. The quantitative separation of purines, pyrimidines, and nucleosides by paper chromatography. *J Biol Chem* **175**, 315-332 (1948).

2. Korlach, J. & Turner, S.W. Going beyond five bases in DNA sequencing. *Curr Opin Struct Biol* **22**, 251-261 (2012).

3. Shatkin, A.J. Capping of eukaryotic mRNAs. *Cell* **9**, 645-653 (1976).

4. Ghosh, A. & Lima, C.D. Enzymology of RNA cap synthesis. *Wiley Interdiscip Rev RNA* **1**, 152-172 (2010).

5. Post-transcriptional processing generates a diversity of 5'-modified long and short RNAs. *Nature* **457**, 1028-1032 (2009).

6. Juhling, F. et al. tRNAdb 2009: compilation of tRNA sequences and tRNA genes. *Nucleic Acids Res* **37**, D159-162 (2009).

7. Machnicka, M.A. et al. MODOMICS: a database of RNA modification pathways--2013 update. *Nucleic Acids Res* **41**, D262-267 (2013).

8. Bodi, Z., Button, J.D., Grierson, D. & Fray, R.G. Yeast targets for mRNA methylation. *Nucleic Acids Res* **38**, 5327-5335 (2010).

9. Clancy, M.J., Shambaugh, M.E., Timpte, C.S. & Bokar, J.A. Induction of sporulation in Saccharomyces cerevisiae leads to the formation of N6-methyladenosine in mRNA: a potential mechanism for the activity of the IME4 gene. *Nucleic Acids Res* **30**, 4509-4518 (2002).

10. Zhong, S. et al. MTA is an Arabidopsis messenger RNA adenosine methylase and interacts with a homolog of a sex-specific splicing factor. *Plant Cell* **20**, 1278-1288 (2008).

11. Bodi, Z. et al. Adenosine methylation in Arabidopsis mRNA is associated with the 3' end and reduced levels cause developmental defects. *Front Plant Sci* **3**, 48 (2012).

12. Hongay, C.F. & Orr-Weaver, T.L. Drosophila Inducer of MEiosis 4 (IME4) is required for Notch signaling during oogenesis. *Proc Natl Acad Sci U S A* **108**, 14855-14860 (2011).

13. Horowitz, S., Horowitz, A., Nilsen, T.W., Munns, T.W. & Rottman, F.M. Mapping of N6-methyladenosine residues in bovine prolactin mRNA. *Proc Natl Acad Sci U S A* **81**, 5667-5671 (1984).

14. Dominissini, D. et al. Topology of the human and mouse m6A RNA methylomes revealed by m6A-seq. *Nature* **485**, 201-206 (2012).

15. Meyer, K.D. et al. Comprehensive analysis of mRNA methylation reveals enrichment in 3' UTRs and near stop codons. *Cell* **149**, 1635-1646 (2012).

16. Motorin, Y., Lyko, F. & Helm, M. 5-methylcytosine in RNA: detection, enzymatic formation and biological functions. *Nucleic Acids Res* **38**, 1415-1430 (2010).

17. Squires, J.E. et al. Widespread occurrence of 5-methylcytosine in human coding and non-coding RNA. *Nucleic Acids Res* **40**, 5023-5033 (2012).

18. Rai, K. et al. Dnmt2 functions in the cytoplasm to promote liver, brain, and retina development in zebrafish. *Genes Dev* **21**, 261-266 (2007).

19. Martinez, F.J. et al. Whole exome sequencing identifies a splicing mutation in NSUN2 as a cause of a Dubowitz-like syndrome. *J Med Genet* **49**, 380-385 (2012).

20. Khan, M.A. et al. Mutation in NSUN2, which encodes an RNA methyltransferase, causes autosomal-recessive intellectual disability. *Am J Hum Genet* **90**, 856-863 (2012).

21. Abbasi-Moheb, L. et al. Mutations in NSUN2 cause autosomal-recessive intellectual disability. *Am J Hum Genet* **90**, 847-855 (2012).

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

23. Blanco, S. et al. The RNA-methyltransferase Misu (NSun2) poises epidermal stem cells to differentiate. *PLoS Genet* **7**, e1002403 (2011).

24. Frye, M. & Watt, F.M. The RNA methyltransferase Misu (NSun2) mediates Myc-induced proliferation and is upregulated in tumors. *Curr Biol* **16**, 971-981 (2006).

25. Frye, M. et al. Genomic gain of 5p15 leads to over-expression of Misu (NSUN2) in breast cancer. *Cancer Lett* **289**, 71-80 (2010).

26. Okamoto, M. et al. Frequent increased gene copy number and high protein expression of tRNA (cytosine-5-)-methyltransferase (NSUN2) in human cancers. *DNA Cell Biol* **31**, 660-671 (2012).

27. Chan, C.T. et al. Reprogramming of tRNA modifications controls the oxidative stress response by codon-biased translation of proteins. *Nat Commun* **3**, 937 (2012).

28. Becker, M. et al. Pmt1, a Dnmt2 homolog in Schizosaccharomyces pombe, mediates tRNA methylation in response to nutrient signaling. *Nucleic Acids Res* **40**, 11648-11658 (2012).

29. Durdevic, Z. et al. Efficient RNA virus control in Drosophila requires the RNA methyltransferase Dnmt2. *EMBO Rep* **14**, 269-275 (2013).

30. Sato, S., Wong, S.K. & Lazinski, D.W. Hepatitis delta virus minimal substrates competent for editing by ADAR1 and ADAR2. *J Virol* **75**, 8547-8555 (2001).

31. Chen, S.H. et al. Apolipoprotein B-48 is the product of a messenger RNA with an organ-specific in-frame stop codon. *Science* **238**, 363-366 (1987).

32. Powell, L.M. et al. A novel form of tissue-specific RNA processing produces apolipoprotein-B48 in intestine. *Cell* **50**, 831-840 (1987).

33. Skuse, G.R., Cappione, A.J., Sowden, M., Metheny, L.J. & Smith, H.C. The neurofibromatosis type I messenger RNA undergoes base-modification RNA editing. *Nucleic Acids Res* **24**, 478-485 (1996).

34. Mukhopadhyay, D. et al. C-->U editing of neurofibromatosis 1 mRNA occurs in tumors that express both the type II transcript and apobec-1, the catalytic subunit of the apolipoprotein B mRNA-editing enzyme. *Am J Hum Genet* **70**, 38-50 (2002).

35. Rosenberg, B.R., Hamilton, C.E., Mwangi, M.M., Dewell, S. & Papavasiliou, F.N. Transcriptome-wide sequencing reveals numerous APOBEC1 mRNA-editing targets in transcript 3' UTRs. *Nat Struct Mol Biol* **18**, 230-236 (2011).

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

37. Ofengand, J. Ribosomal RNA pseudouridines and pseudouridine synthases. *FEBS Lett* **514**, 17-25 (2002).

38. Matera, A.G., Terns, R.M. & Terns, M.P. Non-coding RNAs: lessons from the small nuclear and small nucleolar RNAs. *Nat Rev Mol Cell Biol* **8**, 209-220 (2007).

39. King, T.H., Liu, B., McCully, R.R. & Fournier, M.J. Ribosome structure and activity are altered in cells lacking snoRNPs that form pseudouridines in the peptidyl transferase center. *Mol Cell* **11**, 425-435 (2003).

40. Motorin, Y. & Helm, M. tRNA stabilization by modified nucleotides. *Biochemistry* **49**, 4934-4944 (2010).

41. Charette, M. & Gray, M.W. Pseudouridine in RNA: what, where, how, and why. *IUBMB Life* **49**, 341-351 (2000).

42. Newby, M.I. & Greenbaum, N.L. Sculpting of the spliceosomal branch site recognition motif by a conserved pseudouridine. *Nat Struct Biol* **9**, 958-965 (2002).

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

44. Djebali, S. et al. Landscape of transcription in human cells. *Nature* **489**, 101-108 (2012).

45. Ferre-D'Amare, A.R. Protein synthesis: Stop the nonsense. *Nature* **474**, 289-290 (2011).

46. Saikia, M., Fu, Y., Pavon-Eternod, M., He, C. & Pan, T. Genome-wide analysis of N1-methyl-adenosine modification in human tRNAs. *RNA* **16**, 1317-1327 (2010).

47. Basavappa, R. & Sigler, P.B. The 3 A crystal structure of yeast initiator tRNA: functional implications in initiator/elongator discrimination. *EMBO J* **10**, 3105-3111 (1991).

48. Anderson, J., Phan, L. & Hinnebusch, A.G. The Gcd10p/Gcd14p complex is the essential two-subunit tRNA(1-methyladenosine) methyltransferase of Saccharomyces cerevisiae. *Proc Natl Acad Sci U S A* **97**, 5173-5178 (2000).

49. Voigts-Hoffmann, F. et al. A methyl group controls conformational equilibrium in human mitochondrial tRNA(Lys). *J Am Chem Soc* **129**, 13382-13383 (2007).

50. Yi, C. & Pan, T. Cellular dynamics of RNA modification. *Acc Chem Res* **44**, 1380-1388 (2011).

51. Kramer, G.F., Baker, J.C. & Ames, B.N. Near-UV stress in Salmonella typhimurium: 4-thiouridine in tRNA, ppGpp, and ApppGpp as components of an adaptive response. *J Bacteriol* **170**, 2344-2351 (1988).

52. Daffis, S. et al. 2'-O methylation of the viral mRNA cap evades host restriction by IFIT family members. *Nature* **468**, 452-456 (2010).

53. Zust, R. et al. Ribose 2'-O-methylation provides a molecular signature for the distinction of self and non-self mRNA dependent on the RNA sensor Mda5. *Nat Immunol* **12**, 137-143 (2011).

54. Dong, H. et al. 2'-O methylation of internal adenosine by flavivirus NS5 methyltransferase. *PLoS Pathog* **8**, e1002642 (2012).

55. Ghildiyal, M. & Zamore, P.D. Small silencing RNAs: an expanding universe. *Nat Rev Genet* **10**, 94-108 (2009).

56. Yu, B. et al. Methylation as a crucial step in plant microRNA biogenesis. *Science* **307**, 932-935 (2005).

57. Shumyatsky, G., Wright, D. & Reddy, R. Methylphosphate cap structure increases the stability of 7SK, B2 and U6 small RNAs in Xenopus oocytes. *Nucleic Acids Res* **21**, 4756-4761 (1993).

58. Xhemalce, B., Robson, S.C. & Kouzarides, T. Human RNA methyltransferase BCDIN3D regulates microRNA processing. *Cell* **151**, 278-288 (2012).

59. Chan, C.T. et al. A quantitative systems approach reveals dynamic control of tRNA modifications during cellular stress. *PLoS Genet* **6**, e1001247 (2010).

60. Wilkinson, C.R., Bartlett, R., Nurse, P. & Bird, A.P. The fission yeast gene pmt1+ encodes a DNA methyltransferase homologue. *Nucleic Acids Res* **23**, 203-210 (1995).

61. Wu, G., Xiao, M., Yang, C. & Yu, Y.T. U2 snRNA is inducibly pseudouridylated at novel sites by Pus7p and snR81 RNP. *EMBO J* **30**, 79-89 (2011).

62. Jia, G. et al. N6-methyladenosine in nuclear RNA is a major substrate of the obesity-associated FTO. *Nat Chem Biol* **7**, 885-887 (2011).

63. Jia, G., Fu, Y. & He, C. Reversible RNA adenosine methylation in biological regulation. *Trends Genet* **29**, 108-115 (2013).

64. Kellner, S., Burhenne, J. & Helm, M. Detection of RNA modifications. *RNA Biol* **7**, 237-247 (2010).

65. Motorin, Y., Muller, S., Behm-Ansmant, I. & Branlant, C. Identification of modified residues in RNAs by reverse transcription-based methods. *Methods Enzymol* **425**, 21-53 (2007).

66. Behm-Ansmant, I., Helm, M. & Motorin, Y. Use of specific chemical reagents for detection of modified nucleotides in RNA. *J Nucleic Acids* **2011**, 408053 (2011).

67. Hiley, S.L. et al. Detection and discovery of RNA modifications using microarrays. *Nucleic Acids Res* **33**, e2 (2005).

68. Saikia, M. et al. A systematic, ligation-based approach to study RNA modifications. *RNA* **12**, 2025-2033 (2006).

69. Dai, Q. et al. Identification of recognition residues for ligation-based detection and quantitation of pseudouridine and N6-methyladenosine. *Nucleic Acids Res* **35**, 6322-6329 (2007).

70. Mardis, E.R. A decade's perspective on DNA sequencing technology. *Nature* **470**, 198-203 (2011).

71. Wang, Z., Gerstein, M. & Snyder, M. RNA-Seq: a revolutionary tool for transcriptomics. *Nat Rev Genet* **10**, 57-63 (2009).

72. Iida, K., Jin, H. & Zhu, J.K. Bioinformatics analysis suggests base modifications of tRNAs and miRNAs in Arabidopsis thaliana. *BMC Genomics* **10**, 155 (2009).

73. Ebhardt, H.A. et al. Meta-analysis of small RNA-sequencing errors reveals ubiquitous post-transcriptional RNA modifications. *Nucleic Acids Res* **37**, 2461-2470 (2009).

74. Li, M. et al. Widespread RNA and DNA sequence differences in the human transcriptome. *Science* **333**, 53-58 (2011).

75. Lin, W., Piskol, R., Tan, M.H. & Li, J.B. Comment on "Widespread RNA and DNA sequence differences in the human transcriptome". *Science* **335**, 1302; author reply 1302 (2012).

76. Pickrell, J.K., Gilad, Y. & Pritchard, J.K. Comment on "Widespread RNA and DNA sequence differences in the human transcriptome". *Science* **335**, 1302; author reply 1302 (2012).

77. Kleinman, C.L. & Majewski, J. Comment on "Widespread RNA and DNA sequence differences in the human transcriptome". *Science* **335**, 1302; author reply 1302 (2012).

78. Testa, S.M., Disney, M.D., Turner, D.H. & Kierzek, R. Thermodynamics of RNA-RNA duplexes with 2- or 4-thiouridines: implications for antisense design and targeting a group I intron. *Biochemistry* **38**, 16655-16662 (1999).

79. Schaefer, M., Pollex, T., Hanna, K. & Lyko, F. RNA cytosine methylation analysis by bisulfite sequencing. *Nucleic Acids Res* **37**, e12 (2009).

80. Booth, M.J. et al. Quantitative sequencing of 5-methylcytosine and 5-hydroxymethylcytosine at single-base resolution. *Science* **336**, 934-937 (2012).

81. Yu, M. et al. Base-resolution analysis of 5-hydroxymethylcytosine in the mammalian genome. *Cell* **149**, 1368-1380 (2012).

82. Cantara, W.A. et al. The RNA Modification Database, RNAMDB: 2011 update. *Nucleic Acids Res* **39**, D195-201 (2011).

83. Sakurai, M., Yano, T., Kawabata, H., Ueda, H. & Suzuki, T. Inosine cyanoethylation identifies A-to-I RNA editing sites in the human transcriptome. *Nat Chem Biol* **6**, 733-740 (2010).

84. Seitz, H., Ghildiyal, M. & Zamore, P.D. Argonaute loading improves the 5' precision of both MicroRNAs and their miRNA* strands in flies. *Curr Biol* **18**, 147-151 (2008).

85. Song, C.X., Yi, C. & He, C. Mapping recently identified nucleotide variants in the genome and transcriptome. *Nat Biotechnol* **30**, 1107-1116 (2012).

86. Gowda, M. et al. Genome-wide characterization of methylguanosine-capped and polyadenylated small RNAs in the rice blast fungus Magnaporthe oryzae. *Nucleic Acids Res* **38**, 7558-7569 (2010).

87. Khoddami, V. & Cairns, B.R. Identification of direct targets and modified bases of RNA cytosine methyltransferases. *Nat Biotechnol* (2013).

88. Guelorget, A. & Golinelli-Pimpaneau, B. Mechanism-based strategies for trapping and crystallizing complexes of RNA-modifying enzymes. *Structure* **19**, 282-291 (2011).

89. King, M.Y. & Redman, K.L. RNA methyltransferases utilize two cysteine residues in the formation of 5-methylcytosine. *Biochemistry* **41**, 11218-11225 (2002).

90. Redman, K.L. Assembly of protein-RNA complexes using natural RNA and mutant forms of an RNA cytosine methyltransferase. *Biomacromolecules* **7**, 3321-3326 (2006).

91. Sugimoto, Y. et al. Analysis of CLIP and iCLIP methods for nucleotide-resolution studies of protein-RNA interactions. *Genome Biol* **13**, R67 (2012).

92. Eid, J. et al. Real-time DNA sequencing from single polymerase molecules. *Science* **323**, 133-138 (2009).

93. Flusberg, B.A. et al. Direct detection of DNA methylation during single-molecule, real-time sequencing. *Nat Methods* **7**, 461-465 (2010).

94. Song, C.X. et al. Sensitive and specific single-molecule sequencing of 5-hydroxymethylcytosine. *Nat Methods* **9**, 75-77 (2012).

95. Fang, G. et al. Genome-wide mapping of methylated adenine residues in pathogenic Escherichia coli using single-molecule real-time sequencing. *Nat Biotechnol* **30**, 1232-1239 (2012).

96. Clark, T.A., Spittle, K.E., Turner, S.W. & Korlach, J. Direct detection and sequencing of damaged DNA bases. *Genome Integr* **2**, 10 (2011).

97. Saletore, Y. et al. The birth of the Epitranscriptome: deciphering the function of RNA modifications. *Genome Biol* **13**, 175 (2012).

98. Branton, D. et al. The potential and challenges of nanopore sequencing. *Nat Biotechnol* **26**, 1146-1153 (2008).

99. Clarke, J. et al. Continuous base identification for single-molecule nanopore DNA sequencing. *Nat Nanotechnol* **4**, 265-270 (2009).

100. Li, W.W., Gong, L. & Bayley, H. Single-molecule detection of 5-hydroxymethylcytosine in DNA through chemical modification and nanopore analysis. *Angew Chem Int Ed Engl* **52**, 4350-4355 (2013).

101. Nivala, J., Marks, D.B. & Akeson, M. Unfoldase-mediated protein translocation through an alpha-hemolysin nanopore. *Nat Biotechnol* **31**, 247-250 (2013).

102. Ayub, M. & Bayley, H. Individual RNA base recognition in immobilized oligonucleotides using a protein nanopore. *Nano Lett* **12**, 5637-5643 (2012).

103. Cerf, A., Alava, T., Barton, R.A. & Craighead, H.G. Transfer-printing of single DNA molecule arrays on graphene for high-resolution electron imaging and analysis. *Nano Lett* **11**, 4232-4238 (2011).

104. Tanaka, H. & Kawai, T. Partial sequencing of a single DNA molecule with a scanning tunnelling microscope. *Nat Nanotechnol* **4**, 518-522 (2009).

105. Cerf, A., Cipriany, B.R., Benitez, J.J. & Craighead, H.G. Single DNA molecule patterning for high-throughput epigenetic mapping. *Anal Chem* **83**, 8073-8077 (2011).

CHAPTER 5

CONCLUSIONS AND FUTURE DIRECTIONS

Overview

In the introductory chapter of this dissertation (Chapter 1), we briefly discussed the emerging field of RNA epitranscriptome, with more emphasis on m^5C modification and m^5C-RMT modifier enzymes. As previewed, in spite of their biological importance, little is known about the scope, targets and processes involving many RNA modifying enzymes, including RNA cytosine methyltransferases, largely due to the absence of techniques to precisely map the modifications within RNA molecules and/or accurately link the modifier enzymes to their precise substrates. This indicates the importance of employing the existing, and developing novel innovative techniques to discover the RNA epitranscriptome as a whole, with high resolution and accuracy.

In the second chapter we introduced an improved version of the existing RNA bisulfite sequencing techniques and presented a comprehensive high-resolution RNA cytosine methylome of mouse embryonic fibroblasts (MEFs) in wild type and Dnmt2 null mice. In that chapter we also discussed caveats of RNA bisulfite sequencing techniques and their possible false positives and negatives to be considered in experimental designs and data interpretation steps, indicating the need for development of complementary robust techniques to increase both the resolution and accuracy of the m^5C site calls.

In keeping, in the third chapter, we introduced a substantially innovative technique; Aza-IP, for identification of the direct targets and modified bases of m^5C-RMTs in a single experiment and with high accuracy.[1] We presented the obtained results with the Aza-IP technique for m^5C-RMTs from two different families, representing both 'single cysteine' (DNMT2) and 'two-cysteine' (NSUN2) type enzymes. Aza-IP revealed all of the known target tRNAs and precise target sites of DNMT2 and NSUN2, as well as hundreds of novel and previously unappreciated NSUN2 target sites within those tRNAs. We also reported several

known and novel ncRNA targets of NSUN2 with the precise targets, validated, for selected target sites, with conventional RNA bisulfite sequencing.

In Chapter 4, we presented an overview of all classical, recently developed, and currently available techniques, as well as potential technical strategies to be applied in future for RNA epitranscriptome profiling. In that overview we discussed how strategies similar to both of RNA bisulfite sequencing and Aza-IP can become adapted to explore the distribution and functions of other modifications beside m^5C in diverse RNA species.

Here in this chapter we present the future directions of this work to address fundamental questions of the field regarding m^5C and m^5C-RMTs.

Future directions

Isolating the direct RNA substrates of all human m^5C-RMTs

About ten different human m^5C-RNA methyltransferases (m^5C-RMTs) have been identified so far; NSUN1, 2, 3, 4, 5 (A, B & C), 6 and 7, and DNMT2.[2] Beside NSUN2 and DNMT2 all the other human m^5C-RMTs are remained largely uncharacterized. Here we aim to provide a comprehensive catalogue of the target RNA molecules and the target cytosine bases of all human m^5C-RMTs in HeLa cells, and also provide tissue-specific information about the target profiles of the human m^5C-RMT homologues in mice and/or zebrafish. This will provide a critical foundation for the entire field for future studies on this important class of enzymes. Based on these results we also aim to propose and test some candidate functional targets for m^5C-RMTs and RNA cytosine methylation in our models.

Target profiling of all other human m^5C-RMTs by Aza-IP

Our success with the Aza-IP technique using m^5C-RMTs from two different families, representing both 'single cysteine' (DNMT2) and 'two-cysteine' (NSUN2) type enzymes, strongly suggests that Aza-IP may serve as a general technique for enrichment of direct targets of any cytosine methyltransferase that utilizes a similar catalytic mechanism. Here we will apply Aza-IP for target profiling of other human m^5C-RMTs. We will explore the known and candidate NSUN family methyltransferases NSUN1, 3, 4, 5(A,B,C), 6 and 7. Similar to our previous procedure with DNMT2 and NSUN2, V5-tagged full-length enzymes will be expressed in HeLa cells using lentiviral expression systems, with application of the same 5-aza-C concentration and exposure time. As before, we will perform two biological replicates and an IgG control. The USeq[3] and VarScan[4] packages will then be used to find the enriched RNAs and the exact target cytosines, respectively. This work will provide a comprehensive catalogue of the target sites of m^5C-RMTs in HeLa cells.

Adduct-IP for tissue-specific RNA methylation analysis

A clear and important issue in RNA methylation analysis is to determine the repertoire of targets for m^5C-RMTs in different cell types and contexts. In principle, Aza-IP should be extendable to models systems, such as the mouse, which harbors orthologs of all human m^5C-RMTs. Here, one could administer 5-aza-C to mice, perform Aza-IP with a polyclonal antibody created/verified for IP, and follow the prior regimen to reveal targets. Here, pilot experiments would be performed involving various doses of 5-aza-C, dissection of various tissues of interest, the use of mass spectrometry to quantify the 5-aza-C:cytosine ratio in those tissues, and pilot Aza-IP experiments performed in a multiplexing format. It seems reasonable to initially pilot the

technique with DNMT2 or NSUN2 (as targets are already known) to test and calibrate the system before moving on to m^5C-RMTs with unknown targets. We consider this approach reasonable and feasible and likely to work given sufficient effort. However we have elected to consider this line of experimentation as an 'alternative approach' rather than the main approach to determine tissue-specific targets, and provide two reasons. The first involves a slight concern regarding the likely extent of 5-aza-C incorporation into RNA in a mouse model. During transcription 5-aza-C is randomly (and nonpreferentially) incorporated in place of cytosine. Notably, our Aza-IP regimen involves the growth of HeLa cells in media with relatively low levels of 5-aza-C (3-5μM to avoid toxicity). Although the precise 5-aza-C:cytosine ratio in HeLa RNAs isolated from this regimen has not been determined, it is almost certainly very low, and this ratio sets an upper limit on the proportion of target RNAs that can be covalently linked to the tested m^5C-RMT. This is coupled to concerns about bioavailability and stability of the compound in certain tissues. Together, although these issues are surmountable, optimization will likely require a lengthy pilot phase and large numbers of mice to provide sufficient material. The second reason for relegating Aza-IP to an 'alternative approach' is an interest in pursuing an alternative (below).

To circumvent these limitations we propose to take advantage of an intrinsic feature of the catalysis mechanism of 'two-cysteine' m^5C-RMTs to test another novel technique: Adduct-IP. Nine out of the ten known human m^5C-RMT are from the NSUN methyltransferase family, which utilize two cysteines in their catalytic domains. As explained in previous chapters, to conduct methylation, a particular cysteine (from motif VI), attacks the base and forms the covalent RMT-RNA intermediate, driving the methylation of the C5 position of cytosine. For completion of the methylation cycle and enzyme release, the other cysteine (from motif IV) triggers the reversal of the covalent connection and release of the base from the enzyme.[2, 5] Published work (both in-

vitro and in-vivo) has shown that mutation of the motif IV cysteine will result in formation of stable covalently connected RMT-RNA adducts.[6, 7] Here we aim to take the advantage of this characteristic feature in developing the novel "Adduct-IP" technique for isolation of the direct targets of m^5C-RMTs in tissues of model organisms (Figure 5.1).

Adduct-IP in mice

To perform Adduct-IP, we will generate transgenic mice expressing an epitope-tagged version of cysteine-mutant m^5C-RMTs under the control of Tet-On system.[8] Rapid expression of the mutant enzymes will be triggered by doxycycline treatment, and the animals will be sacrificed at the proper time (determined by pilot experiments involving NSUN2, see below, but likely after 1-3 days). We will pilot Adduct-IP with NSUN2, as it has been studied extensively in mammals, with gain of function linked to cancer,[9-11] and loss of function linked to infertility,[12] intellectual disability and mental retardation.[13-15] Given these links, it seems obvious to include the brain and testes in the initial panel of tissues, and if interesting new RNAs emerge, it can motivate the sorting of cells from these tissues to find cell-type specific RNA targets. The dissected tissues will be subjected to IP with proper replicates and controls (again, calibrated initially by NSUN2). The RNA will be released from the enzyme using the fragmentation protocol (as with Aza-IP) and the RNA will be subjected to sequencing. Here, we expect to capture virtually all of the tissue-specific direct targets of specific m^5C-RMTs. However, because we do not use 5-aza-C, further analysis and experiments are required to determine the exact target cytosines within the enriched RNAs (beginning with focused/directed RNA bisulfite sequencing on enriched targets, likely in a multiplexed MiSeq format). We anticipate Adduct-IP to provide valuable in-vivo insights about the target specificity of m^5C-RMTs in different tissues.

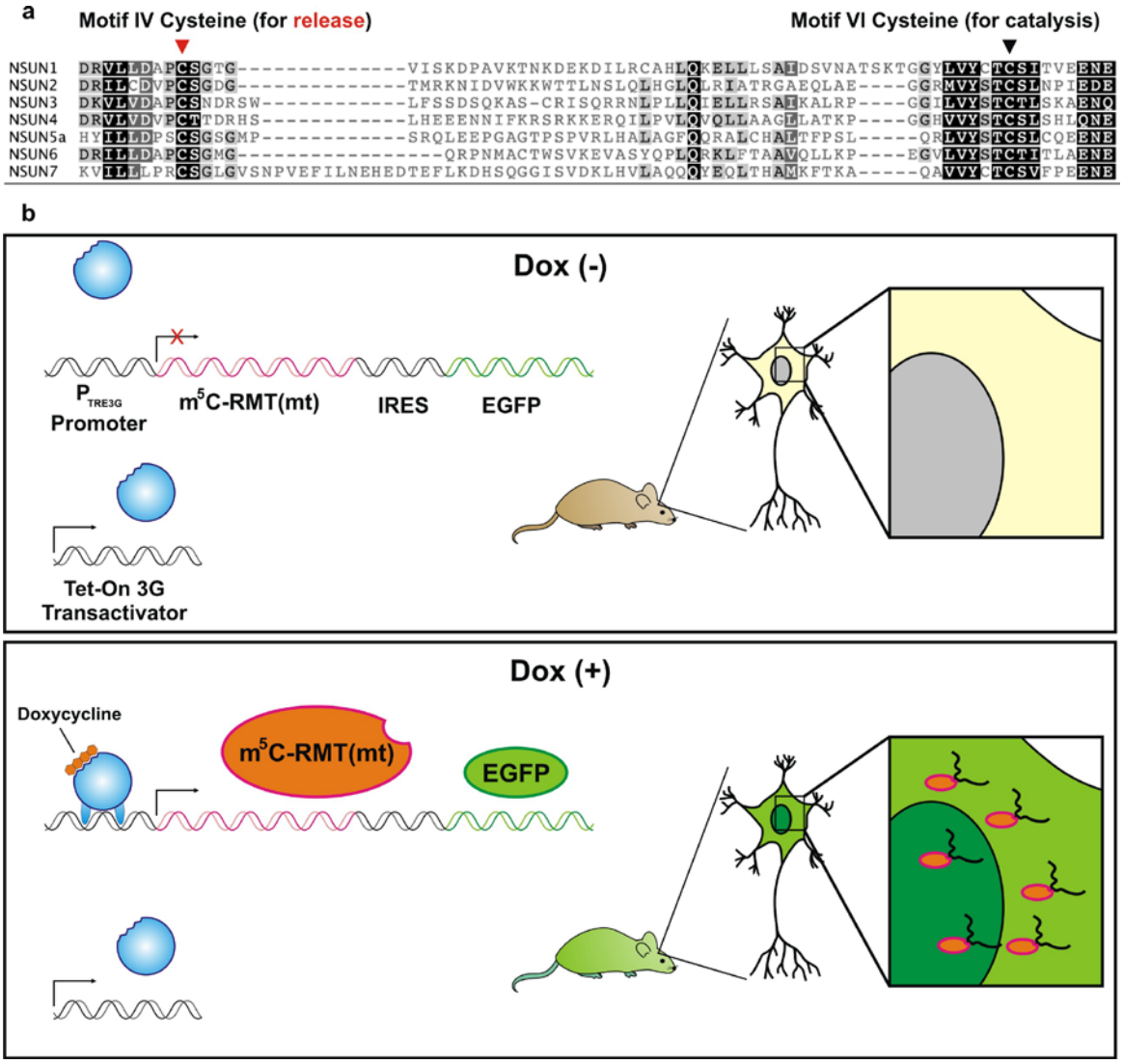


Figure 5.1 | Principle of the proposed Adduct-IP technique. a, Protein alignment of the motifs IV and VI of the human NSUN family (NSUN1-7) shows the conservation of the cysteine residues for catalysis (from motif VI), and release of the enzyme from the base after completion of the methylation (from motif IV). **b**, Adduct-IP experimental design, enabling the regulated expression of the release mutant enzyme (m^5C-RMT(mt)) under the control of the Tet-ON system, and covalent attachment of the enzyme (orange) to the RNA (black wavy line, bottom right).

Beyond NSUN2, prioritization of additional m^5C-RMTs and the target tissues chosen for testing by Adduct-IP will be determined by two main criteria: phenotypic information and expression patterns. Phenotypic information will involve gain or loss of function phenotypes both from published work and from results of this project, and the tissue-specific expression of m^5C-RMTs will be determined by qRT-PCR and/or western blot analysis. In addition we anticipate that the GO (Gene Ontology) term analysis of the target RNAs isolated by Aza-IP in HeLa cells, may provide preliminary information about the underlying pathways, functions and cellular components, helpful in rational selection of target tissues for Adduct-IP experiments. Furthermore, as previewed above, if interesting new RNAs emerge from a tissue, it will motivate the sorting of cells from that tissue to possibly reveal cell-type specific RNA targets, which can motivate focused functional experiments.

Adduct-IP in zebrafish

An additional area of interest is RNA cytosine methylation in early embryogenesis and organ development. However, obtaining sufficient cell and tissue types from the mouse will likely be quite difficult. However, we consider zebrafish an excellent alternative. First, zebrafish harbor all of the NSUN family members (except NSUN7) present in mammals. Second, thousands of embryos can be obtained and easily manipulated (injected) in a few hours providing enough material for the downstream experiments. Also zebrafish embryonic and developmental cycles are rapid, with organ development occurring within 72 hours. Importantly, the University of Utah and the Huntsman Cancer Institute are leaders in modeling diseases such as cancer in the zebrafish, and have more than twenty independent laboratories conducting experiments in the zebrafish. The institution has made a major investment in

zebrafish facilities, which can provide the equipment and expertise that will allow us to quickly and efficiently adapt the Adduct-IP technique for capturing the possible targets of m^5C-RMT homologues in zebrafish. To quickly accomplish this, *in-vitro* transcribed mRNAs encoding a V5-tagged version of mutant m^5C-RMTs will be injected into fish embryos and the Adduct-IP protocol will be applied at the desired developmental stage. Again, prioritization of enzymes, tissues and developmental stages will be made based on the insights provided by the results from previous steps. If we obtain interesting results using the injection approach, we can easily make transgenic fish using tissue- or cell-type specific promoters that will allow a more sophisticated application of Adduct-IP.

Studying the roles of cytosine methylation in mRNAs and ncRNAs.

The ultimate goal of this project is to provide insights about the possible functional roles of RNA cytosine methylation in living organisms. In the first phase we intend to generate a comprehensive catalogue of the direct targets of all human m^5C-RMTs in HeLa cells, supplemented with tissue-, cell- and/or developmental stage-specific information about a prioritized subset of human m^5C-RMTs homologues in mice and fish. This will provide a foundation for proposing and evaluating evidence-based hypotheses regarding the functional roles of RNA cytosine methylation and methyltransferases. Based on such results, the generation of knockout mice and morphant fish lacking specific m^5C-RMTs will be prioritized. In addition the nature of the targets - interesting small and long noncoding and/or coding RNA targets - will be examined in functional contexts to determine the impact of methylation on protein interactions, stability and/or activity of these RNAs.

Determining the roles of m^5C-RMTs in pathophysiology

Out of the 10 known human m^5C-RMTs, knockout mice are available for 2 of them (Dnmt2 and Nsun2). Dnmt2-/- mice do not show any obvious phenotype in normal condition[16] although the involvement of DNMT2 homologues in stress response in fruit fly[17] and fission yeast[18] suggests similar function for the enzyme in human and mouse. Additionally Dnmt2 has recently been shown to function in immune response against RNA viruses in fruit fly.[19] Nsun2-/- mice are viable, but both genders show cyclic alopecia in older animals due to interrupted epidermal differentiation programs and males are sterile through an unknown mechanism[12]. In addition, mice with mutations in Nsun7 display male sterility linked to sperm motility defects, possibly due to defects in mitochondrial structure and/or function[20]. Due to the remarkable phenotypes linked to Nsun2 and Nsun7, we will pursue targeted experiments to identify the functional contribution of methylation to the specific target RNAs identified in the first phase of this project, in the etiology of such defects. In addition, based on the results obtained in the first phase, generation of knockout mice lacking any other members of the NSUN family methyltransferases will be prioritized. Prioritization will also involve additional considerations/opportunities. For example, NSUN6 has not been characterized in any organism, but uniquely (in the NSUN proteins) bears an additional domain: a special pseudouridine synthase and archaeosine transglycosylase (PUA) domain[21]. This domain has not been studied in NSUN6, and we consider this an opportunity to test its possible involvement in targeting or regulation of NSUN6. Interestingly the normalized microarray data (according to RefExA; Reference database for gene Expression Analysis (Laboratory for Systems Biology and Medicine at RCAST, The University of Tokyo)) report expression restricted to the testes in human.

Another method for prioritizing the generation of knockout mice models is the application of morpholino technology (stable, injectible, antisense oligonucleotides that block

RNA translation or splicing) in zebrafish to block the expression of functional m^5C-RMT homologues in zebrafish. We consider this the fastest and easiest screening tool to explore which enzymes are likely important for development of embryos. This preliminary information will then be used for rational selection of enzymes for generation of knockout mice or zebrafish lines for long-term and more thorough analysis. The knockout or morphant animals will be evaluated for several characteristics (viability, morphology, fertility, developmental defects, behavior etc.). Recently, zebrafish researchers at the Huntsman Cancer Institute have been developing bioinformatic methods to analyze RNA-seq datasets from early zebrafish embryos, methods which can detect the loss of RNAs linked to particular developmental pathways, signaling pathways and tissues. The intended use of this system is for drug screens, but was developed using known morpholinos that impaired known pathways and tissues. One can easily adapt this system to identify pathways or tissues affected by the loss of a particular m^5C-RMT.

Exploring functional models for RNA cytosine methylation

Considerable effort has been expended by others in understanding the target sites and roles of cytosine methylation in tRNAs and rRNAs[2]. Indeed, roles for RNA cytosine methylation have been established, including affecting tRNA folding and stability in normal and stress conditions[2, 22]. Here, Aza-IP[1] promises to greatly expand the known repertoire, and based on our results with NSUN2, we expect from several to hundreds of targets per enzyme, and we will almost certainly encounter a wide variety of RNA types. As targets cannot be predicted and could span the full spectrum of RNAs, the key issue is prioritization. RNAs will be prioritized based on biological impact of the loss of methylation (assessed above), 'interest' in the particular target RNA (admittedly subjective, but will focus on impact (see below)), the

likelihood that the methylation(s) affects a critical function (is it located at/near a catalytic or regulatory site?), and pragmatic issues such as the ease of testing the former issues. We are most interested in the possibility that methylation will regulate the catalytic activity of an RNA, its interactions with interesting regulatory protein partners, or movement to or activity within a particular compartment and the likelihood that our efforts will inform and impact an important problem in basic science or the pathophysiology observed in human disease.

References

1. Khoddami, V. & Cairns, B.R. Identification of direct targets and modified bases of RNA cytosine methyltransferases. *Nat Biotechnol* (2013).

2. Motorin, Y., Lyko, F. & Helm, M. 5-methylcytosine in RNA: detection, enzymatic formation and biological functions. *Nucleic Acids Res* **38**, 1415-1430 (2010).

3. Nix, D.A., Courdy, S.J. & Boucher, K.M. Empirical methods for controlling false positives and estimating confidence in ChIP-Seq peaks. *BMC Bioinformatics* **9**, 523 (2008).

4. Koboldt, D.C. et al. VarScan: variant detection in massively parallel sequencing of individual and pooled samples. *Bioinformatics* **25**, 2283-2285 (2009).

5. King, M.Y. & Redman, K.L. RNA methyltransferases utilize two cysteine residues in the formation of 5-methylcytosine. *Biochemistry* **41**, 11218-11225 (2002).

6. Redman, K.L. Assembly of protein-RNA complexes using natural RNA and mutant forms of an RNA cytosine methyltransferase. *Biomacromolecules* **7**, 3321-3326 (2006).

7. Sugimoto, Y. et al. Analysis of CLIP and iCLIP methods for nucleotide-resolution studies of protein-RNA interactions. *Genome Biol* **13**, R67 (2012).

8. Markusic, D. & Seppen, J. Doxycycline regulated lentiviral vectors. *Methods Mol Biol* **614**, 69-76 (2010).

9. Okamoto, M. et al. Frequent increased gene copy number and high protein expression of tRNA (Cytosine-5-)-methyltransferase (NSUN2) in human cancers. *DNA Cell Biol* **31**, 660-671 (2012).

10. Frye, M. et al. Genomic gain of 5p15 leads to over-expression of Misu (NSUN2) in breast cancer. *Cancer Lett* **289**, 71-80 (2010).

11. Frye, M. & Watt, F.M. The RNA methyltransferase Misu (NSun2) mediates Myc-induced proliferation and is upregulated in tumors. *Curr Biol* **16**, 971-981 (2006).

12. Blanco, S. et al. The RNA-methyltransferase Misu (NSun2) poises epidermal stem cells to differentiate. *PLoS Genet* **7**, e1002403 (2011).

13. Abbasi-Moheb, L. et al. Mutations in NSUN2 cause autosomal-recessive intellectual disability. *Am J Hum Genet* **90**, 847-855 (2012).

14. Khan, M.A. et al. Mutation in NSUN2, which encodes an RNA methyltransferase, causes autosomal-recessive intellectual disability. *Am J Hum Genet* **90**, 856-863 (2012).

15. Martinez, F.J. et al. Whole exome sequencing identifies a splicing mutation in NSUN2 as a cause of a Dubowitz-like syndrome. *J Med Genet* **49**, 380-385 (2012).

16. Goll, M.G. et al. Methylation of tRNAAsp by the DNA methyltransferase homolog Dnmt2. *Science* **311**, 395-398 (2006).

17. Schaefer, M. et al. RNA methylation by Dnmt2 protects transfer RNAs against stress-induced cleavage. *Genes Dev* **24**, 1590-1595 (2010).

18. Becker, M. et al. Pmt1, a Dnmt2 homolog in Schizosaccharomyces pombe, mediates tRNA methylation in response to nutrient signaling. *Nucleic Acids Res* (2012).

19. Durdevic, Z. et al. Efficient RNA virus control in Drosophila requires the RNA methyltransferase Dnmt2. *EMBO Rep* **14**, 269-275 (2013).

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

21. Perez-Arellano, I., Gallego, J. & Cervera, J. The PUA domain - a structural and functional overview. *FEBS J* **274**, 4972-4984 (2007).

22. Chan, C.T. et al. Reprogramming of tRNA modifications controls the oxidative stress response by codon-biased translation of proteins. *Nat Commun* **3**, 937 (2012).

APPENDIX A

SUPPLEMENTARY INFORMATION FOR CHAPTER 2

Table A.1: Methylation report for candidate m^5C sites in protein coding genes (mRNAs) from wt and dnmt2-/- datasets

Type	Gene Name	Coordinate	Gene description	Cov.	% meth	Cov.	% meth
Gene	Nus1	Chr10:52137825	nogo-B receptor precursor	91	64.84	110	60.55
3'-UTR	1110067D22Rik	Chr11:20725147	galectin-related protein A	14	7.14	13	23.08
3'-UTR	2310022B05Rik	Chr8:127160378	hypothetical protein LOC69551	56	8.93	33	24.24
Gene	2310047M10Rik	Chr11:68873931	hypothetical protein LOC71923	15	60	21	47.62
Gene	2510039O18Rik	Chr4:147318738	hypothetical protein LOC77034 precursor	532	14.5	684	22.03
Gene	2900010M23Rik	Chr17:27259735	hypothetical protein LOC67267	266	24.81	232	27.71
3'-UTR	5930416I19Rik	Chr6:128308096	Uncharacterized protein C12orf32 homolog	10	40	3	0
3'-UTR	Abhd8	Chr8:73981072	abhydrolase domain-containing protein 8	34	2.94	22	40.91
Gene	Abr	Chr11:76277078	active breakpoint cluster region-related protein	16	50	23	52.17
5'-UTR	Acbd6	Chr1:157405570	acyl-CoA-binding domain-containing protein 6	11	27.27	5	0
Gene	Actg2	Chr6:83477368	actin, gamma-enteric smooth muscle	63	20.63	342	13.16
3'-UTR	Adipor2	Chr6:119304362	adiponectin receptor protein 2	14	7.14	15	26.67
Gene	Aes	Chr10:81028349	amino-terminal enhancer of split	110	40.91	80	36.25
Gene	Aftph	Chr11:20627446	hypothetical protein	13	23.08	10	10
Gene	Ak5	Chr3:152278926	adenylate kinase isoenzyme 5	10	22.22	11	18.18
3'-UTR	Anapc13	Chr9:102536533	anaphase-promoting complex subunit 13	15	6.67	17	25
3'-UTR	Ap1ar	Chr3:127510835	adaptor-related protein complex 1 associated	14	28.57	14	21.43
3'-UTR	Ap3s2	Chr7:87024243	AP-3 complex subunit sigma-2	5	20	13	23.08
Gene	App	Chr16:85013884	amyloid beta A4 protein isoform 1 precursor	793	22.47	410	25.12
3'-UTR	Arf1	Chr11:59025466	ADP-ribosylation factor 1	196	32.14	263	24.33
Gene	Arf3	Chr15:98571471	ADP-ribosylation factor 3	132	41.22	148	40.54
Gene	Arid1b	Chr17:4995476	AT rich interactive domain 1B (SWI1-like)	9	11.11	13	30.77
3'-UTR	Atf5	Chr7:52068324	cyclic AMP-dependent transcription factor ATF-5	14	42.86	6	16.67
3'-UTR	Atf5	Chr7:52068362	cyclic AMP-dependent transcription factor ATF-5	13	23.08	4	50
3'-UTR	Atp2a2	Chr5:122909562	sarcoplasmic/endoplasmic reticulum calcium	22	27.27	13	23.08
Gene	Atp6v0a1	Chr11:100888602	V-type proton ATPase 116 kDa subunit a isoform	30	43.33	20	30

Type	Gene Name	Coordinate	Gene description	Cov.	% meth	Cov.	% meth
3'-UTR	Atp6v1a	Chr16:44086715	V-type proton ATPase catalytic subunit A	21	28.57	16	18.75
3'-UTR	Axl	Chr7:26543418	tyrosine-protein kinase receptor UFO isoform 1	346	18.55	501	21.16
3'-UTR	B4galt2	Chr4:117555866	beta-1,4-galactosyltransferase 2	25	48	25	52
Gene	Bag6	Chr17:35279732	large proline-rich protein BAT3	18	66.67	7	42.86
3'-UTR	Banf1	Chr19:5365005	barrier-to-autointegration factor	90	35.56	121	32.23
Gene	BC018507	Chr13:70745097	mKIAA0947 protein	13	23.08	7	0
5'-UTR	Bdnf	Chr2:109533855	brain-derived neurotrophic factor isoform 1	38	21.05	28	17.86
3'-UTR	Bicc1	Chr10:70387944	Protein bicaudal C homolog 1	7	42.86	16	31.25
Gene	Cad	Chr5:31363290	carbamoyl-phosphate synthetase 2, aspartate	16	18.75	11	27.27
Gene	Ccdc107	Chr4:43508228	coiled-coil domain-containing protein 107	35	71.43	30	86.67
3'-UTR	Ccni	Chr5:93611961	cyclin-I	39	23.08	40	17.5
Gene	Cd109	Chr9:78484404	CD109 antigen precursor	63	36.51	20	25
Gene	Cd109	Chr9:78515082	CD109 antigen precursor	67	59.7	21	66.67
Gene	Cd109	Chr9:78527910	CD109 antigen precursor	67	7.46	24	21.74
3'-UTR	Cdca4	Chr12:114058788	cell division cycle-associated protein 4	16	25	18	38.89
Gene	Cenpb	Chr2:131004566	major centromere autoantigen B	30	33.33	28	71.43
Gene	Cgrrf1	Chr14:47465617	cell growth regulator with RING finger domain	25	40	13	46.15
Gene	Chchd3	Chr6:33010075	coiled-coil-helix-coiled-coil-helix	59	35.59	41	29.27
Gene	Chd3	Chr11:69161295	Chd3 protein	82	6.1	30	20
Gene	Chd3	Chr11:69173788	chromodomain helicase DNA binding protein 3	105	30.48	65	30.77
Gene	Chmp1a	Chr8:125731402	charged multivesicular body protein 1a	240	28.75	253	38.89
Gene	Chpf	Chr1:75472421	chondroitin sulfate synthase 2 isoform b	26	61.54	27	25.93
3'-UTR	Chpf	Chr1:75471517	chondroitin sulfate synthase 2 isoform b	14	28.57	10	11.11
Gene	Chst2	Chr9:95306155	carbohydrate sulfotransferase 2	2	50	12	33.33
Gene	Col1a1	Chr11:94803372	collagen alpha-1(I) chain precursor	729	38.65	476	30.3
Gene	Col1a1	Chr11:94805126	collagen alpha-1(I) chain precursor	565	24.16	485	21.86
Gene	Col4a1	Chr8:11233892	collagen alpha-1(IV) chain precursor	63	28.57	77	27.27
Gene	Col4a5	ChrX:138065880	collagen, type IV, alpha 5	37	78.38	48	66.67

Type	Gene Name	Coordinate	Gene description	Cov.	% meth	Cov.	% meth
Gene	Col5a2	Chr1:45438971	collagen alpha-2(V) chain precursor	126	19.84	163	26.99
Gene	Col6a3	Chr1:92670141	collagen, type VI, alpha 3	1	100	21	66.67
Gene	Col6a3	Chr1:92678637	collagen, type VI, alpha 3	0	0	20	25
Gene	Copg	Chr6:87852347	coatomer subunit gamma isoform 1	140	30.71	148	25
Gene	Copg	Chr6:87852348	coatomer subunit gamma isoform 1	141	22.7	149	17.45
Gene	Copg	Chr6:87852358	coatomer subunit gamma isoform 1	139	23.02	137	18.98
Gene	Copg	Chr6:87852359	coatomer subunit gamma isoform 1	139	21.58	136	18.38
3'-UTR	Cotl1	Chr8:122333604	coactosin-like protein	17	52.94	72	54.17
3'-UTR	Cpd	Chr11:76593325	carboxypeptidase D precursor	8	0	19	26.32
Gene	Crtap	Chr9:114299351	cartilage-associated protein precursor	33	42.42	24	45.83
Gene	Ctnnd1	Chr2:84460581	catenin delta-1 isoform 2	75	10.67	77	28.57
Gene	Cttn	Chr7:151622878	src substrate cortactin	50	12	49	20.41
Gene	Cul5	Chr9:53475259	cullin-5 isoform 1	23	18.18	20	30
Gene	Cul7	Chr17:46788742	uncharacterized protein	17	11.76	10	20
Gene	Cul7	Chr17:46798682	uncharacterized protein	22	22.73	15	13.33
3'-UTR	Dcakd	Chr11:102855858	dephospho-CoA kinase domain-containing protein	10	40	6	16.67
Gene	Ddx54	Chr5:121068727	ATP-dependent RNA helicase DDX54	20	35	20	30
Gene	Dlc1	Chr8:37642668	rho GTPase-activating protein 7 isoform 2	21	23.81	13	0
Gene	Dnajc10	Chr2:80173941	dnaJ homolog subfamily C member 10	68	14.71	78	23.08
Gene	Dnajc3	Chr14:119377347	dnaJ homolog subfamily C member 3 precursor	20	30	10	20
Gene	Donson	Chr16:91688778	protein downstream neighbor of Son	23	21.74	28	14.29
3'-UTR	Ebp	ChrX:7762683	3-beta-hydroxysteroid-Delta(8)	10	40	11	45.45
Gene	Ecm1	Chr3:95538207	extracellular matrix protein 1 precursor	17	41.18	18	38.89
3'-UTR	Eif4ebp1	Chr8:28385714	eukaryotic translation initiation factor	42	30.95	35	25.71
Gene	Emb	Chr13:118056195	embigin precursor	7	0	16	25
Gene	Emd	ChrX:71502521	emerin	141	25.53	274	26.28
Gene	Ergic3	Chr2:155836875	endoplasmic reticulum-Golgi intermediate	189	14.81	145	20.69
Gene	Exosc10	Chr4:147954583	exosome component 10	22	22.73	17	17.65
3'-UTR	Extl3	Chr14:65672410	exostosin-like 3	12	25	15	26.67

Type	Gene Name	Coordinate	Gene description	Cov.	% meth	Cov.	% meth
Gene	Fam108a	Chr10:80048407	abhydrolase domain-containing protein FAM108A	205	14.63	203	22.17
3'-UTR	Fam195b	Chr11:120404476	hypothetical protein LOC192173	13	23.08	23	17.39
Gene	Fam38a	Chr8:125010407	protein PIEZO1	53	18.87	40	20
Gene	Fam38a	Chr8:125026116	protein PIEZO1	10	33.33	7	71.43
Gene	Farsa	Chr8:87384870	phenylalanyl-tRNA synthetase alpha chain	12	33.33	9	22.22
Gene	Fasn	Chr11:120680359	fatty acid synthase	32	18.75	21	38.1
Gene	Fasn	Chr11:120680421	fatty acid synthase	10	40	6	16.67
Gene	Fbln2	Chr6:91218539	fibulin-2 isoform a	1427	79.17	1450	82.88
Gene	Fbn1	Chr2:125128841	fibrillin-1	555	26.49	676	15.56
Gene	Fbn1	Chr2:125147358	fibrillin-1	67	73.13	83	69.88
Gene	Fkbp4	Chr6:128380828	peptidyl-prolyl cis-trans isomerase FKBP4	102	11.76	77	21.05
Gene	Flna	ChrX:71482483	filamin-A	93	19.57	99	27.27
Gene	Flnb	Chr14:8717116	filamin-B	128	18.75	167	23.35
Gene	Frmd4a	Chr2:4524623	FERM domain-containing protein 4A isoform 1	16	0	14	28.57
Gene	Fscn1	Chr5:143722899	fascin	71	57.75	76	53.95
Gene	Ftl1	Chr7:52713562	ferritin light chain 1	290	23.96	165	18.79
3'-UTR	Fzd1	Chr5:4755213	frizzled-1 precursor	21	33.33	27	48.15
Gene	Galnt10	Chr11:57579160	polypeptide N-acetylgalactosaminyltransferase	13	30.77	7	42.86
Gene	Galnt2	Chr8:126819327	polypeptide N-acetylgalactosaminyltransferase 2	247	20.65	286	25.52
Gene	Galnt2	Chr8:126860484	polypeptide N-acetylgalactosaminyltransferase 2	95	22.11	112	13.39
Gene	Galnt2	Chr8:126860488	polypeptide N-acetylgalactosaminyltransferase 2	99	27.27	119	21.85
3'-UTR	Galnt2	Chr8:126868752	polypeptide N-acetylgalactosaminyltransferase 2	14	7.14	22	31.82
Gene	Gas2l1	Chr11:4962229	GAS2-like protein 1 isoform alpha	41	22.5	40	20
Gene	Ghr	Chr15:3269874	growth hormone receptor isoform 1 precursor	19	0	20	25
Gene	Glis3	Chr19:28432378	zinc finger protein GLIS3	14	21.43	26	50
3'-UTR	Glo1	Chr17:30729915	lactoylglutathione lyase	19	21.05	31	9.68
Gene	Gnb2l1	Chr11:48614047	guanine nucleotide-binding protein subunit	1299	29.71	1080	20.09

Type	Gene Name	Coordinate	Gene description	Cov.	% meth	Cov.	% meth
3'-UTR	Gorab	Chr1:165315670	RAB6-interacting golgin	4	50	19	52.63
3'-UTR	Gosr2	Chr11:103539940	Golgi SNAP receptor complex member 2	36	27.78	35	25.71
3'-UTR	Got2	Chr8:98388904	aspartate aminotransferase, mitochondrial	17	47.06	22	40.91
Gene	Gpd2	Chr2:57142306	glycerol-3-phosphate dehydrogenase	34	57.58	25	76
3'-UTR	Grem2	Chr1:176765226	gremlin-2 precursor	2	0	36	38.89
Gene	Gvin1	Chr7:113047862	interferon-induced very large GTPase 1	52	28.85	8	12.5
Gene	Gvin1	Chr7:113304279	interferon-induced very large GTPase 1	50	20	7	14.29
Gene	Hdgf	Chr3:87718567	hepatoma-derived growth factor	234	25.75	212	27.96
Gene	Herc4	Chr10:62780097	probable E3 ubiquitin-protein ligase HERC4	23	8.7	24	20.83
Gene	Hist1h2ai	Chr13:21808490	histone H2A type 1	94	20.21	181	28.18
Gene	Hist1h2ai	Chr13:21808494	histone H2A type 1	88	28.41	178	29.94
Gene	Hist1h2ai	Chr13:21808502	histone H2A type 1	72	29.17	153	30.72
Gene	Hist1h2ai	Chr13:21808503	histone H2A type 1	70	35.71	149	33.56
Gene	Hist1h2ai	Chr13:21808504	histone H2A type 1	70	31.43	141	36.17
Gene	Hist1h2ai	Chr13:21808506	histone H2A type 1	72	27.78	142	30.28
Gene	Hist2h2ab	Chr3:96024038	histone H2A type 2-B	54	15.09	71	28.17
Gene	Hist2h2ab	Chr3:96024042	histone H2A type 2-B	52	21.15	68	25
Gene	Hist2h2ab	Chr3:96024050	histone H2A type 2-B	36	22.22	60	28.33
Gene	Hist2h2ab	Chr3:96024051	histone H2A type 2-B	35	25.71	54	27.78
Gene	Hist2h2ab	Chr3:96024052	histone H2A type 2-B	33	21.21	53	26.42
Gene	Hist2h2ab	Chr3:96024054	histone H2A type 2-B	32	9.38	53	28.3
Gene	Homer3	Chr8:72813338	homer protein homolog 3 isoform 2	112	20.54	72	23.61
Gene	Hspg2	Chr4:137096135	basement membrane-specific heparan sulfate	65	15.38	36	22.22
Gene	Htra1	Chr7:138079898	serine protease HTRA1 precursor	43	34.88	59	35.59
Gene	Igf1r	Chr7:75148812	insulin-like growth factor 1 receptor	166	22.89	86	22.35
Gene	Igf2r	Chr17:12891204	cation-independent mannose-6-phosphate receptor	432	36.81	444	27.03
Gene	Igf2r	Chr17:12891205	cation-independent mannose-6-phosphate receptor	429	48.6	445	39.19
3'-UTR	Igf2r	Chr17:12876061	cation-independent mannose-6-phosphate receptor	24	33.33	33	30.3
Gene	Igfbp6	Chr15:101979684	insulin-like growth factor-binding protein 6	1	0	28	46.43

Type	Gene Name	Coordinate	Gene description	Cov.	% meth	Cov.	% meth
3'-UTR	Il13ra1	ChrX:33710599	interleukin-13 receptor subunit alpha-1	7	28.57	15	20
Gene	Il6st	Chr13:113270418	Putative uncharacterized protein	26	53.85	36	75
Gene	Ints1	Chr5:140235242	integrator complex subunit 1	30	16.67	15	33.33
3'-UTR	Ints3	Chr3:90195457	integrator complex subunit 3	15	26.67	9	33.33
Gene	Jund	Chr8:73223528	transcription factor jun-D	11	45.45	6	33.33
Gene	Katnb1	Chr8:97619573	katanin p80 WD40-containing subunit B1	42	66.67	43	65.12
Gene	Kdelc2	Chr9:53198635	KDEL (Lys-Asp-Glu-Leu) containing 2 protein	25	12	13	30.77
Gene	Klhl26	Chr8:72979480	kelch-like protein 26 isoform 2	6	16.67	11	36.36
3'-UTR	Lamp2	ChrX:35774191	lysosome-associated membrane glycoprotein 2	101	16.16	82	36.59
3'-UTR	Litaf	Chr16:10960142	lipopolysaccharide-induced tumor necrosis	9	55.56	12	25
Gene	Loxl3	Chr6:82987541	lysyl oxidase homolog 3 precursor	78	23.08	110	44.55
Gene	Loxl3	Chr6:82999219	lysyl oxidase homolog 3 precursor	50	24	95	39.36
Gene	Lrp1	Chr10:126997169	prolow-density lipoprotein receptor-related	23	60.87	22	63.64
Gene	Lrpprc	Chr17:85176387	leucine-rich PPR motif-containing protein	20	0	16	25
3'-UTR	Lrrc15	Chr16:30269503	leucine-rich repeat-containing protein 15	4	25	13	46.15
3'-UTR	Lrrc15	Chr16:30272784	leucine-rich repeat-containing protein 15	3	66.67	10	50
3'-UTR	Lrrc15	Chr16:30272785	leucine-rich repeat-containing protein 15	3	33.33	10	50
Gene	Lrrc45	Chr11:120582033	leucine-rich repeat-containing protein 45	17	52.94	5	60
Gene	Lta4h	Chr10:92943517	leukotriene A-4 hydrolase	21	38.1	16	31.25
Gene	Mapk1ip1l	Chr14:47930401	MAPK-interacting and spindle-stabilizing	21	23.81	12	16.67
Gene	Marcks	Chr10:36856328	myristoylated alanine-rich C-kinase substrate	20	30	10	0
Gene	Mardh5	Chr19:37285270	E3 ubiquitin-protein ligase MARCH5 isoform 3	10	30	13	0
5'-UTR	Masp1	Chr16:23520645	mannan-binding lectin serine protease 1	142	16.2	63	22.22
3'-UTR	Metap1	Chr3:138122647	methionine aminopeptidase 1	46	10.87	44	20.45
3'-UTR	Mlec	Chr5:115594247	malectin precursor	253	25.3	178	23.6
3'-UTR	Mlec	Chr5:115595374	malectin precursor	34	23.53	17	23.53
3'-UTR	Mpdu1	Chr11:69470585	mannose-P-dolichol utilization defect 1 protein	21	4.76	11	27.27

Type	Gene Name	Coordinate	Gene description	Cov.	% meth	Cov.	% meth
Gene	Msl1	Chr11:98657418	male-specific lethal 1 homolog	10	40	11	36.36
Gene	Mtap1b	Chr13:100205214	microtubule-associated protein 1B	35	20	20	30
3'-UTR	Mtf1	Chr4:124526963	metal regulatory transcription factor 1	34	20.59	35	17.14
Gene	Mvp	Chr7:134130532	major vault protein	23	30.43	6	33.33
3'-UTR	Myef2	Chr2:124913772	sodium/potassium/calcium exchanger 5 precursor	11	9.09	13	30.77
3'-UTR	Nap1l4	Chr7:150700445	nucleosome assembly protein 1-like 4	40	25	48	20.83
Gene	Ncln	Chr10:80952623	nicalin precursor	35	8.57	15	20
3'-UTR	Necap2	Chr4:140622641	adaptin ear-binding coat-associated protein 2	41	17.07	27	33.33
3'-UTR	Nedd4	Chr9:72596630	E3 ubiquitin-protein ligase NEDD4	114	15.79	92	28.26
3'-UTR	Nek9	Chr12:86641347	serine/threonine-protein kinase Nek9	30	50	14	57.14
3'-UTR	Nfe2l1	Chr11:96678920	nuclear factor erythroid 2-related factor 1	12	25	7	0
Gene	Nfix	Chr8:87296194	nuclear factor 1 X-type isoform 3	9	66.67	17	70.59
Gene	Nid2	Chr14:20571674	nidogen-2 precursor	28	89.29	10	90
3'-UTR	Nme4	Chr17:26228949	nucleoside diphosphate kinase, mitochondrial	21	23.81	14	7.14
Gene	Nmt2	Chr2:3201573	N-myristoyltransferase 2	31	12.9	28	21.43
Gene	Nono	ChrX:98637056	non-POU domain-containing octamer-binding	59	15.25	52	25
Gene	Nuak1	Chr10:83903014	NUAK family SNF1-like kinase 1	19	31.58	7	42.86
Gene	Pacs2	Chr12:114288095	phosphofurin acidic cluster sorting protein 2	18	22.22	18	33.33
Gene	Pcdha4-g	Chr18:37645376	protocadherin alpha 4-gamma	14	57.14	16	93.75
Gene	Pcdha4-g	Chr18:37966649	protocadherin alpha 4-gamma	19	26.32	10	20
3'-UTR	Pfkl	Chr10:77450778	6-phosphofructokinase, liver type	12	25	11	18.18
Gene	Pitpna	Chr11:75433756	phosphatidylinositol transfer protein alpha	97	19.59	71	32.39
Gene	Pitrm1	Chr13:6554876	presequence protease, mitochondrial precursor	150	29.33	137	27.01
3'-UTR	Plbd2	Chr5:120934131	putative phospholipase B-like 2	11	45.45	5	20
Gene	Plec	Chr15:76006893	plectin isoform 1a	18	22.22	16	6.25
Gene	Plec	Chr15:76008546	plectin isoform 1a	4	25	11	30
Gene	Plec	Chr15:76008556	plectin isoform 1a	4	0	12	50
Gene	Plec	Chr15:76008566	plectin isoform 1a	3	33.33	11	72.73
Gene	Plec	Chr15:76008577	plectin isoform 1a	5	40	10	30
Gene	Plec	Chr15:76011740	plectin isoform 1a	16	25	8	0

Type	Gene Name	Coordinate	Gene description	Cov.	% meth	Cov.	% meth
Gene	Plec	Chr15:76012154	plectin isoform 1a	9	77.78	11	36.36
Gene	Plec	Chr15:76012164	plectin isoform 1a	8	75	11	54.55
Gene	Plec	Chr15:76012188	plectin isoform 1a	11	0	13	23.08
Gene	Plec	Chr15:76018177	plectin isoform 1a	22	9.09	12	25
Gene	Plekhg2	Chr7:29155352	pleckstrin homology domain-containing family G	21	38.1	14	28.57
Gene	Plod1	Chr4:147305292	procollagen-lysine,2-oxoglutarate 5-dioxygenase	32	46.88	55	56.36
Gene	Plxnb2	Chr15:88991468	Similar to plexin B1	22	45.45	14	64.29
3'-UTR	Pmepa1	Chr2:173050276	transmembrane prostate androgen-induced protein	23	13.04	21	23.81
3'-UTR	Poldip3	Chr15:82957356	polymerase delta-interacting protein 3	155	42.58	160	54.09
Gene	Ppp1cc	Chr5:122619025	serine/threonine-protein phosphatase PP1-gamma	20	10	12	25
3'-UTR	Psma4	Chr9:54805706	proteasome subunit alpha type-4	34	67.65	70	71.43
3'-UTR	Psmb7	Chr2:38443660	proteasome subunit beta type-7 precursor	64	40.62	58	46.55
Gene	Psmd7	Chr8:110105103	26S proteasome non-ATPase regulatory subunit 7	30	20	23	30.43
Gene	Psme4	Chr11:30712067	proteasome activator complex subunit 4	46	28.26	36	33.33
Gene	Ptdss1	Chr13:67073461	phosphatidylserine synthase 1	32	53.12	48	63.83
3'-UTR	Ptplad1	Chr9:64835631	3-hydroxyacyl-CoA dehydratase 3	19	15.79	18	27.78
Gene	Ptprn	Chr1:75248328	receptor-type tyrosine-protein phosphatase-like	48	54.17	63	63.49
Gene	Ptprn	Chr1:75254850	receptor-type tyrosine-protein phosphatase-like	30	23.33	33	27.27
3'-UTR	Ptprn	Chr1:75243938	receptor-type tyrosine-protein phosphatase-like	11	54.55	6	66.67
Gene	Pttg1ip	Chr10:77055720	pituitary tumor-transforming gene 1	120	14.17	88	32.95
Gene	Pttg1ip	Chr10:77055752	pituitary tumor-transforming gene 1	36	17.14	38	39.47
3'-UTR	Purb	Chr11:6372307	purine rich element binding protein B (Purb)	6	0	14	21.43
3'-UTR	Rai14	Chr15:10499343	ankycorbin	8	12.5	33	27.27
3'-UTR	Rb1	Chr14:73597011	retinoblastoma-associated protein	30	13.33	24	25
Gene	Rpl34	Chr3:130432050	60S ribosomal protein L34 isoform 1	274	28.47	276	13.82
Gene	Rps6ka4	Chr19:6904554	ribosomal protein S6 kinase alpha-4	86	20.24	45	8.89
3'-UTR	Rspo3	Chr10:29173572	R-spondin-3 precursor	17	25	5	0
Gene	Sae1	Chr7:16973041	SUMO-activating enzyme subunit 1	75	33.78	47	27.66

Type	Gene Name	Coordinate	Gene description	Cov.	% meth	Cov.	% meth
Gene	Sae1	Chr7:16973058	SUMO-activating enzyme subunit 1	22	57.14	16	37.5
3'-UTR	Sec61a1	Chr6:88453783	protein transport protein Sec61 subunit alpha	65	56.92	40	80
3'-UTR	Sepn1	Chr4:134095447	selenoprotein N, 1	27	22.22	5	0
Gene	Setbp1	Chr18:79283442	SET-binding protein	13	53.85	3	66.67
Gene	Setd3	Chr12:109403360	SET domain-containing protein 3	22	9.09	23	26.09
Gene	Sgce	Chr6:4641572	epsilon-sarcoglycan isoform 1	27	23.08	17	35.29
3'-UTR	Sgpl1	Chr10:60561833	sphingosine-1-phosphate lyase 1	31	22.58	43	20.93
3'-UTR	Sgta	Chr10:80507347	small glutamine-rich tetratricopeptide	34	50	18	66.67
3'-UTR	Sirt7	Chr11:120480109	NAD-dependent deacetylase sirtuin-7	15	26.67	8	25
3'-UTR	Slc23a2	Chr2:131881811	sodium-dependent vitamin C transporter SVCT2	12	16.67	10	20
Gene	Slc25a4	Chr8:47294488	ADP/ATP translocase 1	239	39.75	268	31.2
3'-UTR	Slc35a3	Chr3:116373862	UDP-N-acetylglucosamine transporter	10	40	17	29.41
Gene	Slc38a10	Chr11:119966810	sodium-coupled neutral amino acid	83	26.51	62	20.97
Gene	Slc39a1	Chr3:90055623	zinc transporter ZIP1	147	58.5	138	65.22
Gene	Smpd1	Chr7:112704250	uncharacterized protein	141	22.7	107	26.17
Gene	Snora21	Chr11:97643136	60S ribosomal protein L23	1100	24.82	1156	13.49
3'-UTR	Sorcs2	Chr5:36360777	VPS10 domain-containing receptor	12	83.33	10	100
Gene	Spag9	Chr11:93975571	C-Jun-amino-terminal kinase-interacting protein	117	20.51	70	32.86
Gene	Srp68	Chr11:116122186	signal recognition particle 68 kDa protein	19	15.79	16	25
3'-UTR	Stom	Chr2:35171342	erythrocyte band 7 integral membrane protein	52	23.08	73	15.07
3'-UTR	Stt3b	Chr9:115152204	dolichyl-diphosphooligosaccharide--protein	20	25	20	10
3'-UTR	Surf4	Chr2:26776217	surfeit locus protein 4	530	33.96	648	43.06
Gene	Syne2	Chr12:77068376	nesprin-2	11	27.27	10	20
Gene	Taf10	Chr7:112892449	transcription initiation factor TFIID subunit	10	40	7	28.57
3'-UTR	Tagln	Chr9:45737989	transgelin	19	31.58	41	0
3'-UTR	Tardbp	Chr4:147999265	TAR DNA-binding protein 43 isoform 4	111	16.67	87	29.89
5'-UTR	Tcf12	Chr9:71959546	transcription factor 12	17	52.94	10	40
3'-UTR	Tcn2	Chr11:3817359	transcobalamin-2 precursor	25	60	29	62.07

Type	Gene Name	Coordinate	Gene description	Cov.	% meth	Cov.	% meth
Gene	Tgfb3	Chr12:87410750	transforming growth factor beta-3 preproprotein	33	36.36	50	26
Gene	Tgm2	Chr2:157943597	protein-glutamine gamma-glutamyltransferase 2	21	9.52	28	25
Gene	Tm9sf1	Chr14:56255332	transmembrane 9 superfamily member 1 precursor	16	12.5	13	23.08
5'-UTR	Tmem126a	Chr7:97605648	transmembrane protein 126A	35	37.14	37	29.73
Gene	Tmem132a	Chr19:10934155	transmembrane protein 132A precursor	20	10	10	40
Gene	Tmem132a	Chr19:10939958	transmembrane protein 132A precursor	15	20	2	0
3'-UTR	Tmem161a	Chr8:72706281	transmembrane protein 161A precursor	12	8.33	12	25
Gene	Tmem198b	Chr10:128239114	hypothetical protein LOC73827	13	0	25	24
Gene	Tmem2	Chr19:21881978	transmembrane protein 2	12	33.33	11	9.09
Gene	Tpp1	Chr7:112897424	tripeptidyl-peptidase 1 precursor	62	67.74	39	69.23
Gene	Tpst2	Chr5:112737342	protein-tyrosine sulfotransferase 2	16	12.5	19	52.63
Gene	Tpst2	Chr5:112738760	protein-tyrosine sulfotransferase 2	13	23.08	10	20
Gene	Trap1	Chr16:4045974	heat shock protein 75 kDa, mitochondrial	15	20	22	13.64
Gene	Trim2	Chr3:83994921	tripartite motif-containing protein 2	11	54.55	8	50
Gene	Trp53i13	Chr11:77321783	tumor protein p53-inducible protein 13	26	34.62	15	53.33
3'-UTR	Tspan9	Chr6:127911773	tetraspanin-9	12	33.33	10	20
3'-UTR	Tspan9	Chr6:127913494	tetraspanin-9	20	0	19	31.58
3'-UTR	Tspan9	Chr6:127913495	tetraspanin-9	20	5	20	25
Gene	Ttc7b	Chr12:101738357	tetratricopeptide repeat domain 7B	12	25	8	12.5
Gene	Ttll12	Chr15:83421911	tubulin--tyrosine ligase-like protein 12	25	20	36	16.67
Gene	Txndc12	Chr4:108520832	tubulin--tyrosine ligase-like protein 12	26	26.92	32	25
3'-UTR	Ube2z	Chr11:95910580	ubiquitin-conjugating enzyme E2 Z	18	5.56	12	25
3'-UTR	Ubiad1	Chr4:147810211	ubiA prenyltransferase domain-containing protein	13	46.15	9	22.22
Gene	Ubqln1	Chr13:58293280	ubiquilin-1 isoform 1	68	8.82	77	22.08
Gene	Uqcrfs1	Chr13:30637014	cytochrome b-c1 complex subunit Rieske	8	12.5	12	25
Gene	Use1	Chr8:73891239	vesicle transport protein USE1 isoform3	288	25.09	253	11.07
Gene	Use1	Chr8:73893072	vesicle transport protein USE1 isoform3	107	32.71	114	25.66

Type	Gene Name	Coordinate	Gene description	Cov.	% meth	Cov.	% meth
Gene	Usp38	Chr8:83537771	ubiquitin carboxyl-terminal hydrolase 38	17	5.88	18	33.33
3'-UTR	Wbp2	Chr11:115940271	WW domain-binding protein 2	26	46.15	4	25
Gene	Wdr18	Chr10:79423028	Putative uncharacterized protein	18	22.22	6	16.67
3'-UTR	Wdr48	Chr9:119834545	WD repeat-containing protein 48	14	28.57	13	23.08
Gene	Wdr61	Chr9:54575404	WD repeat-containing protein 61 isoform a	51	13.73	47	21.74
3'-UTR	Wwc2	Chr8:48913727	protein WWC2	33	33.33	17	29.41
3'-UTR	Xpo6	Chr7:133245523	Ran-binding protein 20	11	27.27	6	0
Gene	Xylt2	Chr11:94531703	xylosyltransferase 2	15	6.67	10	20
Gene	Zbtb39	Chr10:127179602	zinc finger and BTB domain-containing protein	12	25	8	0
3'-UTR	Zdhhc20	Chr14:58453211	probable palmitoyltransferase ZDHHC20	48	22.92	35	20
Gene	Zfhx3	Chr8:111473335	zinc finger homeobox protein 3	22	22.73	21	19.05
3'-UTR	Zfp275	ChrX:70601270	Zinc finger protein 275 isoform 1	25	44	21	38.1
Gene	Zmat3	Chr3:32259939	zinc finger matrin-type protein 3	69	10.14	65	23.08
3'-UTR	Znrf1	Chr8:114145994	E3 ubiquitin-protein ligase ZNRF1 isoform a	5	40	10	50
Gene	Zzef1	Chr11:72610012	hypothetical protein	10	10	22	22.73

APPENDIX B

SUPPLEMENTARY INFORMATION FOR CHAPTER 3

SUPPLEMENTARY INFORMATION

Identification of direct targets and modified bases of RNA cytosine methyltransferases

Vahid Khoddami and Bradley R Cairns*

HHMI, Department of Oncological Sciences, Huntsman Cancer Institute
University of Utah School of Medicine, Salt Lake City, UT, USA

*Corresponding author. HHMI, Department of Oncological Sciences, Huntsman Cancer Institute, University of Utah School of Medicine, Salt Lake City, UT 84112, USA. Tel.: +1 801 585 1822; Fax: +1 801 585 6410; E-mail: brad.cairns@hci.utah.edu

SUPPLEMENTARY RESULTS AND DISCUSSION:

SUPPLEMENTARY FIGURES:

SUPPLEMENTARY TABLES:

Nature Biotechnology: doi:10.1038/nbt.2566

SUPPLEMENTARY RESULTS AND DISCUSSION:

Supplementary Result 1: Examining the substrate specificity of human DNMT2

To identify candidate novel DNMT2 mRNA substrates, we compared V5-DNMT2- to V5-DsRed-Aza-IP datasets, which revealed ~60 mRNAs as statistically enriched (FDR < 0.01, Log2(Test/Control) > 1.5). We then filtered to remove false positives that represent RNAs upregulated simply due to V5-DNMT2 overexpression, which were determined via mRNA-seq of polyA-selected RNA of the same 5-azaC-treated (input) samples (before we performed Aza-IP). This yielded 23 candidates, with further filtering (requiring >3 RPKM) limiting the list to 6 candidates (**Supplementary Data set 2**). These candidates were then screened for our transversion signature, which was clearly evident in keratin 18 (KRT18) mRNA and also one of its pseudogenes. For KRT18 itself, the C to G transversion signature resided at one base near the 5' end in 63% of the reads (**Supplementary Fig. 1**). To test for a DNA sequence variant at that position, we inspected our RNA-seq reads of our 5-azaC-treated input samples (before Aza-IP), which revealed a cytosine at that position (517 total reads in the combined datasets: 513 cytosine, 4 guanosine), effectively ruling out a single nucleotide polymorphism (SNP) at that location.

To test whether KRT18 RNA can serve as a DNMT2 substrate in vitro, we performed an in vitro methyltransferase (MTase) assay (**Supplementary Fig. 2-5**). Here, we utilized purified DNMT2 and three substrates: a KRT18 RNA fragment (75 bases) centered on the candidate target cytosine, a KRT18 RNA with adenosine in place of the candidate target cytosine (a site-specificity control), and $tRNA^{Asp}$ itself as our positive control (**Supplementary Table 3**). All three substrates were produced by in-vitro transcription and purified. Notably, we observed high methylation of our positive control $tRNA^{Asp}$, moderate methylation of the experimental/wild-type KRT18 RNA fragment, but no detectable methylation of our cytosine-substituted (adenosine) control KRT18 fragment (**Supplementary Fig. 5**). These results provide support for KRT18 mRNA as a target for hDNMT2, but one that is extremely poor relative to authentic tRNAs.

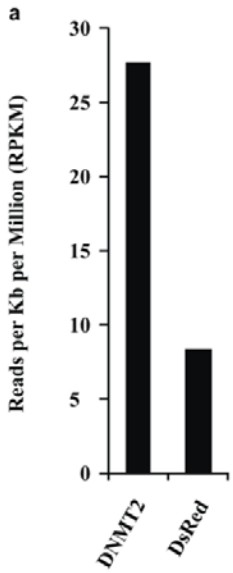
We recognized that only the three known tRNA targets ($tRNA^{Asp}$, $tRNA^{Gly}$ and $tRNA^{Val}$), but not tRNA non-targets, bear a CpG in the stem-loop junction of their anticodon stem-loop (**Supplementary Table 1**). Performing in-silico RNA folding on a truncated 55-mer KRT18 RNA, the most stable structure obtained (ΔG=-24.50 kcal/mol) predicted a cloverleaf tRNA-like structure that placed the candidate cytosine (in the CpG context) at the stem-loop junction, analogous to position C38 in the anticodon loop of the DNMT2 target tRNAs, such as $tRNA^{Asp}$. The 75-base fragment also folded into a similar structure, though alternative structures were also produced (**Supplementary Fig. 6**).

Finally, to test whether a CpG at a tRNA stem-loop junction is sufficient for DNMT2 activity, we made replacements in non-target tRNAs ($tRNA^{Ala}$ and $tRNA^{Pro}$) to impose CpG contexts at that position (**Supplementary Table 3**). However, no activity was observed with hDNMT2 on these modified tRNA substrates in our in-vitro MTase assay (**Supplementary Fig. 7**). Taken together, our studies suggest that having a CpG in a stem loop junction of an anticodon stem-loop is a required, but not sufficient element, for substrate specificity by DNMT2.

SUPPLEMENTARY FIGURES

Supplementary Figure 1: Enrichment of the KRT18 mRNA in DNMT2-Aza-IP dataset and the C>G transversion signature

a



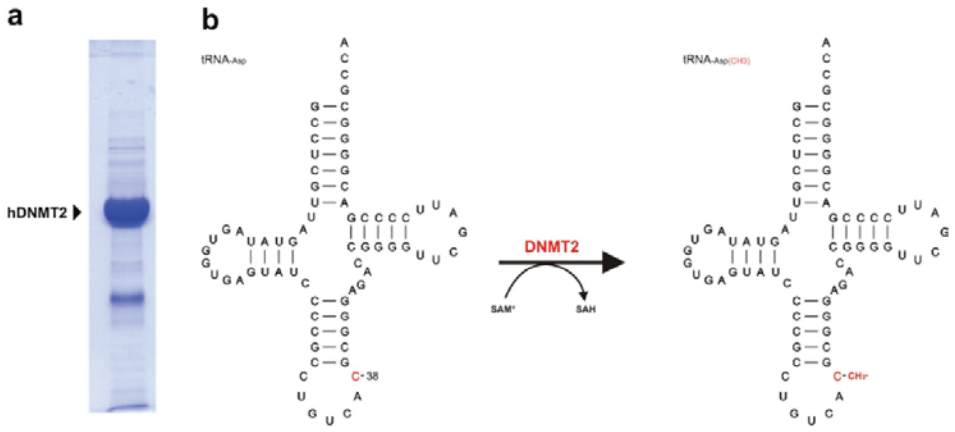
b

Target RNA	Coordinate	Sample	Total	A (%)	C (%)	G (%)	T (%)
KRT18-mRNA	chr12:53,343,029	**DNMT2**	**32**	**2 (6)**	**9 (28)**	**20 (63)**	**1 (3)**
		DsRed	0	0 (0)	0 (0)	0 (0)	0 (0)

Supplementary Figure 1: Enrichment of the KRT18 mRNA in DNMT2-Aza-IP dataset and the C>G transversion signature. (a) KRT18-mRNA shows about 3.3 fold enrichment in the DNMT2 Aza-IP over DsRed Aza-IP dataset. (b) Near the 5' end of KRT18 mRNA, a C>G transversion signature was observed in the 63% of the reads in DNMT2-Aza-IP dataset.

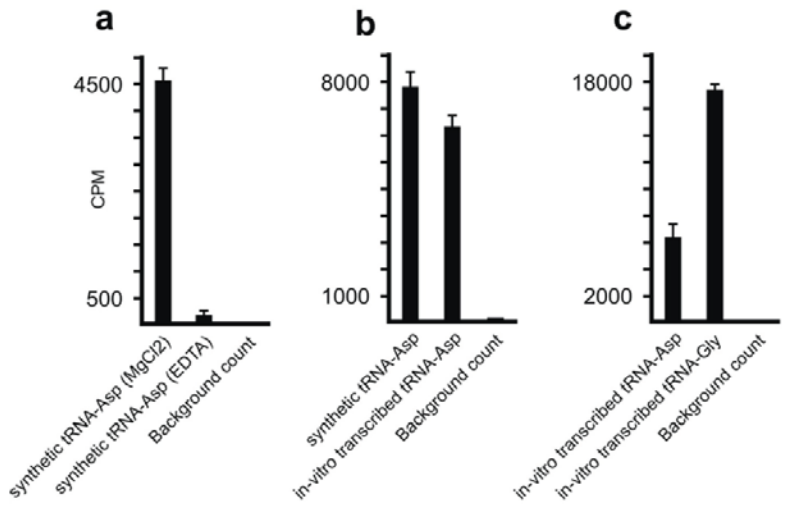
Nature Biotechnology: doi:10.1038/nbt.2566

Supplementary Figure 2: Components in the DNMT2 in-vitro methyltransferase assay (MTase).



Supplementary Figure 2: Components in the DNMT2 in-vitro methyltransferase assay (MTase). (a) Coomassie stained gel of purified His-tagged human DNMT2 protein, expressed in and purified from *E. Coli* (see Supplementary Methods). (b) Simple schematic of experimental set up for DNMT2 in-vitro methyltransferase (MTase) assay. Unmodified synthetic (or in-vitro transcribed) substrate tRNA is incubated with purified human DNMT2 enzyme in the presence of tritium-labeled methyl-donor (S-Adenosyl methionine; SAM) for ~5hrs at 37°C. DNMT2 transfers the tritium labeled methyl group to the fifth position of the target cytosine (here C38). After RNA clean-up the level of tritium labeled methyl group incorporation is measured by scintillation counter and reported as counts per minute (CPM); see Supplementary Methods.

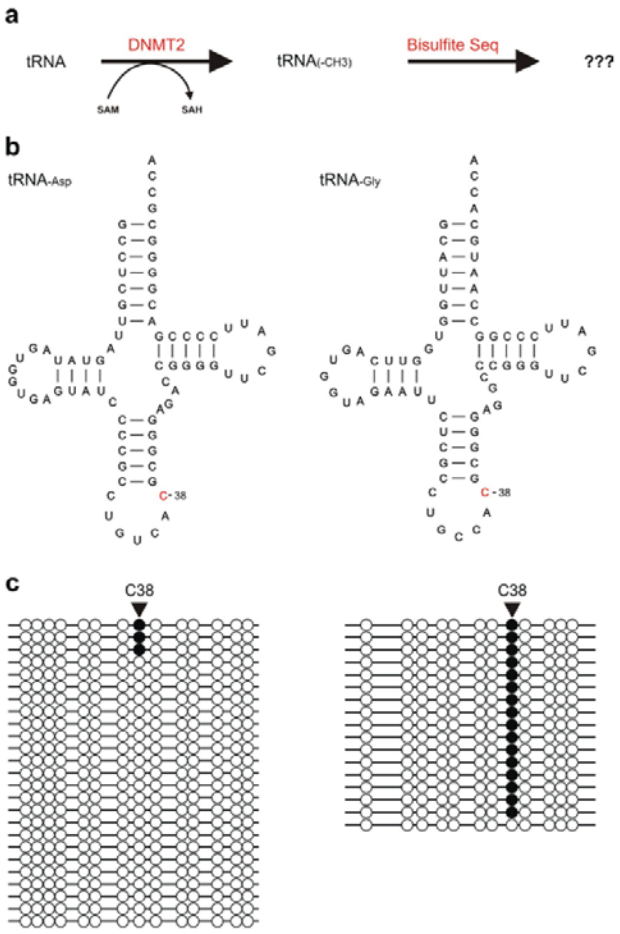
Supplementary Figure 3: Optimization of DNMT2 in-vitro MTase assay



Supplementary Figure 3: Optimization of DNMT2 in-vitro MTase assay. For assay conditions, see Supplementary Methods. (a) Replacing the EDTA (as done previously[1]) with MgCl2 (as done previously[2]) greatly increases the methyl-transfer rate by DNMT2 on $tRNA^{Asp}$. (b) In-vitro transcribed $tRNA^{Asp}$ is as effective as a substrate as synthetic $tRNA^{Asp}$ and was used as the positive control in subsequent assays. (c) DNMT2-MTase assay reveals robust methylation of $tRNA^{Gly}$, an additional DNMT2 tRNA target. Previous work has suggested requirements for DNMT2 action on tRNAs, including additional prior base modifications[1] or assisted folding[2]. Our experiments (a-c), along with the bisulfite sequencing in Supplementary Figure 3, demonstrate that tRNA substrates can be efficiently methylated without the need for prior base modifications[1], or transcribing the tRNA along with a self-cleaving hammerhead ribozyme[2]. Here, however, we note that our work does not rule out the possibility that tRNA modifications, or the use of the hammerhead ribozyme construct for producing the substrate tRNA, might still increase the methylation efficiency. (Note: The amount of enzyme in (b) and (c) is twice the amount used in (a) and the signal is proportionally higher). (CPM: Counts Per Minute).

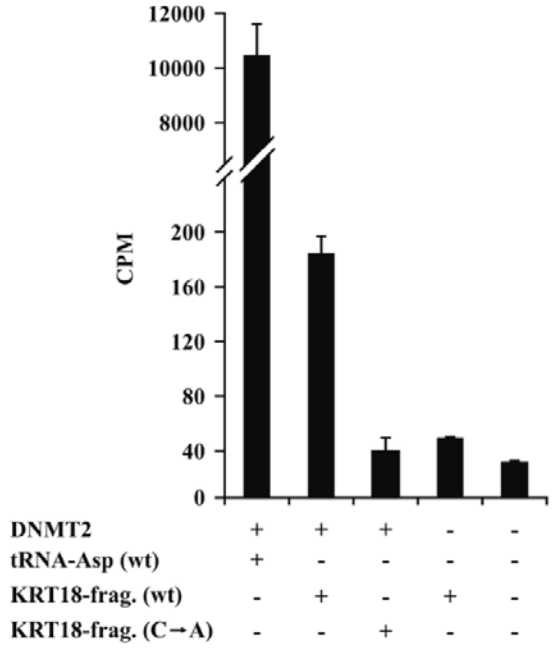
Nature Biotechnology: doi:10.1038/nbt.2566

Supplementary Figure 4: DNMT2 MTase assay coupled with PCR-based Bisulfite Sequencing



Supplementary Figure 4: DNMT2 MTase assay coupled with PCR-based Bisulfite Sequencing. (a) In-vitro methylation of the substrate tRNA (using the non-radioactive SAM) coupled with PCR-based bisulfite sequencing precisely reveals the single methylated cytosine. (b) Cloverleaf structure of $tRNA^{Asp}$ and $tRNA^{Gly}$ with the target cytosines (C38) in red color. (c) Cartoon representation of the bisulfite sequencing results. Open circles show converted cytosines after bisulfite treatment and closed circles show un-converted (methylated) cytosines residues. This shows that C38 residues in $tRNA^{Asp}$ (left) and in $tRNA^{Gly}$ (right) are the only target cytosines for DNMT2 enzyme in these two tRNA molecules under these conditions. In addition, this result along with the scintillation assay (**Supplementary Figure 2c**) shows that under these conditions $tRNA^{Gly}$ is a much better substrate for DNMT2 compared to $tRNA^{Asp}$, with almost complete methylation solely at C38.

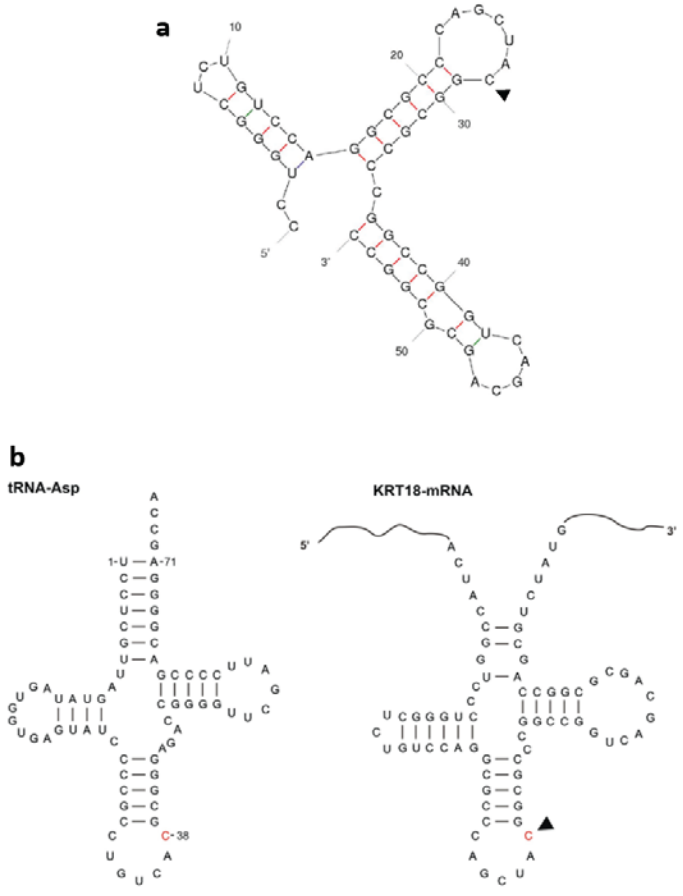
Supplementary Figure 5: DNMT2 dependent methylation of the candidate cytosine in the tRNA-like structure of KRT18 mRNA



Supplementary Figure 5: DNMT2 dependent methylation of the candidate cytosine in the tRNA-like structure of KRT18 mRNA. Human DNMT2 can methylate the tRNA-like structure of KRT18-mRNA (KRT18-frag) in-vitro, and the methylation is blocked by mutating the candidate residue. However, the methylation rate for KRT18-mRNA is much lower than the methylation rate for $tRNA^{Asp}$ control.

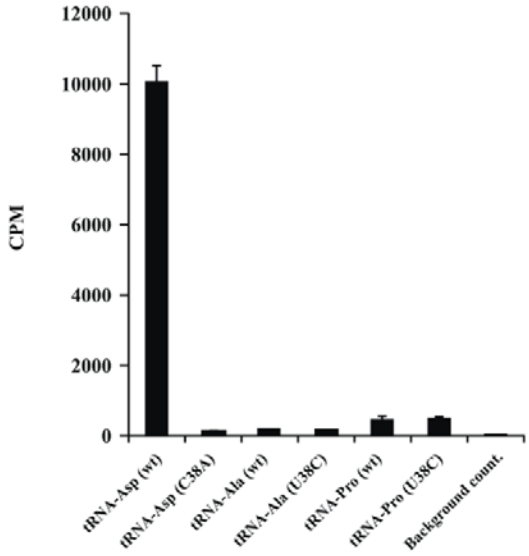
Nature Biotechnology: doi:10.1038/nbt.2566

Supplementary Figure 6: Mfold reveals a structural similarity of the KRT18 mRNA candidate target to known DNMT2 target tRNAs



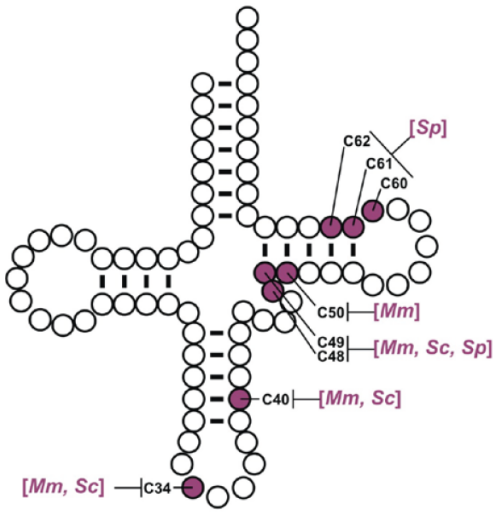
Supplementary Figure 6: Mfold reveals a structural similarity of the KRT18 mRNA candidate target to known DNMT2 target tRNAs (a) In-silico structural analysis of a 55-base fragment, consisting of the candidate cytosine residue (C28) flanked by 27 bases on either side using Mfold, a popular web-based tool for RNA folding analysis[3], predicted a stable cloverleaf tRNA-like structure (ΔG=-24.50 kcal/mol). The predicted structure places the candidate cytosine (in the CpG context) at the stem-loop junction, analogous to position C38 in the anticodon loop of the DNMT2 target tRNAs. (b) Manually adjusted structure shows that KRT18-mRNA can form a regional cloverleaf tRNA-like structure similar to DNMT2 target tRNAs such as $tRNA^{Asp}$. Note that only 75 bp of the KRT18-mRNA are shown from the mRNA. (Arrowheads point to the candidate cytosine).

Supplementary Figure 7: DNMT2 MTase assay on wt and mutant tRNA^{Asp}, tRNA^{Ala} and tRNA^{Pro}



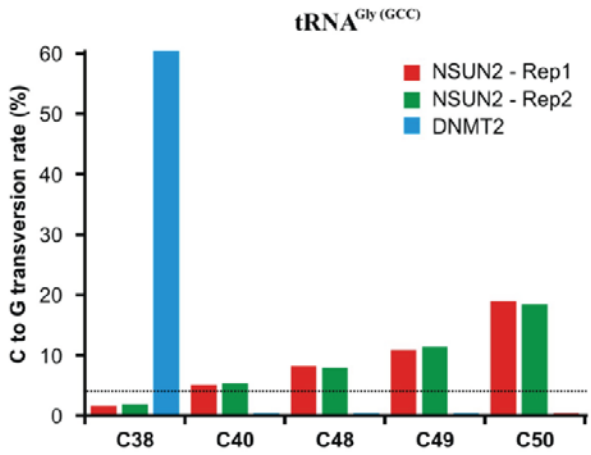
Supplementary Figure 7: DNMT2 MTase assay on wt and mutant tRNA^{Asp}, tRNA^{Ala} and tRNA^{Pro}. To test the hypothesized sufficiency of CpG context for defining the target specificity of DNMT2 in a tRNA structure, the in-vitro transcribed wild-type (wt) and mutant tRNA^{Ala} and tRNA^{Pro} were tested in the DNMT2 MTase assay using both of positive (wt) and negative (C38A) tRNA^{Asp} controls. The mutant tRNA^{Ala} and tRNA^{Pro} were chosen from the tRNA panel because making the CpG context required minimum changes in these two tRNAs, as they already have guanosine in the stem region. Therefore only the uracil at position 38 needed to be changed to cytosine (U38C) to make a CpG context. The MTase assay shows that DNMT2 enzyme does not recognize the cytosine residues in the CpG contexts generated in mutant tRNA^{Ala} and tRNA^{Pro}. This suggests that CpG context is required but not sufficient to define the in-vitro substrate specificity of DNMT2.

Supplementary Figure 8: Known NSUN2 target sites in mouse, budding yeast and fission yeast



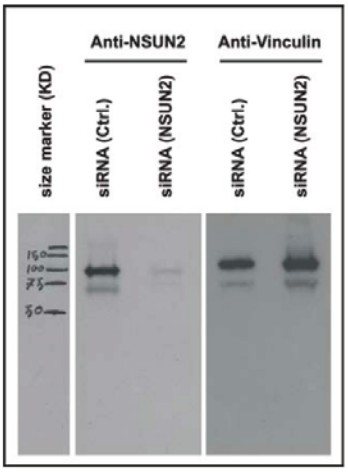
Supplementary Figure 8: Known NSUN2 target sites in mouse, budding yeast and fission yeast. A 'standardized' tRNA, summarizing the currently known NSUN2 target cytosines in the mouse (Mus musculus, Mm), budding yeast (Saccharomyces cerevisiae, Sc), and fission yeast (Schizosaccharomyces pombe, Sp)[4-6]. See text and references for details, as only particular tRNA types have been tested in each organism.

Supplementary Figure 9: Comparison of C>G transversion rates at single vs. multiple target sites



Supplementary Figure 9: Comparison of C>G transversion rates at single vs. multiple target sites. Comparison of C>G transversion rates at sole DNMT2 target site (C38) and multiple NSUN2 target sites (C40, C48, C49 & C50) for $tRNA^{Gly(GCC)}$. The horizontal dotted line shows the 4% transversion cut-off.

Supplementary Figure 10: RNAi-mediated hNSUN2 knockdown verification



Supplementary Figure 10: RNAi-mediated hNSUN2 knockdown verification. After two consecutive rounds of siRNA knockdown (first with 60pmol for day 1-3 and followed by 120pmol for day 4-6) with either of NSUN2 siRNA (siGENOME Human NSUN2 siRNA – SMARTpool (M-018217-01-0005)) or control siRNA (siGENOME Non-Targeting siRNA Pool #1 (D-001206-13-05)) 60 µg of protein from each sample was loaded on the SDS-PAGE. Protein extracts were evaluated by western blotting for knockdown efficiency by immuno-blotting with hNSUN2 polyclonal (Proteintech-20854-1-AP) antibody (middle panel) and then re-probed with hVinculin monoclonal antibody (Sigma-V9131) (right panel). The size marker (in KD) is provided in the left panel).

SUPPLEMENTARY TABLES

Supplementary Table 1: CpG context in the stem-loop junction of anticodon stem loop of *H. sapiens*, *M. musculus*, *A. thaliana* and *D. melanogaster* tRNAs

a.a*	*H. sapiens***	*M. musculus*	*A. thaliana*	*D. melanogaster*
Ala	TG	TG	TG	TG
Arg	AT	AT (AG)	AT/AA/AG	AT (AG)
Asn	AC (CT/AT)	AC	AC	AC
Asp	CG (TG)	CG	CG	CG
Cys	AT	AT	AT	AT
Gln	AT	AT	AT	AT
Glu	CC	CC	CC	CC
Gly	CG (AG)	CG (AG)	CG (AG)	CG (AG)
His	CC	CC	CC	CC
Ile	AC (AT)	AC (AT)	AC (AG)	AC (AT)
Leu	CT/TC (AT/TT)	CT/TC (AT/TT)	TT (CT)	CA/TC (CT/TT)
Lys	AT	AT	AT	AT
Met	AC (TC)	AC (TC)	AC (TC)	AC (TC)
Phe	AT	AT	AT	AT
Pro	TG	TG	TG	TG
SeC	AC (AT)	AC (AT)	--	AT
Ser	AT	AT	AT (AG)	AT
Thr	AA	AA	AG (AA)	AA
Trp	AT	AT	AT	AT
Tyr	AT	AT	AT	AT
Val	CG	CG	CA (CG)	CG

* Only the most prevalent dinucleotide compositions are reported for each tRNA type for simplicity. The less frequent dinucleotides have been reported within parenthesis and the very rare ones can be found in the database.

** There are 2 (out of 49) annotated human $tRNA^{Ala}$ and one (out of 35) annotated human $tRNA^{Arg}$ that show CpG dinucleotide in the stem-loop junction of anticodon stem loop. However, that of the $tRNA^{Arg}$ has been annotated as $tRNA^{Gly}$ in the UCSC genome database (http://genome.ucsc.edu/).

Supplementary Table 1: CpG context in the stem-loop junction of anticodon stem loop of *H. sapiens*, *M. musculus*, *A. thaliana* and *D. melanogaster* tRNAs. Comparing the aligned tRNA sequences (extracted from Genomic tRNA Database; http://gtrnadb.ucsc.edu/[7]) shows that mainly the three known DNMT2 tRNA targets (Asp, Gly and Val), have cytosine at position 38 in a CpG context in the stem-loop junction of their anticodon stem loops.

Supplementary Table 2: Oligonucleotide sequences for making the lentiviral expression vectors

Oligo Name	Sequence	Application
V5(DNMT2/DsRed)-F	ATTCCTCGAGCACCATGGGTAAGCCTATCCCTAACCCTCTCCTCGGTCTC	2nd PCR
hDNMT2-F	CTAACCCTCTCCTCGGTCTCGATTCTACGGAGCCCCTGCGGGTGCTGGAG	1st PCR
hDNMT2-R	GAATCGGATCCATCTTATTCATATAAGATTTTGATTAGTTTAGCTACTACATGCACGTTG	1st and 2nd PCR
DsRed-F	CTAACCCTCTCCTCGGTCTCGATTCTACGGCCTCCTCCGAGAACGTCATC	1st PCR
DsRed-R	GAATCGGATCCATCCTACAGGAACAGGTGGTGGCG	1st and 2nd PCR
V5-NSUN2-F	TATCACCGGTGCCACCATGGGTAAGCCTATCCCTAACCCTCTCCTCGGTCTC	2nd PCR
hNSUN2-F	CTAACCCTCTCCTCGGTCTCGATTCTACGGGGCGGCGGTC	1st PCR
hNSUN2-R	ATTAGCTAGCTCACCGGGGTGGATGGAC	1st and 2nd PCR

Supplementary Table 3: Sequences of RNA substrates used in the MTase assay

Name	Sequence (5'->3')
wt-tRNA^{Asp} *	UCCUCGUUAGUAUAGUGGUGAGUAUCCCCGCCUGUCACGCGGGAGACCGGGGUUCGAUUCCCCGACGGGGAGCCA
wt-tRNA^{Asp} (U1G, A71C)	GCCUCGUUAGUAUAGUGGUGAGUAUCCCCGCCUGUCACGCGGGAGACCGGGGUUCGAUUCCCCGACGGGGCGCCA
mt-tRNA^{Asp} (U1G, A71C & C38A)	GCCUCGUUAGUAUAGUGGUGAGUAUCCCCGCCUGUCAAGCGGGAGACCGGGGUUCGAUUCCCCGACGGGGCGCCA
wt-tRNA^{Gly}	GCAUUGGUGGUUCAGUGGUAGAAUUCUCGCCUGCCACGCGGGAGGCCCGGGUUCGAUUCCCGGCCAAUGCACCA
wt-tRNA^{Ala}	GGGGGUGUAGCUCAGUGGUAGAGCGCGUGCUUAGCAUGCACGAGGCCCCGGGUUCAAUCCCCGGCACCUCCACCA
mt-tRNA^{Ala} (U38C)	GGGGGUGUAGCUCAGUGGUAGAGCGCGUGCUUAGCACGCACGAGGCCCCGGGUUCAAUCCCCGGCACCUCCACCA
wt tRNA^{Pro}	GGCUCGUUGGUCUAGGGGUAUGAUUCUCGCUUAGGGUGCGAGAGGUCCCGGGUUCAAAUCCCGGACGAGCCCCCA
mt-tRNA^{Pro} (U38C)	GGCUCGUUGGUCUAGGGGUAUGAUUCUCGCUUAGGGCGCGAGAGGUCCCGGGUUCAAAUCCCGGACGAGCCCCCA
wt-KRT18-frag. (A1G)**	GCUACCGGUCCCUGGGCUCUGUCCAGGCGCCCAGCUACGGCGCCCGGCCGGUCAGCAGCGCGGCCAGCGUCUAUG
mt-KRT18-frag. (A1G & C38A)	GCUACCGGUCCCUGGGCUCUGUCCAGGCGCCCAGCUAAGGCGCCCGGCCGGUCAGCAGCGCGGCCAGCGUCUAUG

* This tRNA made synthetically and is the exact in-vivo tRNA^{ASP} sequence.

** This is the tRNA-like structure fragment of the KRT18 mRNA and the nucleotide number one is the first nucleotide of the fragment not the actual mRNA.

Supplementary Table 4: ssDNA templates and primer sets used to prepare dsDNA substrates for in-vitro transcription using T7 RNA polymerase

Name	Sequence (5'->3') *
tRNA^{Asp}	
wt tDNA^{Asp}	GCCTCGTTAGTATAGTGGTGAGTATCCCCGCCTGTCACGCGGGAGACCGGGGTTCGATTCCCCGACGGGGCGCCA
mt tDNA^{Asp} (C38A)	GCCTCGTTAGTATAGTGGTGAGTATCCCCGCCTGTCAAGCGGGAGACCGGGGTTCGATTCCCCGACGGGGCGCCA
Asp-IT-F(T7)	AAGCTTAATACGACTCACTATAGCCTCGTTAGTATAGTGGTG
Asp-IT-R	TGGCGCCCCGTCGGGGAATC
tRNA^{Gly}	
wt tDNA^{Gly}	GCATTGGTGGTTCAGTGGTAGAATTCTCGCCTGCCACGCGGGAGGCCCGGGTTCGATTCCCGGCCAATGCACCA
Gly-IT-F(T7)	AAGCTTAATACGACTCACTATAGCATTGGTGGTTCAG
Gly-IT-R	TGGTGCATTGGCCGGGAATCG
KRT18-mRNA-fragment	
wt KRT18-frag.	CGACTCACTATAGCTACCGGTCCCTGGGCTCTGTCCAGGCGCCCAGCTACGGCGCCCGGCCGGTCAGCAG
mt KRT18-frag. (C38A)	CGACTCACTATAGCTACCGGTCCCTGGGCTCTGTCCAGGCGCCCAGCTAAGGCGCCCGGCCGGTCAGCAG
KRT18-frag-F(T7)	GAGCGTAATACGACTCACTATAGCTACCGGTCCCT
KRT18-frag-R	CATAGACGCTGGCCGCGCTGCTGACCGGCC
tRNA^{Ala}	
wt tDNA^{Ala}	GGGGGTGTAGCTCAGTGGTAGAGCGCGTGCTTAGCATGCACGAGGCCCCGGGTTCAATCCCCGGCACCTCCACCA
mt-tDNA^{Ala} (U37C)	GGGGGTGTAGCTCAGTGGTAGAGCGCGTGCTTAGCACGCACGAGGCCCCGGGTTCAATCCCCGGCACCTCCACCA
Ala-IT-F(T7)	AAGCTTAATACGACTCACTATAGGGGGTGTAGCTCAGTGG
Ala-IT-R	TGGTGGAGGTGCCGGGGA
tRNA^{Pro}	
wt tDNA^{Pro}	GGCTCGTTGGTCTAGGGGTATGATTCTCGCTTAGGGTGCGAGAGGTCCCGGGTTCAAATCCCGGACGAGCCCCCA
mt-tDNA^{Pro} (U37C)	GGCTCGTTGGTCTAGGGGTATGATTCTCGCTTAGGGCGCGAGAGGTCCCGGGTTCAAATCCCGGACGAGCCCCCA
Pro-IT-F(T7)	AAGCTTAATACGACTCACTATAGGCTCGTTGGTCTAGGGG
Pro-IT-R	TGGGGGCTCGTCCGGGATTTGA

* The T7 promoter sequence is underlined and in blue color. The target/candidate residues are in red color.

Supplementary Table 5: Bisulfite specific primer sets used for validation of MTase assay

Primer name	Sequence (5'->3')
BS-IVT-Asp-F	CCCATACTCACTAACACCCCATCAA
BS-IVT-Asp-R	GGTTGGGATGAGGTTTTGTTAGTATAGTGG
BS-IVT-Gly-F	CCCATACTCACTAATACATTAACCAAAAATCAAACC
BS-IVT-Gly-R	GGTTGGGATGAGGTATTGGTGG

Supplementary Table 6: Bisulfite specific primer sets used for validation of NSUN2 ncRNA targets in RNAi knockdown experiment

Primer name	Sequence (5'->3')
BS-RPPH1-F	GTTTTAATAGGGTTTTTTTTGAGTTTTGGGGAGGTGAGTTTTTAG
BS-RPPH1-R	TAAATCTATTCCAAACTCCAACAAAAAAAACATCCACCAAACC
BS-RPPH1-F-nested	GGAGGTGAGTTTTTAGAGAATGGGGTTTTGTGTG
BS-RPPH1-R-nested	CCACCAAACCCCTCCCCAAAAAACAAAATC
BS-SCARNA2-F	TTGTGAAGTTTTTTTGGGGTGTTGTGTAGTGAGG
BS- SCARNA2-R	AAAAACAAACCCAACCTCATCTATCTAATCAATTCATCACTTCTA
BS- SCARNA2-F-nested	GTTGTGTAGTGAGGTTTTTTAGGTGGTGGTTATGTTG
BS- SCARNA2-R-nested	CTAATCAATTCATCACTTCTAAACACCAAACCCCACA
BS-VTRNA1-1-F	GGTTGGTTTTAGTTTAGTGGTTATTTTGATAG
BS- VTRNA1-1-R	AAAAAAACTAAAAAACACCCACAAATCTC
BS-tRNA-GlyGCC-F	GGGATGAGGTATTGGTGGTTTAGTGGTAG
BS-tRNA-GlyGCC-R	ACCCATACTCACTAATACATTAACCAAAAATCAAACCC
BS-tRNA-GlyGCC-F-nested	GGATGTATTGGTGGTTTAGTGGTAGAATTTTTGTTTG

REFERENCES:

1. Goll, M.G. et al. Methylation of tRNAAsp by the DNA methyltransferase homolog Dnmt2. *Science* **311**, 395-398 (2006).
2. Jurkowski, T.P. et al. Human DNMT2 methylates tRNA(Asp) molecules using a DNA methyltransferase-like catalytic mechanism. *RNA* **14**, 1663-1670 (2008).
3. Zuker, M. Mfold web server for nucleic acid folding and hybridization prediction. *Nucleic Acids Res* **31**, 3406-3415 (2003).
4. Tuorto, F. et al. RNA cytosine methylation by Dnmt2 and NSun2 promotes tRNA stability and protein synthesis. *Nat Struct Mol Biol* **19**, 900-905 (2012).
5. Becker, M. et al. Pmt1, a Dnmt2 homolog in Schizosaccharomyces pombe, mediates tRNA methylation in response to nutrient signaling. *Nucleic Acids Res* (2012).
6. Motorin, Y. & Grosjean, H. Multisite-specific tRNA:m5C-methyltransferase (Trm4) in yeast Saccharomyces cerevisiae: identification of the gene and substrate specificity of the enzyme. *RNA* **5**, 1105-1118 (1999).
7. Chan, P.P. & Lowe, T.M. GtRNAdb: a database of transfer RNA genes detected in genomic sequence. *Nucleic Acids Res* **37**, D93-97 (2009).

Nature Biotechnology: doi:10.1038/nbt.2566