5-methylcytidine has a complex, context-dependent role in RNA

Andrew Mark Shafik

April 2017

A thesis submitted for the degree of
Doctor of Philosophy

The Australian National University

Originality Statement

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

Signed …….

Date …….

Acknowledgements

I am overwhelmed. I have finally completed this milestone, three and a half years of my life culminating in this body of work. The path hasn't been easy but I have finally reached the end. Of course, this would not have been possible without the support of many people.

To my supervisor, Thomas Preiss, thank you for reading the thesis and the stimulating discussions.

To Maurits Evers, thank you so much for the collaboration, for the bioinformatics support concerning the metagene analyses. This gave another dimension to the thesis. Thanks for reading my thesis, for the many valuable discussion we had, and for the coffee breaks.

To Ulrike Schumann, thank you for your support and guidance. I really appreciate it. Thanks for reading my thesis and your positive feedback.

To Yalin Liao and Eloisa Pagler, thanks for making me feel welcome and helping me to find my way around the lab. Thank you for being the voice of reason when times were tough.

To Tennille Sibbritt, thanks for the support early on, and showing me some techniques in the lab.

To Brian Parker thanks for bioinformatics support concerning the AGO2-m^5C association.

To other members of the lab, Rina Soetanto, Chikako Ragan and Nikolay Shirokikh thanks for the lab get-togethers and coffee breaks.

To my Dad, Mum and sister. Many thanks for your unfailing and undying support. Thank you for being a soundboard when things got tough. I couldn't have done this without you all.

Abstract

Ribonucleic acid (RNA) metabolism processes and function are affected by specific RNA sequence motifs, the ability of RNA to form secondary structure and assemble into ribonucleoprotein (RNP) complexes (Dandekar & Bengert, 2002). As these aspects are likely to be affected by nucleoside modifications, it is important to document the extent and function of post-transcriptional modifications present in these molecules. Thus, there is an increasing focus on exploring the incidence and biological relevance of post-transcriptional marks in RNA.

This thesis is aimed at defining a role for the post-transcriptional modification, 5-methylcytidine (m^5C) particularly in mRNAs. Bisulfite treatment, which allows for the nucleotide-specific identification of m^5C, coupled with next generation sequencing allowed for the detection of m^5C sites transcriptome-wide at single-nucleotide resolution. This revealed candidate m^5C sites in many RNA biotypes including both non-coding and coding RNA in cervical cancer cells (HeLa). In mRNAs, m^5C candidate sites were observed in both untranslated regions and in the coding sequence, suggesting multiple roles for m^5C. Global analyses revealed features of the m^5C modification in nuclear-encoded mRNAs. m^5C sites were shown to be relatively enriched in the 5′ UTR, to be associated with high GC content, low minimum free energy structures, and weakly translated mRNAs. Furthermore, experimental and bioinformatic approaches suggested a role for m^5C in regulating mRNA translation by recruiting specific m^5C-binding proteins, which were identified through a RNA bait pulldown approach. m^5C was also suggested to function in stabilising target mRNA. Bioinformatic and experimental approaches indicated m^5C might physically block the destabilising effect of Argonaute 2/microRNA binding. In this way, m^5C would protect the transcript from degradation. Also, further work identified the m^5C modification in mitochondrial (mt) rRNA, where a function of m^5C is discussed, in novel structural positions in mt-tRNAs, but not in mt-mRNAs. Altogether, this thesis suggests that m^5C has a complex, context-dependent role in RNA.

Abbreviations used in this thesis

Ψ	Pseudouridine
5-aza-IP	5-azacytidine-mediated RNA immunoprecipitation
AdoMet	S-adenosyl L-methionine (SAM)
AGO2	Argonaute 2
ALKBH	α-ketoglutarate-dependent dioxygenase alkB homolog
bp	Base pairs
BR	Broad range
BSA	Bovine serum albumin
bsRNA-seq	Bisulfite RNA sequencing
cDNA	Complementary DNA
CDS	Coding sequence
CIMS-miCLIP	Cross-linking-induced mutation sites miCLIP
CITS-miCLIP	Cross-linking-induced truncation miCLIP
CLIP	Crosslinking and immunoprecipitation
Co-IP	Co-immunoprecipitation
dH_2O	RNase, DNase free water
dsDNA	Double stranded DNA
DNA	Deoxyribonucleic acid
DNase	Deoxyribonuclease
DNMT	DNA methyltransferase
DTT	Dithiothreitol
ECL	Enhanced chemiluminescence
EDTA	Ethylenediamine tetracetic acid
EZR	Ezrin
f^5C	5-formylcytosine
FBS	Fetal bovine serum
FDR	False discovery rate
FTO	Fat mass and obesity associated protein
GO	Gene ontology
hr	Hours
HEK293T	Human embryonic kidney 293 cells
HeLa	Human cervical cancer cell line

HITS-CLIP	High-throughput sequencing of RNA isolated by crosslinking immunoprecipitation
hm^5C	5-hydroxymethylcytosine
HuR	Human antigen R
IGV	Integrated Genomics Viewer
IP	immunoprecipitation
KD	knockdown
kDa	Kilodalton
LSU	Large subunit
m^5C	5-methylcytidine
m^1A	N1-methyladenosine
m^6A	N6-methyladenosine
LC-MS/MS	Liquid chromatography mass spectrometry
lncRNA	Long non-coding RNA
M	Molar
MeRIP	Methyl-RNA immunoprecipitation
METTL3	Methyltransferase like 3
METTL14	Methyltransferase like 14
miCLIP	Methylated individual-nucleotide-resolution crosslinking and immunoprecipitation
min	Minute
miRNA	MicroRNA
ml	Millilitre
mM	Millimolar
mRNA	Messenger RNA
mt	Mitochondrial
MTase	Methyltransferase
MYC	Myelocytomatosis Viral Oncogene Homolog
NaOAc	Sodium acetate
ncRNA	Non-coding RNA
ng	Nanograms
NGS	Next generation sequencing
nM	Nanomolar
Nop	Nucleolar protein
NSUN	NOP2/Sun domain family, member

nt	Nucleotide
NTC	Non-targeting control
P body	Processing body
PAR-CLIP	Photoactivatable ribonucleoside-enhanced CLIP
PBS	Phosphate buffered saline
PBST	Phosphate buffered saline + Tween20
PCR	Polymerase chain reaction
Poly(A)	Polyadenylation
pre-miRNA	Precursor miRNA
pri-miRNA	Primary miRNA
R-Luc	Renilla luciferase
RBP	RNA binding protein
RCMT	RNA (cytosine-5)-methyltransferase
RDX	Radixin
Ribo-Seq	Ribosome profiling sequencing
RIPA	Radioimmunoprecipitation assay buffer
RISC	RNA-induced silencing complex
RNP	Ribonucleoprotein
RNA	Ribonucleic acid
RNA-seq	RNA sequencing
RNAi	RNA interference
RNase	Ribonuclease
RPKM	Reads per kilobase of transcript per million mapped reads
rRNA	Ribosomal RNA
RT	Reverse transcriptase
RT-qPCR	Real time quantitative PCR
SD	Standard deviation
SDS-PAGE	Sodium dodecyl sulfate polyacrylamide gel electrophoresis
sec	Seconds
siRNA	Small interfering RNA
SNP	Single nucleotide polymorphism
snRNA	Small nuclear RNA
snoRNA	Small nucleolar NRA
SSU	Small subunit
TET	Ten-eleven translocase

TRDMT1	tRNA aspartic acid methyltransferase 1
tRNA	transfer RNA
µl	Microlitres
UV	Ultraviolet
UTR	Untranslated region
V	Volts
vtRNA	Vault RNA
WTAP	Wilms tumor 1 associated protein
w/v	Weight per volume
YTDHF	YTH domain family

List of Figures

Figure 1.1: Schematic illustrating the cap and internal modifcations known in mRNAs. 3
Figure 1.2: Schematics illustrating methods to detect m^6A (immunoprecipitation) and m^5C (bsRNA-seq, 5-Aza-IP, miCLIP). 8
Figure 1.3: Schematic illustrating 'writers', 'readers' and 'erasers' of the m^5C modification. 12
Figure 1.4: Function of m^6A and m^5C in ncRNA (taken from my review: Shafik *et al*., 2016). 19
Figure 1.5: Schematic outlining the various roles of m^6A in mRNA translation (adapted from my review: Schumann *et al*., 2016). 22
Figure 1.6: Role of m^1A in mRNA translation. 24
Figure 1.7: Examples where m^5C methylation has both a repressive and enhancing affect on translation in mRNAs 27
Figure 3.1: Human mitochondrial genome 55
Figure 3.2: BLAST search of human mitochondrial genome revealing a number of highly similar sequences to nuclear encoded genes 56
Figure 3.3: Quality control measures to ensure purity of mitoplast isolation. 58
Figure 3.4: Illumina bsRNA-seq coverage across the mitochondrial transcriptome…...60
Figure 3.5: Integrated Genomics Viewer (IGV) screenshot of the called 16S m^5C sites in the Illumina bsRNA-seq experiment 64
Figure 3.6: Integrated Genomics Viewer (IGV) screenshot showing the called candidate m^5C site in cytochrome c oxidase I mRNA (chrM:6057). 65
Figure 3.7: Quality control measures to ensure successful knockdown of NSUN4..... 68
Figure 3.8: MiSeq results expectedly showing no methylation in the negative control. 70
Figure 3.9: MiSeq results showing methylation at expected positions in the positive control $tRNA^{Asp(GUC)}$ 71
Figure 3.10: Primer design strategy. 72
Figure 3.11: Pooled amplicons (pre-library preparation) and final libraries (post-library preparation) for each sample were electrophoresed on a 2% agarose gel 75
Figure 3.12: MiSeq result validating m^5C site C841 in 12S mt-rRNA and showing it to be NSUN4 controlled. 76
Figure 3.13: MiSeq result showing no methylation at the candidate site in 16S mt-rRNA 77
Figure 3.14: Identified m^5C sites in mitochondrial (mt) tRNAs 78

Figure 3.15: Schematic illustrating the proposed role for m^5C in mt-12S rRNA (adapted with modifications from Metodiev *et al*., 2014). 83
Figure 3.16: Secondary structure prediction of a section of helix 44 in 12S rRNA..... 86
Figure 3.17: Schematic representation of the domain structures present in NSUN2 and NSUN4. 89
Figure 4.1. Overlapping HeLa AGO2 and HuR footprint densities with candidate m^5C sites in the CDS and 3≠ UTR. 93
Figure 4.2: Candidate m^5C sites distributed across quartiles and overlapped with AGO2 HITS-CLIP HeLa data 95
Figure 4.3: m^5C sites are depleted in conserved 7-mer seeds over CDS and 3≠ UTR in HeLa. 96
Figure 4.4: Depletion of miRNA seed match density at all methylation levels 98
Figure 4.5: Evidence of efficient AGO2 and NSUN2 immunoprecipitation. 99
Figure 4.6: Optimisation of the AGO2 co-IP experiment 101
Figure 4.7: Optimisation of co-immunoprecipiptation – GW182 associates with AGO2. 102
Figure 4.8: Confirmation that α-tubulin associates with NSUN2. 103
Figure 4.9: NSUN2 interacts with AGO2 in an RNase-sensitive manner. 105
Figure 4.10: Predicted secondary structures of candidate RNA baits using RNAfold. 107
Figure 4.11: Optimisation of the RNA pulldown experiment 109
Figure 4.12: NSUN2 interacts with both baits, GAPDH does not 110
Figure 4.13: Scatterplot showing number of identified peptides associated with each of the baits using the NSUN2 knockdown lysate 112
Figure 4.14: Western blotting confirming mass spectrometry results 114
Figure 5.1: m^5C sites identified in tRNAs 129
Figure 5.2: Example of incorrectly assigning an m^5C site 131
Figure 5.3: Pie chart showing the distribution of m^5C candidate sites within defined mRNA regions 132
Figure 5.4: The distribution of m^5C methylated RNAs categorised by methylation stoichiometry. 133
Figure 5.5: Enrichment/depletion plot of m^5C sites across mRNA regions……..137
Figure 5.6: Absolute spatial arrangement of m^5C and non-m^5C in the 5≠ UTR and CDS. 139
Figure 5.7: GC content of mRNAs with and without the m^5C modification. 143

Figure 5.8: Relatively enriched sequence motifs identified around the m^5C site using MEME and DREME 145
Figure 5.9: Minimum free energies (MFEs) of m^5C-containing transcripts compared to unmethylated transcripts by mRNA region. 146
Figure 5.10: mRNA translation efficiency scores of m^5C containing mRNAs compared to mRNAs completely lacking m^5C across the transcript 149
Figure 5.11: A comparison of translation efficiencies of the most structured m^5C containing mRNAs compared to the most structured mRNAs devoid of m^5C. 150
Figure 5.12: Methylation rate (i.e. number of transcripts containing m^5C compared to the number of transcripts without m^5C) per each mRNA region. 151

List of Tables

Table 1.1: RCMT subfamily and known substrates of NSUN enzymes……....14
Table 2.1: Real time primers used in this thesis 39
Table 2.2: List of bisulfite PCR primers used in this thesis 43
Table 3.1: Identified candidate m^5C sites in mitochondrial RNA from the SOLiD bsRNA-seq transcriptome-wide mapping. 61
Table 3.2: Identified candidate m^5C sites in mitochondrial RNA from Illumina bsRNA-seq transcriptome-wide mapping. 63
Table 4.1: Differentially identified proteins showing statistical significance at p-value < 0.01 when incubating both baits with wild type protein lysate 113
Table 4.2: Differentially identified proteins showing statistical significance at p-value < 0.01 when incubating both baits with NSUN2 knockdown lysate 113
Table 5.1: Read statistics of HeLa 1 and 2 bsRNA-seq libraries 126
Table 5.2: Mapping statistics of HeLa 1 and 2 bsRNA-seq libraries 126
Table 5.3: Gene ontology analysis of Illumina candidate m^5C sites in mRNAs..... 135

Table of Contents

Originality Statement II
Acknowledgements III
Abstract IV
Abbreviations used in this thesis V
List of Figures IX
List of Tables XII
Table of Contents XIII
Chapter I Introduction 1
1.1 Extent and distribution of the post-transcriptional modifications m^6A, m^1A and m^5C 4
1.1.1 Nuclear encoded RNA 4
1.1.1.1 m^6A 4
1.1.1.2 m^1A 5
1.1.1.3 m^5C 5
1.1.2 Mitochondrial RNA 6
1.2 'Writers', 'Readers' and 'Erasers' 9
1.2.1 m^6A 10
1.2.2 m^1A 13
1.2.3 m^5C 13
1.3 Function of the m^6A, m^1A and m^5C modifications in non-coding and coding RNA. 16
1.3.1 Non-coding RNA 16
1.3.1.1 m^6A 15
1.3.1.1.1 MALAT1 16
1.3.1.1.2 miRNA processing 17
1.3.1.2 m^1A 18
1.3.1.3 m^5C 18
1.3.1.3.1 Vault RNAs 18
1.3.1.3.2 XIST and HOTAIR long non-coding RNAs 18
1.3.2 Coding RNA 20
1.3.2.1 m^6A 20
1.3.2.2 m^1A 23
1.3.2.3 m^5C 24
1.3.2.3.1 Role for m^5C in mRNA stability 25
1.3.2.3.2 Role for m^5C in mRNA translation 26
1.4 Aims 28
Chapter II Materials and Methods 29

2.1 Materials 30
2.1.1 Chemicals, reagents and consumables 30
2.1.2 Kits 32
2.1.3 Antibodies 32
2.1.4 Oligonucleotides 33
2.1.5 siRNAs 33
2.2 Methods 34
2.2.1 Unpublished SOLiD bsRNA-seq dataset **34**
2.2.1.1 Alignment strategy and identification of candidate m^5C sites 34
2.2.2 Illumina bsRNA-seq dataset **35**
2.2.2.1 Alignment strategy and identification of candidate m^5C sites 35
2.2.3 Maintaining HeLa cells **36**
2.2.4 Identifying m^5C sites in mitochondrial RNA **36**
2.2.4.1 Mitochondrial and mitoplast isolation 36
2.2.4.2 siRNA transfection of HeLa cells 37
2.2.4.3 RNA extraction and precipitation 37
2.2.4.4 DNase treatment of RNA 38
2.2.4.5 cDNA synthesis 38
2.2.4.6 RT-qPCR 38
2.2.4.7 Whole cell lysis (protein isolation) 40
2.2.4.8 Protein quantification using BCA assay 40
2.2.4.9 Western blotting 40
2.2.4.10 Densitometry of western blots (ImageJ analysis) 41
2.2.5 Validation of m^5C candidate sites **41**
2.2.5.1 Bisulfite conversion of RNA 41
2.2.5.2 Bisulfite PCR 42
2.2.5.3 Amplicon sequencing library preparation for MiSeq 42
2.2.5.4 Quality control after library preparation 44
2.2.5.5 Quantification of libraries 44
2.2.5.6 Analysis of amplicon sequencing 44
2.2.5.7 Generating coverage plots and heatmaps in R programming language 44
2.2.6 Molecular Function of m^5C **45**
2.2.6.1 Overlapping HITS-CLIP data and candidate m^5C sites 45
2.2.6.2 Overlapping conserved miRNA target sites with candidate m^5C sites 45
2.2.6.3 Co-immunoprecipitation of AGO2 and NSUN2 45
2.2.6.4 RNA pulldown approach 46
2.2.7 Investigating features of m^5C methylation in mRNA **47**
2.2.7.1 Gene ontology analysis 47

2.2.7.2 RNAModR 47
Overview of datasets used in this thesis 49
Chapter III Identifying the mitochondrial m^5C methylome 52
3.1 Preamble **53**
3.1.1 Mitochondrial genome 54
3.2 Results **54**
3.2.2 Candidate mitochondrial m^5C sites 59
3.2.2.1 SOLiD bsRNA-seq 59
3.2.2.2 Illumina bsRNA-seq 62
3.2.3 Validating candidate m^5C sites 66
3.2.3.1 Knockdown of NSUN4 in HeLa cells 66
3.2.3.2 Verification of efficient bisulfite conversion 66
3.2.3.3 PCR amplification and library preparation 69
3.2.3.4 Validation of candidate m^5C sites using MiSeq 73
3.2.4 m^5C sites in human mitochondrial tRNAs 74
3.3 Discussion **74**
3.3.1 A validation tool is essential 74
3.3.1.1 Validation approach 80
3.3.2 Possible role for m^5C in mt-RNAs 80
3.3.3 NSUN4 targets human 12S mitochondrial rRNA 81
3.3.4 Function of post-transcriptional modifications in 12S mt-rRNA 81
3.3.5 Role of m^5C in 12S rRNA 82
3.3.6 NSUN4 target recognition 84
3.3.7 NSUN4 likely does not methylate human mitochondrial 16S rRNA 85
3.3.8 Protein domains of NSUN2 and NSUN4 87
3.4 Future Directions – Outlook/open-ended questions **87**
Chapter IV m^5C-associated proteins mediate m^5C function 90
4.1 Preamble **91**
4.2 Results **91**
4.2.1 Potential role for m^5C in mRNA stability 91
4.2.1.1 Argonaute 2 footprints correlate with m^5C candidate sites 92
4.2.1.2 miRNA depletion at m^5C candidate sites 94
4.2.1.3 Is AGO2 involved in m^5C formation? 97
4.2.1.3.1 Specificity and Efficiency of the immunoprecipitation 97
4.2.1.3.2 Optimisation of co-immunoprecipitation 97
4.2.1.3.3 Evidence for co-immunoprecipitation 100
4.2.1.3.4 AGO2 interacts with NSUN2 in an RNase sensitive manner 100
4.2.2 Potential role for m^5C in mRNA translation 104
4.2.2.1 m^5C-binding proteins 104

4.2.2.1.1 Selecting the RNA bait 104
4.2.2.1.2 Optimising the RNA bait pulldown 106
4.2.2.2 Candidate m^5C 'readers' 108
4.3 Discussion 115
4.3.1 AGO2 and m^5C sites 115
4.3.2 m^5C binding proteins 116
4.3.2.1 RNA bait experiment 116
4.3.2.2 NSUN2 is a possible m^5C 'reader' 117
4.3.2.3 Ezrin and Radixin are possible m^5C 'readers' 119
4.3.2.4 Ezrin/Radixin– a role in promoting translation of m^5C-containing mRNAs 120
4.4 Future Directions 121
Chapter V Metagene analyses 122
5.1 Preamble 123
5.2. Results 125
5.2.1 Summary of the Illumina bsRNA-seq HeLa cell datasets 125
5.2.2 Extent of m^5C modification in the human transcriptome 127
5.2.3 Gene ontology analysis of m^5C containing mRNAs 134
5.2.4 Spatial localisation of m^5C sites within mRNAs 134
5.2.4.1 Distribution of m^5C sites across mRNA regions 136
5.2.4.2 Distribution of m^5C sites within each mRNA region 136
5.3 Investigating the function of m^5C in mRNAs 138
5.3.1 m^5C is associated with GC rich sequences 140
5.3.2 Sequence Motifs 141
5.3.3 m^5C is associated with low minimum free energy structures 144
5.3.4 m^5C is associated with poorly translated mRNAs 147
5.4 Discussion 152
5.4.1 Transcriptome-wide map of m^5C sites in HeLa cells 152
5.4.2 Methylation level 153
5.4.3 m^5C in mRNAs 154
5.4.4 m^5C and mRNA translation rates 154
5.5 Future Directions – Outlook/open-ended questions 155
Chapter VI General Discussion and Future Directions 157
6.1 General Discussion 158
6.2 Future Directions 162
Chapter VII Appendix 164
7.1 Sequence for non-humanised *R-Luc in vitro* transcript. 165
7.2 Heatmaps showing candidate mitochondrial m^5C sites not to be methylated 166
7.3 List of RNA binding proteins identified following mass spectrometry of methylated and unmethylated RNA baits 167

7.4 Scatterplot showing number of peptides associated with the methylated and unmethylated baits when incubating with wild type lysate……..167
7.5 RNAModR example code..... 168
7.6 The bash script for the alignment of Illumina bsRNA-seq reads. 169
7.7 List of identified m^5C sites identified in the Illumina bsRNA-seq experiment..... 170
7.8 Full list of gene ontology terms obtained using the complete set of m^5C sites identified in mRNAs from the Illumina bsRNA-seq experiment 170
7.9 Spatial distribution of m^5C sites in 3' UTR. 170
References 171

Chapter I

Introduction

This Introduction features work from two review articles that I published during my PhD: Schumann *et al.* (2016) and Shafik *et al.* (2016).

Currently, more than 140 post-transcriptionally modified nucleosides have been reported in RNA (Cantara *et al.*, 2011; Machnicka *et al.*, 2013). Detection of post-transcriptional modifications in abundant RNA biotypes including transfer (t) RNA, ribosomal (r) RNA and small nucleolar RNA (snoRNA) has been a major focus of the field. However, recent studies, driven by technological advances in next generation sequencing (NGS), have further ventured into identifying modifications in less abundant RNA types such as messenger (m) RNAs and long non-coding (lnc) RNAs – forming the new research area of 'epitranscriptomics' (Frye *et al.*, 2016). Early on, though, mRNA cap modifications were identified. Such modifications are represented by N7-methylguanosine (m^7G) at the 5′ cap (Rottman *et al.*, 1974), N6-methyl-2′-O-methyladenosine (m^6Am) at the first position of the 5′ terminus (Wei *et al.*, 1976) and 2′-O-methylated nucleosides (Nm = Am, Cm, Um, Gm) on the first two positions of the 5′ terminus (Wei *et al.*, 1976) (Fig. 1.1). Recently, there has been an intense effort to globally map internal mRNA modifications. At the moment, four internal post-transcriptional modifications have been identified and mapped transcriptome-wide in mRNAs: 5-methylcytidine (m^5C), N6-methyladenosine (m^6A), N1-methyladenosine (m^1A), pseudouridine (Ψ), (Fig. 1.1) forming the 'epitranscriptomics' research area (Saletore *et al.*, 2012). Although not strictly a modification, A-to-I editing also occurs in mRNAs and has been mapped transcriptome-wide (Levanon *et al.*, 2004); C-to-U editing has also been observed in mRNAs.

Of course, there may be novel post-transcriptional modifications in mRNAs. Indeed, present mining of recent experiments aimed at elucidating proteins that bind poly(A) RNA (termed interactome studies) have demonstrated the scope of novel post-transcriptional modifications in mRNAs. In this thesis, mining of a HeLa cell interactome (Castello *et al.*, 2012), has revealed the possibility of novel post-transcriptional modifications in mRNAs. That is, RNA modifying enzymes previously unknown to target mRNAs were identified in the interactome, suggesting that these enzymes may modify mRNAs. Four RNA modifying enzymes were identified, responsible for the N6-dimethyladenosine ($m^6{}_6A$), dihydrouridine, 5-methyluridine (m^5U), and N2, N2- dimethylguanosine ($m^2{}_2G$); these modifications have not been described in mRNAs. Developing methods to identify these novel post-transcriptional modifications will provide a greater insight into the complexity of the transcriptome,

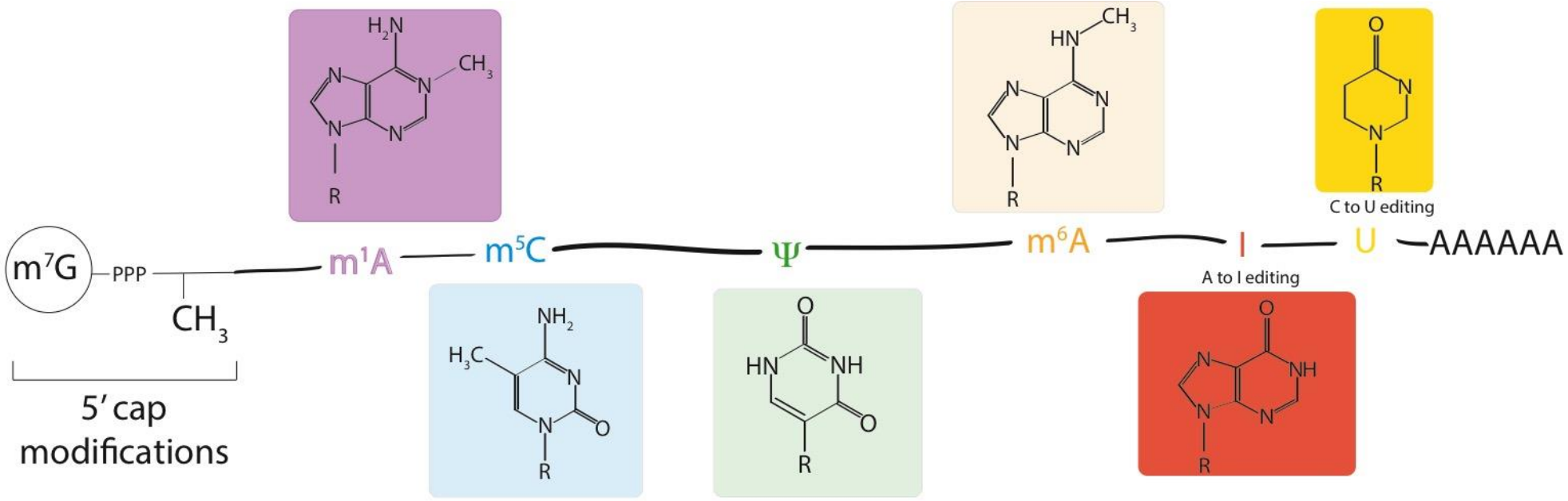


Figure 1.1: Schematic illustrating the cap and internal modifications known in mRNAs.

Schematic shows post-transcriptional modifications detected in mRNAs. A-I editing is not strictly a modification but also occurs in mRNAs. m^6A is enriched around the stop codon. m^1A and m^5C are enriched in the 5≠ UTR.

Cap modifications: m^7G = 7-methylguanosine; CH_3 = 2÷-O-methylation

Internal modifications: m^1A = N1-methyladenosine; m^5C = 5-methylcytidine; Ψ = pseudouridine; I = inosine; m^6A = N6-methyladenosine

with the possibility of revealing an interplay between the various post-transcriptional modifications. For example, is a single mRNA targeted for multiple modifications? How are these modifications spatially localised? What is the affect on mRNA fate?

1.1 Extent and distribution of the post-transcriptional modifications m^6A, m^1A and m^5C

1.1.1 Nuclear encoded RNA

1.1.1.1 m^6A

The first transcriptome-wide m^6A-seq maps identified a prevalent, dynamic (reversible) modification present in thousands of mRNAs in mammalian, yeast and plant cells (Dominissini *et al.*, 2012; Meyer *et al.*, 2012; Bodi *et al.*, 2015; Luo *et al.*, 2014). The approach used to identify this modification is based on the immunoprecipitation (IP) of fragmented, poly(A)$^+$ RNA using m^6A-specific antibodies prior to NGS (MeRIP-seq, m^6A-seq) (Dominissini *et al.*, 2012; Meyer *et al.*, 2012) (Fig. 1.2A). These studies identified in excess of 12,000 m^6A peaks residing in both mRNAs and ncRNA. However, improvements in the bioinformatics analyses of published HEK293T MeRIP-seq data by (Meyer *et al.*, 2012) resulted in the detection of 24,281 m^6A peaks, with 46% of the peaks mapping to mRNAs and 12.1% of the peaks identified in lncRNAs (Cui *et al.*, 2015). N6-methyladenosine has also been reported in tRNAs (Saneyoshi *et al.*, 1969), rRNA (Noon *et al.*, 1998) and snoRNAs (Bringmann & Luhrmann, 1987).

These studies showed that the m^6A modification in mRNAs typically resides within a GGAC consensus motif. These m^6A maps also indicated that the m^6A modification is present throughout all regions of mammalian mRNAs, being somewhat concentrated near the 5′ end, likely due to the prevalence of cap-proximal m^6Am, as the m^6A antibodies cannot differentiate between the m^6A and m^6Am modifications. More striking, however, is the enrichment of m^6A sites around the stop codon and 3′ untranslated region (UTR).

The most recent m^6A mapping has allowed for the detection of m^6A at single-nucleotide resolution rather than the peak data generated through the aforementioned techniques. Linder *et al.* (2015) employed a novel approach whereby RNA is UV cross-linked to m^6A antibodies. This results in either truncation at the m^6A site (termed cross-linking-

induced truncation or CITS-miCLIP), or a specific C-T transversion in the cytosine adjacent to the m^6A site (termed cross-linking-induced mutation sites CIMS-miCLIP). The nature of the CIMS approach dictates that only m^6A sites preceding a cytosine can be identified i.e. the m^6A sites that are located within the GGAC consensus sequence. On the other hand, the CITS approach is not defined by such a bias. The study by Linder *et al*. (2015) identified a total of 9,536 candidate sites using CIMS-miCLIP, whereas 6,543 putative m^6A sites were identified by employing CITS-miCLIP.

1.1.1.2 m^1A

N1-methyladenosine has been identified in total RNA (Dunn, 1961), rRNA (Srivastava & Gopinathan, 1987) and tRNA (El Yacoubi *et al.*, 2012; Kirchner & Ignatova, 2015) for decades. However, it was only recently, by implementing a similar approach to previously mentioned m^6A-seq, that Dominissini *et al.* (2016) identified approximately 15% and 35% of mouse and human mRNAs respectively to contain at least one m^1A peak. This equates to 7,154 peaks identified in 4,151 coding and 63 non-coding genes. Over 70% of these genes harbour a single site located in the start codon upstream of the first splice site. Another independent study by Li *et al.* (2016), using the same m^1A-seq method, identified 901 m^1A peaks from 600 genes in mRNA and noncoding RNA using HEK293T cells. Interestingly, both studies noted that m^1A is preferentially located in structured 5′ UTRs around canonical translation initiation sites or alternative translation initiation sites.

1.1.1.3 m^5C

Squires *et al.* (2012) were the first to couple transcriptome-wide NGS with bisulfite treatment in a technique termed bisulfite RNA sequencing (bsRNA-seq). Bisulfite treatment allows for the detection of m^5C sites, because unmethylated cytosines are chemically converted to uracil following the treatment, whereas m^5C remains unchanged, being resistant to the treatment (Fig. 1.2B). Motorin *et al.* (2010) pointed out m^5C is prevalent in tRNAs, at structural positions 48, 49, 50, and to a lesser extent in rRNA. Squires *et al*. (2012) confirmed 21 of the 28 previously reported m^5C sites in human tRNAs, suggesting robustness in the bsRNA-seq method. Additionally, thousands of candidate m^5C sites in both mRNAs and ncRNAs were uncovered by (Squires *et al.*, 2012). In total, these authors identified 10,275 candidate sites in both coding and ncRNA. Interestingly, they further identify 234 novel candidate sites in

tRNAs. They also showed that m^5C sites are enriched at 5′ and 3′ UTRs of mRNAs, compared to expected levels.

Other methods of m^5C detection are centered on immunoprecipitation (IP) techniques: 5-Azacytidine IP (5-Aza-IP) identifies m^5C-containing RNA, by immunoprecipitating the methyltransferase (MTase) of interest, which is irreversibly bound to the suicide inhibitor 5-azacytidine (Fig. 1.2C). This is first incorporated in place of cytosine into nascent RNA (Khoddami & Cairns, 2013). This approach revealed 544 and 75 candidate m^5C sites in tRNAs and ncRNAs, respectively, but did not identify sites in mRNAs. This technique effectively traps the methyltransferase to its RNA target; over-expression of the methyltransferase of interest is required. Despite this, as mRNAs constitute only 1-5% of the total RNA population, and sites in tRNAs and rRNAs are plentiful, this approach may not be sensitive enough to detect sites in mRNAs. A similar approach, termed methylation individual-nucleotide-resolution crosslinking and immunoprecipitation (miCLIP) (Fig. 1.2D), uses purification of an overexpressed, affinity-tagged mutant RNA:m^5C MTase. A mutation in the catalytic domain of the MTase causes it to be trapped to the RNA following binding. m^5C-containing RNA fragments are then isolated by immunoprecipitation of the MTase, followed by NGS (Hussain *et al.*, 2013). Using stringent parameters, this approach identified 83, 37 and 25 candidate m^5C sites in tRNAs, ncRNAs, and mRNAs respectively.

1.1.2 Mitochondrial RNA

Studies presenting transcriptome-wide maps of m^1A, m^6A and m^5C were not necessarily interested in mitochondrial RNA. Such works have focused on detecting modifications in mRNAs and therefore use poly(A)$^+$ RNA as the input material. However, poly(A) RNA is expected to include mitochondrial mt-tRNA and mt-rRNA, because these RNAs are known to contain poly(A) tracts (Perlman *et al.*, 1973; Nagaike *et al.*, 2008; Slomovic *et al.*, 2006). Therefore it is possible that some of the detected sites represent modifications in these RNAs.

In this thesis, mining of m^5C and m^6A transcriptome-wide datasets revealed candidate sites in mt-mRNAs. To date, however, these modifications have not specifically been reported in mt-mRNAs. In 12S mt-rRNA, m^5C and 4-methylcytidine (m^4C), and also N6-dimethyladenosine ($m^6{}_2A$) have been identified (Rorbach & Minczuk, 2012). A recent comprehensive study aimed at identifying the full complement of post-

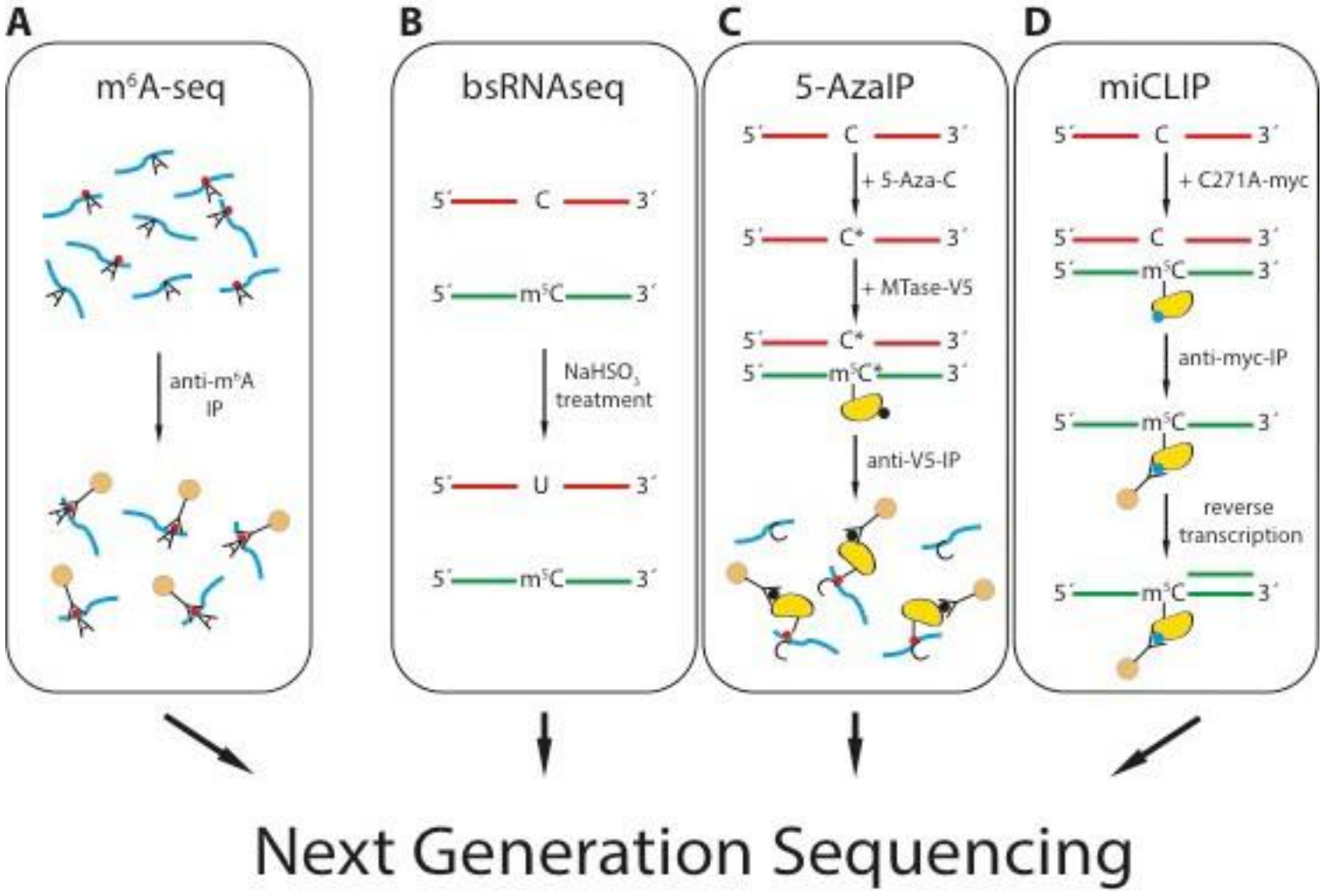

A
m6A-seq
anti-m6A IP
B
bsRNAseq
5´ C 3´
5´ m5C 3´
NaHSO3 treatment
5´ U 3´
5´ m5C 3´
C
5-AzaIP
5´ C 3´
+ 5-Aza-C
5´ C* 3´
+ MTase-V5
5´ C* 3´
5´ m5C* 3´
anti-V5-IP
D
miCLIP
5´ C 3´
+ C271A-myc
5´ C 3´
5´ m5C 3´
anti-myc-IP
5´ m5C 3´
reverse transcription
5´ m5C 3´
Next Generation Sequencing

Figure 1.2: Schematics illustrating methods to detect m^6A (immunoprecipitation) and m^5C (bsRNA-seq, 5-Aza-IP, miCLIP).

(A) Workflow of m^6A-seq to identify m^6A sites transcriptome-wide. This technique specifically immunoprecipitates m^6A containing mRNA, following poly(A) enrichment and fragmentation (to ~ 100 nucleotides) of input RNA. Libraries are constructed using the immunoprecipitated RNA and an input control, then subjected to next generation sequencing (NGS). m^6A peaks are detected relative to an input control. m^1A was detected using a similar methodology (m^1A-seq).

(B) Workflow of bsRNAseq protocol to identify m^5C sites transcriptome-wide. Total RNA is depleted of rRNA and then enriched for poly$(A)^+$ RNA. Bisulfite treatment converts unmethylated cytosines to uracil, whereas methylated cytosines are resistant to the treatment. Fragmentation of the RNA also occurs following bisulfite treatment. Fragmented RNA (50-120 nt in length) is sequenced using next-generation sequencing.

(C) 5-Azacytidine is incorporated in nascent RNA in place of cytidine. The methyltransferase of interest is tagged with V5 and irreversibly binds to the 5-azacytidine in mRNAs, which acts as a suicide inhibitor. Methylated mRNAs are then purified by an anti-V5 immunoprecipitation and subjected to next generation sequencing.

(D) A mutation in NSUN2 (C271A) causes the enzyme to be irreversibly bound to its RNA target. A myc tag is attached to NSUN2; an anti-myc immunoprecipitaion allows NSUN2 targets to be identified through next generation sequencing.

transcriptional modifications in bovine mt-tRNAs was performed using mass spectrometry (Suzuki & Suzuki, 2014). A plethora of modifications were detected. Of particular note, m^1A was identified at position 9 and 58 in many mt-tRNAs, whilst m^5C was also identified in the same structural positions as in nuclear encoded tRNAs.

Generally, technical challenges associated with the identification of modification types, specific positions of modification sites, and the enzymes responsible for the modifications have hampered progress, especially in mt-RNA. For example, over 200 modifications have been reported in eukaryotic cytosolic rRNAs and 30 in bacterial rRNAs, but, only nine modified nucleosides have been identified in mt-rRNAs (Piekna-Przybylska, Decatur, *et al.*, 2008).

Identifying post-transcriptional modifications in mitochondrial RNA, and ultimately understanding their function, is important as these modifications are expected to be involved in RNA stability, regulating RNA abundance and in RNA processing. Some evidence already exists suggesting that post-transcriptional modifications of mitochondrial rRNAs and tRNAs are required for their proper function (Decatur & Fournier, 2002; Gustilo *et al.*, 2008). The thesis attempts to focus this gap in the current knowledge.

1.2 'Writers', 'Readers' and 'Erasers'

To understand the function of post-transcriptional modifications it is important to characterise the RNA binding proteins (RBPs) that 'write', 'read' or 'erase' the modification. The 'writer' enzymes, that generate the modification, have been identified for m^5C and m^6A, although m^1A 'writer/s' are yet to be identified. Interestingly, these 'writer' enzymes tend to perform other functions in addition to their MTase activity. An important task is to understand the impacts, if any, of these other functions on their MTase function. The 'reader' proteins, that selectively bind the modification and mediate a function, have already been described for the m^6A modification; however, 'readers' still remain to be identified for the m^5C and m^1A modifications. The 'eraser' enzymes, that catalyse the removal of the post-transcriptional modifications, have been identified for both m^6A, m^1A, and possibly for m^5C. Figure 1.3 shows a

schematic summarising the 'writers', 'readers', and 'erasers' for the m^5C, m^6A and m^1A modifications. Understanding the dynamics between writing, reading and erasing methylation should be the emphasis of future work.

1.2.1 m^6A

Identifying 'writers', 'readers' and 'erasers' for m^6A in RNA has been the subject of intense research over the last few years, as reviewed in (Blanco & Frye, 2014, Meyer & Jaffrey, 2014). Methyltransferase Like 3 (METTL3) is known to function as an m^6A 'writer' in complex with methyltransferase like 14 (METTL14). Wilms tumor 1 Associated Protein (WTAP), has also been shown to interact with METTL3 and METTL14. Ping *et al.* (2014) showed that WTAP is required for localisation of the complex into nuclear speckles, and for methyltransferase activity *in vivo*.

The METTL3/14 complex modifies a large proportion of m^6A sites found within the consensus sequence RRm^6ACH (R = A/G and H = A/C/U) (Dominissini *et al.*, 2012; Meyer *et al.*, 2012). However, some m^6A sites in rRNA, tRNA and small nuclear (sn) RNA, are not located within the GGACH consensus motif, suggesting that other RNA:m^6A MTases remain to be discovered.

Collectively, the RNA:m^6A MTases and demethylases are responsible for a range of developmental phenotypes; aberrant function has been linked to developmental and neurological disorders in humans (Blanco & Frye, 2014). For example, the m^6A writer METTL3 and WTAP were knocked out in zebrafish embryos using antisense morpholinos (MO). Embryos injected with *WTAP* MO displayed multiple developmental defects at 24 hours post-fertilisation including smaller head, eyes and smaller brain ventricle while embryos injected with *METTL3* MO were only modestly affected (Ping *et al*., 2014). Also, downregulation of MTA (the METTL3 homolog in plants) caused a significant reduction of m^6A levels on a global scale and in several mRNAs critical for plant development. Phenotypically, this manifests as abnormal leaf development (Shen *et al*., 2016). m^6A has also been shown to control cell fate transition in mammalian embryonic stem cells. m^6A marks transcripts for faster turnover allowing the expression of new gene expression networks (Batista *et a*l., 2014). Finally, in the brain, m^6A levels were shown to be low during embryogenesis, whilst a dramatic increase is observed in adult brains (Meyer *et al*., 2012). This indicates that m^6A may function in neuronal maturation and is required for normal development of the adult

brain.

METTL3 belongs to a family of 33 enzymes all with annotated methyltransferase activity. However, the other METTL enzymes do not catalyse m^6A formation. This is unlike the case with the m^5C 'writers' (see below), that all catalyse m^5C synthesis in RNA. Interestingly METTL3, and NSUN enzymes, both contain a S-Adenosyl methionine (SAM)-dependent domain and require the SAM cofactor for activity.

Research into the molecular function of m^6A in mRNAs has been the focus of many recent works, and it is clear that there is no simple consensus mechanism. N6-methyladenosine is implicated in a wide range of context-dependent functions. Accumulating evidence indicates the involvement of m^6A in mRNA splicing, stability and translation, (reviewed in (Meyer & Jaffrey, 2014). Early on, Dominissini *et al.* (2012) identified the human YTH domain family of proteins as m^6A 'readers'. Since then, the functions mediated by these 'reader' proteins have been elucidated. YTHDF3, YTHDC1, and Hu-antigen R (HuR) have also been identified as other m^6A 'readers' (Dominissini *et al.*, 2012; Wang, Lu, *et al.*, 2014; Xu *et al.*, 2014). Two m^6A demethylases (i.e. 'erasers'), fat mass and obesity associated protein (FTO) and α-ketoglutarate-dependent dioxygenase alkB homolog 5 (AlkBH5), have recently been identified. Jia *et al.* (2011) show that FTO knockdown experiments resulted in increased levels of m^6A, whilst overexpression studies yielded a decrease in m^6A levels. Also, Zheng *et al.* (2013) show that Alkbh5-deficient male mice have increased m^6A levels in mRNA compared to wild-type mice. These studies suggest AlkBH5 and FTO function as m^6A demethylases on mRNA. Furthermore, AlkBH5 was found, through *in vitro* protein-RNA binding assays, to have a mild preference for m^6A within the consensus motif and did not target m^6A in rRNA *in vivo* (Zheng *et al.*, 2013). Altogether this suggests that other m^6A demethylases remain to be discovered.

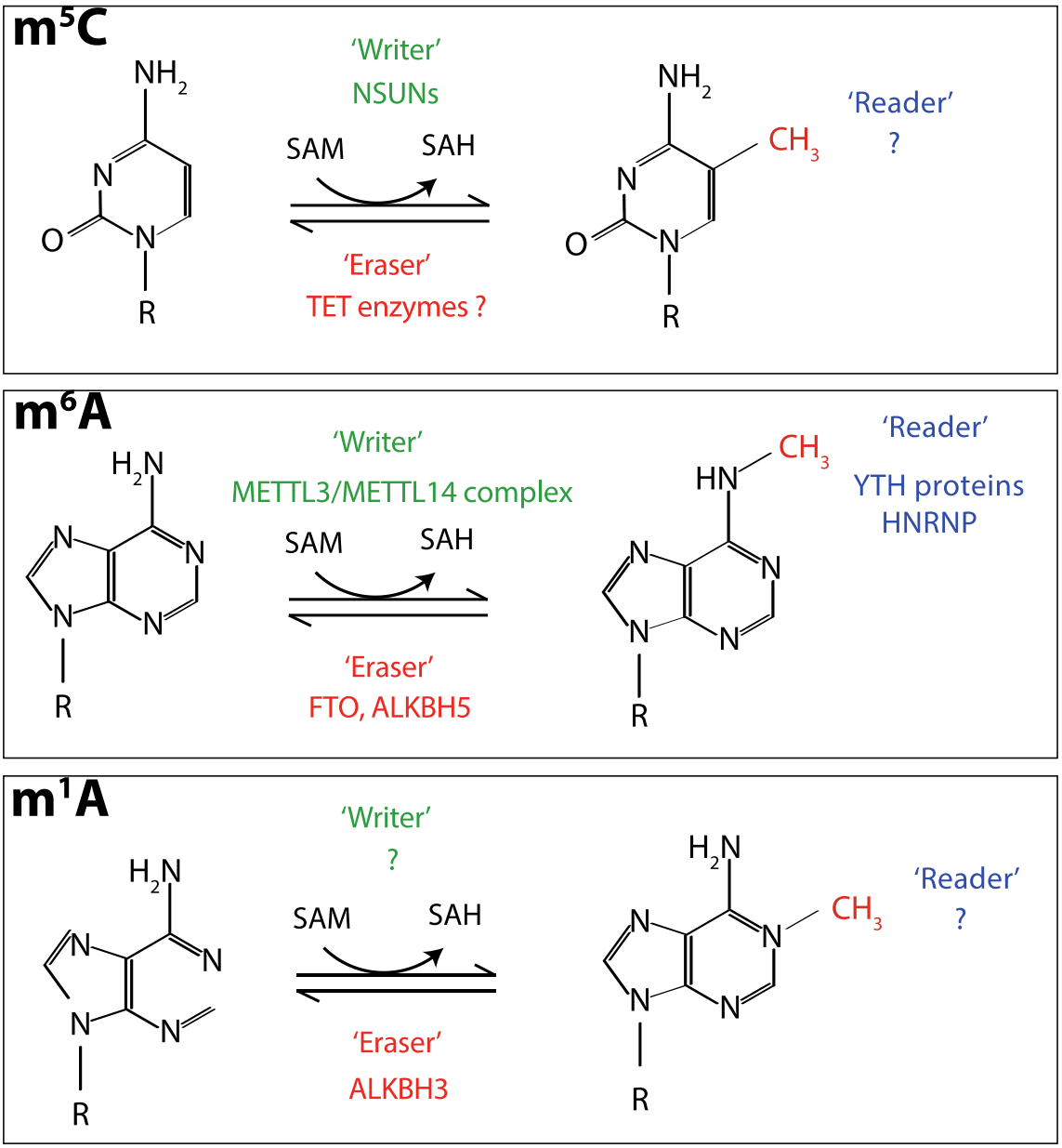


Figure 1.3: Schematic illustrating 'writers', 'readers' and 'erasers' of the m^5C, m^6A and m^1A modification.

The NSUN family of enzymes is known to function as 'writers' of the m^5C modification. METTL3/METTL14 catalyses the formation of m^6A, whilst the 'writer/s' of m^1A are still unknown. The S-adenosyl-L-methionine (SAM) cofactor is required for methylation. S-adenoysl-L-homocysteine (SAH) is formed by the demethylation of SAM. The ten-eleven translocase (TET) enzymes are expected to catalyse the demethylation of m^5C. FTO and ALKBH5 are responsible for the demethylation of m^6A; ALKBH3 has been shown to reverse m^1A. 'Reader' proteins have not yet been identified for m^5C or m^1A. The YTH domain proteins specifically recognise m^6A, executing a number of functions.

1.2.2 m^1A

Dominissini *et al.* (2016) and Li *et al*. (2016) have independently mapped m^1A candidate sites transcriptome-wide. Both studies identified AlkBH3 as an m^1A demethylase, however no 'writer' or 'reader' proteins have yet been identified for this modification.

1.2.3 m^5C

In eukaryotes, RNA m^5C methylation is catalysed by two known classes of RNA methyltransferases: tRNA Aspartic Acid Methyltransferase 1 (TRDMT1), and NOP2/Sun RNA Methyltransferase Family Members 2, 3, 4, and 6 (NSUN2, NSUN3, NSUN4 and NSUN6) (Goll *et al.*, 2006; Frye & Watt, 2006; Metodiev *et al.*, 2014; Haag *et al.*, 2015). TRDMT1 was originally regarded to function as a DNA MTase, however, now it is thought to act primarily as an RNA methyltransferase (Goll *et al.*, 2006; Jurkowski *et al.*, 2008; Schaefer *et al.*, 2010). The full range of RNA targets for both TRDMT1 and NSUN2 are yet to be identified. Currently, TRDMT1 is known only to target C38 in aspartyl tRNA, where Schaefer *et al.* (2010) provided evidence that TRDMT1-mediated methylation inhibits stress-induced cleavage of modified tRNAs.

Identifying targets of NSUN methyltransferases has been the emphasis of many recent studies. The NSUN family comprises of seven RNA:m^5C methyltransferases (RCMTs). Features of these enzymes include conserved domains required for methyltransferase activity, RNA binding, cofactor binding and catalysis. This suggests that all NSUN proteins have the ability to function as MTases, although some are yet to be experimentally confirmed for methyltransferase activity. RCMTs have been organised into subfamilies based on sequence divergence. Reid *et al.* (1999) identified eight subfamilies RCMT1-8, with RCMT 2, 7, 8 representing eukaryotic MTases (i.e. there are no prokaryotic orthologues), whilst the other five subfamilies are strictly prokaryotic (i.e. there are no eukaryotic orthologues). RCMT2 includes the *Saccharomyces cerevisiae* MTase Nop2 (nucleolar protein 2) and its mammalian orthologue, the proliferation-associated nucleolar antigen P120/NSUN1 (NOP2/Sun domain family, member 1), which are involved in rRNA processing and large ribosomal subunit assembly (Hong *et al.*, 1997). Nop2 is known to mediate m^5C at C2870 in 25S rRNA (Sharma *et al.*, 2013). RCMT8 includes the *S. cerevisiae* methyltransferase Rcm1 (known to target C2278 in 25S rRNA) (Sharma *et al.*, 2013), and its mammalian orthologue, NSUN5A (NOP2/Sun domain family, member 5A). Prototypic members of

the RCMT7 subfamily are the *S. cerevisiae* Ncl1/Trm4 (nuclear protein 1/tRNA methyltransferase 4) (Motorin and Grosjean, 1999) and its mammalian orthologue NSUN2 (NOP2/Sun domain family, member 2), also known as MISU (Myc-induced SUN-domain-containing protein) (Brzezicha *et al.*, 2006, Frye & Watt, 2006). Other members include the *S. cerevisiae* methyltransferase Nop8, and its mammalian orthologue NSUN4 (NOP2/Sun domain family, member 4), which are important for ribosomal biogenesis in mitochondria (Metodiev *et al.*, 2014; Santos *et al.*, 2011). Table 1.1 summarises the known eukaryotic substrates and the RCMT subfamily of the NSUNs.

Table 1.1: RCMT subfamily and known substrates of NSUN enzymes.

	RCMT subfamily	**Known substrate**
NSUN1	2	28S rRNA (C4447)
NSUN2	7	tRNA, rRNA, mRNA, ncRNA
NSUN3	7	Wobble position of mt-tRNA^{Met}
NSUN4	7	12S mt-rRNA (C841)
NSUN5	8	28S rRNA (probable)
NSUN6	NSUN6	C27 in tRNA$^{Cys/Thr}$
NSUN7	8	Unknown

NSUN2 is known primarily to target positions C48-50 in many tRNAs (Squires *et al.*, 2012; Motorin & Grosjean, 1999; Auxilien *et al.*, 2012; Tuorto *et al.*, 2012), but has also been shown to target C34 (Brzezicha *et al.*, 2006) and C40 in humans (Khoddami & Cairns, 2013). NSUN2 has also been reported to be responsible for m^{5}C methylation in many coding and ncRNAs (Hussain *et al.*, 2013). NSUN4 is exclusively mitochondrial, and is expected to have a much narrower substrate range than NSUN2 (see Chapter III). Metodiev *et al.* (2014) showed that NSUN4 targets a single site in mouse mitochondrial 12S rRNA, which has a role in ribosome assembly. Recently, NSUN6 has been confirmed to function as a RNA m^{5}C methyltransferase. Haag *et al.* (2015) showed that NSUN6 mediates m^{5}C methylation at C72 in tRNA^{Cys} and tRNA^{Thr} that have the CCA addition at the 3≠ end, suggesting NSUN6-mediated modification occurs at a late stage in tRNA biogenesis. Parallel recent studies by Nakano *et al.* (2016); Van Haute *et al.* (2016) and Haag, Sloan *et al.* (2016) have revealed that human NSUN3 methylates C34 in mitochondrial tRNA^{Met}, which is then further catalysed to 5-formylcytosine (f^{5}C) by an as yet unknown enzyme. These authors also showed that the

f^5C modification is crucial in protein synthesis for the successful translation of the AUA codon to methionine. The NSUN family contains a further three proteins (NSUN1, NSUN5, and NSUN7), which are also predicted to methylate RNA based on sequence conservation of domains required for methyltransferase activity Motorin *et al.* (2010). However they have yet to be confirmed for methyltransferase activity in mammals. Sharma *et al.* (2013) have found that yeast NOP2 (equivalent to human NSUN1) and yeast Rcm1 (equivalent to human NSUN5) target two separate sites in yeast 25S rRNA. Schosserer *et al.* (2016) also show that NSUN5 methylates 25S rRNA in *C. elegans.* Interestingly, the equivalent rRNA in humans (28S) harbours two m^5C sites at analogous positions to those in the yeast 25S rRNA; however it remains unknown as to whether NSUN1 and NSUN5 are responsible for those methylated sites in 28S rRNA.

Several NSUN proteins are critically involved in many diseases. For example, mutations in NSUN2 cause autosomal-recessive intellectual disability syndromes (Abbasi-Moheb *et al.*, 2012). NSUN2 has also been implicated in caner cell proliferation, being a direct target of MYC (myelocytomatosis viral oncogene homolog) (Frye & Watt, 2006). Blanco *et al.* (2011) provide evidence linking NSUN2 in balancing stem cell self-renewal and differentiation, following the observation that NSUN2 knockout mice are small in size. Furthermore, NSUN5 is deleted in Williams-Beuren syndrome (Doll & Grzeschik, 2001), and a connection between a mutation in the NSUN7 gene and male sterility has been reported (Harris *et al.*, 2007). Also, TRDMT1-dependent cytosine methylation of $tRNA^{Asp(GUC)}$ was shown to be inhibited by the anticancer drug 5-azacytidine (Schaefer, Hagemann, *et al.*, 2009). These observations suggest that cytosine methylation in RNA is important in the control of cell growth and its differentiation. Importantly, however, it remains to be discovered if/how the loss of RNA methylation contributes to the pathology of these diseases/processes.

Specific m^5C 'reader' proteins, which may mediate m^5C functions, have not yet been identified. Work in this thesis contributes to our knowledge in this space.

Fu *et al.* (2014) have recently demonstrated *in vitro* and *ex vivo* that the ten-eleven translocase (TET) enzyme family, which have previously been identified as m^5C DNA demethylases, can also target human RNA. This results in the oxidative demethylation of m^5C to 5-hydroxymethylcytidine (hm^5C). Furthermore, an independent study by

Huber *et al.* (2015) also confirmed the presence of hm^5C derived from m^5C by employing an *in vivo* isotope-tracing approach. This allowed for the detection of isotope-labelled m^5C and any directly formed derivatives through LC-MS/MS techniques. Notably, the hm^5C modification was observed at a 40-fold increase in the poly(A)-enriched HEK293T samples compared to total RNA, suggesting that m^5C undergoes dynamic regulation in mRNAs and/or certain lncRNAs. However, it will be important to map specific sites of hm^5C in RNA to gain functional insight. Furthermore, hm^5C was identified in plants, nematodes and mammalian cells, indicating that hm^5C is a common, conserved modification of RNA. Delatte *et al.* (2016) also showed that TET enzymes catalyse the formation of hm^5C from m^5C in poly(A) RNAs in *Drosophila*. Using an antibody specific for hm^5C, the authors map hm^5C transcriptome-wide. Their analysis indicated that hm^5C is present within the coding sequence of many mRNAs in a CU rich consensus motif. The authors also found that high hm^5C content is associated with the later fractions of a sucrose gradient, suggesting that hm^5C in mRNA functions to promote mRNA translation.

1.3 Function of the m^6A, m^1A and m^5C modifications in non-coding and coding RNA

1.3.1 Non-coding RNA

1.3.1.1 m^6A

1.3.1.1.1 MALAT1

The lncRNA, metastasis associated lung adenocarcinoma transcript 1 (MALAT1) has functions in regulating gene transcription and splicing (Gutschner *et al.*, 2013). Recently, two near stoichiometric m^6A methylation at positions A2515 and A2577 within the GGACU consensus context, were identified in MALAT1 (Liu *et al.*, 2013). Secondary structure predictions of MALAT1 RNA segments housing these modified sites showed that both sites reside in hairpin stems. The authors showed that the m^6A methylation in this context resulted in the destabilistion of the hairpin stem, which allowed access of heterogeneous nuclear ribonucleoprotein C (HNRNPC) to a U-rich tract that was located opposite to the m^6A site identified at position A2577 (Fig. 1.4A). This observation prompted global HNRNPC footprinting, which identified thousands of m^6A switches in human transcriptome; Liu *et al.* (2015) showed that m^6A-induced HNRNPC binding, affected the abundance and alternative splicing of target mRNAs.

Unlike with m^5C and m^1A, m^6A sites are located within clear sequence and structural constraints allowing for the identification of m^6A 'reader' proteins. This has translated in much progress towards understanding the molecular function of m^6A in mRNAs. Indeed, m^6A function is thought to be context-dependent and to act either by directly recruiting binding proteins ('readers') or indirectly, through affecting local RNA structure. The YTH proteins YTHDF1-3 and YTHDC1, as well as HNRNPA2B1, have been identified as direct m^6A 'readers', affecting translation efficiency, mRNA half-life and mRNA splicing events.

1.3.1.1.2 miRNA processing

MicroRNAs (miRNAs) are small ncRNAs, usually 21–25 nt in length. They function in the regulation of gene expression through complementary binding to target mRNAs, resulting in RNA cleavage, translation inhibition or RNA deadenylation (Beilharz *et al.*, 2009; Huntzinger & Izaurralde, 2011). The primary miRNA (pri-miRNA), which can be more than 1000 nt in length, is processed by the microprocessor complex, DGCR8 and Drosha to form a 60-120 nt hairpin known as the precursor miRNA (pre-miRNA). In the cytoplasm, Dicer then cleaves the pre-miRNA to form the mature miRNA (Ha & Kim, 2014).

Alarcon *et al.* (2015) discovered an enrichment of the GGm^6ACU consensus motif at the junction between the hairpin stem and the flanking single-stranded RNA regions of pri-miRNAs from MDA-MB-231 breast cancer cells, whilst a depletion of the motif was observed in pre-miRNA transcripts. Knockdown of the m^6A 'writer', METTL3, resulted in the loss of DGCR8 binding to pri-miRNAs. This resulted in the accumulation of unprocessed pri-miRNAs and consequently the reduction of mature miRNAs (Alarcon *et al.*, 2015). Altogether, these observations suggest a crucial role for m^6A in miRNA biogenesis. Furthermore, this loss of DGCR8 binding also indicated that m^6A methylation provides another layer of regulation in miRNA processing; m^6A allows DGCR8 to differentially recognise pri-miRNAs over mRNAs. mRNAs tend to form secondary structures including short hairpins that resemble pri-miRNAs, so m^6A marks may allow DGCR8 to differentiate pri-miRNAs from mRNAs. To further validate this work, Alarcon *et al.* (2015) performed *in vitro* processing assays, using methylated and non-methylated let-7 pri-miRNA transcripts. This revealed that m^6A-modified pri-let-7e was processed more efficiently into pre-let-7e than the unmethylated pri-let-7e. This experimental evidence further strengthened a role for m^6A in the processing of pri-

miRNAs into pre-miRNAs (Fig. 1.4B). As altered miRNA and METTL3 expression have been reported in various cancers, the study also provided evidence linking these two cancer markers.

1.3.1.2 m^1A

A function of m^1A in ncRNA has not been described.

1.3.1.3 m^5C

The recent bsRNA-seq, 5-Aza-IP, miCLIP studies have each reported numerous m^5C candidate sites in ncRNAs. The role of m^5C in some examples have been studied in detail revealing roles in ncRNA processing, and in affecting RNA-protein interactions.

1.3.1.3.1 Vault RNAs

Vault RNAs (vtRNA) are polymerase III transcripts, typically in the range of 100 nt, contained in ribonucleoprotein complexes with no known function (Persson *et al.*, 2009). Both miCLIP and Aza-IP experiments have identified m^5C sites in several vtRNAs (Hussain *et al.*, 2013; Khoddami & Cairns, 2013). Interestingly, Hussain *et al.* (2013) show that NSUN2 methylates position C70 in vtRNA1-1 and egulates its processing into small miRNA-like regulatory RNAs.

1.3.1.3.2 XIST and HOTAIR long non-coding RNAs

Amort *et al.* (2013) identified m^5C sites in the lncRNAs X-inactive specific transcript (XIST) and HOX antisense intergenic RNA (HOTAIR). By using bisulfite amplicon sequencing, the authors were able to interrogate different regions of the two lncRNAs for the presence of m^5C. Amort *et al.* (2013) detected m^5C sites in regions of functional importance in both lncRNAs; these regions are known to interact with chromatin-modifying complexes.

The 5´ proximal region of HOTAIR has been reported to interact with histone methyltransferase polycomb-repressive complex (PRC2), whilst the 3´ end of HOTAIR has been shown to bind the histone demethylase complex, LSD1/CoREST/REST (Bhan & Mandal, 2015). Although a function for the m^5C in HOTAIR was not determined, it is positioned near the LSD1 binding region in HOTAIR (at C1683) and was shown to be stiochiometerically methylated in several different human cell lines (Amort *et al.*, 2013),

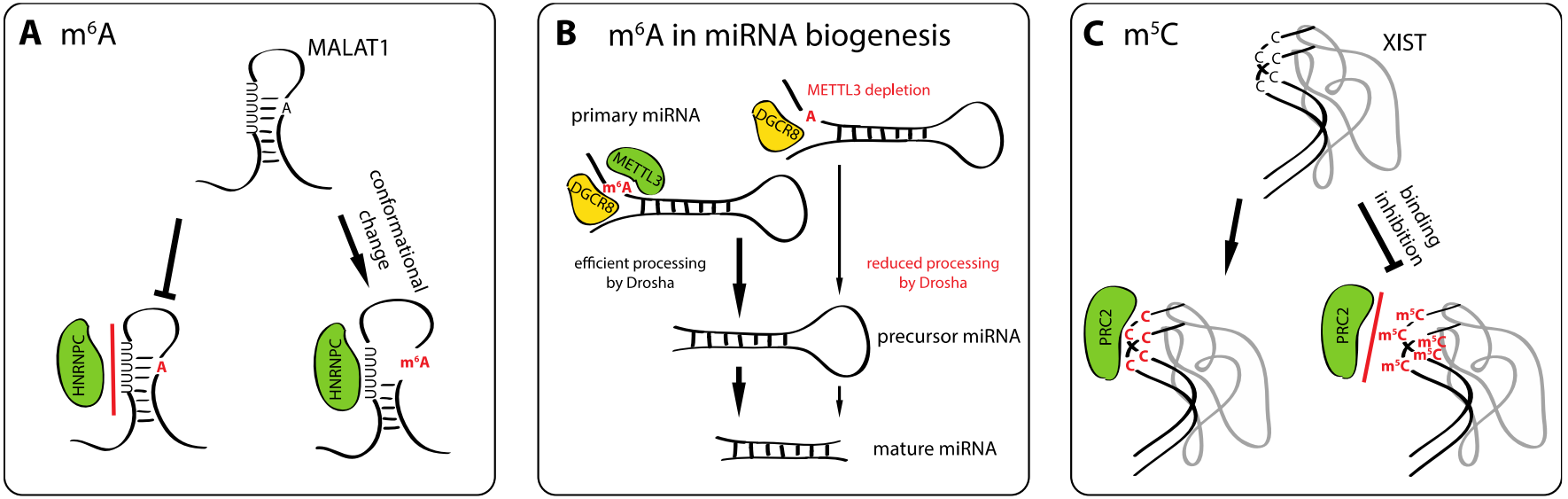


Figure 1.4: Function of m^{6}A and m^{5}C in ncRNA (taken from my review: Shafik *et al*., 2016).

(A) An example of an m^{6}A structural switch. The presence of m^{6}A in the lncRNA MALAT1 induces a conformational change in the RNA allowing HNRNPC to bind the U-tract on the opposite arm of the hairpin to where m^{6}A is located.

(B) m^{6}A at the base of the primary miRNA hairpin allows for efficient processing by Drosha, resulting in increased mature miRNA compared to unmethylated primary miRNAs.

(C) Five clustered sub-stiochiometric m^{5}C sites were identified in the lncRNA, XIST. The presence of m^{5}C inhibits binding of PRC2. In this way, m^{5}C is involved in epigenetic processes by regulating the interaction between the lncRNA and a chromatin modifying protein.

suggesting a mandatory rather than a regulatory function of the m^5C methylation.

XIST is a component of the regulatory locus termed X-inactivation center. XIST is known to transcriptionally silence one of the two X-chromosomes in female placental mammals (Wutz, 2011). XIST has also been shown to bind the PRC2 complex via its 5´ proximal A region (Zhao *et al.*, 2008). Amort *et al.* (2013) identified, within the A region of XIST, five sub-stoichiometric, clustered m^5C sites (C701, C702, C703, C711 and C712) . Of note, these m^5C sites were detected in female human HEK293T cells, however, corresponding sites in mouse XIST did not show any methylation. Furthermore, only about 20% of XIST molecules were targeted for modification. This suggested a regulatory function of the m^5C methylation. *In vitro*, the m^5C sites were shown to function in blocking PRC2 binding to human XIST. Overall, evidence present by Amort *et al.* (2013) suggests the presence of m^5C in XIST, and also possibly in HOTAIR, is involved in epigenetic processes, by regulating the interactions between the lncRNAs and chromatin modifying protein complexes (Fig. 1.4C).

1.3.2 Coding RNA

1.3.2.1 m^6A

N6-methyladenosine 'readers' have diverse roles in affecting splicing, export, translation and turnover of mRNAs. For example, by recruiting the splicing factor SRSF3, the nuclear m^6A 'reader' YTHDC1 favours a particular splice variant by including m^6A methylated exons (Xiao *et al.*, 2016). Another YTH protein, YTHDF1, was recently shown to enhance translation of a subset of m^6A-containing mRNAs (Wang *et al.*, 2015). Briefly, through an YTHDF1 PAR-CLIP experiment, the authors showed a significant correlation between HeLa m^6A sites and YTHDF1 binding regions. Furthermore, upon YTHDF1 or METTL3 knockdown, ribosome profiling revealed a reduction in translation of mRNAs that contained an m^6A site and an overlapping YTHDF1 footprint. Adding further weight to this observation, YTHDF1 was experimentally determined to interact with ribosomes and translation factors, of particular note was eukaryotic initiation factor 3 (eIF3). (Wang *et al.*, 2015) postulated that YTHDF1 functioned to promote translation, likely by directly binding eIF3 (Fig. 1.5A).

In the cytoplasm, YTHDF2 binds to mRNAs containing 3´ UTR-localised m^6A and targets those mRNAs to processing bodies for degradation (Wang, Lu, *et al.*, 2014).

Interestingly, PAR-CLIP experiments have revealed many m^6A-containing mRNAs are targeted by both YTHDF1 and YTHDF2 (e.g. SON and CREBP mRNAs). Regarding these shared mRNA targets, Wang *et al.* (2015) observed that YTHDF1 binds to the mRNA, earlier in its life cycle, compared to YTHDF2. This suggests that the two YTH proteins function together to regulate translation efficiency and mRNA decay. A clear advantage of this cooperative functionality could be a mechanism to ensure stable protein levels. For example, RNA methylation may be a layer of regulation that determines the rate of translation to decay to maintain protein quantities, despite varying mRNA levels. That is, increasing mRNA decay when transcripts are abundant, and increasing translation when transcripts are scarce.

Most recently, (Lin *et al.*, 2016) have showed that m^6A 'writer', METTL3, also functions as an m^6A 'reader'. The authors show that its novel reading ability results in the increased translation of a subset of m^6A-containing mRNAs. Interestingly, (Lin *et al.*, 2016) show it is METTL3, and not the other components of the complex, that function as a 'reader' (Fig. 1.5B). A proportion of METTL3 was identified in the cytoplasm of human lung cancer cell lines, where it is expected to perform its 'reader' function. METTL3 knockdown did not affect RNA abundance, however, the proteins encoded by a subset of m^6A-containing transcripts were reduced e.g. EGFR and TAZ proteins showed decreased levels. This suggested that METTL3 may affect protein abundance at the level of mRNA translation. To investigate this possibility, polysome gradient and immunoprecipitation experiments were performed. The former showed that METTL3 was associated with polysomes. The immunoprecipiation experiments revealed an interaction between the translation initiation machinery (e.g. eIF4E, CBP80, eIF3) and METTL3. To confirm METTL3 reading functionality, tethered-function and domain deletion analyses showed that the catalytic activity of METTL3, in the C-terminus, was distinct from its m^6A 'reader' role, which is dependent upon METTL3's N-terminus; this 'reading' activity was shown to be independent of YTHDF1, YTHDF2, METTL14 or WTAP. More specifically, Lin *et al.* (2016) showed that the direct tethering of METTL3 to the 3´ UTR of different reporters only resulted in enhanced translation when the initiation factors were present. Enhanced translation was not observed in the assay where an element on the reporter directly recruited the ribosome, allowing translation in the absence of initiation factors. This suggested that METTL3 recruits eIF3 to mRNAs to enhance their translation (Fig. 1.5B). Altogether, these observations strongly suggest a role for METTL3 in promoting the translation of a subset of m^6A-containing mRNAs.

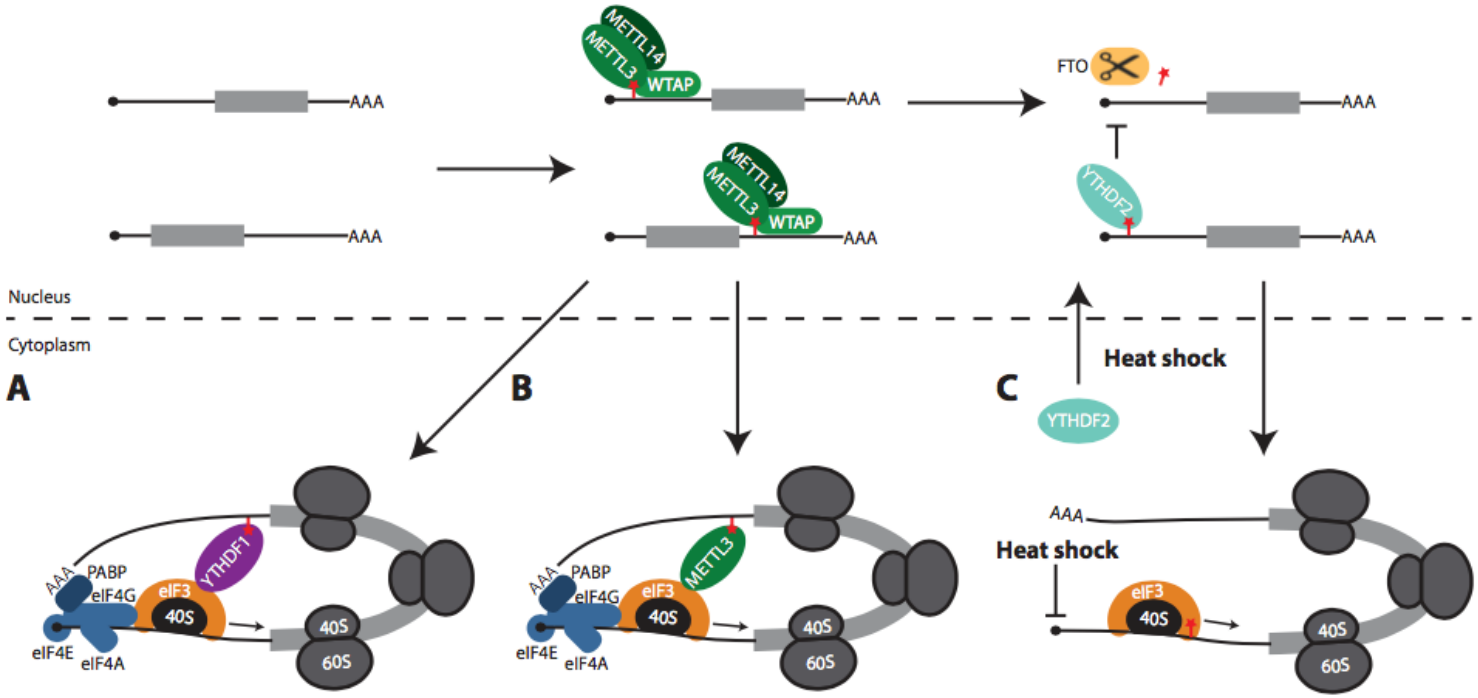


Figure 1.5: Schematic outlining the various roles of m^{6}A in mRNA translation (adapted from my review: Schumann *et al.*, 2016).

(A and B) A subset of m^{6}A sites localised near the stop codon are recognised by either (A) YTHDF1 (Wang *et al.*, 2015) or (B) METTL3 (Lin *et al.*, 2016) in the cytoplasm. Both 'reader' proteins have been shown to interact with initiation factor eIF3, resulting in enhanced translation.

C) Heat shock stress inhibits cap-dependent translation. In response to heat-shock YTHDF2 is relocated from the cytoplasm to the nucleus. There it binds 5′ UTR localised m^{6}A sites protecting them from demethylation by FTO (Zhou *et al.*, 2015). In the cytoplasm, eIF3 then directly recognises m^{6}A of those mRNAs, resulting in cap-independent translation (Meyer *et al.*, 2015).

It is expected, and to some extent shown, that either METTL3 or YTHDF1, in complex with eIF3, would operate primarily by recognising m^6A sites near the stop codon and in the 3´ UTR. eIF3 binding to the 5´ end of the mRNA would then result in enhanced translation through the closed loop model. However, 5´ UTR localised m^6A sites may also function in promoting translation with Zhou *et al.* (2015) and Meyer *et al.* (2015) leading efforts in this space. Zhou *et al.* (2015) showed upon heat shock of mouse embryonic fibroblasts, an apparent increase in m^6A methylation in the 5´ UTRs of stress-induced mRNAs was evident compared to untreated fibroblasts. Mechanistically, this observation was explained by the relocation of YTHDF2 to the nucleus from the cytoplasm. Here, YTHDF2 functioned to block demethylation by FTO, resulting in modified mRNAs being exported into the cytoplasm. Meyer *et al.* (2015) then demonstrated that 5´ UTR m^6A sites are able to directly bind eIF3 resulting in cap-independent translation of those mRNAs (Fig. 1.5C). This is important, as heat shock generally causes a reduction in cap-dependent translation, this mechanism bypasses the need for cap proteins, allowing for translation of stress-induced mRNAs.

1.3.2.2 m^1A

Interestingly, m^1A-contaning genes showed relatively increased translation efficiencies compared to non-m^1A containing genes, suggesting that m^1A may enhance translation (Dominissini *et al.*, 2016). As m^1A is positively charged under physiological pH, it is probable that m^1A can affect mRNA structure. This may be a possible mechanism by which m^1A enhances mRNA translation (Fig. 1.6). Examples of m^1A function in specific mRNAs have not yet been reported.

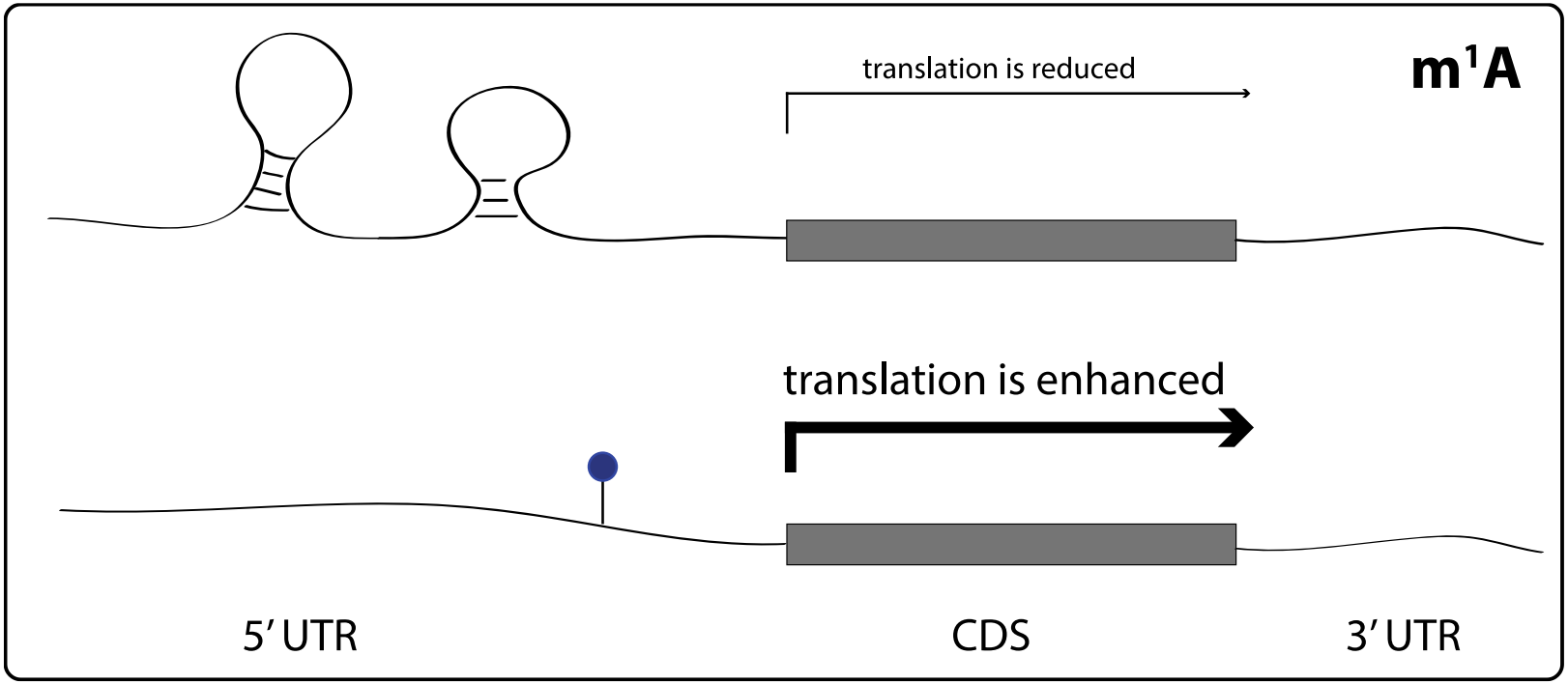


Figure 1.6: Role of m^1A in mRNA translation.

mRNAs with strong secondary structure in their 5′ UTR are normally weakly translated. m^1A is enriched in mRNAs near the translation initiation sites of structured mRNAs. It is suggested that the introduction of m^1A, relaxes this strong secondary structure resulted in enhanced translation of the mRNA (Dominissini *et al*., 2016).

1.3.2.3 m^5C

The role of m^5C in mRNAs remains largely elusive. Direct effects of m^5C on RNA structure are expected to be minimal, as m^5C does not affect the base-pairing properties. The position of the m^5C site within the mRNA will likely dictate its function, perhaps by attracting a binding protein or by affecting the binding or functionality of other regulatory factors nearby. No factor that specifically binds to or recognises m^5C in RNA has yet been reported, although m^5C is suspected of having a role in mRNA stability and mRNA translation.

1.3.2.3.1 Role for m^5C in mRNA stability

A recent report suggests that NSUN2-mediated methylation in the 3´ UTR of the p16 mRNA confers stability (Zhang *et al.*, 2012). The authors showed that NSUN2 knockdown in HeLa cells resulted in ~77% decrease in p16 mRNA levels, whilst overexpression of NSUN2 caused a concomitant increase in p16 mRNA. However no change in p16 pre-mRNA was observed following knockdown or overexpression of NSUN2. Cumulatively, this suggested that NSUN2 may regulate the half-life of p16 mRNA. Furthermore, using reporter assays, Zhang *et al.* (2012) showed that modification of p16 mRNA protected the transcript from localisation into processing bodies. The authors also consolidated these results by testing whether methylation of p16 mRNA would affect the binding of destablising interactions of AU-rich element RNA binding protein 1 (AUF1), HuR and AGO2/RISC which are known to target p16 3´ UTR. They employed ribonucleoprotein immunoprecipitaiton assays to show that modified p16 3´ UTRs do not associate with HuR, AUF1 and Ago2/RISC. The caveat to these findings is that the authors, using *in vitro* methylation assays, showed methylation targeted position A988 in the 3´ UTR, however NSUN2 is known to target cytosine and not adenosine.

Other evidence for a role of m^5C in mRNA stability is derived from computational analyses described by (Squires *et al.*, 2012). They observed that AGO2 footprints were significantly enriched just upstream of NSUN2-controlled m^5C sites in the coding sequence (CDS) and 3´ UTR of mRNAs, whilst Pumilio 2 footprints, another RISC protein, did not correlate with m^5C sites in mRNA. This suggested a role for m^5C in the miRNA mechanism of post-transcriptional gene regulation.

1.3.2.3.2 Role for m^5C in mRNA translation

Recent reports have implicated m^5C methylation in affecting mRNA translation. Tang *et al*. (2015) show that NSUN2 targets the 5´ UTR of p27 mRNA. *p27* is an inhibitor of cyclin-dependent kinase (CDK) activity and is required to arrest growth in senescent cells. Using an *in vitro* translation assay, Tang *et al*. (2015) demonstrated that translation of *p27* mRNA was repressed by m^5C methylation at C64. The authors provide no mechanistic details as to how the methylation contributes to the reduced translation of *p27*. Experimentally, they observed increased p27 levels and reduced CDK1 protein levels upon knockdown of NSUN2 in human diploid fibroblasts, whilst the opposite effect when NSUN2 is overexpressed. As HuR or CUGBP1 are known to bind the *p27 5′* UTR resulting in repressed translation (Kullmann *et al.*, 2002; Zheng & Miskimins, 2011), the authors investigated whether there was a link between HuR or CUGBP1 binding with NSUN2 methylation of the mRNA. They found the repressive effect of NSUN2 methylation is independent of the effect elicited by HuR or CUGBP1 binding. Furthermore, a sister study has revealed that NSUN2 methylation of CDK1 3´ UTR enhanced translation of CDK1 mRNA (Xing *et al.*, 2015). The authors show that NSUN2 regulated expression of CDK1 plays a role in cell division cycle (Fig. 1.7).

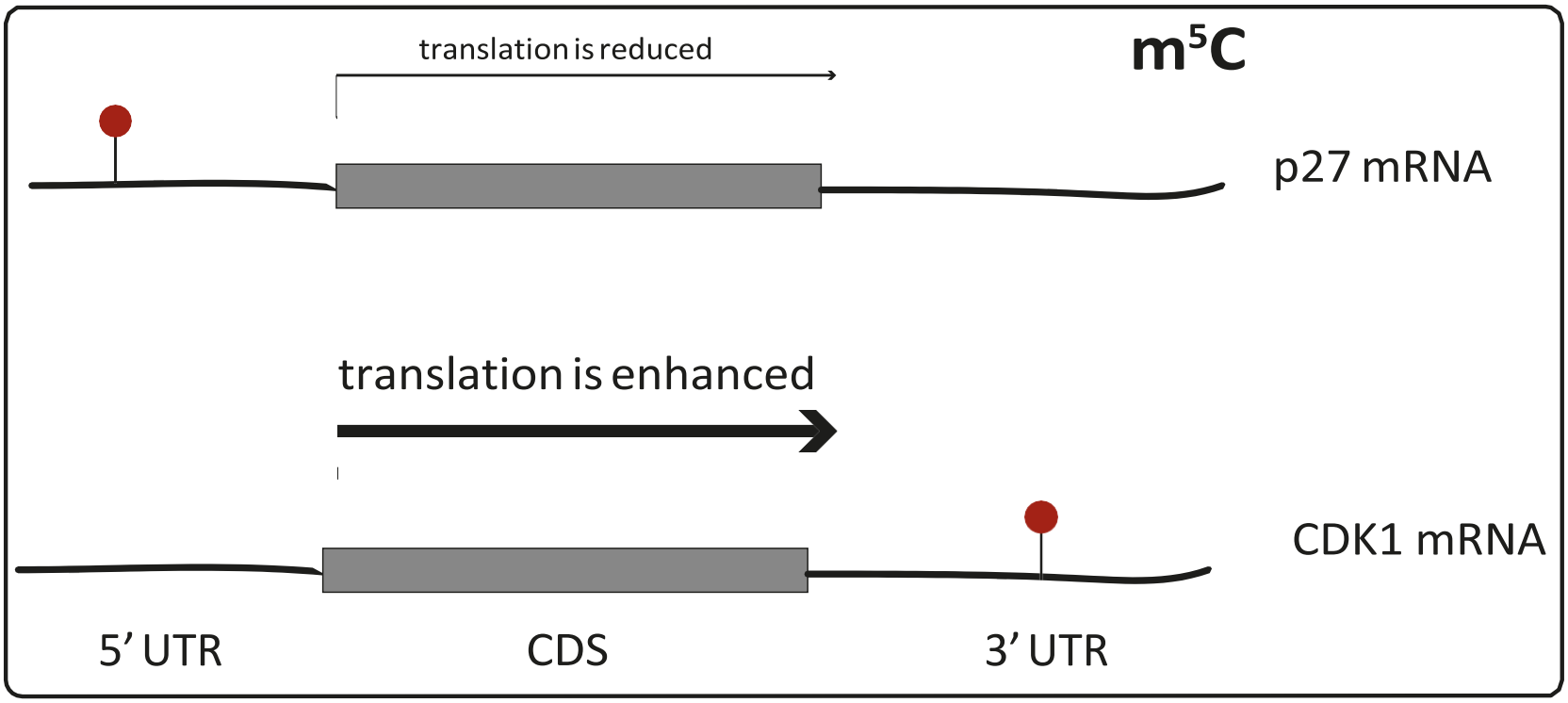


Figure 1.7: Examples where m^5C methylation has both a repressive and enhancing affect on translation in mRNAs.

NSUN2-mediated m^5C methylation of p27 in the 5′ UTR results in reduced translation (Tang *et al*., 2015), whereas m^5C methylation in the 3′ UTR of CDK1 results in enhanced translation (Xing *et al*., 2015). A mechanism for how m^5C regulates translation in these cases has not been given.

1.4 Aims

The overarching aim is to determine the extent and function of m^5C particularly in mRNAs. The approach is both experimental and computational.

Specifically the thesis is aimed at:

1) determining the m^5C methylome in mitochondrial RNA, and the responsible 'writer' enzymes;

2) determining m^5C-binding proteins and their possible role; and

3) performing metagene analyses on a high confidence set of m^5C sites to investigate the role of m^5C function on mRNA translation.

Chapter II

Materials and Methods

2.1 Materials

2.1.1 Chemicals, reagents and consumables

Chemicals, reagents and consumables used in this thesis are tabulated below.

Chemical/reagent/consumable	**Manufacturer**
Digitonin	Abcam (Cambridge, UK)
EZ-Vision® Three DNA Dye	Amresco (OH, USA)
DTT	Amresco
Glycine	Amresco
Agencourt® AMPure XP	Beckman Coulter (CA, USA)
Micro Bio-Spin™ Chromatography Columns	Bio-Rad (CA, USA)
Skim milk powder	Coles (VIC, Australia)
T-75 flasks	Corning (Tewksbury, MA, USA)
6-well plates	Corning
Axygen® 8-strip thin wall PCR tube strips and caps	Corning
1.5 ml tubes	Eppendorf (Hamburg, Germany)
InstantBlue Coomassie Blue stain	Expedeon (CA, USA)
Sets of dNTPs (100 mM each)	Fisher Biotec, (WA, Australia)
Amersham Hybond™ PVDF transfer membrane	GE Healthcare (BKM, UK)
Whatman™ chromatography paper	GE Healthcare
Biotinylated RNA baits	GeneLink Inc. (New York, USA)
Glycogen (5 mg/ml)	Life Technologies (CA, USA)
Lipofectamine RNAiMax	Life Technologies
Tris-HCl pH 8.0	Life Technologies
3M sodium acetate (NaOAc) (pH 5.5)	Life Technologies
UltraPure™ DNase/RNase-Free Distilled Water (dH_2O)	Life Technologies
RNaseOUT™	Life Technologies
TURBO™ DNase	Life Technologies

First strand buffer	Life Technologies
Fast SYBR® Green Master Mix	Life Technologies
L-glutamine	Life Technologies
0.25% Trypsin- EDTA	Life Technologies
Dulbecco's phosphate buffered saline (PBS)	Life Technologies
DMEM	Life Technologies
Fetal Bovine Serum	Life Technologies
Novex® TBE Sample Buffer	Life Technologies
Novex® TBE Gels 10%	Life Technologies
NuPAGE® MES SDS Running Buffer	Life Technologies
NuPAGE® LDS Sample Buffer	Life Technologies
NuPAGE® Novex® 4-12% Bis-Tris Gels	Life Technologies
Opti-MEM® Reduced Serum Media	Life Technologies
Platinum® Taq DNA Polymerase	Life Technologies
Superscript® III Reverse Transcriptase	Life Technologies
Magnetic Dynabeads® Protein G	Life Technologies
Cytobuster™ Protein Extraction Reagent	Merck Millipore (MA, USA)
ReBlot Plus Strong Antibody Stripping Solution 10x	Merck Millipore
EMSURE® methanol for analysis, EMSURE® isopropanol for analysis	Merck Millipore
EMSURE® ethanol absolute for analysis	Merck Millipore
BSA 100x	New England Biolabs (MA, USA)
Tris base	Promega (WI, USA)
Complete protease inhibitor cocktail tablets (EDTA-free)	Roche Applied Science (IN, USA)
Minisart® syringe filters	Sartorius AG (Gottingen, Germany)
Ethidium bromide	Sigma-Aldrich (MO, USA)
Glycerol	Sigma-Aldrich
Chloroform	Sigma-Aldrich

Tween® 20	Sigma-Aldrich
Hydroquinone	Sigma-Aldrich
Sodium metabisulfite	Sigma-Aldrich
Tri Reagent®	Sigma-Aldrich
NaOH	Sigma-Aldrich
KCl	Sigma-Aldrich
Sodium Citrate	Sigma-Aldrich
RNase A/T1	Thermo Fisher Scientific
Nunc® cell culture dish (15 cm)	Thermo Scientific (MA, USA)
$MgCl_2$	VWR International (PA, USA)
NaCl	VWR International
Sucrose	VWR International

2.1.2 Kits

Kit	**Manufacturer**
TruSeq DNA LT Sample Prep Kit v2, MiSeq Reagent Kit v2 (300 cycles)	Illumina (CA, USA)
Library Quantification Kit-Illumina/Universal	Kapa Biosystems (MA, USA)
Qubit® dsDNA BR Assay Kit	Life Technologies
Wizard® SV Gel and PCR Clean-Up System	Promega (WI, USA)
Pierce™ BCA Protein Assay Kit	Thermo Scientific (MA, USA)
QiaQuick® Gel Extraction Kit	Qiagen (Hilden, Germany)

2.1.3 Antibodies

Antibody	**Manufacturer**	**Catalogue Number**
Rabbit anti-Radixin	Abcam (Cambridge, UK)	ab52495
Rabbit anti-Ezrin	Abcam	ab41672
Mouse anti-histone H3	Abcam	ab10799
Mouse anti-AGO2	Abcam	ab57113
Mouse anti-GAPDH	Abcam	ab8245
Polyclonal rabbit anti-mouse immunoglobulin	Agilent Technologies (CA, USA)	P0260

Rabbit anti-GW182	Bethyl Laboratories (TX, USA)	A302-329A
Goat anti-Rabbit IgG antibody, peroxidase conjugated	Merck Millipore (Darmstadt, Germany)	AP132P
Mouse anti-ACO2	Mitosciences (OR, USA)	MS793
Rabbit anti-NSUN4 polyclonal	Protientech (IL, USA)	16320-1-AP
Rabbit anti-NSUN42 polyclonal	Protientech	20854-1-AP
Normal rabbit IgG	Santa Biotechnology (TX, USA)	sc-2027
Normal mouse IgG	Santa Biotechnology	Sc-2025
Monoclonal anti α-tubulin antibody, produced in mouse	Sigma-Aldrich (MO, USA)	T9026

2.1.4 Oligonucleotides

Primers used throughout this thesis were purchased from Integrated DNA Technologies (IA, USA). Primers were purified by standard desalting.

2.1.5 siRNAs

siRNAs were purchased from Dharmacon (CO, USA).

siRNA	**Targeted sequences**	**Target gene**
siGENOME Human NSUN2 siRNA SMARTpool	5´-UAAGAAAGAUGGCGUGUGU-3´ 5´-GAGUUGGUAUCCUGAAGAA-3´ 5´-GAGCGAUGCCUUAGGAUAU-3´ 5´-GGAGACGGCACUAUGAGAA-3´	NSUN2
siGENOME Human NSUN4 siRNA SMARTpool	5´-GACCGCCACUCCCUUCAUG-3´ 5´-ACACUAGCGUUGCUUCAGA-3´ 5´-AUGGUGCACUGGUCAAUAA-3´ 5´-CCGCGGAGACATCGATATA-3´	NSUN4
siGENOME non-targeting control pool #1	5´- UAGCGACUAAACACAUCAA-3´ 5´- UAAGGCUAUGAAGAGAUAC-3´ 5´- AUGUAUUGGCCUGUAUUAG-3´ 5´- AUGAACGUGAAUUGCUCAA-3´	N/A

2.2 Methods

2.2.1 Unpublished SOLiD bsRNA-seq dataset

RNA was extracted from HeLa cells transfected with NSUN2, TRDMT1, DNMT1 siRNAs or NTC control using Tri Reagent®. Genomic DNA was removed using TURBO™ DNase. 40 µg of total RNA was depleted of rRNA using the RiboMinus™ Eukaryote Kit according to the manufacturer's protocol. 2-2.5 µg of rRNA-depleted RNA was treated with bisulfite as described below. Library preparation was performed as per the manufacturer's protocol. Briefly, the bisulfite-converted RNA was first end-polished using Tobacco Acid Pyrophosphatase and T4 Polynucleotide Kinase. Specific adaptors were then ligated to RNA, followed by conversion to cDNA using specific SOLiD reverse transcription primers. A 50 -120 nt insert size was chosen, followed by PCR amplification and purification of the libraries by ethanol precipitation. The libraries were then sequenced using the SOLiD 5500 XL platform (Life Technologies, Carlsbad, CA, USA).

2.2.1.1 Alignment strategy and identification of candidate m^5C sites

Read mapping and m^5C site identification was performed by Dr Brian Parker (scripts will be made available on Github under "parker-lab"). Multi- mapping reads were discarded. Uniquely mapped reads were defined by employing the Bowtie parameters –best and –strata, where reads with the least number of mismatches are retained. Additional criteria included setting the n, l, e parameters as follows: -n 3 -1 35 -e 200, where three mismatches in colour-space (-n) are allowed in the first 35 bases (-1) of the 50 nt SOLiD sequencing reads. The remaining bases may have any number of mismatches as long as their total Phred score is below 200 (-e). These values were chosen to allow for more mismatches in the less reliable 3´ end of the read. Also it should be noted that the SOLiD reads are in colour space, where two mismatches in colour space equates to a single mismatch in base space. Overall, the -e parameter was chosen to allow for an equivalent of three mismatches in base space.

To decrease the false-positive rate, manual filtering steps were taken. Firstly, candidate m^5C sites that overlapped with either single nucleotide polymorphisms (SNPs), from the dbSNP database (Sherry *et al.*, 2001), or editing sites, from the DARNED database, were removed (Kiran & Baranov, 2010). Also

candidate m^5C sites that showed C/T coverage less than 95% were also removed due to the high sequencing read error rate or the presence of as yet unannotated. That is, the called site may actually coincide with a SNP that has not been identified yet.

2.2.2 Illumina bsRNA-seq dataset

Read mapping and detection of m^5C sites were performed by Dr Maurits Evers. Library preparation for the Illumina sequencing was performed by Beijing Genomics Institute (BGI), China according to the Illumina protocol and guidelines. Briefly, Ribo-Zero Gold was used to remove cytoplasmic and mitochondrial rRNA, before library construction. RNA was then bisulfite treated (see below) which resulted in the chemical fragmentation of the RNA. The RNA was then DNase treated and random primed for reverse transcription. For library preparation, the Illumina TruSeq Stranded Total RNA Sample Preparation Kit was used. First strand synthesis and second strand synthesis (using dUTP instead of dTTP) was then performed, followed by end repair, A-tailing, adaptor ligation and finally PCR amplification.

The average insert size of chosen for the libraries was approximately 150 bp. The libraries were sequencing using the Illumina platform; paired end sequencing, 50 bp read length was selected.

2.2.2.1 Alignment strategy and identification of candidate m^5C sites

Reads are quality and adapter trimmed (minimum trailing/leading Phred score = 20, average quality = 15, minimum read length = 36). Bowtie2 is used for the alignment, where only one mismatch per 20 nt seed is allowed. The total alignment efficiency (uniquely mapped + multi-mapped reads) is 50% and 74% for the HeLa 1 and HeLa 2 libraries respectively. Following alignment, the two HeLa libraries were pooled and methylated cytosines were identified according to the following parameters: 1) a minimum of 10 reads, 2) at least 10% methylation rate, and 3) overlap with annotated sequence features (e.g. mRNA exons, lncRNAs, miRNAs, snoRNAs, etc). To obtain a more stringent (high confident) list of m^5C sites, two filters were imposed. First, candidate m^5C sites which showed a C:T/G:A coverage of less than 80% of the total coverage was discarded (in total 9 sites). Second, clustered sites (i.e. ≥ 3 successive m^5C methylation calls) were also discarded, as they most likely are artifacts due to non-conversion resulting from strong secondary structure (in total 85 sites).

2.2.3 Maintaining HeLa cells

HeLa cells were cultured in Dulbecco's Modified Eagle's Medium (DMEM) supplemented with 2 mM L-glutamine and 5% fetal bovine serum (FBS). The cells were passaged when 90-100% confluency was reached. A T-75 flask (flask with a 75 cm^2 surface area) was seeded with approximately 2 million cells. 48 hours later cells were passaged: The cells were washed with 5 ml pre-warmed Phosphate Buffered Saline (PBS), then 2 ml of pre-warmed 0.25% Trypsin-EDTA was added and the flask was incubated at 37°C for 5 min, which resulted in the dissociation of the cells. The cells were pipetted to avoid clumping, and the trypsin was inactivated by addition of 8 ml of the supplemented DMEM. HeLa cells were maintained at 37°C with 5% CO_2.

2.2.4 Identifying m^5C sites in mitochondrial RNA

2.2.4.1 Mitochondrial and mitoplast isolation

The protocol for mitochondrial isolation was adapted from (Mercer *et al.*, 2011). Briefly, mitochondria were isolated from 2 x 10^7 HeLa cells (a 15 cm dish). Cells were washed in PBS and harvested (800 x *g* for 5 min at 4 °C), then resuspended in 8 ml/dish of ice-cold homogenisation buffer (250 mM sucrose, 10 mM NaCl, 1.5 mM $MgCl_2$, and 10 mM Tris-HCl pH 7.6). Cells were homogenised on ice, 5 times with a loose glass homogeniser, and then 10 times with a tight glass homogeniser (Cat No. D9063 Sigma Dounce tissue grinder set, 7 ml). The suspension was then centrifuged at 1,300 x *g* for 5 min and the supernatant retained. The supernatant was centrifuged at 12,000 x *g* for 15 min to sediment the mitochondria. The pellet from this spin contains mitochondria and was kept, the supernatant represented the cytosol fraction. The pellet containing mitochondria was resuspended in 1 ml 250 mM sucrose in $T_{10}E_{20}$ buffer (10 mM Tris-HCl, and 1 mM EDTA pH 7.5) and centrifuged again at 12,000 x *g* for 15 min. The supernatant was discarded. The pellet was resuspended in 100 µl 250 mM sucrose in $T_{10}E_{20}$ buffer.

Mitoplasts were prepared from a crude mitochondrial isolation using the non-ionic detergent, digitonin. ~0.1 mg of digitonin was added per mg of mitochondrial protein and incubated for 15 min on ice. Mitoplasts were isolated by spinning at 15, 000 x *g* for 10 min, the supernatant was saved for subsequent analyses. Mitoplasts were then resuspended in 250 mM sucrose and $T_{10}E_1$ buffer (10 mM Tris-HCl, 1 mM EDTA, pH

7.6) and incubated for 30 min at 25°C with 0.1 mg/ml RNase A, followed by washing twice in 1 ml of 250 mM sucrose and $T_{10}E_{20}$ buffer. RNA or protein is isolated as below and analysed for purity by RT-qPCR and western blotting respectively.

2.2.4.2 siRNA transfection of HeLa cells

HeLa cells were reverse transfected with siGENOME Human siRNA SMARTpools (ThermoFisher) using Lipofectamine® RNAiMAX (Invitrogen) over a 6-day protocol. Reverse transfection is where freshly passaged cells are added to pre-plated transfection complexes. Transfections were performed at time 0 hr and 72 hr. 5 µl Lipofectamine® RNAiMAX or 150 pmoles siRNA was incubated at room temperature with 250 µl Opti-MEM transfection medium for 5 min. Following the incubation the two mixtures were combined and incubated at room temperature for 20 min to allow siRNA/Lipofectamine® RNAiMAX complexes to form. The 500 µl of the complex was then added drop-wise to a 6-well plate seeded at a density of 1.5 x 10^5 HeLa cells. 2 ml of supplemented DMEM was added to each well for a final volume of 2.5 ml. The cells were maintained in an incubator at 37°C with 5% CO_2. 72 hr post the first transfection, the cells were passaged and seeded at the same density as before for a second round of transfection. 144 hr post-initial transfection RNA was isolated by adding 1 ml Tri Reagent® directly to PBS washed cells. Protein was isolated by adding 300 µl Cytobuster™ Protein Extraction directly to PBS washed cells. The transfections were performed in triplicate.

2.2.4.3 RNA extraction and precipitation

Media was aspirated from adherent HeLa cells, washed with PBS before 1 ml of Tri Reagent® was directly added to the cells and incubated for 5 min at room temperature. 200 µl of chloroform was added to the 1 ml Tri Reagent®, and the suspension, was mixed vigorously for 15 sec and incubated at room temperature for 3 min. The emulsion was centrifuged at 12,000 x *g* for 15 min at 4°C, after which the aqueous phase was transferred to a clean 1.5 ml Eppendorf tube containing 1 µl glycogen. The RNA was precipitated by adding 500 µl isopropanol and allowing to incubate at room temperature for 10 min, followed by centrifugation at 12,000 x *g* for 10 min at 4°C. The supernatant is removed and 1 ml of cold 75% ethanol was added, the tube was vortexed briefly. Another centrifugation at 7,500 x *g* was performed for 5 min at 4°C, the supernatant

was then removed. The RNA pellet was air dried in a biohazard cabinet for 5 – 10 min and resuspended in 50 µl of dH_2O. For purification purposes, 0.1 volumes of 3 M NaOAc, 3 volumes of 100% ethanol and 1 µl glycogen was added to the isolated RNA and precipitated overnight at -80°C. The following day, the suspension was centrifuged at 17,000 x *g* for 30 min at 4°C and the RNA pellet washed with 75% ethanol as before. The pellet was finally resuspended in 25 µl dH_2O.

2.2.4.4 DNase treatment of RNA

Genomic DNA was removed from isolated RNA using TURBO™ DNase as per the manufacturer's protocol. Briefly, 2 µg total RNA in 17 µl dH_2O was treated with 1 µl TURBO™ DNase in 1x TURBO™ DNase Buffer (2 µl) and incubated at 37°C for 30 min. TURBO™ DNase was removed by adding 3 µl TURBO™ DNase Inactivation Reagent and incubated at room temperature for 5 min. The tubes were then spun for 1.5 min and the supernatant transferred to a clean tube. RNA was stored at - 80°C.

2.2.4.5 cDNA synthesis

RNA was reverse transcribed into cDNA using Superscript® III Reverse Transcriptase (RT) according to the manufacturer's protocol. Briefly, 1 µl of 20x random primer mix, 1 µl of 1 mM dNTP mix, and 1 µl of dH_2O were added to 1 µg of DNase treated RNA (in 10 µl) and incubated at 65°C for 5 min, followed by incubation on ice for 2 min. For reverse transcription, 1 µl of 0.1 M DTT, 1 µl of Superscript® III RT, 1 µl of RNaseOUT™, 4 µl of First-strand Buffer were then added and incubated for 50°C for 90 min, then 75°C for 15 min. For the -RT control the 1 µl of RT was replaced with dH_2O. For bisulfite PCR, 200 ng cDNA was diluted 1:10 in dH_2O. For RT-qPCR, 1 µg cDNA was diluted 1:10 in dH_2O.

2.2.4.6 RT-qPCR

Forward and reverse primers were designed and confirmed for uniqueness using PrimerBlast (Ye *et al.*, 2012). Primers were designed to span exon-exon junctions to avoid amplification of genomic DNA (real time primers used for RT-qPCR are listed in Table 2.1). RT-qPCR was performed with Fast SYBR® Green Master Mix in a 384-well plate format on the Applied Biosystems 7900HT Fast Real-Time PCR System (CA, USA). Each reaction was performed in triplicate using 2 µl diluted 1:10

Table 2.1: Real time primers used in this thesis

Real time PCR primers		
Gene	**Forward primer (5′– 3′)**	**Reverse primer (5′–3′)**
NSUN2	TGGAATAAACGTCAGCCAAA	GGGCTCAGCTGTGTGCTT
NSUN4	TTTGGAGATCTTTGGCCATC	GCTGCTCCAGCTTAGCACTT
COX3	CTCACTATCTGCTTCATCCG	CCCTCATCAATAGATGGAGAC
12S	TAGATACCCCACTATGCTTAGC	CGATTACAGAACAGGCTCC
GAPDH	TTGCCATCAATGACCCCTTCA	CGCCCCACTTGATTTTGGA
c-myc	TCAAGAGGTGCCACGTCTCC	TCTTGGCAGCAGGATAGTCCTT
α-tubulin	GTCTGCTCCTGTCGCCTTC	CCGTGTTCCAGGCAGTAGAG
eIF6	GCATGTGTGTGGGGAACA	CGCCTAATCTGCACTGTGTC

cDNA mixed with 8 µl master mix, which consisted of 5 µl Fast SYBR® Green Master Mix, 0.5 µl of each primer (final concentration of 5 µM) and 2 µl dH_2O per reaction. The PCR reaction conditions: initial denaturation at 95°C for 20 sec, 40 cycles of 95°C for 1 sec and the specific annealing temperature for 20 sec, followed by melt curve analysis. PCR products were run on a 2% agarose gel at 100 V for 30 min to confirm correct amplicon size.

2.2.4.7 Whole cell lysis (protein isolation)

Cells were washed with ice cold PBS. An appropriate volume of Cytobuster™ Protein Extraction Agent was added according to the manufacturer's guidelines, and incubated at room temperature for 5 min. The cell debris was scraped using a cell scraper. The cell suspension was transferred to a clean Eppendorf tube and centrifuged at 16,000 x *g* for 5 min at 4°C. The supernatant was transferred to a clean tube and stored at -80°C.

2.2.4.8 Protein quantification using BCA assay

The Pierce™ BCA Protein Assay Kit Protein was used to quantify protein lysates as per the manufacturer's protocol. Briefly, eight BSA protein standards (0-2000 µg/ml) were prepared; dilutions were made with the Cytobuster™ reagent. 25 µl of each standard and 25 µl of each protein sample were pipetted in duplicate in a 96-well plate. 200 µl working reagent (50:1, Reagent A:B) was added to each well, mixed at low speed for 30 sec and incubated at 37°C for 30 min. Absorbance was measured at 562 nm on the FluoSTAR OPTIMA plate reader (BMG Labtech, Ortenberg, Germany), after the plate down had cooled to room temperature.

2.2.4.9 Western blotting

To 10 µg of protein, 7 µl of 4x LDS Sample Buffer and 2.8 µl of 0.1 M DTT were added, and the volume was made to 28 µl with dH_2O. The sample was denatured at 70°C for 10 min, then chilled on ice for 2 min. The protein sample was loaded into a denaturing, gradient gel (NuPAGE® Novex® Bis- Tris 4-12%) and typically run at 180 V for 50 min. Proteins were transferred onto a PVDF membrane in transfer buffer (25 mM Tris, 192 mM Glycine, 20% (v/v) methanol, pH ~8.3) at 100 V for 1 hr at 4°C. The membrane was blocked with 5% skim milk in PBST (PBS, 0.2% Tween®20) for 1 hr at room temperature with gentle shaking. The blocked membrane was then incubated with

the primary antibody (at the manufacturer's recommended dilution) in 5% skim milk PBST overnight at 4°C with gentle shaking. The membrane was then washed with 1x PBST for 5 min (3 times) then incubated with the secondary antibody (at the manufacturer's recommended dilution) in 5% skim milk PBST for 1 hr at room temperature with gentle shaking. The membrane was then washed again as before. An appropriate amount of enhanced chemiluminescence (ECL) solution was added to cover the membrane and incubated in the dark for 5 min. Membranes were visualised using the ImageQuant LAS 4000 (GE Healthcare, Little Chalfort, BKM, UK).

2.2.4.10 Densitometry of western blots (ImageJ analysis)

The open source image processing software, ImageJ was used to quantify western blot protein bands (Schneider *et al.*, 2012). ImageJ converts TIFF image files from the ImageQuant LAS 4000 into greyscale 8-bit image option. The reference lane was selected using the rectangular selection tool, followed by the other lanes. The "plot lanes" function reveals a graph, where the area under each peak represents the relative intensity of each band. The peak of interest was closed off using the straight-line tool, and the "wand tool" is used to calculate the area under the selected peak. The intensity of each band was normalised to the intensity of the loading control.

2.2.5 Validation of m^5C candidate sites

2.2.5.1 Bisulfite conversion of RNA

For 20 ml of sodium bisulfite solution, 8 ml of 40% (w/v) sodium metabisulfite and 20 µl of 600 µM hydroquinone were added. The pH of the solution was adjusted to 5.1 by addition of 10 M NaOH and the volume adjusted to 20 ml with dH_2O. The final solution was then filtered through a 0.22 µm filter syringe.

2 µg of DNase-treated RNA was spiked in with non-humanised R-Luc *in vitro* transcript at final dilution of 1:1000. The RNA was first denatured by heating at 75°C for 5 min. 100 µl of pre-heated (at 75°C) sodium bisulfite solution was added to the RNA, mixed, then 100 µl mineral oil carefully pipetted on top to prevent evaporation. The tube was then covered in foil and incubated at 75°C for 4 hr. The RNA mixture was passed through Micro Bio-spin 6 chromatography columns (twice) according to the

manufacturer's protocol to remove the salt. 1 volume of 1 M Tris- HCl (pH 9.0) was then added to the RNA, overlaid with 150 µl of mineral oil and incubated for 1 hr at 75°C. The RNA was then precipitated as previously described (see above).

2.2.5.2 Bisulfite PCR

To validate candidate sites, primers were designed to span many converted cytosines. For the R-luc negative control, primers were designed to span uncoverted cytosines. Primers that span uncoverted cytosines will bias toward selecting unconverted sequences. If these amplified R-luc sequences do not contain any detectable m^5C sites, then this indicates that the bisulfite treatment was successful. Also if an m^5C site is validated using primers covering converted cytosines this gives higher confidence that the detected m^5C site is real. The bisulfite PCRs (Chapter III) were performed using diluted cDNA (1:200) and indicated primers (Table 2.2). The PCR reaction included: 1x PCR Buffer, 0.5 µl 10 mM dNTP mix, 0.5 µl of 10 µM of each primer, 0.75 µl of 50 mM $MgCl_2$ and 0.1 µl of Platinum® Taq DNA Polymerase, 1 µl of cDNA. A two-phase "touchdown" PCR program was employed. The first phase used an initial annealing temperature 5°C above the highest primer melting temperature. This was then successively decreased by 1°C per cycle until 5°C below the lowest primer melting temperature was reached. The PCR program: 95°C for 2 min, (Tm + 5°C - Tm - 5°C) cycles at 95°C for 30 sec, Tm + 5°C to Tm - 5°C for 30 sec, 72°C for 15-30 sec. The second phase followed a typical PCR program: 25-40 cycles at 95°C for 30 sec, Tm - 5°C for 30 sec, 72°C for 15-30 sec, and a final extension at 72°C for 5 min.

A small volume of the PCR products was run on 2% agarose gel to confirm expected product. The remaining volume was purified using MinElute PCR Purification Kit according to the manufacturer's instructions and stored at -20°C until library preparation.

2.2.5.3 Amplicon sequencing library preparation for MiSeq

To validate candidate m^5C sites, PCR amplicons (see above) were pooled and sequenced using the MiSeq platform (Illumina, San Diego, CA, USA). Pooled amplicons were quantified using the Qubit dsDNA BR Assay Kit according to the manufacturer's protocol. 30 ng of each purified amplicon was pooled, and the final volume was made up to 55 µl with dH_2O. Amplicon sequencing library preparation included end repair, A-tailing, ligation of adaptors and PCR amplification as outlined in the TruSeq DNA LT Sample Prep Kit protocol.

Table 2.2: List of bisulfite PCR primers used in this thesis

Bisulfite PCR primers		
m^5C site (hg19)	**Forward Primer (5′–3′)**	**Reverse Primer (5′–3′)**
chrM:10215 (sense)	GTGTGGTTTTGATTTTATATTTTT	AATCACTCATAAACCAAACTT
chrM:13811 (antisense)	AACAACATTTCCCCCACATC	GGGATTGTGTGGTGTGTGAT
chrM:9575 (sense)	TTAGGAGGGTATTGGTTTTTA	AACCATAAATACCATCAAAAAT
chrM:2644 (sense)	TAGGGATTTGTATGAATGG	ACTACTCAAAAATTAAATTCTAC
chrM:318 (antisense)	CACAACCACTTTCCACACAA	TTTTGTTTTTGGGGTTTGGT
chrM:4714 (antisense)	CTATCCTCTTCAACAATATA	TGTTATGTGAGAAGAAGTAGG
chrM:263 (antisense)	AATAACAATTAAATATCTACACAAC	TTTTGTTTTTGGGGTTTGGT
chrM:8166 (antisense)	TCCCAAACATCTAAACCAAACC	GGATGATGGGTATGAAATTGTG
chrM:16363 (antisense)	CAACAAACCTACCCACCCTTA	TGGGTTTGGAGTGAGGAGAG
chrM:16084 (sense)	TGGGGAAGTAGATTTGGGTA	ACCATACAATATTCATAATAACTAA
chrM:9294 (sense)	AGTTTATGATTTTTAATAGGGGT	CAAAAACCTTTTTAAACAAATAA
chrM:15200 (sense)	AATATTATTTTGAGGGGTTA	TTTAATCCCATTTCATACAAAAA
chrM:11915 (antisense)	CAAACACACTCACAATCACATCAT	TGTTGAGTTTTGTAAGTAGGAGAGTGA
chrM:11494 (antisense)	TCACTACCCAAAAACTATCAAACTC	ATGTGTTTTGTTAGGGGGTTGAG
chrM:1490 (sense)	ACTTATCTCCTCTATATAAAT	TTAAGGGTTGAAGGTGGATTTAG
$tRNA^{Asp(GUC)}$	CTCCCCATCAAAAAATCAA	TTTTTGTTAGTATAGTGGTG
Non-humanised R-luc	CATTATAATTCCTAACAATTT	GAAGAAGGAGAAAAAATGGTT

2.2.5.4 Quality control after library preparation

To ensure successful ligation of adaptors during library preparation, 2 µl of the pooled amplicons (pre-library preparation) and 1 µl of each library were run on a 2% agarose gel at 100 V for 40 min.

2.2.5.5 Quantification of libraries

Libraries were quantified using the Qubit dsDNA BR Assay Kit as per the manufacturer's protocol. The concentration of the libraries was then adjusted to ~50 nM. For accurate quantification, the concentration of the diluted libraries was determining by qPCR using the Illumina Library Quantification Kit according to the manufacturer's protocols.

2.2.5.6 Analysis of amplicon sequencing

Illumina adaptor sequences were trimmed using Trimmomatic (Bolger *et al.*, 2014) in palindromic mode. Sequencing reads were aligned to the reference sequence with Bismark, using Bowtie2 and implementing the parameters bismark --non_directional --bowtie2 (Krueger & Andrews, 2011). The reference sequences consisted of each unconverted amplicon sequence. A single mismatch was allowed in the alignment, not counting the candidate m^5C site at the expected position within the amplicon. The methylation stoichiometry was determined by extracting the total number of cytosines and thymines and the candidate m^5C site position.

2.2.5.7 Generating coverage plots and heatmaps in R programming language

Coverage plots were generated using the plot (…) function in the R statistical computing environment.

Heatmaps were generated using the heatmap (…) function in the gplots package in the R statistical computing environment.

2.2.6 Molecular Function of m^5C

2.2.6.1 Overlapping HITS-CLIP data and candidate m^5C sites

Distances between publicly available HeLa AGO2 HITS-CLIP (Hafner *et al.*, 2010) or HuR PAR-CLIP (Lebedeva *et al.*, 2011) footprint densities and candidate 3´ UTR or CDS m^5C localised sites identified in the SOLiD bsRNA-seq experiment were calculated. The plot summarising this was anchored at candidate m^5C sites in mRNAs indicating distances to footprints in 500 nt flanking regions; introns were computationally removed. Candidate m^5C sites were randomly shuffled to produce the random shuffled negative control distribution plot. The significance of RBP footprint density relative to the random shuffled negative control was determined by performing a permutation test with 100 iterations.

2.2.6.2 Overlapping conserved miRNA target sites with candidate m^5C sites

Conserved 7-mer motifs were divided into two sets: 1) those that were the reverse complement of miRNA seed regions or 2) those that were the complement of miRNA seed regions (negative control) in miRBase v19. Distances between these motifs and m^5C localised 3´ UTR or CDS candidate sites identified in the SOLiD bsRNA-seq experiment were calculated. As before plots summarising this were generated by anchoring at the candidate m^5C site, indicating distances to conserved miRNA target sites within 200 nt flanking regions. As before, introns were computationally removed. Again, as before, candidate m^5C sites were randomly shuffled to generate the random shuffled negative control. The significance of the association between conserved 7-mer motifs and candidate m^5C sites compared to the random shuffled negative control was again determined using a permutation test with 100 iterations (scripts will be made available on Github under “parker-lab”).

2.2.6.3 Co-immunoprecipitation of AGO2 and NSUN2

Whole cell lysate was used directly in immunoprecipiation experiments, or was treated with 2 µg/ml RNase A and 0.6U/ ml RNase T1, 10 min at 37°C prior to use.

Co-immunoprecipitation was performed with 1 mg of Magnetic Dynabeads® Protein G (Life Technologies) incubated with 4 µg of either Anti-AGO2 (Abcam) or Anti-NSUN2 (Proteintech) antibody diluted in 200 µl 0.05% PBST for 1 hr at 4°C with rotating. The supernatant was discarded. 300 µg of HeLa whole cell lysate was then added and

incubated overnight at 4°C with rotation. The supernatant was removed and stored for subsequent analysis. The immunocomplex was washed three times with NT2 buffer (50 mM Tris-HCl pH 7.4, 150 mM NaCl, 1 mM $MgCl_2$, 0.05% Tween20). After each wash the supernatant was discarded. Elution of the immunocomplex was achieved with the addition of 20 µl 0.1M citrate pH 3, 10% DTT, 4X LDS with heating at 70°C for 10 min. The eluate and supernatants were analysed by western blotting.

2.2.6.4 RNA pulldown approach

To ensure the biotinylated RNA baits folded in the correct structure, 1 µg of the RNA baits was denatured by heating at 90°C for 2 min, then chilled on ice for 2 min. The RNA was supplemented with 10 µl RNA structure buffer (10 mM Tris pH 7, 0.1 M KCl, 10 mM $MgCl_2$) and incubated at room temperature for 20 min. 300 µg of protein lysate was added to 700 µl of binding buffer (150 mM KCl, 1.5 mM $MgCl_2$, 10 mM Tris pH 7.5, 0.5 mM DTT, 1x complete protease inhibitor). The RNA baits (2 µg) were then added, and the mixture was incubated at room temperature for 30 min with rotation, before being incubated at 4°C for 2 hr with rotation. The mixtures were then transferred to a 35 mm sterile plate and two doses of UV crosslinking (400mj/cm^2 + 120mj/cm^2) were performed using the Stratagene Stratalinker® UV crosslinker (CA, USA). The streptavidin-magnetic beads (50 µl) were prepared by washing three times with binding buffer. The magnetic beads were isolated using a magnetic rack, and the supernatant was removed by pipetting. The washed beads were then added to the cross-linked lysate and incubated at 4°C for 2 hr with rotation. The supernatant was then removed and stored for subsequent analysis. The magnetic beads were washed twice with binding buffer as before. For elution of bound protein complexes, the beads were incubated with a 60 µl of RNase mixture containing 10 mM Tris-HCl (pH 7.2), 1 mM $MgCl_2$, 40 mM NaCl, 2 µl of RNase A (1 mg/ml), and 3 µl of RNase T1 (0.2 U/µl) at 37°C for 30 min with occasionally shaking. The supernatant eluate (containing the proteins of interest) was then collected as described. 7 µl of 4x LDS, 2.8 µl of 1M DTT was added to 18.2 µl of each eluate and loaded onto a 4-12% Bis-Tris gel. The protein bands were visualised using Coomassie Blue staining. Bands of interest were sliced out using a scalpel and sent off to the BMSF mass spectrometry facility in University of New South Wales (NSW, Australia) for identification. Mass spectrometry results were analysed using Scaffold 4 proteome software (OR, USA).

2.2.7 Investigating features of m^5C methylation in mRNA

2.2.7.1 Gene ontology analysis

The gene ontology (GO) analysis was performed using The DAVID Gene Functional Classification Tool (Huang *et al.*, 2007). GO terms were identified for the list of candidate m^5C sites identified in the Illumina bsRNA-seq experiment relative to a list of background genes (not containing a m^5C site) expressed in the bsRNA-seq.

2.2.7.2 RNAModR

RNAModR was designed as a package within the statistical computing environment R, R Core Team (2016). The package is a set of tools which allows for metagene analyses between a post-transcriptional modification and a genomic feature e.g. mRNA regions, GC content, splice-sites and/or using other data e.g. RNA-binding protein footprint data. RNAModR was developed by Dr Maurits Evers in collaboration with myself.

RNAModR depends on a list of existing R/Bioconductor packages that facilitate the mapping of genomic to transcriptomic loci and provide methods for handling genome and transcriptome annotations and sequences. RNAModR requires BED files with genomic coordinates of a post-transcriptional modification.
We discuss key steps in a typical RNAModR workflow analysing a set of RNA modifications provided as a BED file of genomic loci.

Firstly, RNAModR generates a custom reference transcriptome (function BuildTx()). The function automatically downloads a UCSC RefSeq-based reference genome and annotation, and builds a custom transcriptome. This is achieved by collapsing the multiple transcript variants for a give gene into one with the longest CDS and associated longest 5´/3´ UTRs. Transcripts regions are defined as 5´ UTR, CDS, 3´ UTR, and introns. The position of m^5C candidate sites was visualised across and within the different transcript regions using the function (PlotSectionDistribution(...), PlotSpatialDistribution(...), respectively.

RNAModR was also used in the enrichment analyses in Chapter V. For example, the enrichment or depletion of RNA modifications within a particular transcript section was investigated. Also whether RNA modifications localised with particular transcript

features e.g. the start codon, stop codon was also addressed (Chapter V). Null sites were provided manually as a list based on the unmethylated genes from the bsRNA-seq data. The null sites were comprised of *non*-modified cytosines. Enrichment/depletion of candidate m^5C sites per mRNA region was visualised using the PlotSectionEnrichment(...) function. Enrichment/depletion of candidate m^5C sites within each mRNA region was visualised using the PlotSpatialEnrichment(...) function. Furthermore, determining distances between candidate m^5C sites and RNA-binding protein footprint densities was visualised using the PlotRelDistDistribution (…) function, whilst enrichment/depletion of a particular RBP footprint with candidate m^5C sites compared to unmethylated cytosines was visualised using PlotRelDistEnrichment (…) function. These same commands were also used to determine any association between candidate m^5C sites and exon-exon junctions. Also, the difference in the distribution of GC content around candidate m^5C sites and unmethylated cytosines was evaluated using a two-tailed t-test (function PlotGC(...).

Overview of datasets used in this thesis

Three different transcriptome-wide datasets of m^5C candidate sites in HeLa cells were available as source material for the analyses presented in this thesis. The datasets were generated in the lab over a period of five years, thus representing different developmental stages of high-throughput sequencing technology as well as the associated bioinformatics methods. Two datasets were generated with the now superseded SOLiD technology, which used a sequencing-by-ligation approach that does not directly yield nucleotide sequence. The output instead consists of sequence reads in 'four colour space', which are 'translated' into base space during mapping. Essentially one colour represents two bases; this can be deconvoluted as long as the identity of one of the two bases is known. The most recent dataset was generated with the current Illumina technology, which uses the more conventional sequencing-by-synthesis method. The approaches to read mapping and m^5C candidate site selection were also very different for each of the three datasets.

The first dataset was featured in published work from the lab (Squires *et al.*, 2012). Briefly, a single run of sequencing on a SOLiD version 3 instrument (50 sequencing cycles) was performed on bisulfite treated, rRNA depleted, poly$(A)^+$ RNA. Reads were mapped to a custom human reference transcriptome with SOCS-B, an alignment tool designed to map bisulfite-converted SOLiD colour-space reads (Ondov *et al.*, 2008). Selection criteria of ≥ 10 read coverage, $\geq 20\%$ methylation and ≤ 2 mismatches, were defined as thresholds, revealing a total of 10,530 candidate m^5C sites in both coding and non-coding RNAs.

The second dataset used similar experimental approaches as above, but analysed four separate RNA samples derived from RNAi-treated cells. Specifically, knockdowns were performed of the m^5C: RNA methyltransferases NSUN2 and TRDMT1, with siRNA against the DNA methyltransferase DNMT1 and a non-targeting control (NTC) siRNA serving as controls. This dataset is currently being used to identify target sites of NSUN2 and TRDMT1 in separate work. Libraries were generated from the four RNA samples after rRNA depletion and bisulfite treatment and sequenced on a SOLiD 5500 XL instrument (50 sequencing cycles). The aligner B-SOLANA in combination with Tophat was then used to successively map reads to the *in silico* bisulfite converted human Refseq transcriptome and the *in silico* bisulfite converted hg19 reference genome. The criteria for inclusion of sites included requiring ≥ 5 reads in all samples, $\geq 10\%$ mean methylation across the at least three of the four knockdown samples, a single mismatch, and a FDR $\leq 1\%$, yielding a total of 4,254 candidate m^5C sites in both coding

and ncRNA.

The most recent dataset used two replicates of rRNA-depleted and bisulfite converted HeLa RNA and employed current Illumina sequencing technology. Mapping and downstream bioinformatics analyses of these data are described in Chapter V. Briefly, read alignment was performed using Bismark in conjunction with Bowtie2, allowing for one mismatch in the 20 nt seed. The m^5C selection criteria included requiring ≥ 10% mean methylation, ≥ 10 read coverage and a FDR ≤ 1%. This yielded 847 m^5C candidate sites when mapping to a completely *in silico* converted human hg38 reference genome.

The different sequencing technologies coupled with the different approaches used to identify and threshold m^5C candidate sites will create differences in transcript coverage as well as the type and proportion of false positive and false negative sites for each dataset. Consequently, there is a rather poor overlap between the three datasets in all but the most 'robust' types of sites. For example, all three datasets converge well in detecting m^5C sites within tRNAs, which are an abundant RNA biotype featuring near-quantitative m^5C methylation in a small number of well-defined structural positions (see Chapter V; Fig. 5.1). By contrast, sites in less abundant noncoding RNAs and mRNAs are typically only sub-stoichiometrically methylated and were rarely detected in more than one of the datasets. Nevertheless, a number of such novel and unconventional m^5C sites identified from each dataset have been validated by amplicon-bsRNA-seq (Squires *et al.*, 2012) and in Chapter III).

In this thesis, I have chosen to primarily use the set of m^5C candidate sites identified from the Illumina dataset as these were obtained with current sequencing technology and bioinformatics tools, and selected with the most stringent criteria. This dataset (henceforth referred to as "high-confidence" m^5C set of sites) was used to determine features of m^5C sites and in the metagene analyses presented in Chapter V. Both SOLiD and Illumina datasets were also used to provide a greater initial set of mitochondrial candidate m^5C sites for validation by amplicon-bsRNA-seq (Chapter III).

Chapter III

Identifying the mitochondrial m^5C methylome

3.1 Preamble

Many nucleotides in mitochondrial rRNAs and tRNAs are targeted for post-transcriptional modification. However, the current knowledge surrounding the enzymes responsible for these modifications and the function of these modifications is comparatively limited. For example, there are reports of over two-dozen modified nucleosides in a number of mitochondrial tRNA species, though only the enzymes responsible for five of these modifications have been identified (Suzuki & Suzuki, 2014). Similarly, nine modifications have been detected in both the 12S and 16S mitochondrial rRNAs but only two 'writer' enzymes have been identified (Rorbach & Minczuk, 2012). This is in contrast to the large number of modified nucleosides, and the responsible enzymes, identified in both eukaryotic and bacterial rRNAs (Piekna-Przybylska, Decatur, *et al.*, 2008). However, there is more information known regarding the role of mitochondrial post-transcriptional modifications than the number or the 'writer' enzymes involved. Motorin and Helm (2010) have shown that mt-tRNA modifications are required for their correct function. For example, most of the modifications are thought to afford RNA stability. Furthermore, a few connections between mitochondrial modification defects and disease have been suggested (Kirino & Suzuki, 2005). Also, modifications of mitochondrial RNA may have a role in regulating RNA abundance, as all mitochondrial transcripts originate from the same polycistronic transcript, even though differential RNA abundance is observed.

Thus, as little is known generally regarding post-transcriptional modifications in mitochondrial RNA, the aim of this Chapter was to identify the mitochondrial m^5C methylome profile and the enzymes that mediate the m^5C modification in this context.

Previously, Squires *et al.* (2012) identified a single m^5C site in human 12S mt-rRNA. They also showed that this site was not regulated by either NSUN2 or TRDMT1, the only two confirmed m^5C MTases at the time (Squires *et al.*, 2012). As NSUN3 and NSUN4 exclusively localise to the mitochondria (Rhee *et al.*, 2013; Camara *et al.*, 2011), and possess the domains required for methyltransferase activity, they represent prime candidates to target mt-RNA for m^5C modification. This thesis advances the current knowledge by demonstrating that NSUN4 is responsible for the m^5C modification in human 12S mt-rRNA, and discusses the role of m^5C in this context.

3.1.1 Mitochondrial genome

The human mitochondrial DNA (mtDNA) is a double-stranded, circular genome of 16,569 bp and contains a total of 37 genes coding for two rRNAs, 22 tRNAs and 13 proteins (Fig. 3.1). The two strands are referred to as the heavy (sense) and light (antisense) strands. Mitochondrial (mt) DNA also contains a stretch of approximately 1,100 nt that is non-coding. This region is termed the D-loop region.

The transposition of mtDNA fragments into the nuclear genome during evolution gave rise to the nuclear sequences of mitochondrial origin (NUMTs). For example, both (Parr *et al.*, 2006; Lascaro *et al.*, 2008) report that many human NUMTs have been identified and cover the entire human mitochondrial genome. Interestingly, a human NUMT that covers 90% (14,654 bp) of the human mitochondrial genome has been reported earlier (Mourier *et al.*, 2001). Clearly, the mitochondrial genome shows high sequence similarity to nuclear encoded sequences (Fig. 3.2), which can be problematic when mapping reads to the mitochondrial genome.

3.2 Results

In order to gain an accurate representation of m^5C sites in the mitochondrial transcriptome, it would be necessary to isolate mitochondrial RNA in high purity to perform a specialised mitochondrial bsRNA-seq experiment. Unfortunately, this approach turned out to be unfeasible and an alternative practice was adopted. Mitoplasts (mitochondria stripped of their outer membrane) were isolated with high purity; nuclear contamination and mitochondrial enrichment were assessed at both the RNA level by RT-qPCR (Table 3.1 for primer list) and protein level by western blotting (Fig. 3.3A and B, respectively). However, the low RNA yield obtained from these mitoplasts meant that an excessive amount of cells would be required to yield enough RNA to perform downstream rRNA depletion, followed by bisulfite treatment. To put it into numbers, a 15 cm dish of confluent HeLa cells yielded approximately 1 µg of pure mitoplast RNA. Therefore, 15 dishes would be required per replicate to yield enough RNA. Furthermore, a considerable amount of siRNAs would be required to obtain the NSUN4 KD sample, especially as knockdown of NSUN4 causes significant cell death. Thus, the cost and time associated with such an undertaking did not make sense, in light of existing transcriptome-wide bsRNA-seq datasets, which were mined for information

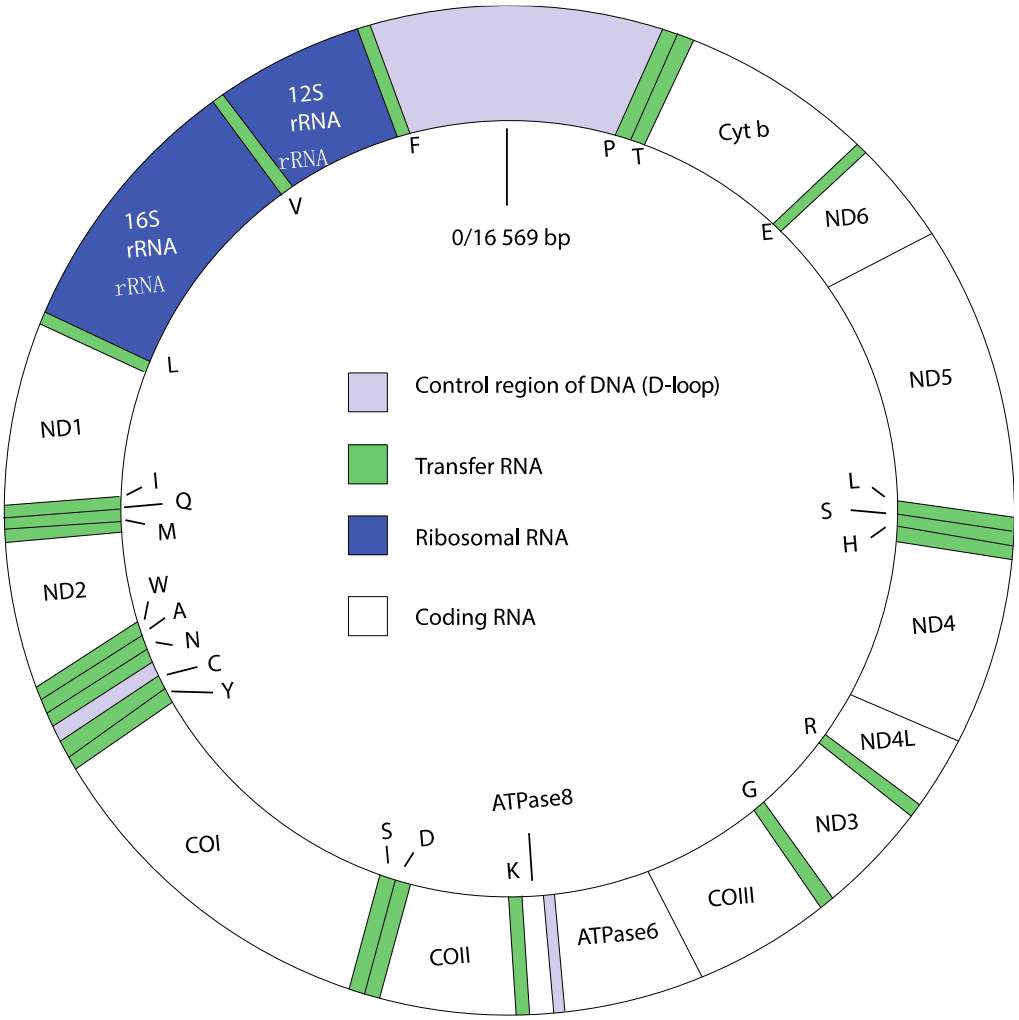


Figure 3.1: Human mitochondrial genome.

Double-stranded, circular molecule of 16,569 bp and contains a total of 37 genes coding for two rRNAs (blue), 22 tRNAs (green) and 13 mRNAs (white). Non-coding regions (D-loop) are coloured in light purple. Single letter code refers to tRNA annotation.

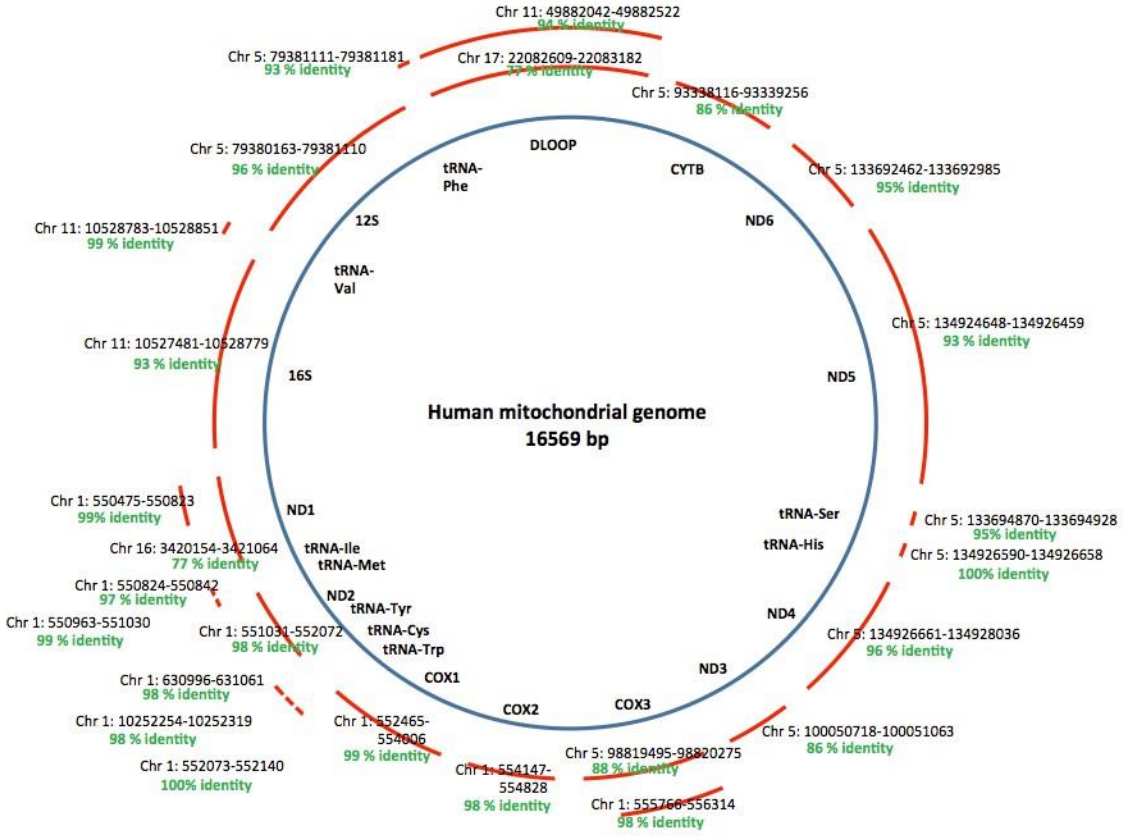
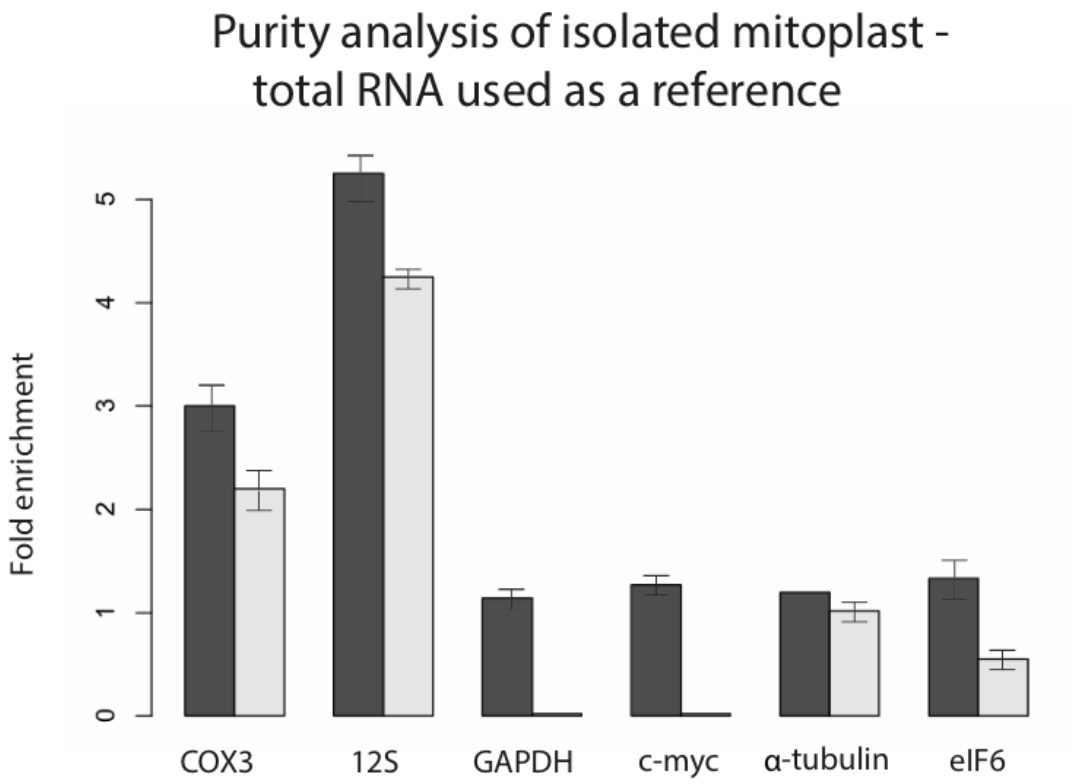


Figure 3.2: BLAST search of human mitochondrial genome revealing a number of highly similar sequences to nuclear encoded genes.

The red lines represent nuclear encoded sequenes (chromosomal location and percentage identity is given) that are similar to the corresponding mitochondrial genes as shown by the labelled blue circle.

(A)


Purity analysis of isolated mitoplast - total RNA used as a reference
Fold enrichment
0
1
2
3
4
5
COX3
12S
GAPDH
c-myc
α-tubulin
eIF6

(B)

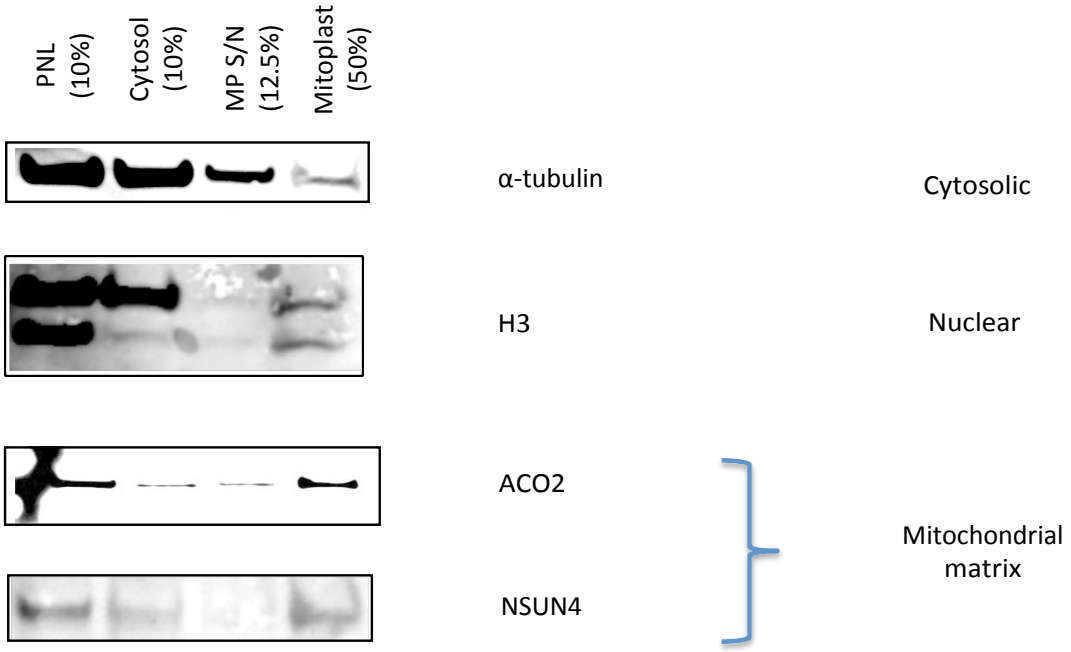

PNL (10%)
Cytosol (10%)
MP S/N (12.5%)
Mitoplast (50%)
α-tubulin
Cytosolic
H3
Nuclear
ACO2
NSUN4
Mitochondrial matrix

Figure 3.3: Quality control measures to ensure purity of mitoplast isolation.

(A) RT-qPCR to assess purity at RNA level of mitoplast isolation. RT-qPCR values were normalised to total RNA. Normalisation was performed by dividing the mRNA expression obtained in the mitoplast fraction by that in the total RNA. COX3 and 12S are mitochondrial RNAs. These are relativley enriched in both the crude mitochondrial and mitoplast isolations compared to total RNA. GAPDH, c-myc, α-tubulin and eIF6 are nuclear encoded genes and are expected to be depleted in the mitoplast isolation. The level of mitoplast purity achieved here was similar to what has been previously described by Mercer *et al*. (2013). The black bars correspond to the crude mitochondrial isolation, the light grey bars represent the mitoplast isolation. The error bars indicate the standard deviation of n = 2 experiments.

(B) Western blots to assess purity at protein level of mitoplast isolation. Cytosolic (α-tubulin) and nuclear (H3) proteins were depleted from mitoplast, whereas mitochondrial matrix proteins (ACO2 and NSUN4) were enriched. The % refers to the amount of total sample that was loaded on the gel. These blots are representative of n = 2 experiments. PNL = post nuclear lysate; MP S/N = mitoplast supernatant.

regarding m^5C sites in mt-RNA. However, with the future development of m^5C detection techniques, optimised for use with less input RNA, a targeted analysis into the mitochondrial m^5C methylome could be undertaken.

These technical limitations meant that, in this thesis, mitochondrial m^5C sites were identified from bsRNAseq transcriptome-wide datasets. Thus, the second SOLiD dataset was revisited, with a view of elucidating mitochondrial m^5C candidate sites. Furthermore, with the availability of the Illumina dataset, a comparison of detected candidate mitochondrial m^5C sites between the two sequencing approaches is presented.

Generally the obtained mitochondrial transcriptome bsRNA-seq coverage from both datasets was low, with some regions exhibiting no coverage (a representative plot is shown in Fig. 3.4). The reduced base complexity, resulting from the bisulfite treatment, presented a challenging problem to map bsRNA-seq reads to the reference using standard short read alignment tools. Further issues arised as mitochondrial sequences have high sequence similarity to nuclear-encoded sequences (Fig. 3.2); therefore a proportion of reads is expected to map to multiple regions and therefore would be discarded. Altogether this delimits the number of m^5C sites that can potentially be detected.

3.2.2 Candidate mitochondrial m^5C sites

3.2.2.1 SOLiD bsRNA-seq

The alignment of the SOLiD reads was executed against two reference genomes: 1) a non-CpG *in silico* bisulfite converted reference sequences (i.e. all cytosines except those in a CpG context are converted, allowing detection of m^5C in CpG clusters) and 2) a completely *in silico* converted reference sequence (i.e. all the cytosines are converted) (see Discussion). When aligning reads to the former, 27 candidate mitochondrial m^5C sites were detected (Table 3.1) at a 1% false discovery rate (FDR). More specifically, nine sites in mt-mRNAs, four sites in mt-rRNAs, ten sites in mt-tRNAs and four sites in the D-loop region were identified. Interestingly, removal of the CpG bias (i.e. mapping to a completely converted reference) resulted in the detection of only a single site in mt-rRNA (chrM:1490) and four sites in four different mt-tRNAs (chrM:626, chrM:3280, chrM:5822, chrM:5850).

In an effort to decrease the false-positive rate of detected mitochondrial m^5C sites, the

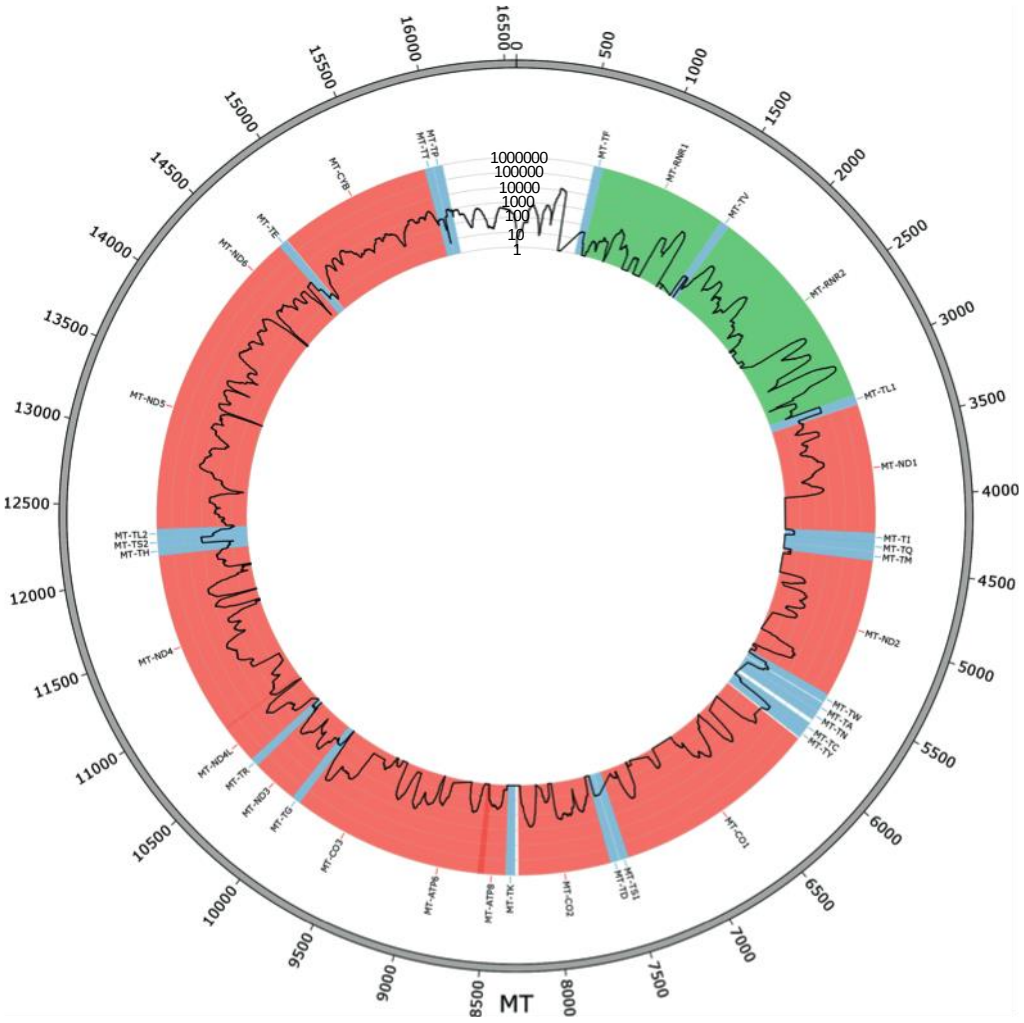


Figure 3.4: Illumina bsRNA-seq coverage across the mitochondrial transcriptome.

The coverage is non-uniform across the mitochondrial transcriptome. The green shaded regions refer to the 12S and 16S mt-rRNA. The orange shaded regions indicate protein-coding genes, and the blue shaded regions show positions of mt-tRNAs, the white region corresponds to the D-loop region. The black trace indicates the sequencing coverage in read number; the outer ring shows the genomic coordinates.

Table 3.1: Identified candidate m^5C sites in mitochondrial RNA from the SOLiD bsRNA-seq transcriptome-wide mapping.

Coordinate (chrM)	Gene Name	RNA Biotype	Meth Rate (%)	Coverage (reads)
263		D-loop	17.85	587
318		D-loop	36.37	26
16084		D-loop	12.79	161
16363		D-loop	40.02	23
1475	12S	rRNA	15.09	2004
1477	12S	rRNA	15.11	1683
1490	12S	rRNA	41.98	187
2644	16S	rRNA	94.64	43
4714	NADH dehyrogenase II	mRNA	25.31	20
8166	Cytochrome c oxidase II	mRNA	10.21	70
9294	Cytochrome c oxidase II	mRNA	12.80	37
9575	Cytochrome c oxidase II	mRNA	8.90	33
10215	NADH dehydrogenase III	mRNA	19.29	12
11494	NADH dehydrogenase IV	mRNA	9.12	21
11915	NADH dehydrogenase IV	mRNA	65.24	80
13811	NADH dehydrogenase V	mRNA	8.30	391
15200	Cytochrome b	mRNA	16.52	10
626	mt-tRNA-Phenylalanine, pos 47	tRNA	18.49	196
3280	mt -tRNA-Leucine 1, pos 50	tRNA	8.73	21
5624	mt-tRNA-Alanine, pos 36	tRNA	20.47	8
5822	mt-tRNA-Cysteine, pos 6	tRNA	69.83	20
5850	mt-tRNA-Tyrosine, pos 47	tRNA	20.70	123
12182	mt-tRNA-Histidine, pos 47	tRNA	69.41	95
12192	mt-tRNA-Histidine, pos 57	tRNA	41.44	19
12207	mt-tRNA-Histidine, pos 73	tRNA	58.59	362
12237	mt-tRNA-Serine 2, pos 30	tRNA	10.40	135
14699	mt-tRNA-Glutamic acid, pos 49	tRNA	13.35	628

The coordinate of the candidate site is given using hg19 reference genome. The methylation rate and coverage is an average across all four SOLiD libraries (NSUN2 knockdown, TRDMT1 knockdown, DNMT1 knockdown and non-targeting control). The highlighted rows represent candidates that were included for validation using MiSeq amplicon sequencing.

original list of 27 candidate m^5C sites was filtered manually for possible artifacts. For example, false positives arising from non-conversion, manifested as a clustering of m^5C sites were removed (except in the case of tRNAs, where sites have been demonstrated to be clustered). Clustered sites, such as the two candidate 12S rRNA m^5C sites (chrM: 1475 and 1477) are likely artifacts due to inefficient bisulfite conversion. This can result from extensive secondary structure, especially present in rRNA, which may hinder the efficient conversion of non-methylated cytosines resulting in false-positive calls.

3.2.2.2 Illumina bsRNA-seq

The Illumina reads were aligned to a completely converted human reference genome (hg38), yielding a total of 23 candidate m^5C sites in mitochondrial RNA, at a FDR of 1%. Eight of these were observed in mt-tRNAs, fourteen sites in the 16S mt-rRNA and one site was present in the mt-mRNA cytochrome c oxidase I (Table 3.2).

Manual filtering of these 23 candidate sites retained only the eight identified sites in mt-tRNAs. The candidate 16S mt-rRNA m^5C sites were discarded because they are an example of clustering due to incomplete conversion (Fig. 3.5). Ribosomal RNA is known to possess strong secondary structure and it is probable that such structure would inhibit the bisulfite conversion resulting in false positives. Also, the site in the cytochrome c oxidase I mRNA might have been falsely identified as a result of clonal read bias; the reads containing a cytosine at the site of interest have the same alignment start and stop points (Fig. 3.6). It is expected that high confidence candidate sites are represented by multiple sequencing reads with different alignment start and end points (i.e. no clonal reads), expect possibly in the case of tRNAs, where clonal reads may be generated as tRNAs are short.

As the candidate m^5C sites in mt-tRNAs were not affected by these issues, they were deemed to be high confident sites and therefore would be ideal candidates for validation. However PCR amplification of these candidates proved difficult, as they are relatively lowly expressed (see e.g. Mercer *et al.,* 2011). As a result mt-tRNA sites were not included for validation by MiSeq amplicon sequencing. Even though the majority of post-filtered candidate sites seemed to be low confidence sites, validation was attempted in order to definitively conclude the methylation status of candidate m^5C sites identified

Table 3.2: Identified candidate m^5C sites in mitochondrial RNA from Illumina bsRNA-seq transcriptome-wide mapping.

Coordinate (chrM)	Gene Name	RNA Biotype	Meth Rate (%)	Coverage (reads)
2867	16S	rRNA	14.31	601
2868	16S	rRNA	10.85	295
2872	16S	rRNA	16.84	95
2877	16S	rRNA	34.09	44
2881	16S	rRNA	32.50	40
2884	16S	rRNA	37.50	40
2889	16S	rRNA	38.10	42
2891	16S	rRNA	35.00	40
2899	16S	rRNA	47.06	34
2900	16S	rRNA	48.48	33
2906	16S	rRNA	50.00	30
2911	16S	rRNA	35.71	28
2912	16S	rRNA	48.15	27
3184	16S	rRNA	27.91	43
6057	Cytochrome c oxidase I	mRNA	15.25	177
624	mt-tRNA-Phenylalanine, pos 47	tRNA	33.33	33
3279	mt-tRNA-Leucine 1, pos 49	tRNA	88.93	271
5822	mt-tRNA-Cysteine, pos 6	tRNA	55.32	47
5849	mt-tRNA-Tyrosine, pos 47	tRNA	31.67	1585
12239	mt-tRNA-Serine 2, pos 32	tRNA	92.22	1003
12241	mt-tRNA-Serine 2, pos 34	tRNA	95.51	1003
12296	mt-tRNA-Leucine 2, pos 30	tRNA	33.33	30
12304	mt-tRNA-Leucine 2, pos 38	tRNA	33.33	30

The coordinate of the candidate site is given using hg38 reference genome. The methylation rate and coverage is an average across the two biological replicates.

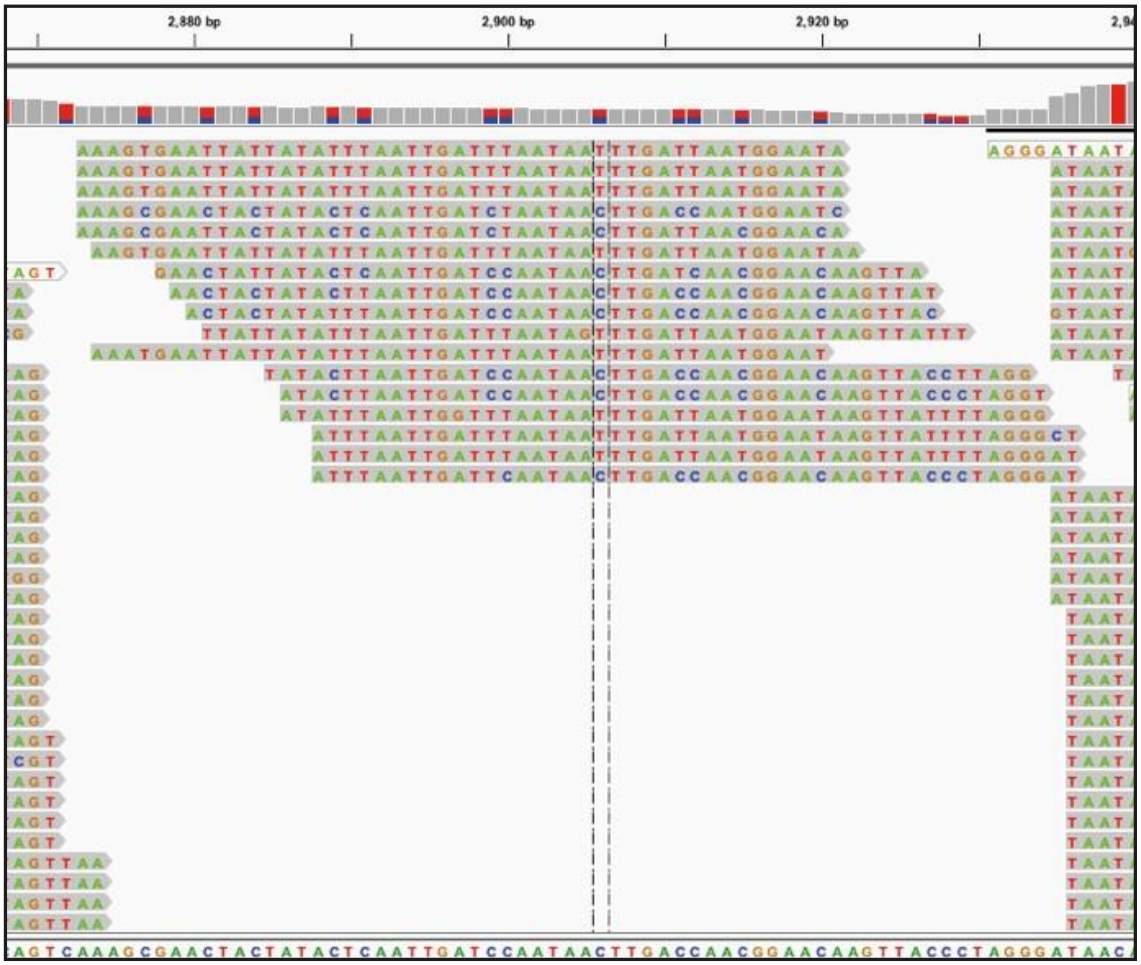


Figure 3.5: Integrated Genomics Viewer (IGV) screenshot of the called 16S m^5C sites in the Illumina bsRNA-seq experiment.

Every cytosine in this 60 bp window is identified as a candidate m^5C site. As strong secondary structure (particularly in rRNA) is known to impede the bisulfite conversion reaction chemistry, this stretch of cytosines is probably incorrectly identified as m^5C sites (i.e. they are false-positives). The top panel shows the genomic coordinates, the underlying panel shows the proportion of base calls at each position (red indicates the thymine base; blue indicates the cytosine base). The panel below shows the sequencing reads. The bottom panel shows the genomic sequence. The cytosine at chrM:2906 was incorrectly identified as an m^5C site.



Figure 3.6: Integrated Genomics Viewer (IGV) screenshot showing the called candidate m^{5}C site in cytochrome c oxidase I mRNA (chrM:6057).

The identified site is indicated by the vertical black dotted lines. This site was falsely identified as an m^{5}C site as a result of clonal read bias. Each of the reads which identify a cytosine at postion chrM:6057 have the same alignment start and end coordinates. This suggests that this is a low confidence called site. The top panel shows the genomic coordinates, the underlying panel shows the proportion of base calls at each position (red indicates the thymine base; blue indicates the cytosine base). The panel below that shows the sequencing reads. The final panel shows the genomic sequence.

in mt-RNA. Fifteen sites, nine of which were within mRNAs, two in rRNAs and four in the D-loop region (highlighted in yellow in Table 3.1) were included for validation using MiSeq amplicon sequencing.

3.2.3 Validating candidate m^5C sites

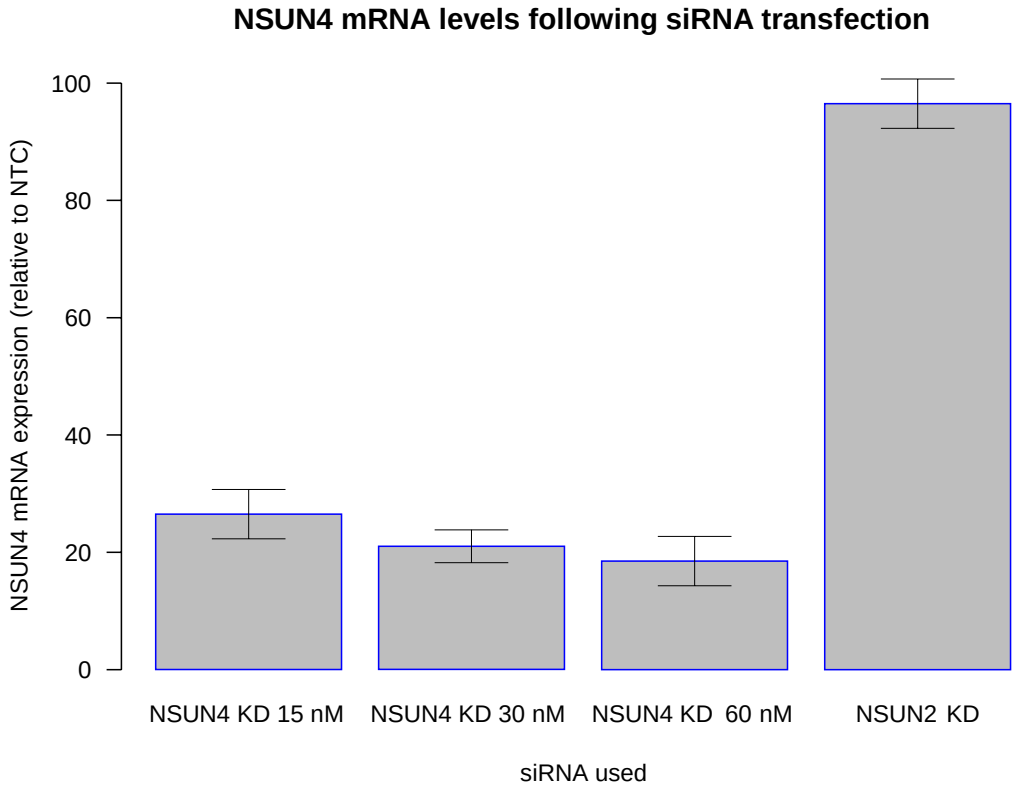
3.2.3.1 Knockdown of NSUN4 in HeLa cells

Efficiency of NSUN4 knockdown was determined at mRNA and protein levels using RT-qPCR (primers shown in Table 2.1) and western blotting techniques, respectively. Optimisation of the siRNA concentration was required to achieve an appropriate reduction in NSUN4 levels, whilst maintaining enough cells to ensure sufficient RNA yield for downstream applications. A siRNA concentration series (15, 30, 60 nM) was tested, which resulted in an mRNA reduction of 75, 80, and 82% respectively. It was concluded that (a) the 30 nM siRNA treatment produced a satisfactory knockdown at the protein level ensuring an observable decrease in methylation of NSUN4 controlled sites, and (b) enough RNA was yielded to perform bisulfite treatment. A greater than 80% depletion in NSUN4 mRNA was achieved (Fig. 3.7A), translating to a greater than 90% reduction in NSUN4 protein levels compared to the NTC and NSUN2 control samples (Fig. 3.7B). NSUN4 mRNA expression was normalised to the HPRT housekeeping gene and is plotted relative to NTC; expression of NSUN4 mRNA in the NSUN2 knockdown sample was also included as a control (Fig. 3.76A). To confirm knockdown at the protein level an ImageJ analysis was performed on the western blot. NSUN4 density values were normalised to the α-tubulin loading control and reported relative to the NTC (Fig. 3.7B).

3.2.3.2 Verification of efficient bisulfite conversion

Following siRNA transfection, total cellular RNA was extracted and bisulfite treated. To determine the efficiency of bisulfite conversion, *in vitro* transcribed non-humanised R-luciferase (R-luc) transcripts were 'spiked-in' following RNA extraction (prior to bisulfite treatment), to be used as a negative control. This transcript lacked m^5C sites and was expected to show near 100% conversion. As a positive control, $tRNA^{Asp(GUC)}$, which harbours two previously validated NSUN2-mediated m^5C sites

(A)


NSUN4 mRNA levels following siRNA transfection
NSUN4 mRNA expression (relative to NTC)
100
80
60
40
20
0
NSUN4 KD 15 nM
NSUN4 KD 30 nM
NSUN4 KD 60 nM
NSUN2 KD
siRNA used

(B)

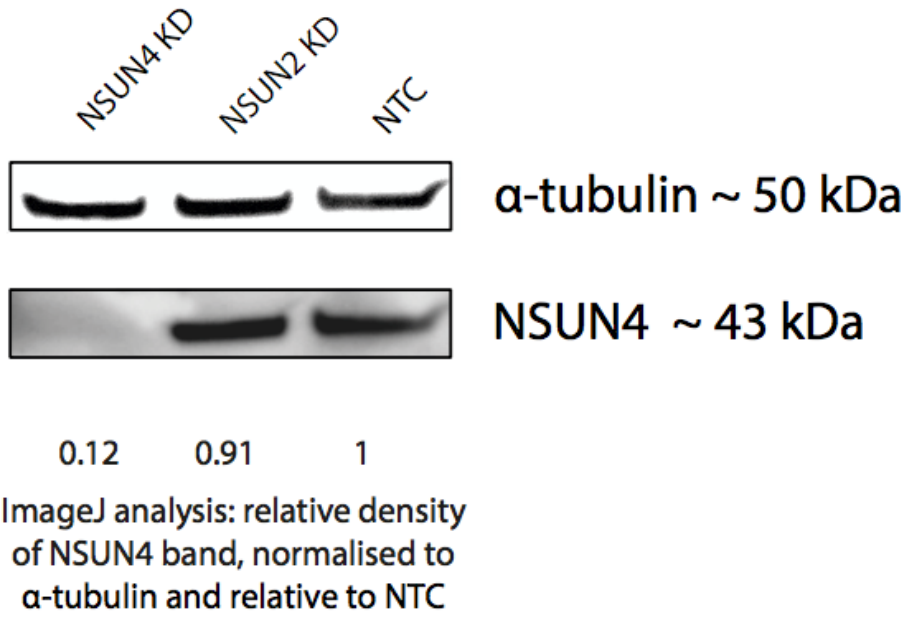

NSUN4 KD
NSUN2 KD
NTC
α-tubulin ~ 50 kDa
NSUN4 ~ 43 kDa
0.12
0.91
1
ImageJ analysis: relative density of NSUN4 band, normalised to α-tubulin and relative to NTC

Figure 3.7: Quality control measures to ensure successful knockdown of NSUN4.

(A) Results of RT-qPCR measuring NSUN4 mRNA levels following different siRNA transfections. The greatest reduction in NSUN4 mRNA is observed when using 60 nM of NSUN4 siRNA but resulted in the least amount of extracted RNA. 30 nM of NSUN4 siRNA was used resulting in an 80% knockdown of NSUN4 at the mRNA level, with enough RNA being extracted for downstream applications. NSUN2 KD is used here as a control, showing that the NSUN4 siRNAs are specific. 30 nM of NSUN2 siRNA was also used. The error bars represent the standard deviation from n = 3 experiments. mRNA expression was normalised to the HRPT housekeeping gene and shown relative to the NTC.

(B) Results of western blotting showing that NSUN4 is absent from the NSUN4 KD sample and present in the NSUN2 KD and NTC controls. α-tubulin is used as a loading control. A representative western blot is shown from n = 3 experiments. An ImageJ analysis of the blot confirms knockdown in the NSUN4 KD sample. The given densities represent the amount of NSUN4 normalised to the α-tubulin loading control and quoted relative to the NTC.

at positions C48 and C49 (at ~90% methylation) and one TRDMT1-mediated site at C38 (~80% methylation) was used (Squires *et al.*, 2012).

MiSeq PCR amplicon sequencing results for the *in vitro* non-humanised R-Luc transcript (see Appendix 7.1 for sequence) revealed a greater than 99% conversion of cytosines in all samples, indicating successful bisulfite conversion for all samples (Fig. 3.8). Positions C48 and C49 in $tRNA^{Asp(GUC)}$ showed an approximate 80–90% conversion in the NSUN2 knockdown sample, whereas in the NSUN4 knockdown and NTC samples, expected methylation levels are observed at both sites (Fig. 3.9). Position C38 shows ~80% methylation across all samples. Altogether, this confirms efficient C:T conversion by bisulfite treatment in all three samples.

In this thesis, heatmaps are used to illustrate the methylation percent of cytosines across a transcript of interest. Each cytosine in the PCR amplified region of the transcript is portrayed as a rectangle. Shading intensity of the rectangles indicates the percentage of methylation according to the given key. Heatmaps were generated using the R programming language.

3.2.3.3 PCR amplification and library preparation

As bisulfite treatment results in fragmentation of the RNA, primers were chosen to amplify a small region (~100–300 nt) containing the candidate site of interest (primers shown in Table 2.2). The primers were designed to span a number of converted cytosines to bias toward amplification of converted sequences (Fig. 3.10). This provides an additional level of confidence that detected m^5C sites, using MiSeq, are real. To minimise PCR bias, PCRs were performed in triplicate and pooled for downstream use. Optimisation steps were performed for each primer pair to obtain a sufficient amount of PCR product. This included increasing PCR cycle numbers, input cDNA amount, primer concentration and annealing temperatures. The triplicate PCR fragments were then combined into one amplicon for MiSeq library preparation (see Chapter II).

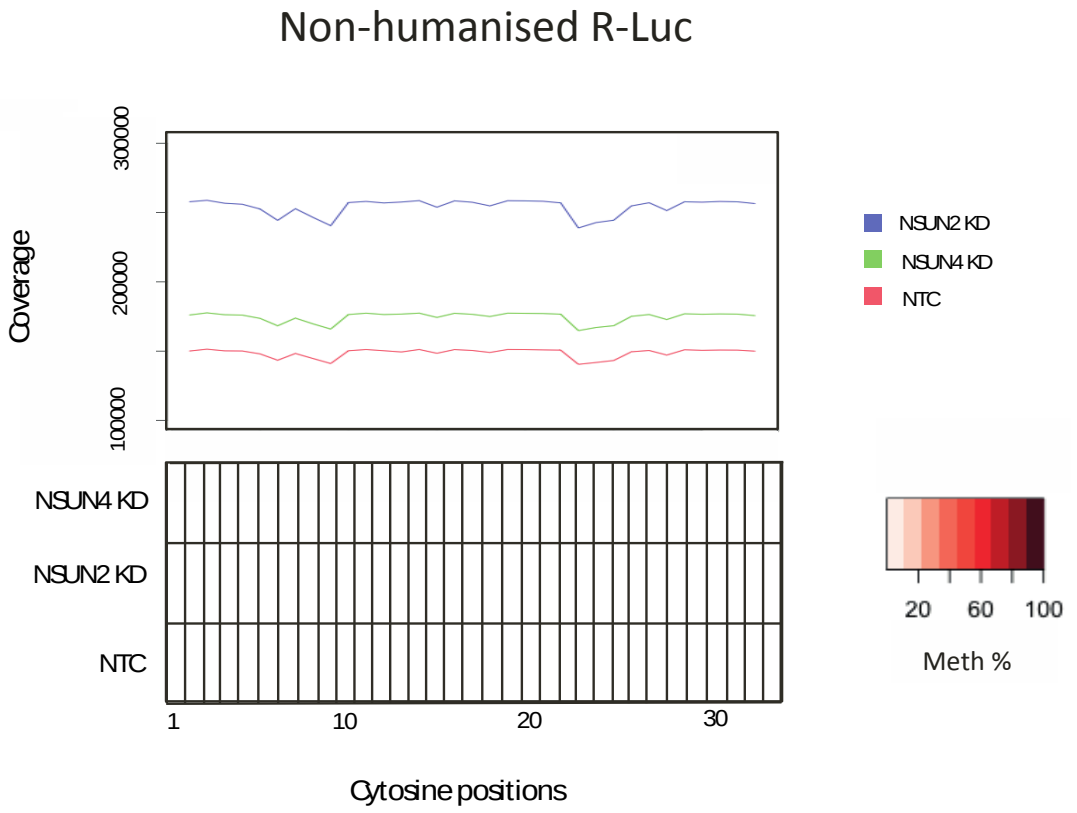


Figure 3.8: MiSeq results expectedly showing no methylation in the negative control.

The top panel displays the obtained coverage for the *in vitro* transcribed non-humanised R-Luc. This serves as the negative control as it does not contain m^5C sites. The bottom panel shows the methylation percentage for each cytosine in the R-Luc sequence. As expected, methylation was not detected anywhere across the R-Luc sequence. Each box represents a cytosine in the amplified region.

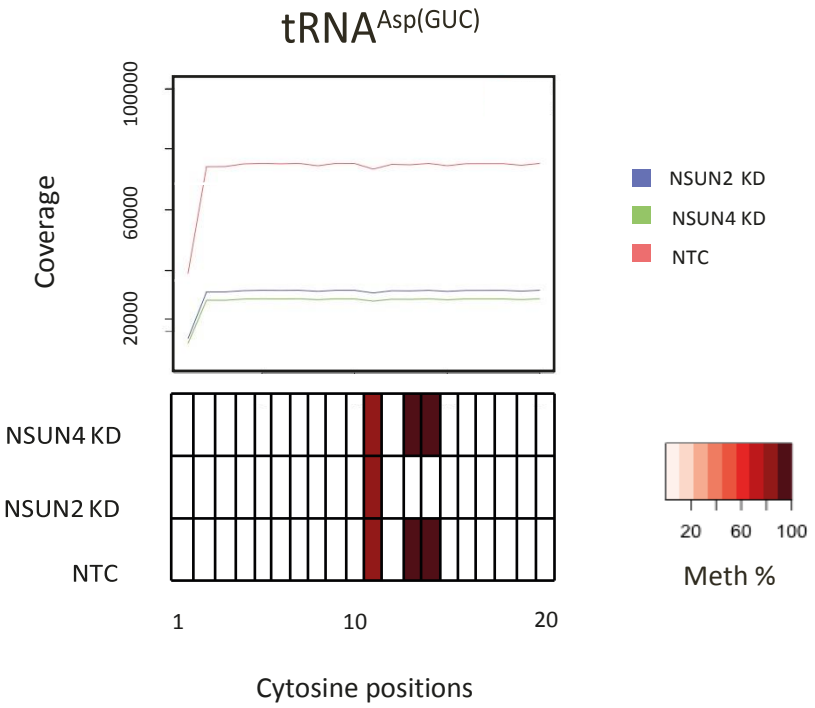


Figure 3.9: MiSeq results showing methylation at expected positions in the positive control tRNA$^{Asp(GUC)}$.

The top panel displays the obtained coverage for tRNA$^{Asp(GUC)}$. This serves as the positive control as it contains three known m^{5}C sites. The bottom panel shows the methylation percentage for each cytosine in the tRNA$^{Asp(GUC)}$ transcript. As expected, methylation was detected at position C38 (indicated as cytosine 11), and positions C48 and C49 (indicated as cytosines 12 and 13). Positions C48 and C49 are known to be controlled by NSUN2 and therefore show converted cytosines in the NSUN2 KD sample. Each box represents a cytosine in the amplified region.

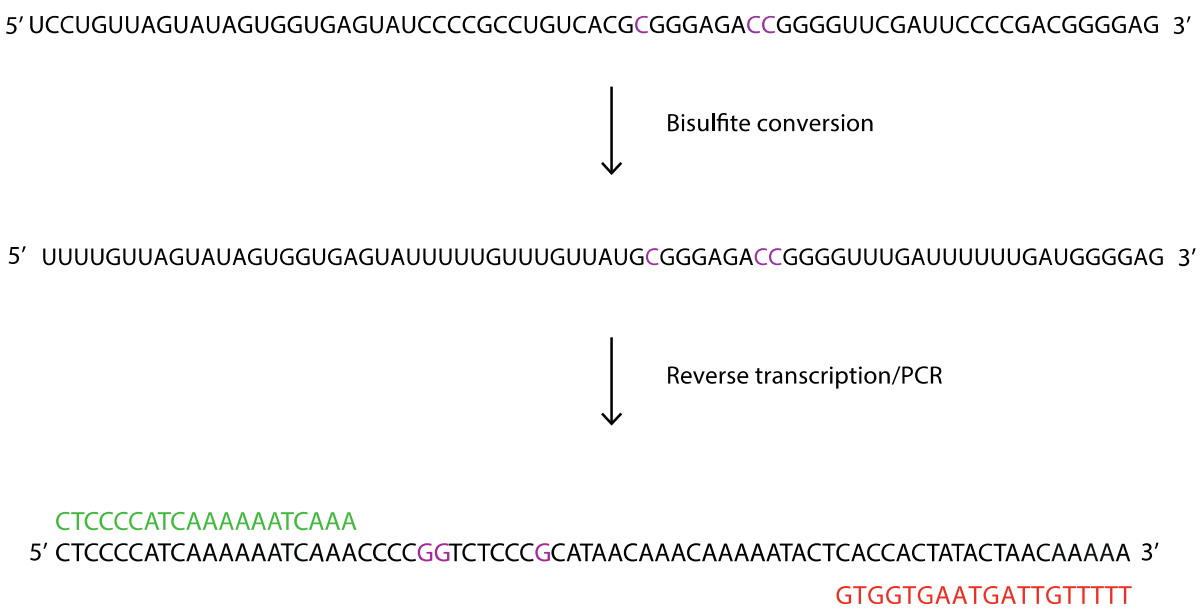


Figure 3.10: Primer design strategy.

Primers designed for the $tRNA^{Asp(GUC)}$ control. The forward primer is indicated in green; the reverse primer is indicated in red. The primers are designed to preferentially amplify converted sequences. The purple bases represent the m^5C sites.

Equal amounts of each amplicon derived from one knockdown experiment were then pooled into a single sample for library preparation. To ensure successful library preparation and to determine the average library size, an aliquot of the amplicons pre- and post-library preparation were electrophoresed (Fig. 3.11). An increase in the size of the DNA fragments following library preparation indicated successful ligation of adaptors. MiSeq library quantification was performed using qPCR and resulted in a cluster density within the optimal range (764 K/mm^2) for the MiSeq run indicating a high proportion of the reads were successfully mapped to the theoretical amplicon sequence. The optimal cluster density for MiSeq V2 is 700 – 1000 K/mm^2 (clusters per square millimetre): As the library is loaded to the flow cell of the sequencer, the single-stranded, adaptor-ligated fragments hybridise to immobilised complementary sequences that cover the flow cell. Each hybridised fragment undergoes rounds of amplification resulting in the many copies of the same fragment, in the same location on the flow cell. This is termed a 'cluster'. Underclustering results in high data quality, however it also results in lower data output. Whilst overclustering may affect run performance, lead to lower Q30 scores, and may introduce sequencing artefacts.

Mapping was carried out using Bismark and 61.5%, 64.2% and 62.4% of the total reads could be mapped in the NSUN2 knockdown, NSUN4 knockdown and NTC libraries, respectively.

3.2.3.4 Validation of candidate m^{5}C sites using MiSeq

Of the fifteen candidate mitochondrial m^{5}C sites included in the MiSeq analysis, a single site in 12S mt-rRNA returned a positive result for cytosine methylation at the expected site. A 40% methylation rate was detected in the 12S mt-rRNA at position C841 in both the NTC and NSUN2 knockdown library samples. This level was reduced to less than 5% in the NSUN4 knockdown library, suggesting that this site is mediated by NSUN4 (Fig. 3.12).

The remaining fourteen candidate m^{5}C sites were shown not to be methylated by MiSeq analysis (see Appendix 7.2), despite high coverage across each amplicon. In particular, the SOLiD candidate site detected in 16S mt-rRNA was shown not to be methylated (Fig. 3.13). This suggests that despite manual filtering of the identified sites, the false positive rate was high.

3.2.4 m^5C sites in human mitochondrial tRNAs

Four m^5C sites, one each in mt-$tRNA^{Leu(UUR)}$, mt-$tRNA^{Phe}$, mt-$tRNA^{Cys}$ and mt-$tRNA^{Tyr}$, were identified in both the SOLiD and Illumina bsRNA-seq experiments. Although, the validation of these sites could not be attempted due to the low abundance of these transcripts (see Rossmanith, 2012), detection of these sites is reproducible suggesting that these sites are real. In fact the site in mt-$tRNA^{Leu(UUR)}$ had already been reported (Suzuki *et al.*, 2011); the other three detected m^5C sites are novel. These three sites are not detected by Suzuki *et al*. (2011) as these tRNAs have different sequences between bovine and human. For example, bovine mt-$tRNA^{Tyr}$ and mt-$tRNA^{Phe}$ do not have a cytosine at the position where an m^5C site is detected in humans. Interestingly, this m^5C site and the sites detected in mt-$tRNA^{Tyr}$ (Fig. 3.14A) and mt-$tRNA^{Phe}$ (Fig. 3.14B) are in a structural position (C48) similar to the NSUN2 controlled sites in nuclear encoded tRNAs. This may suggest similar roles for the m^5C methylation in both contexts. However, the m^5C site in the mt- $tRNA^{Cys}$ is within the stem, very close to the 5≠ end (Fig. 3.14C). To date, no m^5C sites have been reported in this region of a tRNA. This may suggest a novel role for m^5C in this context.

3.3 Discussion

The aim of this Chapter was to identify m^5C sites in mitochondrial transcripts and to determine their 'writers'. Candidate mitochondrial m^5C sites were identified in two independent bsRNA-seq experiments. An m^5C site in human 12S mt-rRNA was shown in this thesis to be regulated by NSUN4 and the function of m^5C in that context is discussed below. A further four m^5C sites, in mt-tRNAs, overlapped between the two datasets, three of which constitute novel sites. Also previous reports suggested the presence of m^5C methylation in 16S mt-rRNA, however, no evidence of this was found in this thesis. Challenges in mapping bsRNA-seq reads stemming from limitations of the bisulfite treatment, and the high sequence similarity between mitochondrial and nuclear encoded sequences, has ultimately translated into a high false-positive rate in detecting mitochondrial m^5C sites.

3.3.1 A validation tool is essential

Identifying mitochondrial m^5C sites was problematic as a result of high sequence similarity between mitochondrial and nuclear encoded sequences. This dramatically reduced the likelihood of obtaining a complete high confidence set of mitochondrial m^5C sites. This is especially true, as the read length was comparatively short at only 50

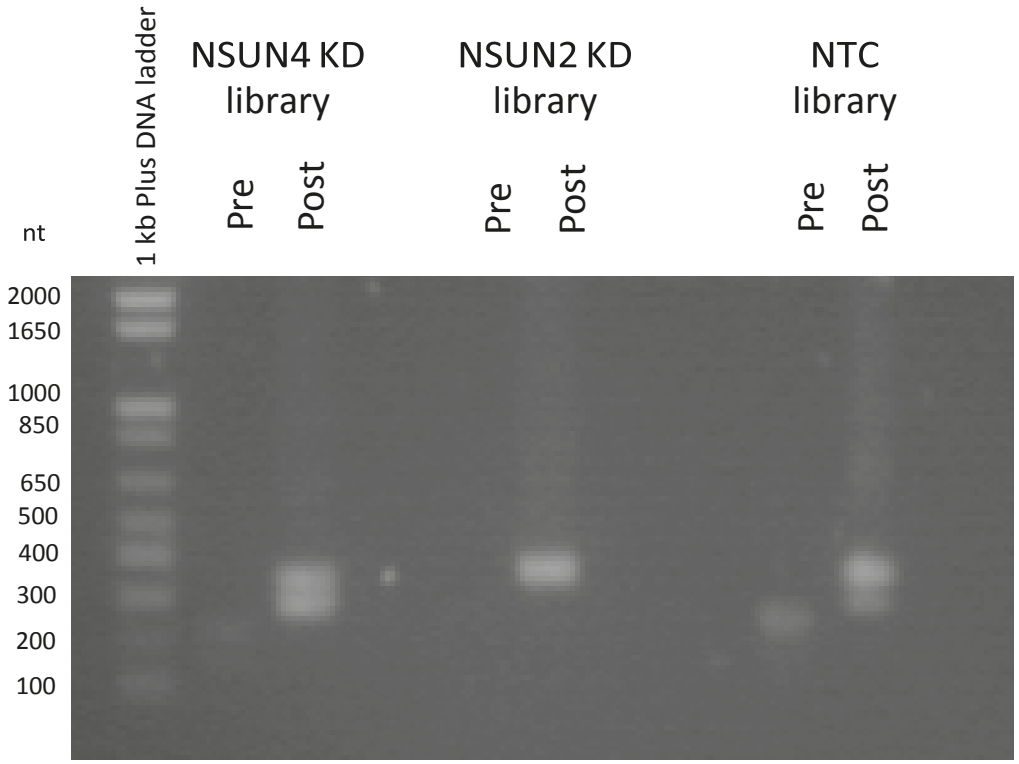


Figure 3.11: Pooled amplicons (pre-library preparation) and final libraries (post-library preparation) for each sample were electrophoresed on a 2% agarose gel.

The average library size was ~200 nt prior to library preparation. After library preparation, the library size was ~350 nt (i.e. including the ligated adaptors). The DNA fragments, following library preparation, were more abundant as a result of the final PCR amplification step in the library preparation protocol. The double band reflects the size of the different RNA species present in the library. The lower band corresponds to tRNAs, whilst the higher size band corresponds to mRNA fragments. The sizes of the 1 kb Plus DNA ladder are indicated in nucleotides (nt). KD = knockdown

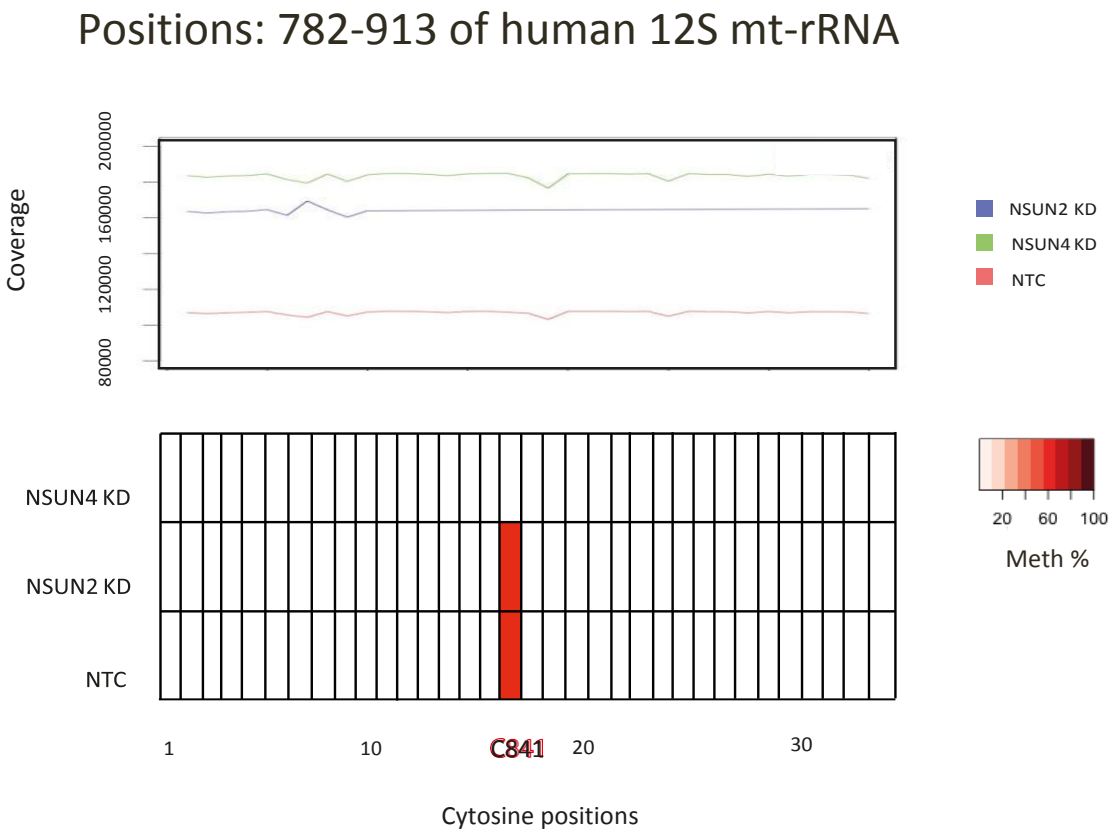


Figure 3.12: MiSeq result validating m^5C site C841 in 12S mt-rRNA and showing it to be NSUN4 controlled.

The top panel displays the obtained coverage across positions: 782-913 of human 12S mt- rRNA. The bottom panel shows a 40% methylation was detected in the NSUN2 KD and NTC samples, whereas it was observed at <5% in the NSUN4 KD sample, suggesting NSUN4 is responsible for methylation at the site. Each box represents a cytosine in the amplified region.

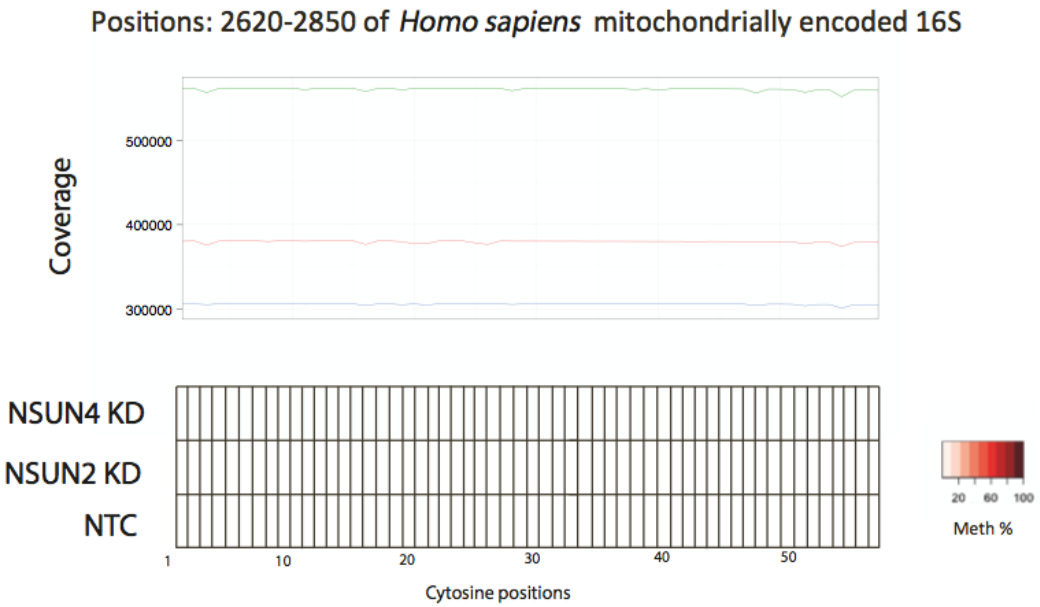


Figure 3.13: MiSeq result showing no methylation at the candidate site in 16S mt-rRNA.

The top panel displays the obtained coverage across positions: 2620-2850 of human 16S mt- rRNA. The bottom panel shows no detected methylation along the amplified region in any of the samples. This site was falsely identified as an m^5C site in the SOLiD™ dataset.

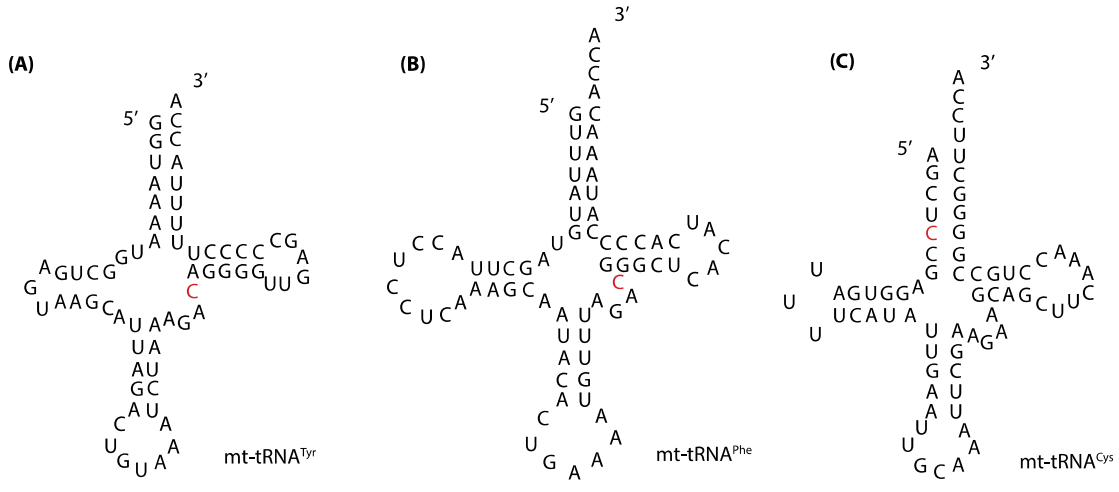


Figure 3.14: Identified m^{5}C sites in mitochondrial (mt) tRNAs.

(A) 2D-structure of mt-tRNA^{Tyr} showing the position of the m^{5}C site in red (chrM:5849). This site is in a structural position similar to NSUN2 controlled m^{5}C sites in nuclear-encoded tRNAs. **(B)** 2D-structure of mt-tRNA^{Phe} showing the position of the m^{5}C site in red (chrM:624). This site is in a structural position similar to NSUN2 controlled m^{5}C sites in nuclear-encoded tRNAs. **(C)** 2D-structure of mt-tRNA^{Cys} showing the position of the m^{5}C site in red (chrM:5822). m^{5}C sites have not been previously been reported in this position.

nt. Indeed, this short read length coupled with the fact that 95% of the mitochondrial transcriptome is highly similar to nuclear encoded sequences (see Fig. 3.2) dictates that many reads actually originating from the mitochondria will also map to other genomic regions and therefore would be discarded. Consequently this would result in reduced sequencing reads covering the mitochondrial transcriptome.

It is important to note that in this current analysis, unique mapping was further hampered as a result of bisulfite treatment of the RNA samples prior to sequencing. This ultimately reduces the overall base complexity of the reads as cytosines are converted to thymines, further increasing the risk of reads being assigned to a 'new area'. This would only exacerbate the problem of assigning reads to a single unique location. In the present analysis, a read is only retained if it has a unique 'best match' genome-wide, and therefore sequencing reads that multi-map to several regions may be assigned to the highest scoring match region or could be discarded. This was done in an attempt to assign reads unambiguously, and to allow for the accurate identification of m^5C sites, especially those residing in mitochondrial transcripts. However this process is not perfect and it is expected that some multi-mapping reads may be randomly assigned to one of any possible locations. These factors may result in low confidence candidate sites or even false positive methylation calls, highlighting the need for a robust validation technique.

Additionally, when mapping, an assumption was made that m^5C sites in RNA would preferentially be located within a CpG context as is the case with m^5C sites in DNA. Therefore, SOLiD™ bsRNA-seq reads were initially aligned to a non-CpG *in silico* bisulfite converted reference genome, yielding a total of 27 sites. However by removing this bias and aligning the sequencing reads to a completely *in silico* bisulfite converted reference genome, only five m^5C sites were observed (one in 12S mt-rRNA, and one each in four mt-tRNAs). This clearly highlights that the chosen mapping and m^5C detection parameters can significantly affect the numbers of identified candidate m^5C sites. Again this reemphasises the importance of validating any candidate sites through a high-throughput approach.

Clearly a combination of the above-mentioned factors would explain the apparent discrepancy between the transcriptome-wide bsRNA-seq and MiSeq results, particularly for mitochondrial sequences. This has also been observed elsewhere where a high

(90%) false positive rate in global m^5C candidate sites detection using bsRNA-seq on RNA extracted from mouse fibroblasts was reported (Khoddami 2013, PhD thesis). This suggests that this may be a feature of identifying m^5C sites using bsRNA-seq and highlights the necessity to perform robust, high-throughput validations following transcriptome-wide detection of m^5C.

3.3.1.1 Validation approach

Previously, the validation of candidate m^5C sites was achieved using Sanger sequencing as the output (see e.g. Squires *et al.*, 2012). However, this technique is quite laborious as it requires the sequencing of multiple clones for accurate stoichiometric validation, and therefore only allows a limited number of samples to be processed. Recently, NGS was developed as a high throughput means to validate candidate m^5C sites (Sibbritt *et al.*, 2016). This approach uses MiSeq NGS technology to determine methylation sites in multiple bisulfite PCR amplicons, delivering fast, high confident validations. One clear advantage of this technique over traditional Sanger sequencing is the higher sequencing depth, with the advantage of processing multiplexed samples. This allows for: 1) an accurate quantification in a targeted approach, resulting in a confident assessment of methylation sites, and 2) the possibility of multiplexing, therefore increasing the number of validations that may be performed. This technique is now the gold standard for validating candidate m^5C sites.

3.3.2 Possible role for m^5C in mt-RNAs

Interestingly, within the mitochondrial transcriptome, differential abundance, especially for tRNAs is observed, even though all transcripts originate from the same polycistronic transcript. This concurs with reports of wide variation in mt-tRNA and mt-mRNA abundance (see Mercer *et al.*, 2011). Furthermore, as all mt-RNAs are transcribed from a single polycistronic precursor, the variation observed in RNA abundance suggests that post-transcriptional mechanisms regulate mt-RNA abundance (Piechota *et al.*, 2006). Thus, there is a scope for the involvement of m^5C (or another post-transcriptional modification) in regulating RNA abundance, for example in marking transcripts for RNA decay (or protection from it). Indeed, it seems that m^5C plays such a role is 12S mt-rRNA (see below).

3.3.3 NSUN4 targets human 12S mitochondrial rRNA

Five RNA modifications have been reported in human 12S mt-rRNA, specifically m^5U429, m^4C839, m^5C841, $m^6{}_2A936$ and $m^6{}_2A937$ (Dubin, 1974; Baer & Dubin, 1980). The present work extends the current knowledge by demonstrating that NSUN4 is responsible for the synthesis of m^5C at C841 in human 12S mt-rRNA. MiSeq validation of the site returned 40% methylation in both the NTC and NSUN2 knockdown libraries, whereas the level of methylation dropped to near 5% in the NSUN4 knockdown library (Fig. 3.12). While this work was in progress, Metodiev *et al.* (2014), using bisulfite sequencing, independently reported that the analogous site in mouse 12S mt-rRNA is methylated by NSUN4 and suggested a function for the m^5C modification in mitoribosome biogenesis (see below). In the present study, having independently validated the 12S site, this gives further weight to the accuracy of the results presented in this Chapter.

Interestingly, Metodiev *et al.* (2014) also detected a nearby m^4C modification, a cytosine modification known to be partially resistant to bisulfite treatment (Schaefer, Pollex, *et al.*, 2009). However, this site was not detected in either the transcriptome-wide mapping or MiSeq data presented in this study. This discrepancy may be explained by technical differences in the "harshness" of the bisulfite treatment protocols used in the two studies.

It should be noted that the validated 12S mt-rRNA site was detected only in the SOLiD but not the Illumina bsRNAseq experiment, because of a lack of sequencing coverage at that site in the latter. This is likely due to the fact that the rRNA depletion kit used in the Illumina experiment targeted nuclear as well as mitochondrial rRNA, whereas the kit used in the SOLiD experiment only targeted nuclear-encoded rRNA. However as ribosomal RNA is the most abundant RNA biotype, and these kits are not completely efficient at its removal, a large number of reads were still mapped to both the 12S and 16S mt-rRNAs.

Generally there is sufficient coverage across the 12S mt-rRNA transcript (Fig. 3.4) to conclude that there are likely no other m^5C sites present in human 12S mt-rRNA.

3.3.4 Function of post-transcriptional modifications in 12S mt-rRNA

The $m^6{}_2A936$, $m^6{}_2A937$ and m^5U429 post-transcriptional modifications in 12S mt-

rRNA are suggested to play critical roles in the stabilisation/assembly of the small ribosomal subunit (SSU) (Rorbach & Minczuk, 2012). In particular, the TFB1M mediated N6-dimethylation modifications at the two adenines have been implicated in stabilising 12S mt-rRNA. Indeed, according to Metodiev *et al.* (2009), a TFB1M knockout mouse showed dramatic reduction in the steady state levels of 12S mt-rRNA, causing a concomitant loss (> 90% decrease) of mitochondrial translation. To date, the function of the m^5U429 modification in 12S mt-rRNA still remains elusive. This is mainly because the enzyme responsible for this modification is yet to be identified. However, the analogous position in *E. coli* 16S rRNA (U788) is present within the 790-loop (Piekna-Przybylska, Przybylski, *et al.*, 2008). In bacteria, U788 is not methylated, however Desai *et al.* (2011) showed that replacing the 790-loop with a tetraloop (4 nt loop) yielded no fully formed 30S small subunits.

3.3.5 Role of m^5C in 12S rRNA

Work in this thesis revealed that only 40% of 12S mt-rRNA molecules are targeted for m^5C modification (i.e. sub-stoichiometric modification rate). This is interesting as Lee *et al.* (2015) showed that in 100% of 12S mt-rRNA molecules the two highly conserved adenines are modified by N6-dimethylation $m^6{}_2A936$, $m^6{}_2A937$ (i.e these adenines are stoichiometrically modified), and are known to stabilise the 12S mt-rRNA in the SSU (Metodiev *et al.*, 2009). The observed sub-stoichiometric methylation level suggests that the m^5C modification may have a modulating function, and therefore is more likely to be involved in regulating the number of ribosomes that are formed. It is conceivable that only mature SSUs that are fully modified (with the m^5C mark) are destined for assembly into mitoribosomes, and those that are not would likely be degraded. Furthermore, in this thesis it is hypothesised that ribosome biogenesis function of NSUN4 may be facilitated by NSUN4 acting as an m^5C 'reader' protein. The NSUN4/MTERF4 complex is known to interact with the large ribosomal subunit (LSU) (Camara *et al.*, 2011; Yakubovskaya *et al.*, 2012). It is speculated in this thesis that NSUN4 brings the LSU and SSU together by binding MTERF4, which is known to bind the LSU, and 'reading' the m^5C mark in the SSU. Of note, Moon and Redman (2014) have postulated that all NSUNs could possibly function as specific m^5C 'reader' proteins. It is suggested that the m^5C mark may serve to regulate ribosome assembly. This proposed mechanism is illustrated in Figure 3.15. Certainly, as mentioned previously, since mt-rRNA, mRNA and tRNA molecules are produced from a common

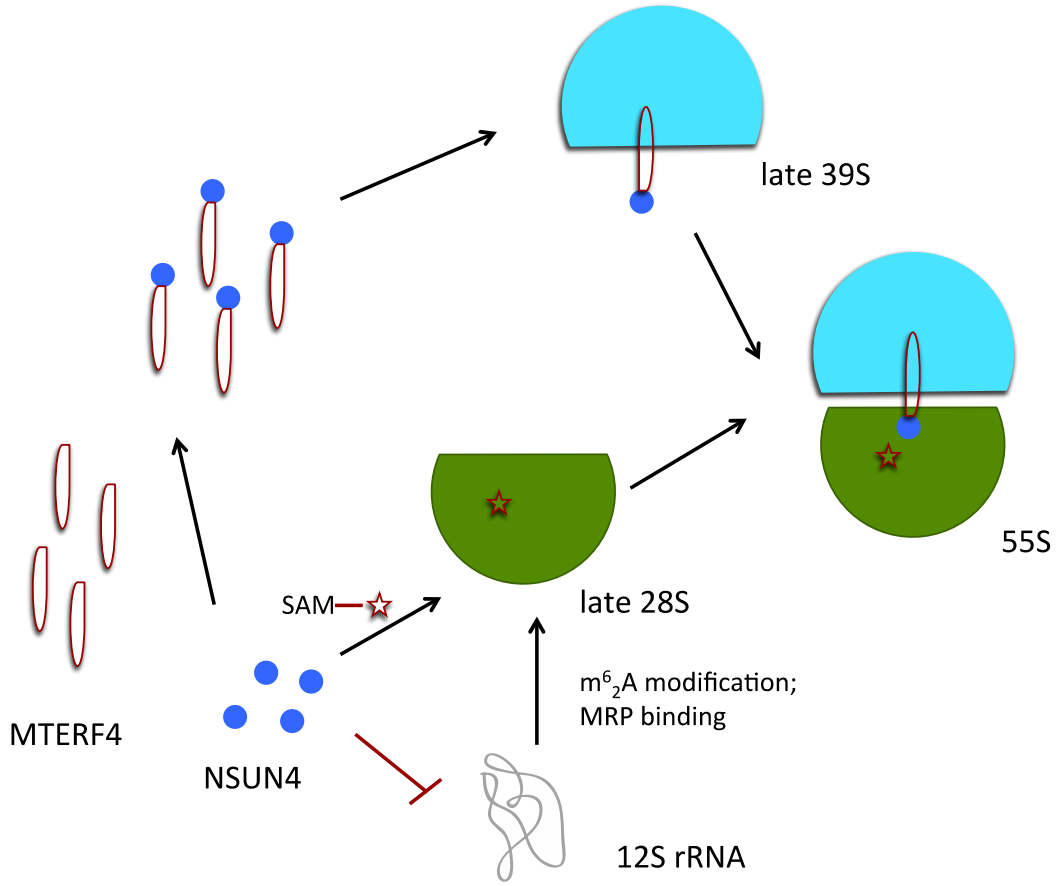


Figure 3.15: Schematic illustrating the proposed role for m^{5}C in mt-12S rRNA (adapted with modifications from Metodiev *et al.*, 2014).

NSUN4 is unable to methylate the free 12S rRNA molecule, but only able when it is assembled in the 28S small subunit (SSU) following m$^6_2$A modification and mitochondrial ribosomal protein (MRP) binding. NSUN4 forms a complex with MTERF4 which targets the 16S rRNA in the 39S large subunit (LSU). NSUN4 in complex with MTERF4 bound to the LSU, brings the SSU and LSU together by recognising m^{5}C modified SSUs, resulting in the 55S ribosome. SAM is the cofactor required for the methylation reaction (see Figure 1.3 for more details).

source, there must be a post-transcriptional mechanism in place to regulate RNA abundance. Thus, the m^5C methylation may function to maintain the correct concentration of mitoribosomes, providing an additional regulatory mechanism to control the levels of mitoribosomes and consequently protein expression.

3.3.6 NSUN4 target recognition

The RNA binding protein MTERF4 is known to interact with NSUN4. It was thought that MTERF4 physically directed NSUN4 to its target for methylation. This was the proposed mechanism to explain why NSUN4 was able to methylate RNA even though it lacked a known RNA binding domain. However, Metodiev *et al.* (2014) showed that methylation of 12S mt-rRNA was still observed in MTERF4 knockout mice, suggesting that NSUN4 is able to independently methylate 12S mt-rRNA. Thus the question arises, how does NSUN4 recognise its target?

The NSUN4 bacterial homolog, YebU, is responsible for the m^5C methylation at cytosine 1407 in the bacterial 30S SSU (Andersen and Douthwaite, 2006). YebU can only mediate methylation when the rRNA is present in the assembled SSU, but not the free rRNA molecule (Andersen & Douthwaite, 2006). It is possible that SSU assembly, results in the conformational change of the rRNA, which then allows NSUN4 to bind. Indeed, Kaushal *et al.* (2014), show that the conformation of helix 44 of 12S mt-rRNA (which harbours the m^5C site) is dramatically different in the mitochondrial SSU compared to the equivalent helix of cytoplasmic SSU.

It is postulated, in this thesis, that NSUN4 generally has a low affinity for RNA, but specifically recognises the particular strong secondary structure associated with the m^5C site in helix 44 of 12S mt-rRNA (Fig. 3.16), where the m^5C modification is present at the bulge-stem junction of a three-nucleotide bulge. It is possible that NSUN4 specifically recognises this structure and sequence motif. This could be similar to a previously identified 38 kDa cellular factor which has little general binding for RNA, but a strong specificity for the bulge and upper stem region of the trans-activation response element (TAR) RNA (Baker *et al.*, 1994). This mechanism for binding would explain why NSUN4 is able to independently methylate the RNA, even though it lacks the ‘canonical’ RNA binding domain present in the other NSUN methyltransferases.

It is suggested that NSUN4's methyltransferase activity is required for its primary function in ribosome biogenesis. That is, it is possible that NSUN4 has utilised its methyltransferase activity as a means to mark small subunits destined for ribosome assembly.

3.3.7 NSUN4 likely does not methylate human mitochondrial 16S rRNA

The Illumina bsRNA-seq data showed a region of clustered sites in the 16S mt-rRNA transcript, and as a result these sites were considered as false positives (Table 3.2). However, a single m^5C site was identified in the 16S mt-rRNA transcript from the SOLiD data (Table 3.1). This candidate site displayed an apparent ~10% conversion (90% methylation rate) across all four SOLiD libraries (NSUN2 knockdown, TRDMT1 knockdown, DNMT1 knockdown and NTC). However, MiSeq amplicon sequencing did not find methylation at the expected candidate site (Appendix 7.2) This apparent contradiction between the SOLiD and the MiSeq result could be attributed to: 1) erroneous mapping resulting from high sequence similarity between 16S mt- rRNA and the nuclear encoded 16S-like mRNA sequence, and 2) low Phred scores at the called site, indicating cytosines were identified at the site with low confidence. Both of these factors likely contributed to the false positive detection of m^5C at this site. This result is a case in point, which clearly highlights the importance of employing a validation approach to assess the accuracy of m^5C detection on a transcriptome-wide scale.

Generally, as sequencing coverage across the 16S mt-rRNA transcript was sufficient, it is concluded that there are likely no m^5C sites present in 16S mt-rRNA.

The finding in this thesis contradicts a study that suggests that NSUN4 targets an unknown site in 16S mt-rRNA. Yakubovskaya *et al.* (2012) performed an *in vitro* methylation assay, which showed a significant incorporation of ^3H-methyl groups into 16S mt-rRNA in the presence of the NSUN4/MTERF4 complex. That work followed on from structural homology modeling which showed a strong similarity between NSUN4 and bacterial methyltransferases known to target the analogous bacterial 16S rRNA (Camara *et al.*, 2011). However, findings presented in this thesis are consistent with those of Metodiev *et al.* (2014) who do not detect any m^5C sites in mouse 16S mt-rRNA. Overall, this result may be unexpected, as MTERF4 is known to direct NSUN4 to the 16S mt-rRNA LSU during ribosome biogenesis. Therefore, as NSUN4 interacts

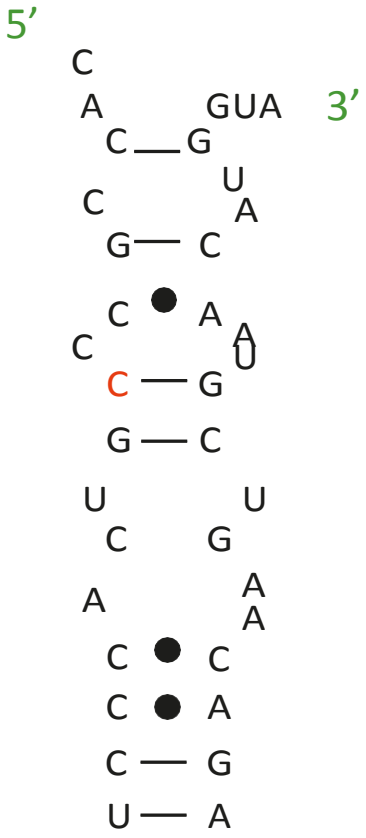


Figure 3.16: Secondary structure prediction of a section of helix 44 in 12S rRNA.

The structure was generated using RNAfold. The red C denotes the m^5C methylated cytosine. The methylated cytosine is present in a bulge stem sturcture which my be a requirement for recognition by NSUN4.

with both subunits, it likely acts as a mediator between the two subunits, bringing them together in the final stage of ribosome assembly.

3.3.8 Protein domains of NSUN2 and NSUN4

Recent work from our laboratory has provided evidence that NSUN2 targets a plethora of different RNA biotypes (Squires *et al.*, 2012). This versatile, multi-substrate specific enzyme has been shown to target coding RNAs and various ncRNAs. This is quite interesting as work presented here suggests that NSUN4 most probably methylates only a single target. This may be a consequence of the fact that these enzymes have evolved along different lines representing phylogenetically different subfamilies. Furthermore, a comparison between the domains present in NSUN2 and NSUN4 may explain differences in substrate specificity. NSUN2 contains five RNA-cytosine methyltransferae (RCMT) motifs which are involved in target cytosine binding, AdoMet/SAM binding and catalysis (Fig. 3.17). NSUN4 only possesses the AdoMet/SAM binding and the two catalysis domains but not the target cytosine binding domain, nor the RNA binding domain (Fig. 3.17), even though, its closest bacterial homologue of NSUN4, YebU, does contain a domain for RNA binding. All NSUN RNA m^5C methyltransferases contain these core RCMT domains, which are thought to provide cytosine-specific RNA methyltransferase activity. Bujnicki *et al.* (2004) show that the catalytic region forms into a Rossmann fold, which is required to bind AdoMet/SAM. NSUN2 also contains a specific 3-domain fingerprint known as NCL1: motif 1 is at the N- terminus; motif 2 is in the catalytic region; and motif 3 is at the C-terminus (Fig. 3.17). These domains are expected to contribute toward RNA recognition and catalysis.

The difference in domain structure between NSUN2 and NSUN4 as highlighted in Figure 3.17, suggests that NSUN4 might have evolved as a specific mt-rRNA methyltransferase. That is, the lack of N- or C-terminal extensions associated with RNA binding, suggests that NSUN4 may have a much tighter substrate range compared to NSUN2. Therefore, it follows that the methylation mark deposited by NSUN4 would be required for NSUN4 to facilitate mitoribosome biogenesis.

3.4 Future Directions – Outlook/open-ended questions

This Chapter has provided insights into the substrate specificity of the mitochondrial NSUN4 m^5C:RNA methyltransferase. Future work should be directed at determining

substrates of the other putative NSUN methyltransferases and understanding the function of m^5C in those contexts. It seems likely that the NSUN methyltransferases are not redundant and each will selectively target a specific set of sites. It would be interesting, therefore, to determine what contributes to the selectively of each enzyme, and whether m^5C, synthesised by different enzymes, is able to mediate a different function. For example, do 'reader' proteins selectively bind m^5C marks put on by a specific enzyme? Also, as most of the NSUN enzymes have known roles in disease, it would be interesting to determine whether their methyltransferase activity is coupled with their role in disease.

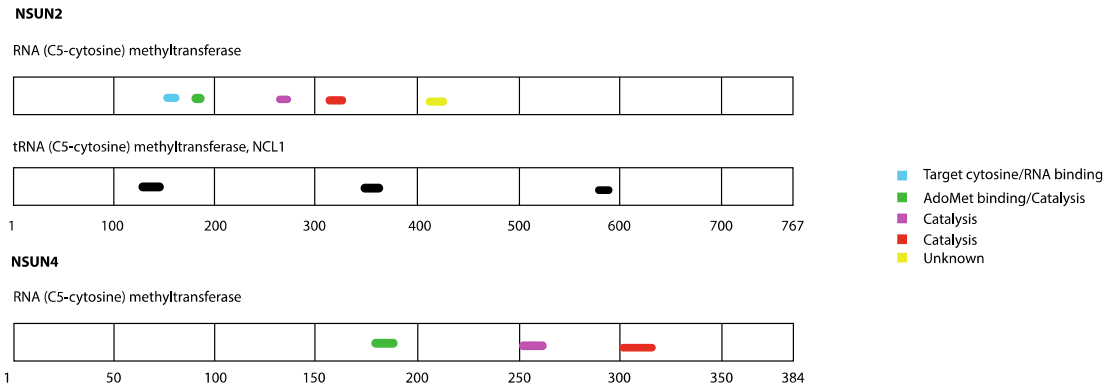


Figure 3.17: Schematic representation of the domain structures present in NSUN2 and NSUN4.

NSUN4 does not contain a RNA recognition domain, but does contain the domains for cofactor (AdoMet/SAM) binding and catalysis. NSUN2 contains the five prominent RNA C5- cytosine methyltransferase domains, but also contains NCL1 domains, which are expected to contribute to RNA recognition and catalysis.

Elements corresponding to particular enzyme domains are coloured: blue = target cytosine/RNA binding, green = AdoMet/SAM (cofactor) binding; red and purple = catalytic domains.

Chapter IV

m^5C-associated proteins mediate m^5C function

4.1 Preamble

Currently m^5C-binding proteins in mRNA have not yet been identified, however such proteins, would be expected to mediate m^5C function. Also, as m^5C is located within all mRNA regions, it is conceivable that the m^5C mark may elicit multiple functions depending on its location within the mRNA. Thus, different m^5C specific RNA binding proteins (RBPs) may recognise different spatially localised m^5C sites and facilitate different functions. Such a scenario would emulate findings with m^6A. The m^6A modification is clearly defined in mRNAs, being present in the 5≠UTR, but also more prominently located at the 3≠end of the mRNA around the stop codon. Indeed, many recent studies have shown that different RBPs are associated with each spatial set of m^6A sites. For example, Meyer *et al.* (2015) recently showed that 5≠UTR localised m^6A sites correlate with eIF3 footprints, resulting in cap-independent translation. Another example is that of Wang, Lu, *et al.* (2014) who showed that YTHDF2 specifically recognises m^6A sites in the 3≠UTR, directing the transcripts to processing bodies for decay.

This Chapter focuses on determining proteins that are associated with m^5C. The presence/absence of the m^5C mark is broadly expected to regulate protein-RNA binding, and ultimately affect mRNA fate. This may be achieved either by 1) directly recruiting proteins (m^5C-binding proteins), analogous to m^5C-binding proteins in DNA and m^6A 'reader' proteins in mRNAs, or 2) by blocking/enhancing protein binding.

In this thesis, both bioinformatic and experimental approaches were employed to determine proteins associated with m^5C. Overall findings in this Chapter suggest that m^5C may be involved in regulating mRNA stability, by blocking Argonaute 2 (AGO2) binding, and in enhancing mRNA translation, by recruiting m^5C-binding proteins.

4.2 Results

4.2.1 Potential role for m^5C in mRNA stability

Previously, Squires *et al.* (2012) reported an association between AGO2 and m^5C candidate sites in the CDS and 3≠UTR. Presently, this analysis was confirmed and

extended using the second SOLiD dataset (see overview of datasets used in this thesis), suggesting m^5C may be involved in the mechanism of miRNA-mediated gene regulation.

It should be noted that the AGO2-m^5C association was not observed with the sites detected from the Illumina dataset. This could be because there is a relatively small number of mRNAs that contain both a high confidence m^5C candidate site and an AGO2 footprint in the same region.

4.2.1.1 Argonaute 2 footprints correlate with m^5C candidate sites

Publicly available HeLa AGO2 HITS-CLIP data (Hafner *et al*., 2010) was cross-referenced with transcriptomic coordinates of candidate m^5C sites identified in the 3≠ UTR or the CDS in the unpublished SOLiD dataset. The same m^5C candidate sites were also interrogated with a publicly available HeLa HuR PAR-CLIP dataset (Lebedeva *et al*., 2011) (see Chapter II). This was performed in order to determine whether the AGO2-m^5C association was specific. The rationale is: HuR is another known RBP that has been shown to increase RNA stability by blocking miRNA targeting (Xu *et al*., 2005). Lastly, to determine the specificity of the observations, the AGO2 and HuR footprints were overlapped with randomly shuffled candidate m^5C sites (negative control) within the same mRNA region. These analyses were performed by Dr Brian Parker (see Acknowledgements).

A 2.8-fold enrichment of AGO2 footprints was observed, on average, 25 nt upstream of 3≠ UTR m^5C candidate sites relative to the random shuffled control ($p \approx 0$, permutation test, 100 repetitions) (Fig. 4.1A). A similar association was noted between AGO2 footprints and CDS m^5C candidate sites. A 2.3-fold enrichment of AGO2 footprints was shown in the CDS, on average, 25 nt upstream of candidate CDS m^5C sites, relative to the random shuffled control ($p \approx 0$, permutation test, 100 repetitions) (Fig. 4.1B). HuR footprints were not enriched near candidate m^5C sites in the 3≠ UTR ($p=0.0910$, permutation test, 100 repetitions) or CDS ($p=0.2727$, permutation test, 100 repetitions) compared to the random shuffled control.

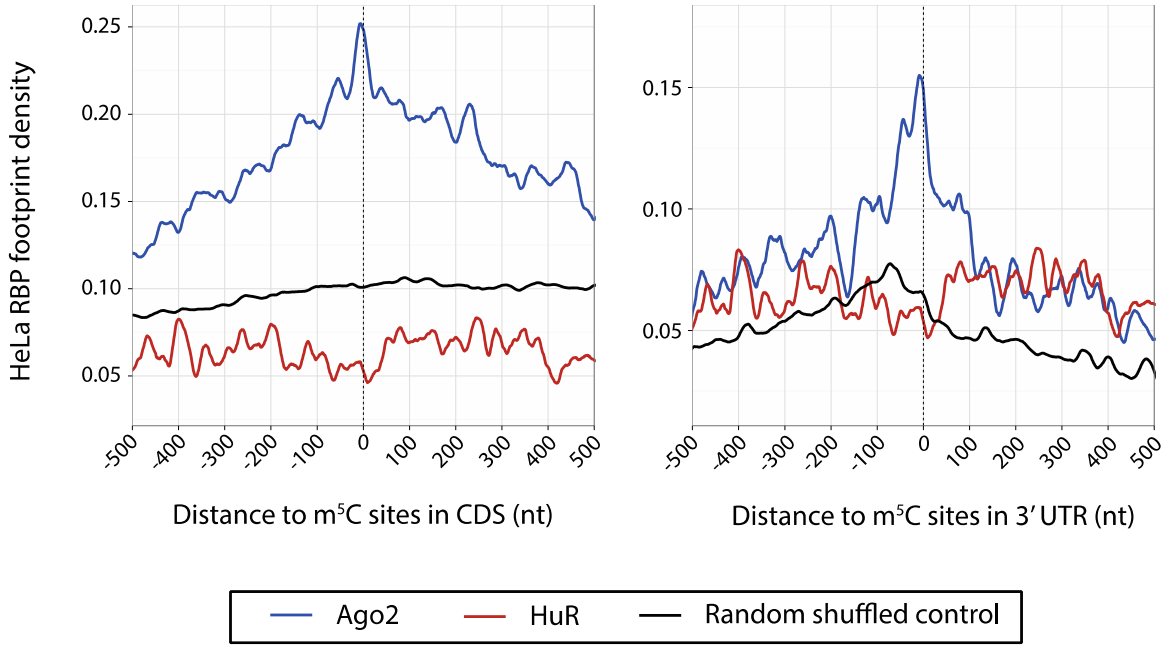


Figure 4.1. Overlapping HeLa AGO2 and HuR footprint densities with candidate m^{5}C sites in the CDS and 3≠ UTR.

Candidate m^{5}C sites detected from SOLiD HeLa bsRNA-seq was cross-referenced with publicly available HeLa AGO2 HITS-CLIP (Hafner *et al.*, 2010) data and HuR PAR-CLIP data (Lebedeva *et al.*, 2011). The spatial distribution summary plots were generated by measuring distances from candidate m^{5}C sites to the centre of AGO2 footprints. The random shuffled control was generated by randomly shuffling candidate m^{5}C sites within the same mRNA region. 500 nt regions flanking the candidate m^{5}C sites are shown. Significant enrichment of AGO2 footprints upstream of candidate m^{5}C sites in the CDS and 3≠ UTR was observed relative to the random shuffled control (p≈0, permutation test, 100 repetitions). HuR footprints did not show a correlation with candidate m^{5}C sites in the CDS or 3≠ UTR.

Furthermore, as an extension to the analysis, it was explored whether there was a dependency between methylation stoichiometry and AGO2 footprint density. To this end, candidate m^5C sites were distributed amongst four sets, each increasing in methylation percentage (0-25%, 25-50%, 50-75%, 75 -100% methylation). Interestingly, highly methylated sites in the 3≠UTR and CDS showed a statistically significant negative relationship with AGO2 footprint densities (Fig. 4.2).

4.2.1.2 miRNA depletion at m^5C candidate sites

The RNA-induced silencing complex (RISC), in which Argonaute proteins are the core subunit, uses miRNAs to target mRNAs through complementary base pairing. This results in mRNA destabilisation or translational repression (Krol *et al.*, 2010). The miRNA recognises its mRNA target primarily through base pairing with its 7 nt 'seed' region (Pasquinelli, 2012). Therefore, as an independent confirmation of the observed AGO2-m^5C association, the connection between density of miRNA seed matches and m^5C sites in the CDS and 3≠UTR was investigated. To this end, moderately to highly conserved 7-mer motifs in CDS and 3′ UTR regions were divided into two subsets. The first contained miRNAs possessing seed matches complementary to the target mRNA (i.e. these could potentially be functional miRNAs), whilst the second contained miRNAs seed matches identical to the mRNA target (i.e. these serve as a negative control).

In accordance with AGO2 footprint densities being observed just upstream of candidate CDS and 3≠UTR m^5C sites, a stark depletion of miRNA seed matches overlapping candidate m^5C sites was observed (Fig. 4.3). The CDS and 3′ UTR showed a striking 49% and 47% depletion of miRNA seed matches at m^5C sites when compared to the non-miRNA seed match distribution, respectively ($p \approx 0$, permutation test, 100 repetitions). Furthermore, as a lower cytosine frequency is apparent within the 3≠UTR of mRNAs (discussed in Chapter V), additional controls were employed in order to ensure these observations were not a result of differences in base composition. Firstly, and most importantly, no depletion in miRNA seed matches was apparent with regards to unmethylated cytosines (Fig. 4.3, solid grey trace). Secondly, no depletion was observed over a control set of randomly shuffled 7-mer motifs, having identical base composition to actual miRNA seeds (Fig. 4.3, green trace).

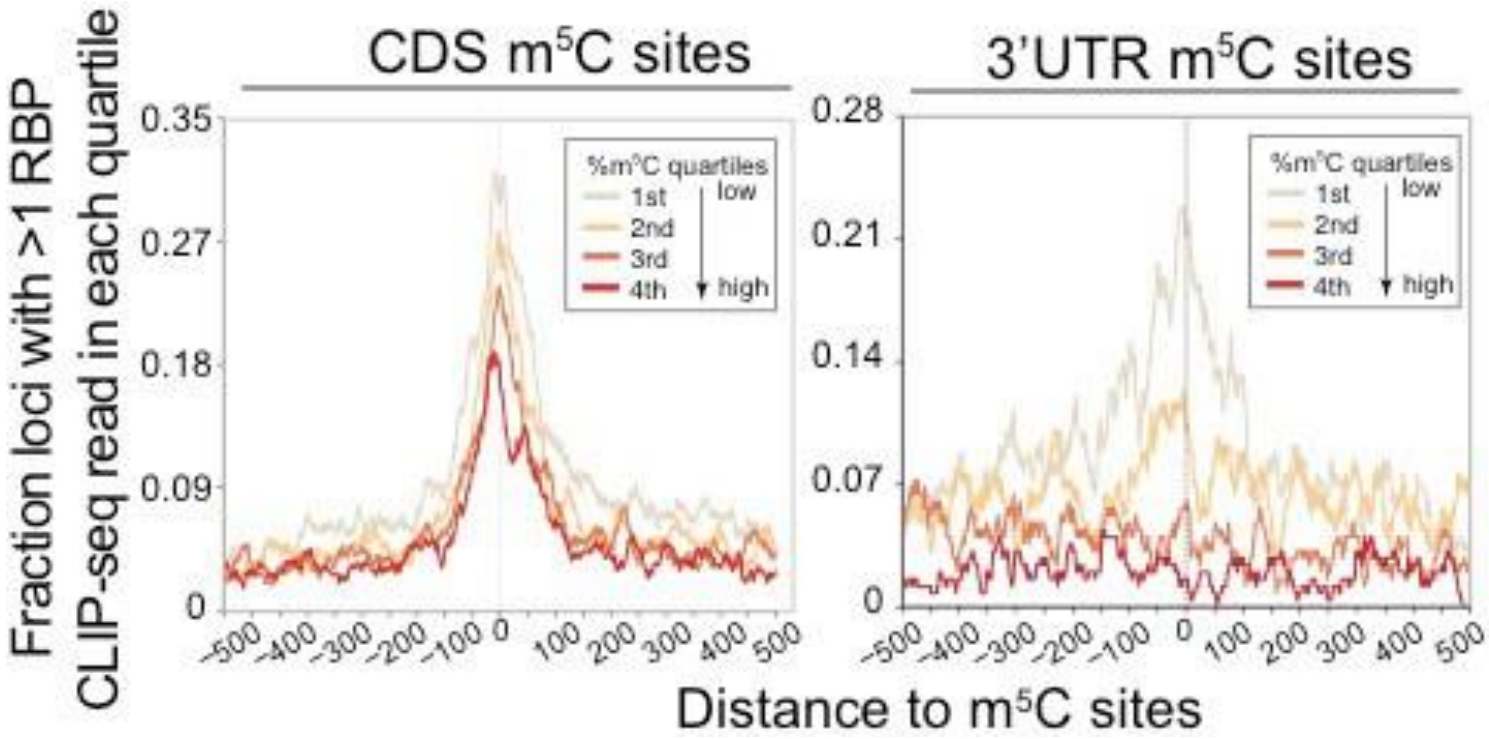


Figure 4.2: Candidate m^{5}C sites distributed across quartiles and overlapped with AGO2 HITS-CLIP HeLa data.

Spatial distribution plots anchored at candidate m^{5}C sites. Candidate m^{5}C sites are distributed across quartiles ranked by methylation percentage (note: introns are included in the CDS plot). Lowly methylated transcript stoichiometries candidates correlate well with AGO2 footprint densities, whilst highly methylated transcript stoichiometries candidates show weak correlation with AGO2 footprint densities. This effect is more pronounced with the 3≠UTR m^{5}C sites than sites localised in the CDS.

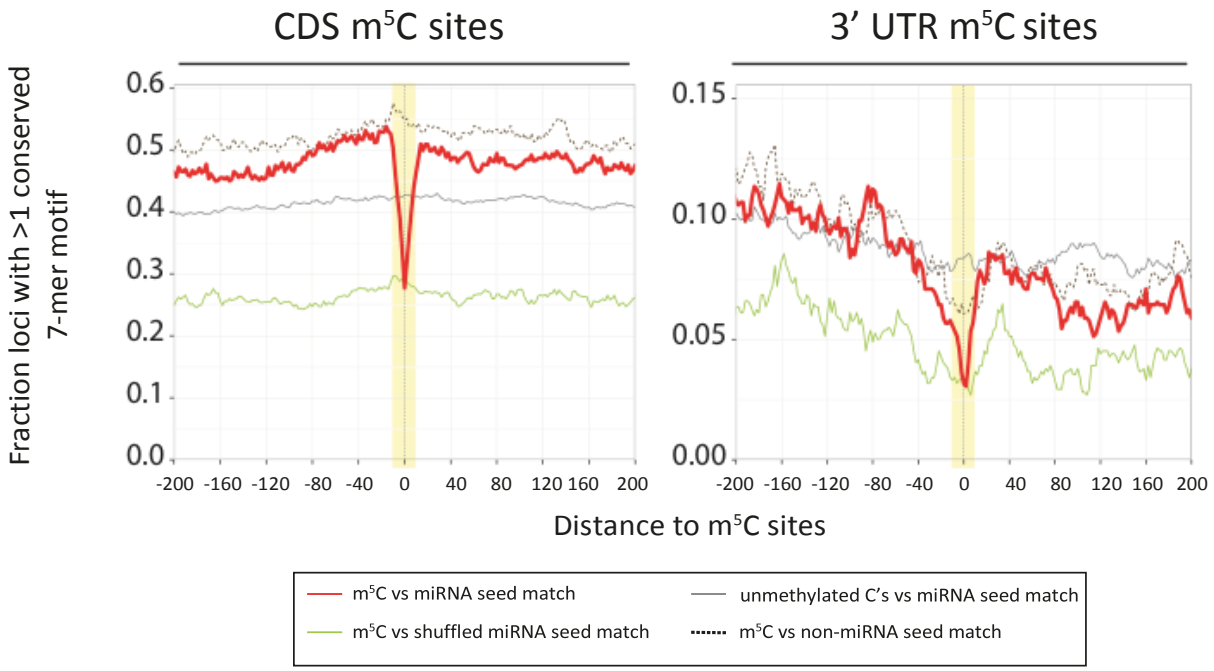


Figure 4.3: m^{5}C sites are depleted in conserved 7-mer seeds over CDS and 3≠ UTR in HeLa.

Spatial distribution plots of conserved 7-mer motif density around m^{5}C sites were generated by anchoring at m^{5}C sites in the CDS and 3≠UTR of mRNAs in HeLa cells. The x-axis shows 100 nt flanking regions around m^{5}C sites and the y-axis shows 7-mer motif density. The red trace shows a depletion of m^{5}C sites from conserved seeds matching the reverse complemented miRNA seed regions from miRBase v19, the grey dotted trace shows no relationship between m^{5}C sites and conserved motifs not matching miRBase v19. The grey solid trace shows no relationship between unmethylated cytosines and conserved motifs not matching miRBase v19. Then green trace shows no the relationship between m^{5}C sites and mono-shuffled miRNA seed regions for the two mRNA regions.

Notably, the depletion of miRNA seed match density at m^5C sites occurred across all methylation levels (Fig. 4.4). Moreover, permutation tests for miRNA seed match enrichment flanking m^5C sites (see Chapter II) showed a significant enrichment of seed matches within 10 to 30 nt upstream of m^5C sites in the CDS (1.37 fold enrichment), but not the 3≠UTR (Fig. 4.4). This finding concurs with the observation that AGO2 footprints are enriched just upstream of CDS m^5C sites.

4.2.1.3 Is AGO2 involved in m^5C formation?

The association between AGO2 footprint densities and m^5C sites may be explained by different scenarios (see Discussion). Presently, the possibility that AGO2/miRNAs may be involved in the formation of m^5C was considered. For example, it is possible that AGO2 may be in complex with NSUN2; i.e. through miRNA-mRNA binding, NSUN2 may be guided to its target for methylation. Thus in order to investigate this possibility, co-immunoprecipiation (co-IP) experiments of AGO2 and NSUN2 were performed.

4.2.1.3.1 Specificity and Efficiency of the immunoprecipitation

To ensure the specificity and efficiency of the AGO2 and NSUN2 co-IPs, positive and negative controls were first performed. AGO2 blotting was performed on AGO2 IP eluate; a depletion of AGO2 was observed in the supernatant (Fig. 4.5A). Similarly, NSUN2 blotting was performed on NSUN2 IP eluate; a depletion of NSUN2 was observed in the supernatant (Fig. 4.5B). This suggests that both the antibody and experimental parameters chosen resulted in efficient immunoprecipitation.

4.2.1.3.2 Optimisation of co-immunoprecipitation

Aspects of the experimental procedure required optimisation to achieve an accurate, reproducible co-IP result.

Different wash strategies were trialed to reduce non-specific proteins binding to the beads or antibody, while maintaining specific interactions between the protein of interest and associated proteins. A high salt buffer proved to be ineffective, with

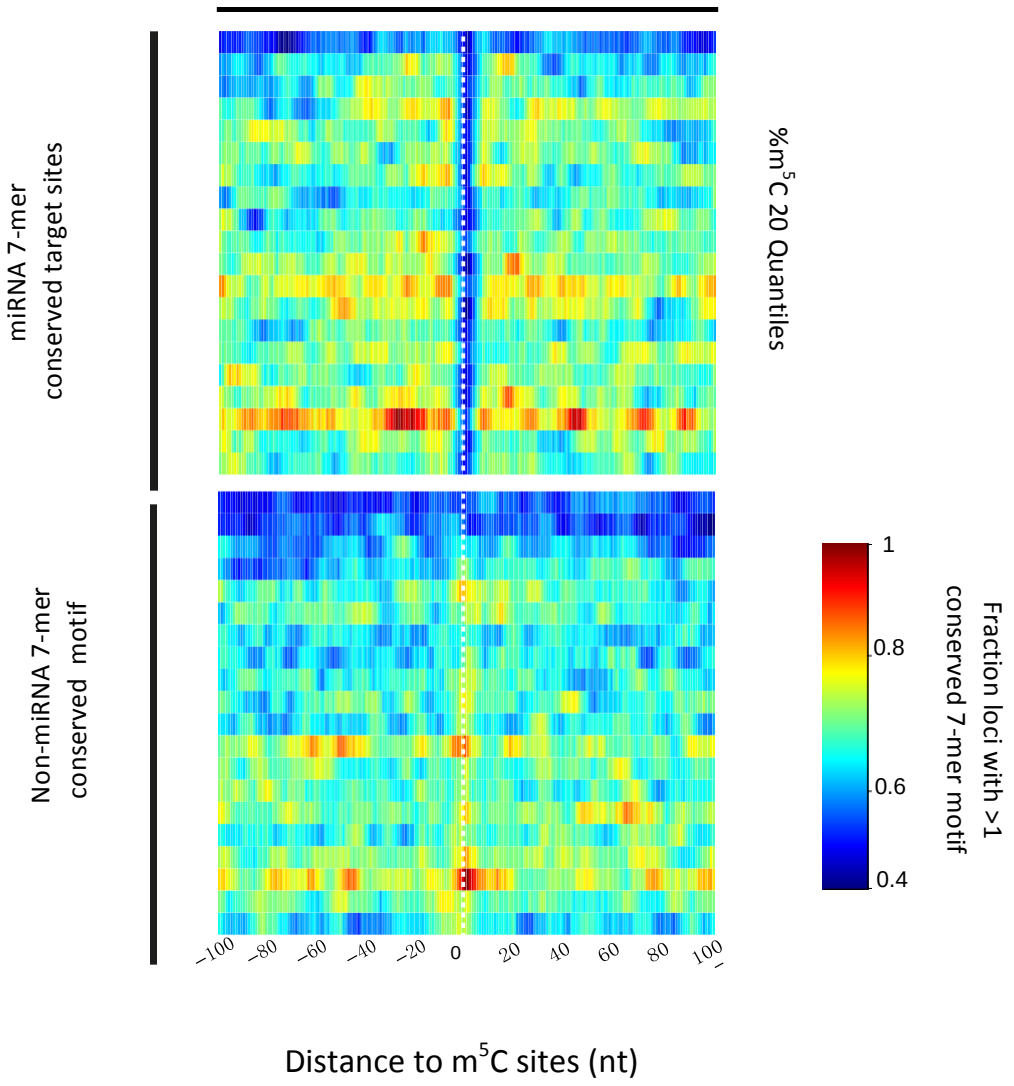


Figure 4.4: Depletion of miRNA seed match density at all methylation levels.

Spatial distribution plots as shown in Figure 4.3, but showing quantiles ranked by methylation percentage (note: introns are included in the CDS plot). Plots are anchored at the m^5C site. A depletion of 7-mer conserved motifs is observed at m^5C sites. There is no such association observed with non-miRNA 7-mer conserved motifs.

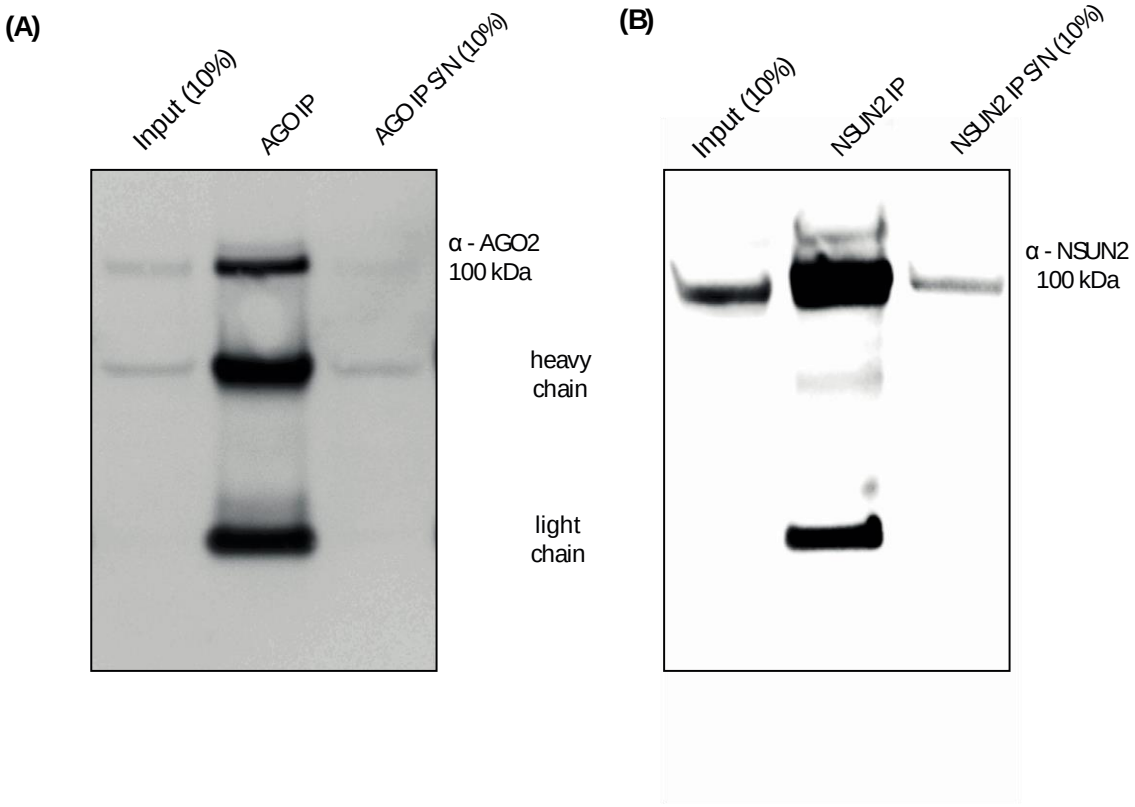


Figure 4.5: Evidence of efficient AGO2 and NSUN2 immunoprecipitation.

(A) AGO2 blotting performed on AGO2 IP eluate and AGO2 S/N. AGO2 is enriched in the AGO IP eluate. The band at approximately 100 kDa corresponds to AGO2.

(B) NSUN2 blotting performed on NSUN2 IP eluate and NSUN2 S/N. NSUN2 is enriched in the NSUN2 IP eluate. The band at approximately 100 kDa corresponds to NSUN2. 100% of the IP eluates are loaded; 10% of the S/Ns are loaded. No other bands, except for those corresponding to the heavy and light chain of the antibody, were observed on the blots.

NOTE: S/N = supernatant (flowthrough); IP = immunoprecipitation

proteins of interest being detected in the IgG negative control eluate (Fig. 4.6). A high percentage detergent wash solution (RIPA buffer) resulted in a clean IgG negative control elaute i.e. proteins of interest did not bind non-specifically to the IgG negative control. This suggests that the RIPA buffer was effective in washing non-specific bound proteins, without affecting the specific protein-protein interactions of interest.

The antibody–protein lysate incubation conditions and times were varied. A trial of different temperatures and times indicated that a 4°C overnight incubation did not negatively affect antibody–protein binding, whilst non-specific binding to the IgG negative control was not observed. This was determined by probing AGO2 IP eluates for GW182, which is known to interact with AGO2. GW182 is recruited to the miRNA target through direct interaction with AGO2 to promote mRNA silencing (Eulalio *et al.*, 2008). The AGO2-GW182 complex was observed following incubation of the antibody-protein lysate at both 1 hr at room temperature and overnight at 4°C (Fig. 4.7). The latter was chosen as this increased the amount of specific binding, with no evidence in background binding in the IgG negative control.

4.2.1.3.3 Evidence for co-immunoprecipitation

To ensure that specific, intact, functional complexes were immunoprecipitated, it was determined whether known interactors of AGO2 and NSUN2 were also precipitated. To this end, the presence of GW182 in the AGO2 IP eluate and its absence in IgG control IP eluate was determined by western blotting (Fig. 4.7). Also, α-tubulin has been reported to directly interact with NSUN2 (Hussain *et al.*, 2009). The presence of α-tubulin in the NSUN2 IP eluate and its absence in IgG control IP eluate was confirmed by western blotting (Fig. 4.8).

4.2.1.3.4 AGO2 interacts with NSUN2 in an RNase sensitive manner

Next, the presence of NSUN2 was probed for in the AGO2 IP eluate and vice versa. This revealed a fraction of endogenous HeLa NSUN2 co-immunoprecipates with AGO2. However subsequent experiments showed this interaction to be RNase-sensitive; therefore AGO2 and NSUN2 are most probably indirectly interacting through a bridging RNA, rather than directly through a physical protein-protein interaction (Fig. 4.9). This result also suggests that AGO2 does not participate in the formation of m^5C

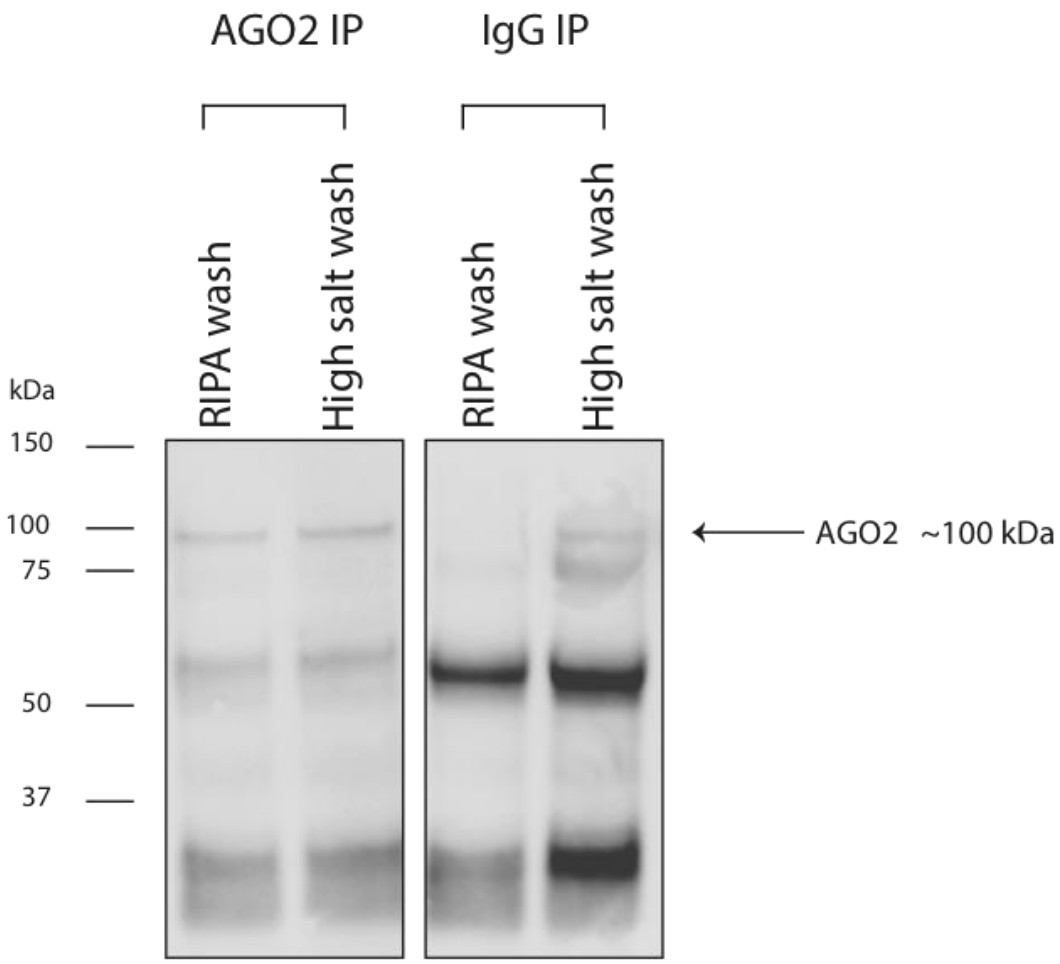


Figure 4.6: Optimisation of the AGO2 co-IP experiment.

A harsh RIPA wash buffer and a mild high salt wash buffer were trialed. Both wash buffers did not affect AGO2-antibody binding, although the RIPA wash resulted in slightly less AGO2 being bound compared to the high salt wash buffer. AGO2 was also detected in the negative control IgG IP when using the high salt wash buffer but not the RIPA buffer.

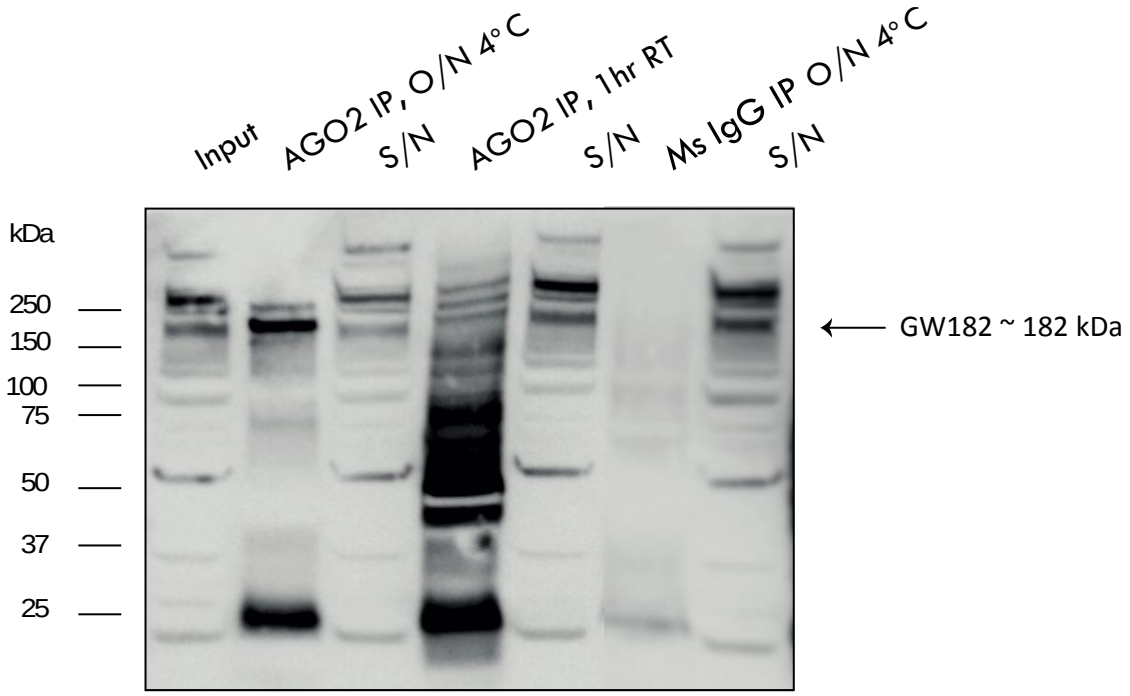


Figure 4.7: Optimisation of co-IP – GW182 associates with AGO2.

GW182 is known to interact with AGO2. GW182 was observed following incubation of the AGO2 antibody with the protein lysate for both 1 hr at room temperature and overnight at 4°C The longer incubation increased specific binding; GW182 is associated with AGO2 and depleted from the AGO S/N. GW182 is not present in the Ms IgG IP, whilst being present in the Ms IgG S/N. 10% of the input is loaded; 100% of the IP eluates are loaded; 10% of the S/N are loaded. The band at approximately 182 kDa corresponds to GW182.

This is representative of n = 3 experiments

NOTE: S/N = supernatant (flowthrough); IP = immunoprecipitation; O/N = overnight; RT = room temperature; Ms = mouse

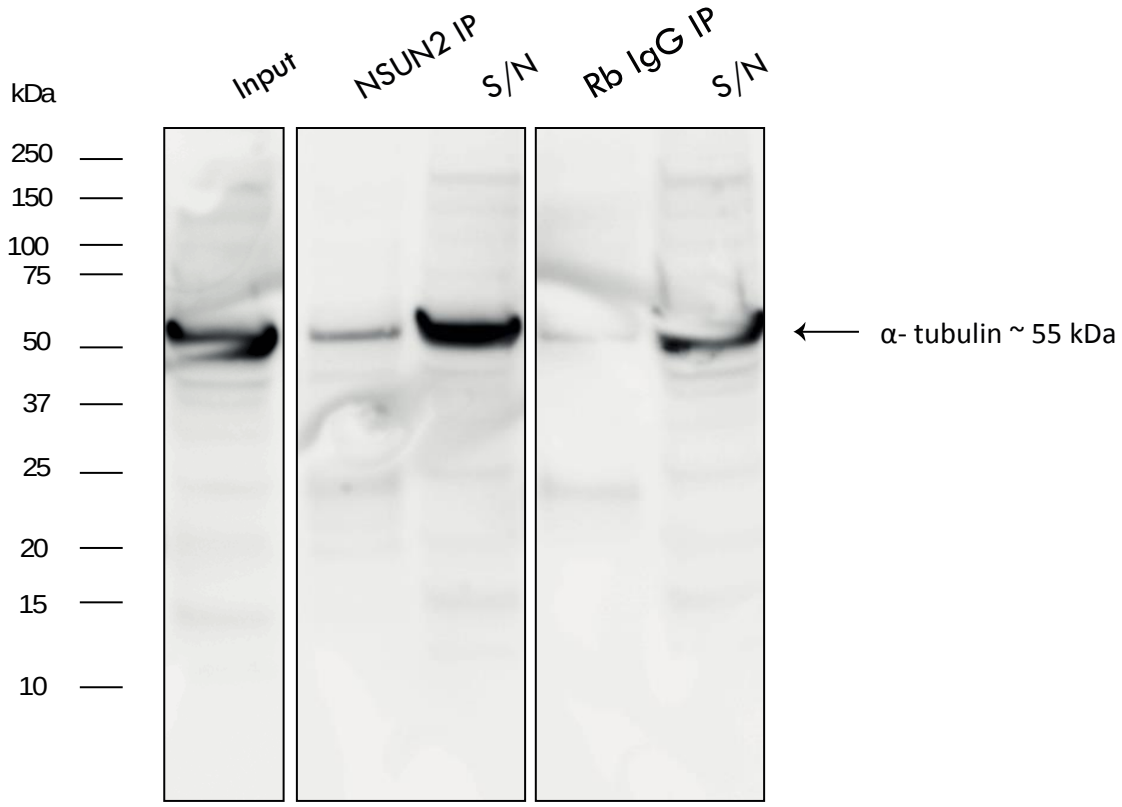


Figure 4.8: Confirmation that α-tubulin associates with NSUN2.

α-tubulin is present in the NSUN2 IP eluate but not present in the Rb IgG IP. α-tubulin but is present in the Rb IgG S/N. 10% of the input is loaded; 100% of the IP eluates are loaded; 10% of the S/N are loaded. The band at approximately 55 kDa corresponds to α-tubulin.

This is representative of n = 3 experiments.
NOTE: S/N = supernatant (flowthrough); IP = immunoprecipitation; Rb = rabbit

by directing NSUN2 to its target. However, the possibility that miRNAs regulate m^5C formation cannot be excluded (see Discussion).

4.2.2 Potential role for m^5C in mRNA translation

4.2.2.1 m^5C-binding proteins

4.2.2.1.1 Selecting the RNA bait

Previous unpublished work provided some indication that m^5C in mRNA may be involved in regulating translation processes (Sibbritt 2015, PhD thesis). In a candidate-based approach, the ratio of methylated to unmethylated mRNA molecules, was determined in five different translation rates. Of the nine chosen mRNAs containing validated m^5C sites, two showed a trend linking the presence/absence of m^5C with translation rate. Interestingly, the presence of m^5C methylation seemed to positively correlate with increasing translation rates for the *RTN3* mRNA i.e. a larger proportion of actively translated *RTN3* transcripts were methylated compared to the pool of weakly translated *RTN3* transcripts. Conversely, an increased translation rate was observed with decreased methylation rate for the *RPS3* mRNA i.e. the pool of weakly translated *RPS3* transcripts contained more methylated transcripts than the pool of highly translated RPS3 transcripts. Thus these m^5C-containing mRNAs represented promising candidates to identify possible proteins that may mediate a role for m^5C in translation in this thesis. Presently, the approach taken, employed two synthetic oligonucleotides (of the same sequence), with and without the m^5C mark, to serve as baits to pulldown potential m^5C-binding proteins. A 5≠ biotin label was added to the baits, in order to allow for efficient isolation of baits using streptavidin beads. Isolated proteins were then identified by mass spectrometry analysis.

As the synthesised RNA bait would be short in length (approximately 40-50 nt), one of the challenges was in selecting a region of the RNA containing the m^5C site that would resemble the actual secondary structure in a physiological setting. This is critically important, as the correct RNA structure may be required for an m^5C 'reader' to recognise its target. Using RNAfold, multiple predicted secondary structures for each candidate were generated. The *RPS3* candidate yielded structures where the m^5C site was placed in different structural motifs (Fig. 4.10A). In contrast the m^5C site in the

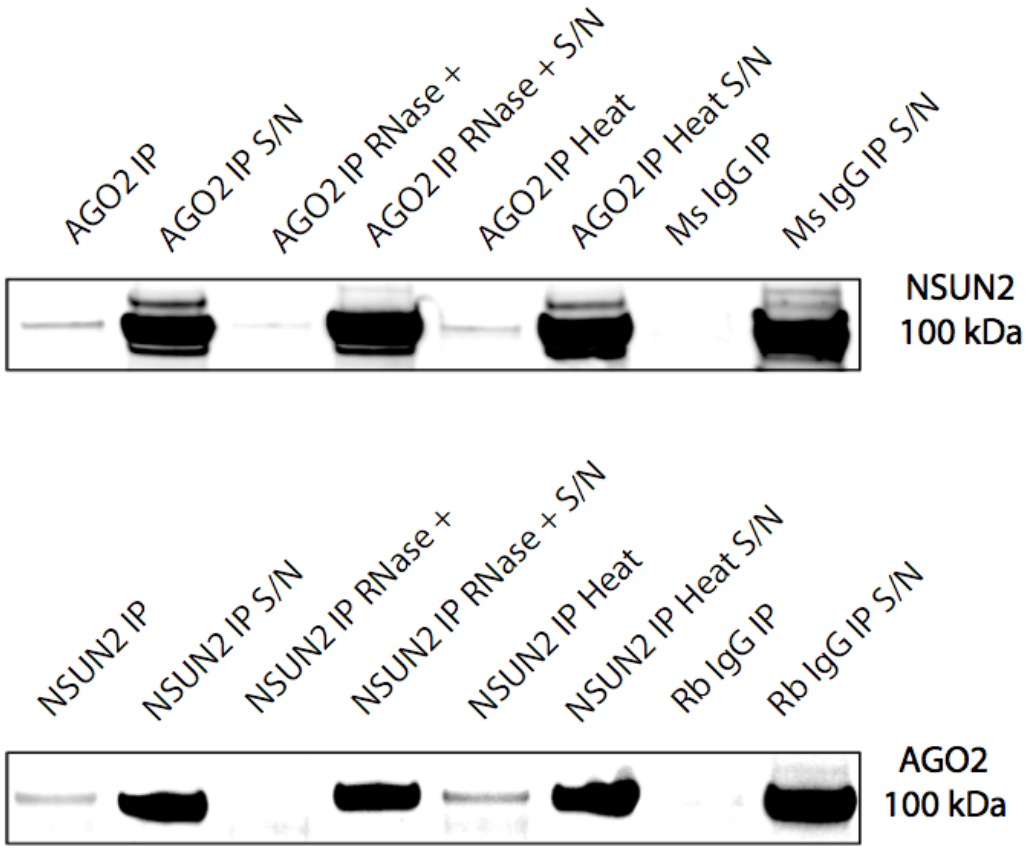


Figure 4.9: NSUN2 interacts with AGO2 in an RNase-sensitive manner.

Top panel: AGO2 IP and mouse IgG IP eluates probed with NSUN2. NSUN2 associates with AGO2, however upon RNase treatment NSUN2 no longer interacts with AGO2 suggesting an indirect interaction between the two proteins. Heat treatment alone (without the addition of RNase) does not dissociate the interaction. The mouse IgG negative control shows no NSUN2 present. 100% of eluate is loaded; 10% of the supernatant is loaded. The band at approximately 100 kDa corresponds to NSUN2.

Bottom panel: NSUN2 IP and rabbit IgG IP eluates probed with AGO2. The reciprocal interaction is also observed i.e. AGO2 associates with NSUN2. Again, upon RNase treatment AGO2 no longer interacts with NSUN2 suggesting an indirect interaction. Heat treatment alone does not dissociate the interaction. The rabbit IgG negative control shows no AGO2 present. 100% of eluate is loaded; 10% of the supernatant is loaded. The band at approximately 100 kDa corresponds to AGO2.

These are representative blots from n = 4 experiments.
NOTE: S/N = supernatant (flowthrough); IP = immunoprecipitation; Ms = mouse; Rb = rabbit

RTN3 candidate was consistently located with the loop of a hairpin (Fig. 4.10B).

Thus, the RNA baits were modelled on the m^5C site in the *RTN3* candidate. The sequence of the bait is
ACTCCCTCTGCCACTATCCCCAGGGAAGGAAAGGCTCCGC,
where the red cytosine denotes the m^5C site.

4.2.2.1.2 Optimising the RNA bait pulldown

Several parameters in the RNA bait pulldown experiment required optimisation.

As the structure of the RNA is expected to contribute to its recognition by the 'reader' enzyme, it is necessary to ensure proper secondary structure formation. To this end the RNA baits were first heat denatured then chilled on ice and finally supplied with RNA structure buffer (see Chapter II) to assist in correct secondary structure formation. This was done in order to achieve a reproducible, accurate set of proteins bound to each of the baits.

Also, as the mRNA is expected to be modified in the nucleus, and therefore could facilitate a function there e.g. in splicing, it was suggested that a nuclear lysate might be suitable. However, as m^5C function is most probably elicited in the cytoplasm, for example, in affecting mRNA translation or stability, a cytoplasmic lysate may provide more physiologically relevant candidate m^5C binding proteins. The Cytobuster™ lysis reagent has been shown to produce 'whole cell lysates' under non-denaturing conditions. Thus this lysis method was preferred as both intact nuclear and cytoplasmic proteins would be present.

A major issue with the pulldown approach was the binding of non-specific proteins to the streptavidin coated magnetic beads. This is a problem as this reduces the resolution of the experiment in regards to detecting proteins of interest in the mass spectrometry. Thus, in an attempt to reduce non-specific binding, the beads were initially blocked with bovine serum albumin (BSA) before incubation with the RNA bait and whole cell lysate. This seemed to only reduce the issue of non-specific binding only slightly (Fig. 4.11). Also as the elution method employed was quite harsh (i.e. everything bound to the beads would be eluted), it was assumed that the BSA signal could 'flood' the mass

(A)

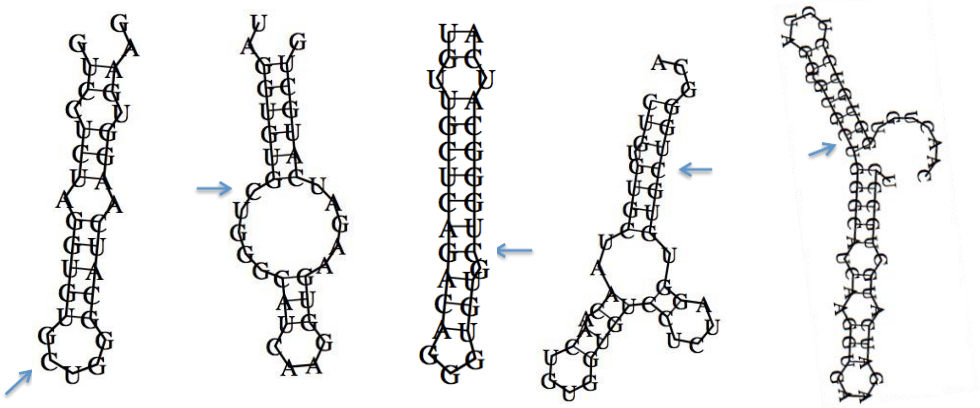
GTCCTCTAGGTGTGCTGGGCATCAAGGTGAAG

TAGGTGTGCTGGGCATCAAGGTGAAGATCATGCTG

TGTTGCTCAGACAGGGTGTGCTGGGCATCA

CUGUGUGCUAACAACTGTGGTGTCCTCTAGGTGTGCTGGGCA

CAACTGTGGTGTCCTCTAGGTGTGCTGGGCATCAAGGTGAAGATCATGCTGCCCT



(B)

ACTCCCTCTGCCACTATCCCCAGGGAAGGAAAGGCTCCGC

TATCCCCAGGGAAGGAAAGGCTCCGCCATTT

GTGGCGGTCACTCCCTCTGCCACTATCCCCAGGGAAGGAA

TCTGCCACTATCCCCAGGGAAGGAAAGGCTCCGCCATTTGGG

CCACTATCCCCAGGGAAGGAAAGGCTC

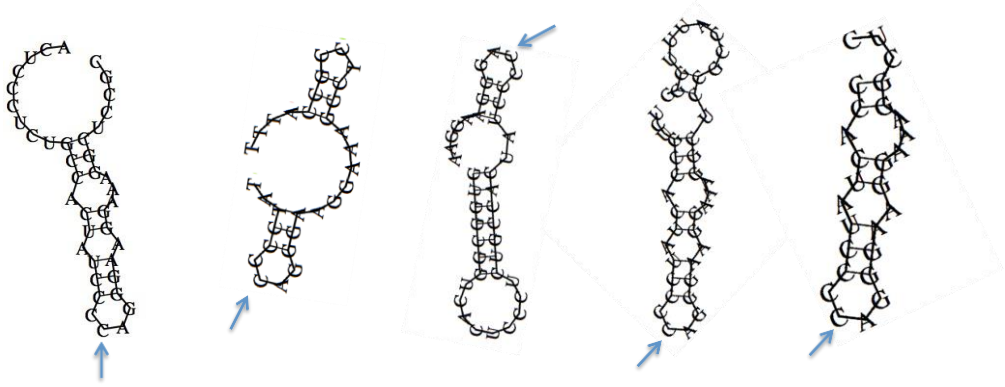


Figure 4.10: Predicted secondary structures of candidate RNA baits using RNAfold.

(A) Predicted secondary structures for the *RPS3* candidate bait were generated using RNAfold, the m^5C site is denoted in red. The m^5C site is present within different structural motifs.

(B) Predicted secondary structures for the *RTN3* candidate bait were generated using RNAfold, the m^5C site is denoted in red. The m^5C site is consistently shown to be located in the loop of a hairpin.

spectrometry and proteins of interest may not be detected. Therefore, this approach to block the beads was abandoned. To resolve this issue, the elution strategy was optimised so that potential m^5C-binding proteins were eluted off the RNA baits, whilst non-specific proteins bound to the beads were not. To this end a more specific elution method was employed, herein termed RNase-assisted elution in order to significantly reduce elution of non-specific proteins bound to the streptavidin beads. This approach uses a combination of RNase A and T1 (see Chapter II), in order to degrade the RNA baits into smaller components. Only proteins bound to the baits would be present in the eluate, and not proteins bound to the beads, thus producing a more specific eluate (Fig. 4.11). Clearly, this specific elution method would increase the sensitivity of the experiment, allowing for the more confident detection of m^5C-binding proteins.

4.2.2.2 Candidate m^5C 'readers'

Following pulldown of the RNA baits, eluates were analysed on a SDS-PAGE gel. It was expected that differences in band patterns between the sample and control bait profiles would constitute m^5C-binding proteins, and would require subsequent follow-up work. Bands were sliced from the SDS-PAGE gel and processed by mass spectrometry (at the Bioanalytical Mass Spectrometry Facility, UNSW) for protein identification.

It was expected that NSUN2 would recognise the unmethylated bait, as this is modelled on a validated NSUN2 target. This would act as a positive control to ensure the chosen RNA bait resembled a relevant, correctly folded NSUN2 target, whilst also indicating that the chosen experimental conditions allowed for protein-RNA interactions to form. Thus, in order to establish whether this was the case, total cell lysates were incubated with the baits, and the eluates were run a SDS-PAGE gel and processed by western blotting to determine the presence or absence of NSUN2 with each of the baits. Western blotting revealed that NSUN2 indeed was associated with the unmethylated RNA bait (Fig. 4.12, top panel). Interestingly, NSUN2 was also observed in the methylated bait eluate, suggesting that NSUN2 might have the capacity to function as a 'reader' of m^5C, in addition to its known function as a 'writer' of m^5C in RNA. The western blot was also probed with the RNA binding protein GAPDH, to serve as a negative control. GAPDH was not detected in either methylated or unmethylated bait eluates (Fig. 4.12, bottom panel), suggesting specificity in the experimental setup.

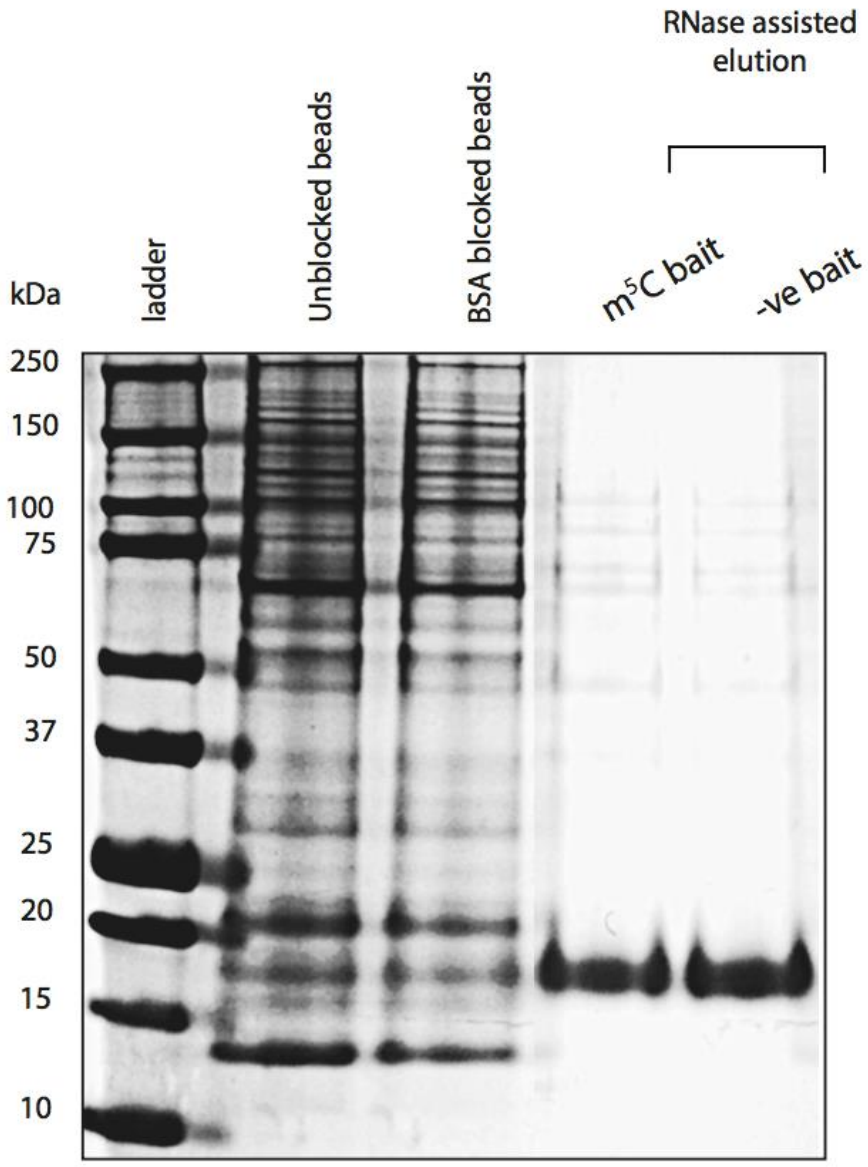


Figure 4.11: Optimisation of the RNA pulldown experiment.

Incubating the unmethylated bait–protein lysate mix with either unblocked beads or BSA blocked beads resulted in similar protein profiles (lanes 2 and 3 respectively). Incubating unblocked beads with the RNA-protein lysate followed by RNase assisted elution resulted in only proteins bound to the RNA bait being eluted (lanes 4 and 5). Non-specific protein binding is evident when using either blocked or unblocked beads (lanes 2 and 3). Coomassie Blue stain was used to visualise the gel. The image of the gel was taken using a LAS4000.

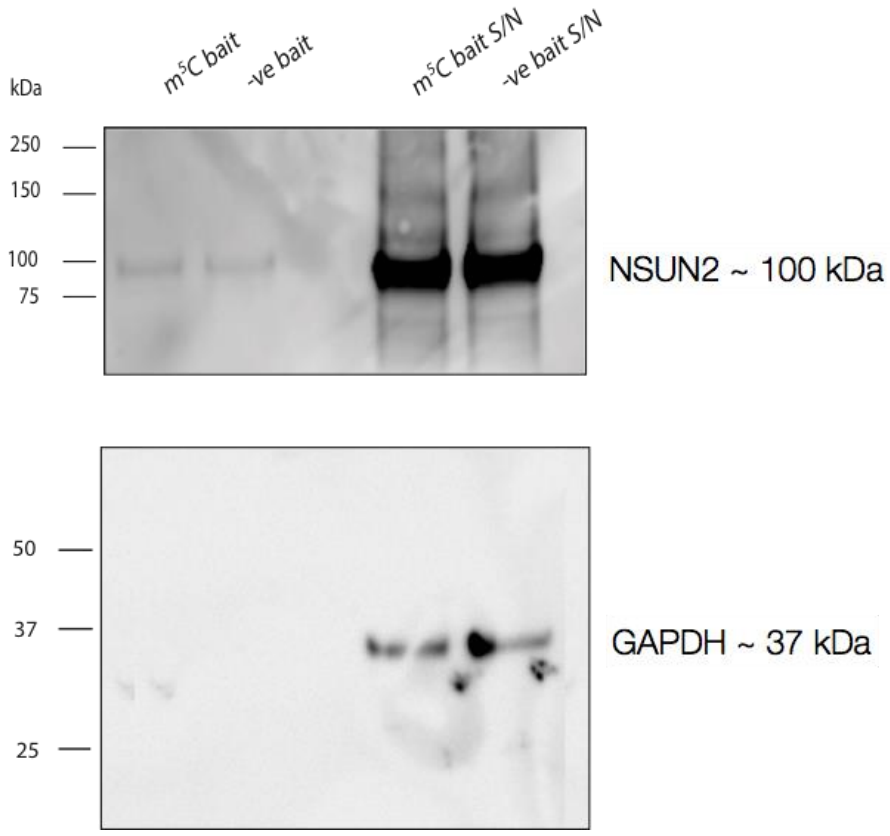


Figure 4.12: NSUN2 interacts with both baits, GAPDH does not.

Both RNA baits are incubated with wild type protein lysate. This blot is representative blot of n = 2 experiments.

Top panel: RNA bait pulldown eluates probed with NSUN2. NSUN2 is present in the methylated and control bait eluates. 100% of eluates loaded; 10% of supernatants loaded. The band at approximately 100 kDa corresponds to NSUN2.

Bottom panel: RNA bait pulldown eluates probed with GAPDH. The RNA binding protein GAPDH is not present in the methylated and control bait eluates, suggesting specificity in the experimental design. 100% of eluates loaded; 10% of supernatants loaded. The band at approximately 37 kDa corresponds to GAPDH.

m^5C bait = methylated bait eluate; -ve bait = unmethylated bait (control) eluate; S/N = supernatant (flowthrough).

A number of RBPs were identified following mass spectrometry of the sample and control bait eluates (a full list of identified peptides/proteins is shown in Appendix 7.3). Two experimental conditions were trialed: 1) incubating both baits with a wild type protein lysate, or 2) incubating both baits with a lysate void of NSUN2 (see Discussion). Figure 4.13 shows a scatter plot of the number of peptides identified with methylated and unmethylated baits when incubating with the NSUN2 knockdown lysate. Appendix 7.4 shows the comparable plot when incubating the baits with wild type lysate. The number of peptides associated with the baits, is used as a measure of protein binding strength to the RNA baits; the greater the number of peptides identified, the stronger the binding to the bait. To determine the proteins that are 'enriched' with the methylated bait, an odds ratio was calculated. This was achieved by calculating the ratio of peptides associated with methylated to unmethylated bait for each protein, and comparing that with the same ratio calculated for all the other identified proteins. The Fisher's exact test was used to test for statistical significance. This analysis was completed using the Scaffold4 software and the output is graphed (Fig. 4.13). The majority of identified peptides bind the RNA non-specifically (Fig. 4.13).

Tables 4.1 and 4.2 show the number of peptides that are enriched with the methylated bait compared to the negative control baits identified from the mass spectrometry analysis. Table 4.1 is based on incubating both baits with the wild type protein lysate (n=1). Table 4.2 is based on incubation of both baits with the NSUN2 knockdown lysate (n=2).

The Radixin/Ezrin proteins (~70 kDa) were identified as potential m^5C-binding proteins, as they were prominently associated with the methylated bait compared to the unmethylated bait in three replicates; this result was validated by western blotting (Fig. 4.14).

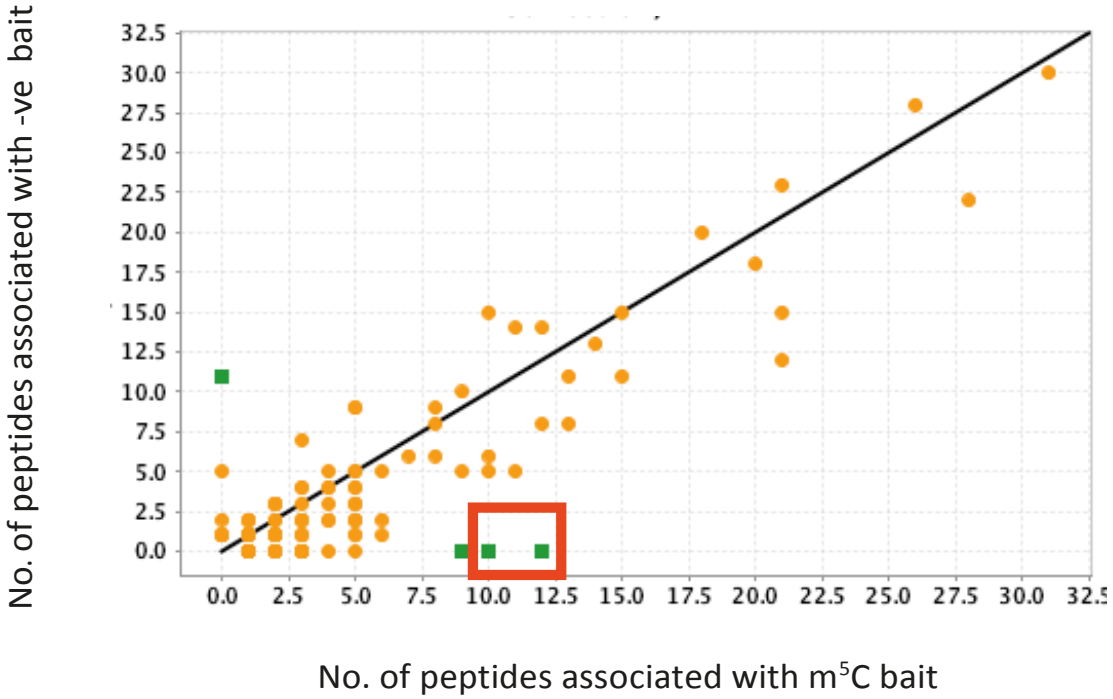


Figure 4.13: Scatterplot showing number of identified peptides associated with each of the baits using the NSUN2 knockdown lysate.

Numbers of peptides associated with each bait from the mass spectrometry analysis is plotted. The x-axis shows the number of peptides associated with the m^5C bait, and the y-axis shows the number of peptides associated with the negative control bait. Proteins specific to the m^5C bait fall on the x-axis and protein specific to the negative control bait fall on the y-axis. Proteins showing an insignificant odds ratio are coloured orange. Proteins with a significant odds ratio are coloured green. Fisher's exact test was applied to determine any significant changes between the odds ratio for a given peptide, compared to the odds ratios of all the other detected peptides. Significance was determined at a p-value ≤ 0.01. The statistical test was performed using Scaffold4 mass spectrometry analysis software. This plot is representative of n = 2 experiments.

Table 4.1: Differentially identified proteins showing statistical significance at p-value < 0.01 when incubating both baits with wild type protein lysate.

Gene symbol	Total # of peptides with m^5C bait	Total # of peptides with control bait
RDX	10	0
EZR	12	0

Table 4.2: Differentially identified proteins showing statistical significance at p-value < 0.01 when incubating both baits with NSUN2 knockdown lysate.

Gene symbol	Total # of peptides with m^5C bait		Total # of peptides with control bait	
	Rep1	Rep2	Rep1	Rep2
RDX	11	11	0	0
EZR	10	13	5	5
ROA2	10	13	0	5

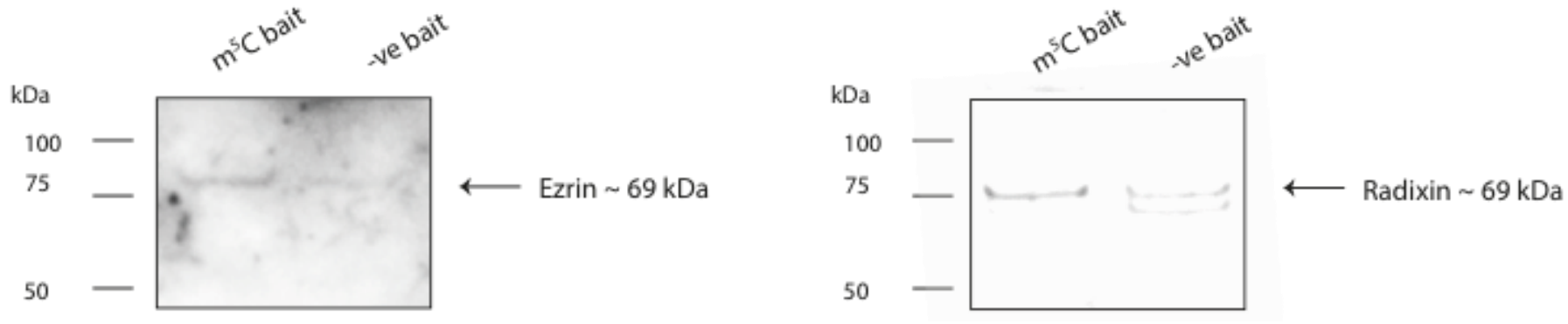


Figure 4.14: Western blotting confirming mass spectrometry results.

Protein eluates from the m^5C bait and negative control baits were probed for Ezrin and Radixin. Both blots show these proteins to be more associated with the m^5C bait than the negative control bait, validating the mass spectrometry result. These are representative blots of n = 2 experiments.

4.3 Discussion

This Chapter provided evidence that the m^5C modification in mRNA may be involved in regulating mRNA stability and mRNA translation. Possible mechanisms describing the role for m^5C in these two contexts is provided below. m^5C function is expected to be mediated by proteins that are associated with m^5C. This Chapter revealed possible m^5C 'reader' proteins and m^5C-associated proteins identified through both experimental and bioinformatics approaches respectively.

4.3.1 AGO2 and m^5C sites

The observation that AGO2 footprints associate just upstream of candidate CDS and 3′ UTR m^5C sites suggests m^5C may function in the recruitment or blocking of AGO2; this would either result in the destabilisation or stabilisation of the mRNA. Possible mechanisms include: 1) the m^5C mark promotes mRNA stability by blocking AGO2 binding. The reverse situation where AGO2 binding blocks methylation is also feasible; 2) the m^5C modification could act as a signal for mRNA degradation by removing secondary structure elements, allowing AGO2 to bind the mRNA. The reverse may also be true in that AGO2 binding may allow NSUN2 access to the mRNA. To determine if m^5C affects stability, the expression of a selection of m^5C-modified mRNAs in both NSUN2 KD and wild type samples was measured previously (Sibbritt 2015, PhD thesis). However, no change in mRNA abundance was observed.

Another possibility to explain the AGO2-m^5C association is that miRNA binding may regulate the formation of the m^5C mark in a subset of cases. For example, AGO2 interacts with NSUN2 to direct the methyltransferase to its binding site through miRNA-mRNA base pairing. Presently, the co-IP experiment suggests this is not the case. On the same note, Chen *et al.* (2015) also explored such a notion in regards to the m^6A modification. They report that miRNAs function in regulating m^6A formation by modulating binding of the m^6A methyltransferase METTL3 to mRNAs. In this way a proportion of m^6A sites are selected. Interestingly, though, Chen *et al*. (2015) show that none of the AGO 1-4 proteins were involved in the miRNA-mRNA mediated m^6A regulation. Thus, it is likely that miRNAs associate with proteins other than AGO to regulate m^6A formation. Given the presence of a large number of RNA binding proteins with unknown functions, determining whether other miRNA binding proteins are involved in the m^6A/m^5C modification remains challenging and requires further

investigation.

Presently, overlapping candidate m^5C sites, categorised by methylation percent, with AGO2 footprints has revealed some functional insight into the observed association. A statistically significant strong negative correlation of AGO2 footprint density was observed with highly methylated sites in the 3′ UTR. This suggests that m^5C in the 3′ UTR may function to block AGO2 binding, ultimately stabilising the mRNA. Conversely, m^6A methylation has been implicated in mRNA destabilisation. Wang, Lu, *et al.* (2014) observe an increased HuR binding to demethylated mRNA extracted from METTL3 knockdown cells when compared to control cells. They suggest that HuR binding acts to increase RNA stability by blocking AGO/miRNA targeting. Upon METTL3 knockdown, they report a ~25-40% decrease in AGO2 binding to a METTL3 target, Igfbp3 3′ UTR (which also contains a HuR binding motif) but little change to Pou5f1 mRNA (which lacks m^6A). They interpret this data by suggesting that the m^6A methylation blocks HuR binding ultimately resulting in the destabilisation of the mRNA.

Overall it seems m^5C may have a role in the miRNA mechanism and more broadly in the post-transcriptional control of gene expression, with a possible role in regulating RNA stability. Supporting this argument, NSUN2 was recently shown to promote the stability of the mRNA encoding the cell cycle regulator CDKN2A/$p16^{INK4}$ (Zhang *et al.*, 2012). *In vitro* methylation assays indicated that recombinant NSUN2 targeted a specific AAC motif in the 3′ UTR region of $p16_{INK4}$ mRNA. Intriguingly, the presence of this motif suggests involvement of m^6A methylation, however this contradicts the known specificity of NSUN2 for m^5C. $p16_{INK4}$ mRNA is known to be destabilised by interactions of AU-rich element RNA binding protein 1 (AUF1), HuR and AGO2 with its 3′ UTR. Zhang *et al.* (2012) showed that NSUN2-mediated methylation prevented the localisation of $p16^{INK4}$ 3′ UTR containing transcripts to processing bodies, as well as the destabilising affects of HuR/AUF1/AGO2 on $p16^{INK4}$ mRNA. Despite this study suggesting that NSUN2 mediates the m^6A modification, the authors describe a particular scenario where mRNA methylation by NSUN2 promotes mRNA stability by preventing the association of $p16_{INK4}$ 3′ UTR with regulatory RBPs.

4.3.2 m^5C binding proteins

4.3.2.1 RNA bait experiment

In this Chapter, an *in vitro* RNA pulldown approach was employed to identify m^5C-binding proteins. This technique involves the mass spectrometry identification of proteins bound to biotinylated RNA baits (with and without the m^5C mark). These baits are chemically modified by the addition of biotin to the 5′ end of the synthetic oligonucleotide. Biotin allows these baits to be captured using a streptavidin antibody following incubation with a protein lysate. This experimental framework is modelled on work by Dominissini *et al.* (2012): they use m^6A baits to characterise m^6A-binding proteins. In a similar fashion, it is expected that proteins could be identified that differentially bind to the m^5C methylated bait. However such an experiment has its drawbacks. For example, the unmethlyated bait may become methylated, and therefore differential RBP binding would no longer be observed between the two bait conditions. Also as m^5C in RNA is thought be a dynamic modification, the reverse is true for the methylated bait. That is, it is also possible that the methylated bait may become unmethylated. To address this apparent complication with the experiment, the baits were incubated with a protein lysate void of NSUN2 in an attempt to ensure the unmethylated bait remained so. Indeed, this approach identified a handful of possible RNA-binding proteins that may specifically recognise m^5C. In doing so these 'readers' may affect the fate of the bound mRNAs, ultimately influencing the outcome of specific cellular processes. It is also important to note that the chosen experimental design allows for the isolation of protein-protein complexes, and therefore proteins not directly interacting with m^5C may also be identified.

4.3.2.2 NSUN2 is a possible m^5C 'reader'

Initially, incubating the baits with total protein lysates revealed that NSUN2 was associated with both RNA baits. This confirmed that the chosen 40 nt long bait and experimental parameters were similar to physiological conditions, allowing RNA-protein interactions to form. However, this suggested that no major difference in RBP profiles between the two cases might be observed. Also, the presence of NSUN2 with the methylated bait can be interpreted through one of two possible scenarios: 1) m^5C is being actively reverted back to cytosine by a demethylase, followed by re-methylation by NSUN2 or 2) NSUN2 may act as a 'reader' protein, not only generating the methylation but also binding to it.

Interestingly, (Moon & Redman, 2014) have previously shown *in vitro* that NSUN2 is

able to form a complex with m^5C modified Aspartyl tRNA. They provide evidence that NSUN2-RNA complex formation requires S-adenosylhomocysteine (AdoHcy) and its removal results in the dissociation of the complex. The authors also show that complex formation is dependent upon AdoHcy concentration and solution pH, where optimal complex formation was achieved in a pH range of 5.5-6.5. However the authors do no provide any insight into the possible function of NSUN2 as a 'reader' protein or provide any information into the m^5C targets that are recognised by NSUN2. This notion that NSUN2 may function as both a 'writer' and 'reader' provides an interesting context into m^5C function. For example, determining the cellular compartment in which NSUN2 functions as a 'reader' will provide a greater insight into m^5C function. NSUN2 is normally localised to the nucleus, presumably where the RNAs are modified. Thus, NSUN2 may then bind the modification following 'writing' in the nucleus and mediate a function. For example, this may suggest a role for m^5C in splicing. Another possibility is that the NSUN2-RNA complex is exported from the nucleus into the cytoplasm where NSUN2 carries out the m^5C function. It is also possible that the modified mRNA is exported out of the nucleus where cytoplasmic NSUN2 may function as a 'reader'. This would be interesting as NSUN2 is prominent in the nucleus during interphase, however is present in the cytoplasm during mitosis (Sakita-Suto *et al.*, 2007). Recognition of m^5C-containing mRNAs by cytoplasmic NSUN2, during mitosis, may suggest a role for m^5C in development. In any case, m^5C sites recognised by NSUN2 may likely be in mRNAs that encode for proteins involved in an NSUN2-related pathway.

Recently, Lin *et al.* (2016) have provided evidence that suggested METTL3, the m^6A 'writer', also functions as an m^6A 'reader'. METTL3 is enzymatically active in complex with METTL14 and the splicing regulator WTAP, which localises the complex to nuclear speckles. By contrast, the authors show that METTL3, but not METTL14 or WTAP, is also found in the cytoplasm of human lung cancer cell lines. METTL3 knockdown reduced mRNA methylation with little effect on their abundance. Nevertheless, the levels of proteins encoded by a subset of m^6A-containing transcripts were reduced, prompting an investigation of METTL3's role in mRNA translation. Domain deletion analyses then showed that the METTL3 N-terminus was required for the enhanced translation of a subset of m^6A-containing mRNAs, and not its C-terminus (catalytic activity domain). Finally, METTL3 was found to co-sediment with polysomes and to interact with the translation initiation machinery, including the cap-

binding proteins eIF4E and CBP80 and the small-ribosomal subunit-associated multi-subunit complex eIF3. The authors suggested that METTL3 recruits eIF3 to a subset of m^6A-containing mRNAs to enhance their translation.

4.3.2.3 Ezrin and Radixin are possible m^5C 'readers'

The RNA bait pulldown experiments yielded a couple of possible candidate m^5C 'reader' proteins. Both Radixin and Ezrin are more prominently associated with the methylated bait than the unmethylated bait (Fig. 4.14). Radixin and Ezrin are members of the ERM (Ezrin-Radixin-Moesin) protein family; the three proteins have high sequence similarity and are known to function in binding actin filaments to the plasma membrane (Fehon *et al.*, 2010). Interestingly, previous studies have demonstrated that cytoskeletal dynamics and translation may be co-regulated (Kim & Coulombe, 2010; Polak *et al.*, 2006). Furthermore, Briggs *et al.* (2012) were the first to demonstrate that ERM proteins are associated with the elevated expression of translation machinery components e.g. eIF4E in cancer cells. The authors also performed polysome gradients and showed that the ERM proteins are detected in free mRNP fractions and fractions corresponding to the 40S, 60S and 80S particles. More specifically Briggs *et al.* (2012) also provide evidence indicating that a proportion of Ezrin can associate with the translation ribonucleoprotein complex, suggesting that Ezrin plays a role in stabilising the translational machinery with the mRNA during translation initiation. Importantly, they show Ezrin is able to directly interact with poly(A)-binding protein 1 (PABP1) and affects mRNA translation of certain mRNAs, however does not participate in active mRNA translation. This interaction between Ezrin and PABP1 may stabilise the mRNA and increase translation through a closed loop model. As Radixin/Ezrin have been identified as potential m^5C-binding proteins, in this thesis, it is possible that they may enhance translation of a subset of m^5C-containing mRNAs.

Also, Radixin has been identified to interact with NSUN2, in high-throughput interactome type experiments. Notably, these experiments also suggest that NSUN2 interacts with eIF3, and Radixin complexes with eEF2 (Chatr-Aryamontri *et al.*, 2015). It is possible then that Radixin/NSUN2 recognises the m^5C mark in a subset of mRNAs, enhancing mRNA translation through an interaction with the translation machinery. From the same interactome studies, Radixin is also suspected of binding HuR, a protein that stabilises mRNA. In this way, Radixin may bind to 3′ UTR m^5C sites and direct HuR binding, ultimately affecting mRNA half-life.

4.3.2.4 Ezrin/Radixin– a role in promoting translation of m^5C-containing mRNAs

As 5′ UTR localised m^5C sites tend to reside in structured regions, resulting in weak mRNA translation, it is conceivable that the m^5C mark functions in enhancing translation of these mRNAs.

As eIF4E binding to the 5′ cap of mRNA is considered to be the rate-limiting step in protein synthesis, increased eIF4E expression could result in the enhanced translation of normally 'weakly translated mRNAs'. Such mRNAs tend to be implicated in oncogenic processes such as cell proliferation, evasion of apoptosis, angiogenesis, and metastasis (Konicek *et al.*, 2008; Silvera *et al.*, 2010). Indeed, increased expression of eIF4E has been shown to enhance translation of mRNAs with a complex, highly structured 5′ UTR (Graff & Zimmer, 2003). Following from this, Briggs *et al*. (2012) specifically show that Ezrin, through an as yet unknown mechanism, results in the enhanced translation of an *in vitro* transcribed RNA containing a 24-nt stem hairpin, resembling a 'weakly translated mRNA'. Consistent with this observation, Ezrin was shown not to affect global protein synthesis at large, but is suggested to affect the translation efficiency of a subset of highly structured mRNAs.

It has been hypothesised that 'weakly translated mRNAs', which possess complex 5′ and 3′ UTRs, are generally maintained as stable mRNAs (Hsieh & Ruggero, 2010). However upon certain cues, the translational machinery is activated resulting in the translation of these complex mRNAs. As it is energetically demanding for the cell to express all proteins constitutively, it is thought that these complex mRNAs undergo translation only under certain circumstances (Livingstone *et al.*, 2010). Furthermore, Hsieh and Ruggero (2010) note this model of translation regulation applies to several oncogenes as examples of 'weakly translated mRNAs'. Interestingly, important oncogenes such as YAP1, FYN contain m^5C in their 5≠UTRs. Furthermore, NSUN2 is expressed at low levels in normal tissues, but is overexpressed in a range of human and mice tumor types including squamous cell carcinoma, colorectal cancer and breast cancer (Fyre & Watt, 2006), whilst Ezrin and Radixin are also upregulated in many cancers. Thus, it is enticing to speculate that in cancer, the increased expression of NSUN2, could lead to the stoichiometric methylation of an oncogenic target, whilst the increased Ezrin/Radixin expression could result in the enhanced translation of the m^5C-containing oncogene, which would normally be characterised by weak translation.

4.4 Future Directions

The Chapter has explored the possibility of RNA binding proteins that specifically recognise m^5C in mRNAs, revealing two possible m^5C 'reader' candidates. Further work is required to determine 1) the specificity of these proteins, 2) the function of these proteins and 3) the mechanism of action.

It is interesting that NSUN2 is responsible that for the synthesis of m^5C in may contexts including tRNAs, different regions of mRNAs, ncRNAs, and therefore m^5C is expected to have diverse roles dependent upon its locality. It is essential to determine which m^5C-binding proteins target which m^5C sites and how this relates to function.

Further work is required to better understand how 'writers', 'readers', and 'erasers' recognise specific m^5C sites. For example, as mentioned earlier, different RBPs may bind m^5C in different locations. How is selectively for particular sites achieved by the different m^5C-binding proteins? Will a 'reader' always affect the same outcome upon mRNA binding, or might this also be context dependent, e.g., on the location of m^5C within mRNA and/or on what other RNA-binding proteins make up the mRNP complex? What are the structural or sequence determinants on the mRNA, and how might recognition be influenced by the sub-cellular context? Are there other proteins in complex with the m^5C-binding proteins that direct the 'readers' to a particular location within the mRNA? Progress in the space will reveal further insight into the function of the m^5C post-transcriptional modifications.

Chapter V

Metagene analyses

5.1 Preamble

The presence of m^5C has long been documented in tRNAs and rRNAs. In tRNAs, m^5C methylation has been reported to affect tRNA stability and protein synthesis. For example, it has been noted by Blanco *et al.* (2014) that the depletion of NSUN2 resulted in the accumulation of 5′ tRNA fragments caused by angiogenin-cleavage. Also, TRDMT1 mediated C38-methylated aspartyl tRNA has been linked to preferential tRNA charging and increased protein synthesis of poly (Asp) sequences. In line with this observation, a TRDMT1 knockout resulted in a 30% reduced charging level of aspartyl tRNA (Shanmugam *et al.*, 2015). In rRNAs, m^5C has been implicated in ribosome biogenesis, ultimately affecting protein synthesis (Metodiev *et al.*, 2014).

It is only recently, thanks to advancements in NGS technology, that m^5C has also been identified in mRNAs; it is important to understand the effects of m^5C on mRNA metabolism. For example, m^5C may affect mRNA stability and/or mRNA translation efficiencies. Furthermore, as RBPs can regulate these processes, it is important to understand the impacts, if any, of m^5C on the binding abilities of RBPs to these molecules.

Information on the function of m^5C in mRNAs is slowly emerging. Indeed, the m^5C modification may exert multiple functions, similar to that observed for N6-methyladenosine (m^6A) (e.g. as reviewed in Schumann *et al.*, 2016). Three independent studies have indicated a role of m^5C in translation using a candidate approach (discussed below). Cumulatively, those works indicate that m^5C can both negatively and positively affect mRNA translation depending on m^5C localisation. For example Hoernes *et al.* (2016) employed an *in vitro* translation assay revealing a 40% reduction in translation upon the introduction of m^5C methylation in bacterial mRNA. However as m^5C has not been detected in bacterial mRNAs as yet, this study may not be physiologically relevant, although it shows the potential scope for further work in defining m^5C function. The second study is by Xing *et al.* (2015) who showed that NSUN2 methylation of *CDK1* mRNA at the 3≠UTR results in increased translation of *CDK1*. Another related study showed that NSUN2 mediated 5≠UTR methylation of *p27* mRNA at cytosine C64 *in vitro* and *ex vivo* (in cells), resulted in the translation repression of p27 (Tang *et al.*, 2015).

In order to study the consequences of the m^5C modification on mRNA regulation/function, there should be a focus on determining a comprehensive, high confidence set of m^5C sites present in mRNAs. Traditionally, bisulfite treatment has been used to discriminate between m^5C and cytosine residues in DNA. This method has been adapted for use with RNA (Schaefer, Pollex, *et al.*, 2009). Bisulfite treatment is considered the gold standard for the detection of m^5C sites (Frye *et al.*, 2016). When coupled with next generation sequencing in a technique known as bsRNAseq, m^5C sites can be identified at single nucleotide resolution throughout the transcriptome. Methylation stoichiometry, that is, the proportion of methylated to unmethylated transcripts, can also be obtained. m^5C methylated cytosines can be differentiated from unmethylated cytosines as the former are resistant to the bisulfite treatment, whereas the latter undergo a chemical transformation, ultimately being converted to uracil. This then allows m^5C sites to be detected by comparing bsRNAseq reads to a reference. In the present study a transcriptome-wide map of m^5C sites was obtained by coupling bisulfite treatment of rRNA depleted HeLa RNA with Illumina NGS.

As there is relatively little known regarding the function of m^5C in mRNAs, in this thesis, there was precedence to investigate m^5C function on a global level, rather than focusing on the function of m^5C in a particular candidate. Such a metagene analysis provides an average view of m^5C methylation across a genomic feature such as an mRNA region, particular sequence and/or structure motifs. These analyses often require a large sample size for increased statistical power. However, as the number of identified m^5C sites in the Illumina bsRNA-seq experiment is small, any observed global features of m^5C methylation will likely have a large effect size, and be real. This would provide an initial insight into m^5C function in mRNA. On the other hand, the advantage of a candidate approach is its ability to specifically show a function for m^5C, although often choosing a suitable or 'representative' candidate m^5C-containing mRNA could be challenging.

This Chapter presents the findings from these metagene analyses, which suggested an association between m^5C methylation with GC rich regions, low minimum free energy structures and weakly translated mRNAs.

5.2. Results

5.2.1 Summary of the Illumina bsRNA-seq HeLa cell datasets

This Chapter is based on the analysis of high confidence m^5C candidate sites detected in the Illumina bsRNA-seq data.

Library preparation and sequencing of the Illumina libraries was conducted by the Beijing Genomics Institute (BGI). Sequencing library processing, read mapping and site identification was performed by Dr Maurits Evers (see Acknowledgements). Together with Dr Maurits Evers, I conceived of the different metagene analyses, which have lead to the development of the publicly available R package RNAModR (manuscript under review) (an example R code is provided in Appendix 7.5).

Two bsRNA-seq Illumina replicates were generated. Overall there was good agreement between both Illumina datasets (HeLa 1 and HeLa 2) in coverage and m^5C site identification. Both libraries were subjected to rRNA depletion followed by bisulfite treatment. Before quality control steps were performed, HeLa 1 and HeLa 2 libraries contained 138,303,098 and 129,966,290 reads respectively (Table 5.1). Quality control steps include adaptor trimming, removal of flanking nucleotides that have a Phred score less than 20, and discarding of reads that were ≤36 nt in length. This reduces the number of reads to 135,111,738 and 128,597,339 for the HeLa 1 and HeLa 2 libraries respectively (Table 5.1). Alignment of the post-quality control reads resulted in 8,878,962 (6.57 %) and 12,995,324 (10.11 %) uniquely aligned reads for the HeLa 1 and HeLa 2 libraries respectively (Table 5.2). These low alignment rates are not unusual due to the reduced base complexity (resulting from bisulfite treatment). Also, reads derived from tRNAs will likely map to multiple regions and be discarded (see below). Altogether these factors account for the small number of mapped reads in this thesis.

The alignment was performed using Bismark in conjunction with Bowtie2, allowing for one mismatch per a 20 nt seed. Bismark was originally developed to specifically map bisulfite-converted sequence reads (Krueger & Andrews, 2011), to identify methylated cytosines in DNA within CpG, CHG and CHH contexts (where H = A, T or C). Later Bismark was extended to identify m^5C sites without the constraints of a specific context (i.e. without CpG bias) (the bash script for the alignment of reads is provided in Appendix 7.6).

Table 5.1: Read statistics of HeLa 1 and 2 bsRNA-seq libraries.

Sample	Read length	No. of reads (pre-QC)	No. of reads (post QC)	% of reads retained
HeLa 1	49	138,303,098	135,111,738	97.69
HeLa 2	49	129,966,290	128,597,339	98.95

Table 5.2: Mapping statistics of HeLa 1 and 2 bsRNA-seq libraries.

Sample	Sequences analysed	Unique alignments	% of reads mapped	Number of unmappable reads	Non-unique alignments
HeLa 1	135,111,738	8,878,962	6.57	66,982,907	59,249,869
HeLa 2	128,597,339	12,995,324	10.11	31,355,266	84,246,749

Following alignment, the two HeLa libraries were pooled and methylated cytosines were identified according to the following parameters: 1) a minimum of 10 reads, 2) >10% methylation and, 3) overlapping with annotated features (e.g. mRNA exons, lncRNAs, miRNAs, snoRNAs, etc).

Reads were mapped to an *in silico* bisulfite-converted hg38 genome, yielding a total of 847 gene-associated m^5C methylation sites. For the rest of this work, these sites are referred to as the set of high confidence m^5C sites.

5.2.2 Extent of m^5C modification in the human transcriptome

As expected, most of the candidate sites were identified in tRNAs, with most eukaryotic tRNAs exhibiting more than one m^5C site. To identify sites in tRNAs, the reads were mapped to an *in silico* bisulfite converted human GRCh38 Ensembl reference. However, as identical copies of tRNA genes are present in numerous locations throughout the human genome, this means that tRNA reads would map with equal fidelity to several different locations throughout the genome (multi-mapping) and ultimately would be discarded. Thus, to overcome this problem, a customised reference was generated containing a single representative sequence for each tRNA isoacceptor. Initially, tRNAs were mapped genome-wide to all tRNA annotations from the genomic tRNA database (Chan & Lowe, 2009), without the CCA addition. A non-templated conserved CCA sequence is added to the 3' end of all mature tRNA molecules. This addition serves as the site of amino acid attachment. This so called 'global mapping' approach, resulted in the detection of previously reported m^5C sites as documented in the literature. Thus, m^5C sites were identified at the structural positions C34, C48, C49 and C50 of many tRNA molecules (Fig. 5.1); these sites are known to be NSUN2 mediated (Edelheit *et al.*, 2013; Motorin *et al.*, 2010; Khoddami & Cairns, 2013; Hussain *et al.*, 2013). C38 m^5C methylation in $tRNA^{Asp}$, a confirmed TRDMT1 site (Goll *et al.*, 2006) was also observed in the present study. However to achieve a more in-depth view of candidate m^5C sites in tRNAs, a strategy to map tRNAs with the *in silico* CCA addition (local tRNA mapping approach) was employed. That is, an *in silico* non-template 3′ CCA addition was added to the reference sequences used. This specifically allows for the mapping of mature tRNA sequences. Using this approach, methylation at C72 in $tRNA^{Thr}$ (Fig. 5.1) was detected. This m^5C site has also been observed in other studies (see e.g. Hussain *et al.*, 2013; Khoddami & Cairns, 2013), however it was only recently that Haag *et al.* (2015) showed that m^5C at C72 in $tRNA^{Thr}$ is mediated by NSUN6.

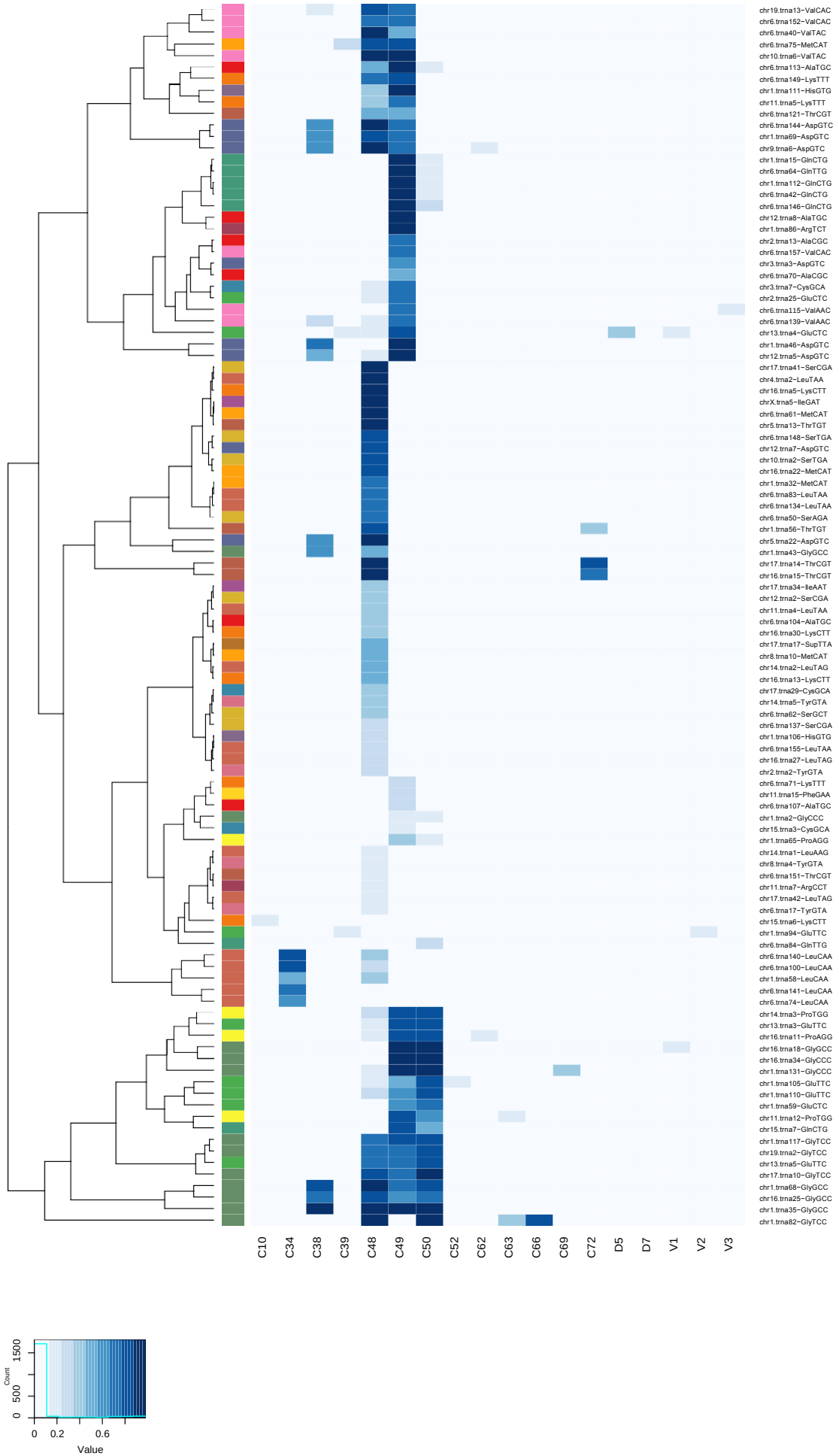

chr19.trna13−ValCAC
chr6.trna152−ValCAC
chr6.trna40−ValTAC
chr6.trna75−MetCAT
chr10.trna6−ValTAC
chr6.trna113−AlaTGC
chr6.trna149−LysTTT
chr1.trna111−HisGTG
chr11.trna5−LysTTT
chr6.trna121−ThrCGT
chr6.trna144−AspGTC
chr1.trna69−AspGTC
chr9.trna6−AspGTC
chr1.trna15−GlnCTG
chr6.trna64−GlnTTG
chr1.trna112−GlnCTG
chr6.trna42−GlnCTG
chr6.trna146−GlnCTG
chr12.trna8−AlaTGC
chr1.trna86−ArgTCT
chr2.trna13−AlaCGC
chr6.trna157−ValCAC
chr3.trna3−AspGTC
chr6.trna70−AlaCGC
chr3.trna7−CysGCA
chr2.trna25−GluCTC
chr6.trna115−ValAAC
chr6.trna139−ValAAC
chr13.trna4−GluCTC
chr1.trna46−AspGTC
chr12.trna5−AspGTC
chr17.trna41−SerCGA
chr4.trna2−LeuTAA
chr16.trna5−LysCTT
chrX.trna5−IleGAT
chr6.trna61−MetCAT
chr5.trna13−ThrTGT
chr6.trna148−SerTGA
chr12.trna7−AspGTC
chr10.trna2−SerTGA
chr16.trna22−MetCAT
chr1.trna32−MetCAT
chr6.trna83−LeuTAA
chr6.trna134−LeuTAA
chr6.trna50−SerAGA
chr1.trna56−ThrTGT
chr5.trna22−AspGTC
chr1.trna43−GlyGCC
chr17.trna14−ThrCGT
chr16.trna15−ThrCGT
chr17.trna34−IleAAT
chr12.trna2−SerCGA
chr11.trna4−LeuTAA
chr6.trna104−AlaTGC
chr16.trna30−LysCTT
chr17.trna17−SupTTA
chr8.trna10−MetCAT
chr14.trna2−LeuTAG
chr16.trna13−LysCTT
chr17.trna29−CysGCA
chr14.trna5−TyrGTA
chr6.trna62−SerGCT
chr6.trna137−SerCGA
chr1.trna106−HisGTG
chr6.trna155−LeuTAA
chr16.trna27−LeuTAG
chr2.trna2−TyrGTA
chr6.trna71−LysTTT
chr11.trna15−PheGAA
chr6.trna107−AlaTGC
chr1.trna2−GlyCCC
chr15.trna3−CysGCA
chr1.trna65−ProAGG
chr14.trna1−LeuAAG
chr8.trna4−TyrGTA
chr6.trna151−ThrCGT
chr11.trna7−ArgCCT
chr17.trna42−LeuTAG
chr6.trna17−TyrGTA
chr15.trna6−LysCTT
chr1.trna94−GluTTC
chr6.trna84−GlnTTG
chr6.trna140−LeuCAA
chr6.trna100−LeuCAA
chr1.trna58−LeuCAA
chr6.trna141−LeuCAA
chr6.trna74−LeuCAA
chr14.trna3−ProTGG
chr13.trna3−GluTTC
chr16.trna11−ProAGG
chr16.trna18−GlyGCC
chr16.trna34−GlyCCC
chr1.trna131−GlyCCC
chr1.trna105−GluTTC
chr1.trna110−GluTTC
chr1.trna59−GluCTC
chr11.trna12−ProTGG
chr15.trna7−GlnCTG
chr1.trna117−GlyTCC
chr19.trna2−GlyTCC
chr13.trna5−GluTTC
chr17.trna10−GlyTCC
chr1.trna68−GlyGCC
chr16.trna25−GlyGCC
chr1.trna35−GlyGCC
chr1.trna82−GlyTCC
C10
C34
C38
C39
C48
C49
C50
C52
C62
C63
C66
C69
C72
D5
D7
V1
V2
V3
Count
1500
500
0
0
0.2
0.6
Value

Figure 5.1: m^5C sites identified in tRNAs.

Each row represents a single tRNA isoacceptor (also indicated by the colouring of the first column), whilst every other column represents the structural positions of cytosines in the tRNA as labelled. The scale in the bottom left hand corner of the image defines the extent of methylation stoichiometry; the darker blue colour is indicative of increasing methylation rate.

High methylation transcript stoichiometry is observed for many tRNAs at the known structural positions C48, C49 and C50. High levels of methylation stoichiometry are also detected at the lesser-defined positions C34, C38 and C72 in various tRNAs.

D5, D7 refer to positions in the D-arm of tRNAs; V1, V2 and V3 refer to positions in the variable region of tRNAs.

Other than these aforementioned sites, no other m^5C sites within nuclear-encoded tRNAs were detected, though candidate m^5C sites within mitochondrial tRNAs were identified (see Chapter III). Overall, these findings confirm previously published work (mentioned above), and thus these tRNA m^5C sites are not included in any subsequent analyses and discussions.

A small number (25) of m^5C sites were further identified in lncRNAs, snRNAs and other ncRNAs.

As the major focus of the current work is to understand the role of m^5C in mRNAs, it was preferable to determine the number of m^5C sites per mRNA segment. As many mRNA transcript variants exist, m^5C sites were mapped to a customised transcriptome. This customised library was compiled by choosing a canonical representative mRNA transcript for each gene, annotated with 5′ UTR, CDS and 3′ UTR designations based on the most recent NCBI RefSeq sequence and annotation database. More specifically, multiple transcript annotation entries per gene were collapsed into one comprising of the longest CDS and the corresponding longest UTRs. The representative transcript for each gene in the complied library may not be the most abundant transcript, or the actual transcript that is targeted for the modification. However, the consensus is that the longest isoform is often the most dominant form, in at least 75% of genes (see e.g. (Tress *et al.*, 2008; Rodriguez *et al.*, 2013). Nevertheless, this approach has the limitation that m^5C sites may be erroneously assigned to an incorrect region of the mRNA. For example, an m^5C site may be located within an intron in a transcript that is not deemed to be the 'representative' transcript. Thus, it is possible, that an intronic region may coincide with an exon in the primary 'consensus' transcript. In this case, the m^5C site will be called as an exonic site, even though it actually sits in an intron (Fig. 5.2).

Mapping m^5C sites to the customised transcriptome revealed 142, 298 and 37 candidate m^5C sites in the 5′ UTR, CDS and 3′ UTR regions respectively, whilst 87 sites were located within introns (Fig. 5.3). These identified sites are used in subsequent metagene analyses in order to elucidate, in part, the function of m^5C in mRNAs.

Information regarding the chromosomal position and methylation stoichiometry for each candidate m^5C site is summarised in Appendix 7.7 (see also Fig. 5.4).

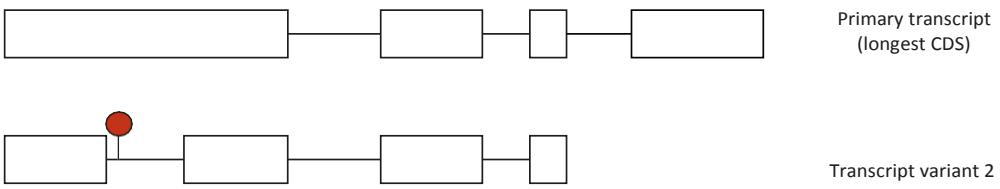


Figure 5.2: Example of incorrectly assigning an m^5C site.

Multiple transcript variants are collapsed into a single consensus representative transcript for subsequent analyses. This is done computationally by choosing the transcript with the longest coding sequence and its corresponding untranslated regions. This method may erroneously assign locality of the m^5C sites. For example, an m^5C site (red circle) resides within an intron in transcript variant 2. However in the consensus transcript chosen in the customised library (primary transcript), the site would correspond to an exon, and therefore its location would be incorrectly noted. Rectangles represent exons and lines represent introns.

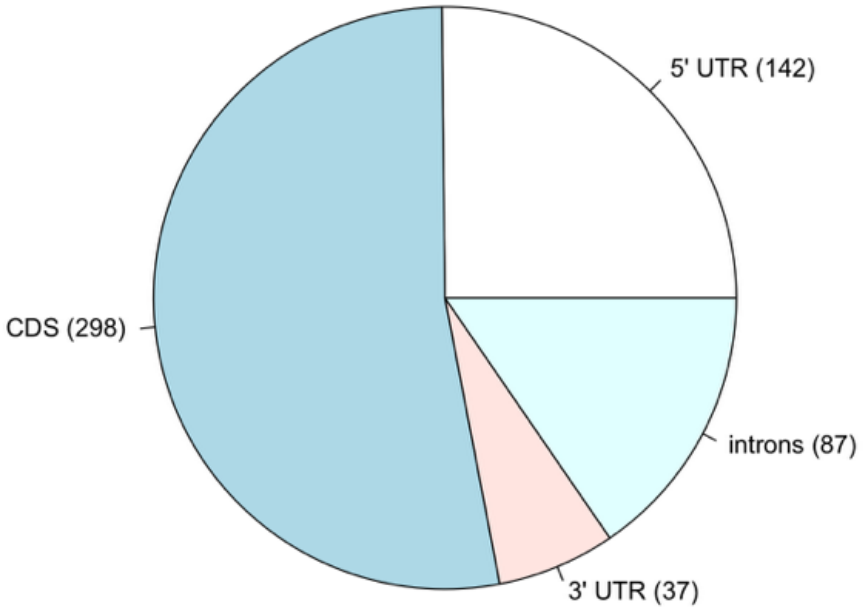


Figure 5.3: Pie chart showing the distribution of m^5C candidate sites within defined mRNA regions.

The majority of m^5C sites (298) are located in the CDS, the 5≠ UTR contains 142 candidate m^5C sites, 87 candidate m^5C sites reside within intronic regions and the 3≠ UTR contains 37 candidate m^5C sites. These sites are used for any subsequent analyses.

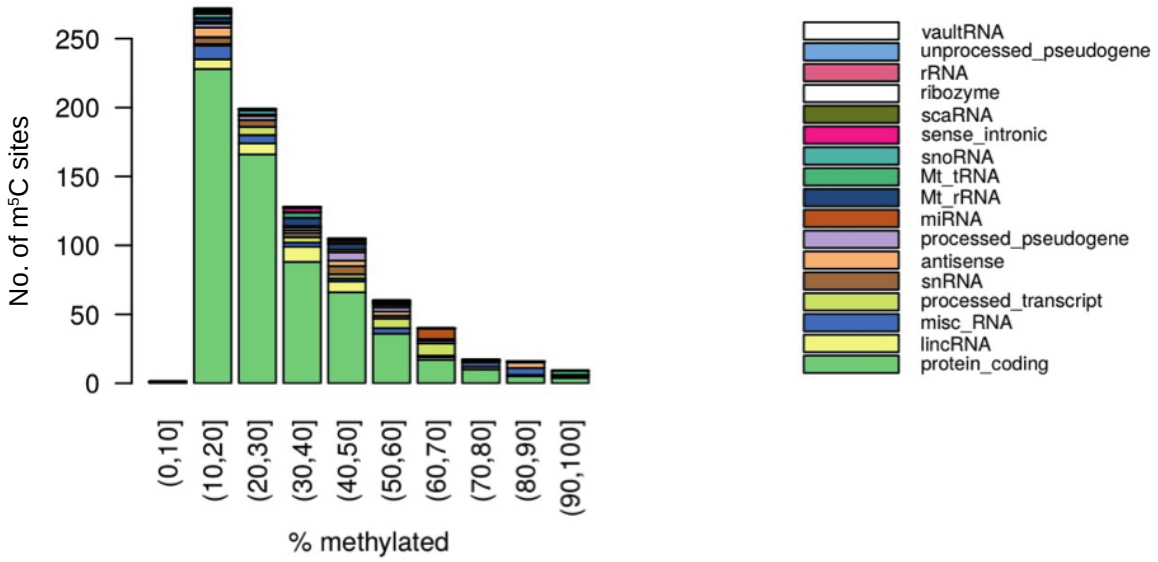


Figure 5.4: The distribution of m^{5}C methylated RNAs categorised by methylation stoichiometry.

Each column represents the number of methylated candidates characterised by RNA biotype. Most candidate sites detected are lowly methylated in the 10-20% bracket. Particular RNA biotypes are either lowly (e.g. snoRNAs) or highly methylated (e.g. miRNAs). Protein coding transcripts are represented in both low and high methylation groups.

5.2.3 Gene ontology analysis of m^5C containing mRNAs

As an initial probe to understand m^5C function, it was investigated whether the proteins translated from m^5C modified RNAs are enriched for a particular biological process or function; this may indirectly reveal a function of m^5C. To this end a gene ontology (GO) analysis was performed on the candidate m^5C containing transcripts using the Database for Annotation, Visualisation and Integrated Discovery (DAVID) (Huang *et al.*, 2009). GO terms are grouped into three ontologies: biological process, molecular function and cellular component (Ashburner *et al.*, 2000). The GO analysis of candidate m^5C sites in HeLa mRNA did not reveal any particular enrichment of complexes or pathways (see Appendix 7.8 for full list of obtained GO terms). However, identified GO terms were diverse and included 'metabolic' and 'cellular' processes and pathways (Table 5.3). Thus, as m^5C is present in transcripts that encode proteins of diverse cellular functions, it is suggested that m^5C may be a critical regulator of many cellular processes. Indeed, previous m^5C and m^6A transcriptome-wide studies, also did not uncover strong enrichments of any particular GO terms for the interrogated m^5C or m^6A-containing mRNAs (Meyer *et al.*, 2012; Dominissini *et al.*, 2012; Squires *et al.*, 2012).

5.2.4 Spatial localisation of m^5C sites within mRNAs

In order to reveal some insight into the functional role of m^5C in mRNAs the spatial localisation of m^5C sites within transcripts was investigated. Enrichment testing was employed to determine whether identified candidate m^5C sites preferentially associated with a certain mRNA region, and whether those sites are specifically distributed within a region compared to unmethylated cytosines. The odds ratio is used to quantify (in terms of an effect-size) the likelihood of observing a methylated to unmethylated cytosine in a given region compared to the other defined regions. In this case, Fisher's exact test was used to assess the statistical significance of an observed odds ratio, conditional on the null hypothesis (odds-ratio = 1) being true.

In the following metagene analyses, the sample, which constitutes the set of m^5C candidate sites, is compared to the null, a set of unmethylated cytosines expressed in the bsRNA-seq experiment. Associations between a genomic feature and the sample or null set of sites, generates the sample and null distributions. A statistical test is applied to assess differences between the sample and null distributions. The choice of null is

Table 5.3: Gene ontology analysis of Illumina candidate m^5C sites in mRNAs.

Term	Number of m^5C sites	% of m^5C sites	p-value
Biological Process			
Metabolic process	138	61.9	1.7E-8
Cellular process	168	75.3	6.4E-8
Molecular Function			
Catalytic activity	101	45.3	1.8E-6
Binding	181	81.2	7.7E-4

critical in order to gain accurate information about the absence/presence of an association with m^5C sites. The choice of null ultimately affects the interpretation of the results. For example, in the present study, random shuffling of cytosines would probably not generate a valid null. This is because cytosines are not uniformly distributed across a transcript; they are more prominently featured in the 5′ UTR, followed by the CDS and finally the lowest proportion of cytosines are generally found in the 3′ UTR of mRNAs (Jaksik & Rzeszowska-Wolny, 2012; Louie *et al.*, 2003). That is, comparing the distribution of m^5C sites with that of shuffled positions would only constitute a test for the uniformity (or lack thereof) of cytosines.

5.2.4.1 Distribution of m^5C sites across mRNA regions

Enrichment/depletion of m^5C sites for each mRNA region is determined by calculating the ratio of m^5C sites to unmethylated cytosines for each mRNA region, and assessing whether that ratio differed significantly from the ratio of m^5C to unmethylated cytosines across the other two regions of mRNA:

$$\frac{\left(\frac{m^5C}{C}\right)_{\text{mRNA region 1}}}{\left(\frac{m^5C}{C}\right)_{\text{mRNA region 2 + mRNA region 3}}} \equiv \text{odds ratio}$$

Fischer's exact test was also used to test the null hypothesis that methylated cytosines and unmethylated cytosines are evenly distributed across the transcript. A Benjamini-Hochberg correction was required to correct for multiple comparisons; the resulting adjusted p-values are indicative of the false discovery rate (FDR).

This analysis revealed an enrichment of m^5C sites compared to unmethylated cytosines in the 5′ UTR relative to the other two regions (odds ratio = 2.316, p value = 4.47E-15). An under-representation of m^5C sites was noted in the 3′ UTR relative to the 5≠ UTR and CDS (odds ratio = 0.381, p value = 3.477E-10). m^5C sites were neither enriched nor depleted in the CDS relative to the 5′ UTR and 3′ UTR (Fig. 5.5).

5.2.4.2 Distribution of m^5C sites within each mRNA region

The distribution of m^5C sites within each region was also explored. This was achieved by determining the ratio of m^5C sites to unmethylated cytosines for a given 20-nt bin, and comparing this with the number of m^5C sites to unmethylated cytosines for the

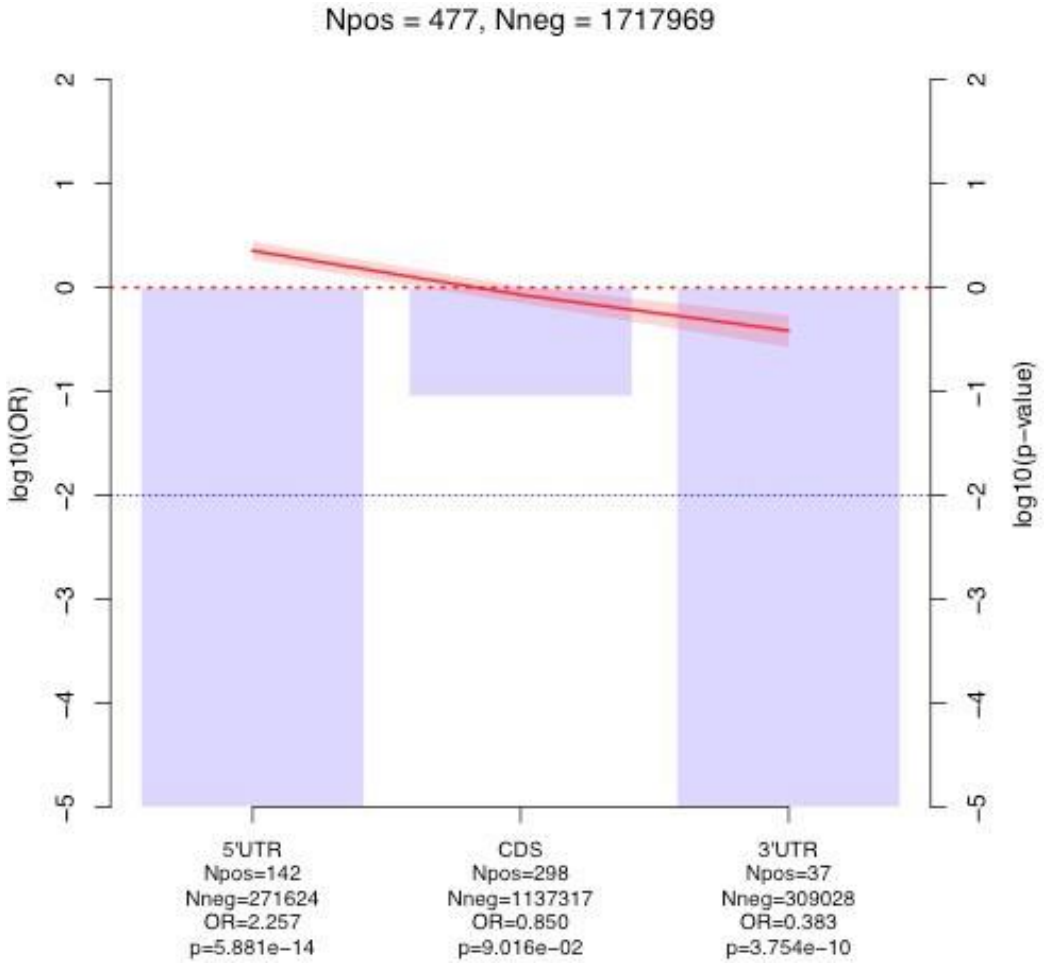


Figure 5.5: Enrichment/depletion plot of m^5C sites across mRNA regions.

The red line represents the $\log_{10}$ odds ratio. The odds ratio is the ratio of methylated to unmethylated cytosines per each mRNA region compared to the same ratio for the other two regions. The red shading on either side of the red line represents the 95% confidence interval. The purple bars represent the $\log_{10}$ (p-value), which is also indicative of the false discovery rate (FDR). The plot shows the enrichment/depletion of m^5C sites across mRNA regions. According to the data, m^5C sites are enriched in the 5′ UTR (red line is above $\log_{10}$ odds ratio of 0), neither enriched nor depleted in the CDS (red line is at $\log_{10}$ odds ratio of 0), and is depleted from the 3′ UTR (red line shows $\log_{10}$ odds ratio is less than 0). Note: p-value is equivalent to FDR.

remaining bins across the mRNA region of interest:

$$\frac{\left(\frac{m^5C}{C}\right)_{bin\ 1}}{\left(\frac{m^5C}{C}\right)_{bin\ 2-20}} \equiv \text{odds ratio}$$

Comparison of the ratios effectively gives the odds ratio, where a value of one corresponds to the null hypothesis i.e. there is no difference between the ratio of m^5C to unmethylated cytosines for the bin of interest compared to the remaining bins in a defined 1000 nt mRNA region. Multiple Fisher's exact test and Benjamini-Hochberg's multiple hypothesis testing correction were used to assess the statistical significance of the observed odds ratios.

Even though a relative enrichment of m^5C sites is observed in the 5′ UTR, there is no significant difference between the distribution of m^5C sites (Fig 5.6A) and unmethylated cytosines (Fig. 5.6C) within the 5′ UTR. Figure 5.6E compares the two ratios, indicating the odds ratio for each bin. It is important to note here that because the average length of the 5′ UTR is short at ~100 – 200 nt (Mignone *et al.*, 2002), there are fewer long 5′ UTRs and therefore there are fewer cytosines present far from the start codon. This produces a tailing off effect when plotting cytosine distribution against UTR length (Fig. 5.6A, C). This is also true when plotting cytosine distribution against CDS length (Fig.B, D).

Within the CDS, a statistically significant enrichment of m^5C sites was observed within the first 40 nt downstream of the start codon compared to the distribution of unmethylated cytosines in the same region of the CDS (Fig. 5.6F) (odds ratio of 3.2 and p value 10^{-3}).

Lastly, the spatial distribution of m^5C sites within the 3′ UTR is not significantly different from the distribution of unmethylated cytosines within the same region (see Appendix 7.9).

5.3 Investigating the function of m^5C in mRNAs

It was considered that, as m^5C is relatively enriched in the 5′ UTR and just downstream of the start codon, these sites might influence translation initiation. This is because certain features of the 5′ UTR, and the beginning of the open reading frame, are important in regulating translation initiation. For example, many factors such as GC

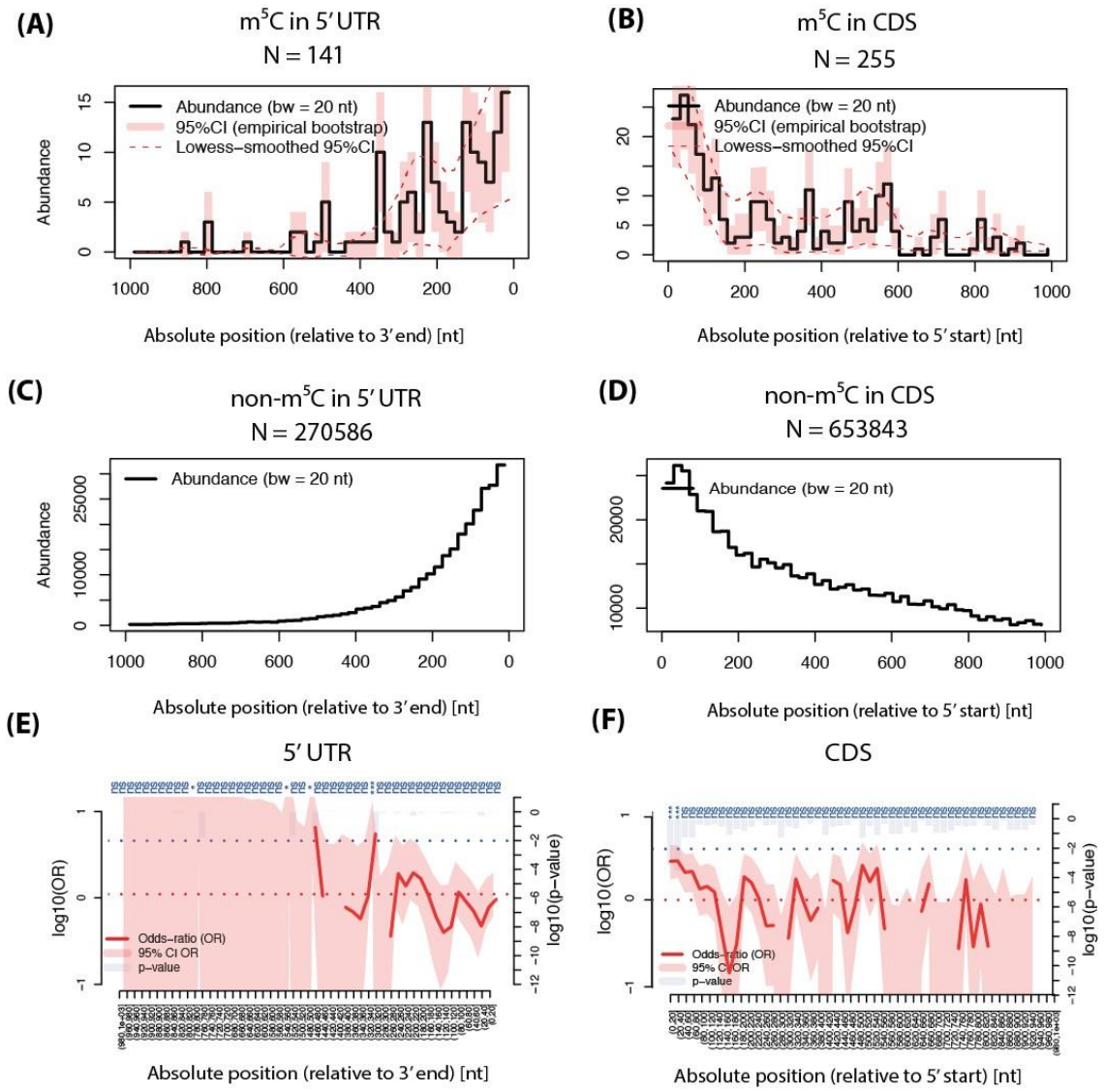


Figure 5.6: Absolute spatial arrangement of m^{5}C and non-m^{5}C in the 5′UTR and CDS.

There is no significant difference between the distributions of the m^{5}C and non-m^{5}C in the 5′ UTR **(E)**. However, m^{5}C sites within the first 40 nt downstream of the start codon (in the CDS) are enriched compared to non-m^{5}C **(F)**. The red shading represents the 95% confidence interval. **(E)** shows the odds ratio from comparing the m^{5}C distribution **(A)** with the non-m^{5}C distribution **(C)** in the 5′ UTR. **(F)** shows the odds ratio from comparing the m^{5}C distribution **(B)** with the non-m^{5}C distribution **(D)** in the CDS.

content, minimum free energy (MFE), 5′ UTR length, have been documented to affect mRNA translation (e.g. Babendure *et al*., 2006; van der Velden and Thomas, 1999; Jansen, 2001; Mignone *et al*., 2002). Therefore, it is possible that m^5C may be associated with some of these features and have an impact on mRNA translation.

In particular, 5′ UTR structure and GC content strongly impact translation initiation efficiency. Indeed, it has been postulated that mRNAs exhibiting low GC content and secondary structure may be more efficiently translated than mRNAs containing high GC content (Barrett *et al.,* 2012). This notion is supported by genome-wide studies that correlate low free energy 5′ UTRs with reduced protein abundances, as compared to less-structured 5′ UTRs (Anderson *et al.*, 2000). Whilst the exact molecular mechanism to explain this observation is not fully understood, it is plausible that less structured 5′ UTRs may permit more efficient 40S ribosomal subunit scanning as it traverses from the 5′ cap to the start codon of the main open reading frame (Mobarak *et al.*, 2000). Therefore, it was investigated whether m^5C candidate sites had a preference for GC content/structure compared to unmethylated cytosines.

5.3.1 m^5C is associated with GC rich sequences

GC content surrounding m^5C sites was calculated within a 21-nt window centered on each methylated or unmethylated cytosine in the three mRNA regions, normalised to the total GC content of each transcript region. This normalisation step accounts for GC content variability across the same mRNA region of different genes. The normalisation also accounts for the difference in GC content between the different mRNA regions (see above). This thesis shows that m^5C sites in the 5′ UTR and CDS are present within high GC rich regions, even when accounting for that fact the GC content in these two mRNA regions is relatively high (see Fig. 5.7).

The null set of sites comprises all unmethylated cytosines (per mRNA region) that are present in transcripts where an m^5C site has been detected. In other words, the null set comprises unmethylated cytosines that are known to be sufficiently “expressed” in the data. The difference in means between the m^5C sites and the null sites was assessed using a two-sided two-sample t-test. This analysis revealed that m^5C sites in the 5′ UTR and CDS are enriched within a local GC rich region by 6.9% (p-value = 3.23E-6) and 17.6% (p-value = 7.10E-33) relative to unmethylated cytosines present in the 5′ UTR and CDS respectively, normalised to total GC content for each region (Fig. 5.7A).

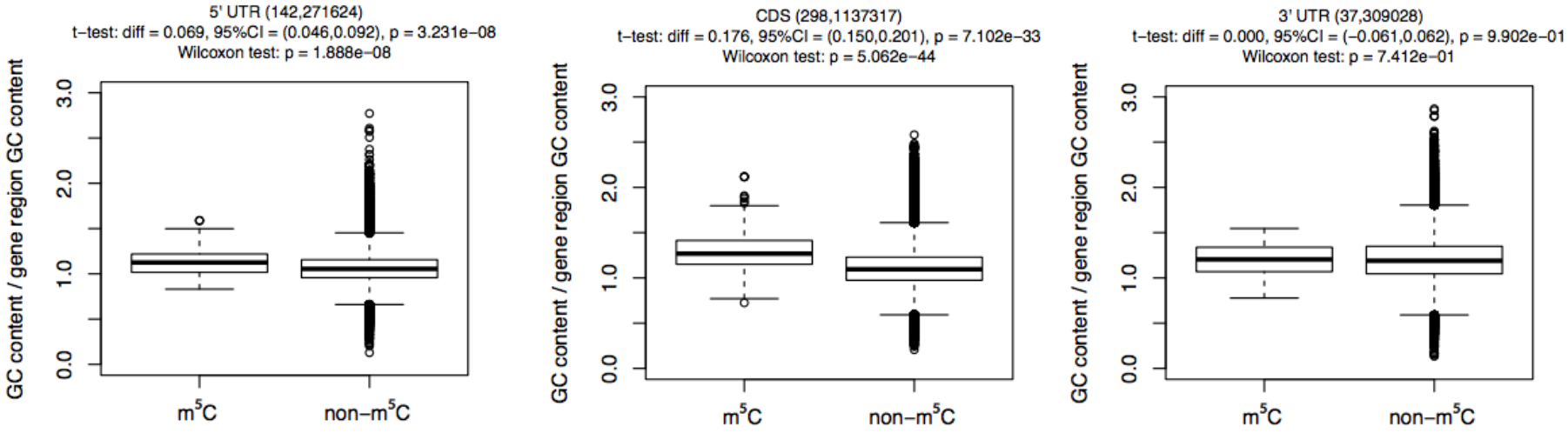
However there was no statistically significant difference in the means between GC content surrounding m^5C candidate sites in the 3′ UTR from unmethylated cytosines in the same region.

Furthermore, as a comparison, the absolute GC content around m^5C sites is also reported. This analysis was performed, as before, but without the normalisation step being applied. Again, only unmethylated cytosines present in genes with a detected m^5C comprise the null set. Similar to before, m^5C sites were present in higher GC regions compared to unmethylated cyotsines. An increase of 10.1%, 16.1% and 6% in GC content surrounding m^5C sites compared to unmethylated cytosines for the 5′ UTR, CDS and 3′ UTR respectively is observed. The difference in the mean between the distribution of m^5C sites and null sites was again assessed using a two-sided two-sample t-test. This revealed the difference in GC content surrounding m^5C sites compared to unmethylated cytosines to be significant in the 5′ UTR (p-value = 8.4E-24), CDS (p-value = 7.8E-84) and 3′ UTR (p-value = 3.4E-03) (Fig. 5.7B). Indeed, the absolute median local GC content around 5′ UTR and CDS m^5C-localised sites is 90% and 80% respectively (Fig. 5.7B), compared to 75% and 60% for unmethylated cytosines in the same regions, suggesting m^5C sites are preferentially located in GC rich regions in the 5′ UTR and CDS compared to unmethylated cytosines in the same region.

5.3.2 Sequence Motifs

Motif detection algorithms did not reveal a specific signature motif around m^5C sites, despite the fact that m^5C sites were observed in GC rich regions. The Multiple Em for Motif Elicitation (MEME) (Bailey *et al.*, 2006) and Discriminative Regular Expression Motif Elicitation (DREME) (Bailey, 2011) algorithms (which are part of the MEME suite) returned GC-rich sequences around the m^5C site, although a distinct sequence motif could not be discerned. Both MEME and DREME use shuffled sequences as the negative control, however the user can also input sequences to form the null set. MEME generally is able to discover more complex motifs and uses statistical modeling to determine the best width and number of occurrences. On the other hand, DREME specifically finds short, ungapped motifs that are relatively enriched compared with background control sequences. As a result, DREME is able to identify short motifs relatively quickly (compared to MEME), and therefore can process larger datasets.

(A)


5' UTR (142,271624)
t−test: diff = 0.069, 95%CI = (0.046,0.092), p = 3.231e−08
Wilcoxon test: p = 1.888e−08
CDS (298,1137317)
t−test: diff = 0.176, 95%CI = (0.150,0.201), p = 7.102e−33
Wilcoxon test: p = 5.062e−44
3' UTR (37,309028)
t−test: diff = 0.000, 95%CI = (−0.061,0.062), p = 9.902e−01
Wilcoxon test: p = 7.412e−01
GC content / gene region GC content
0.0
1.0
2.0
3.0
m5C
non−m5C

(B)

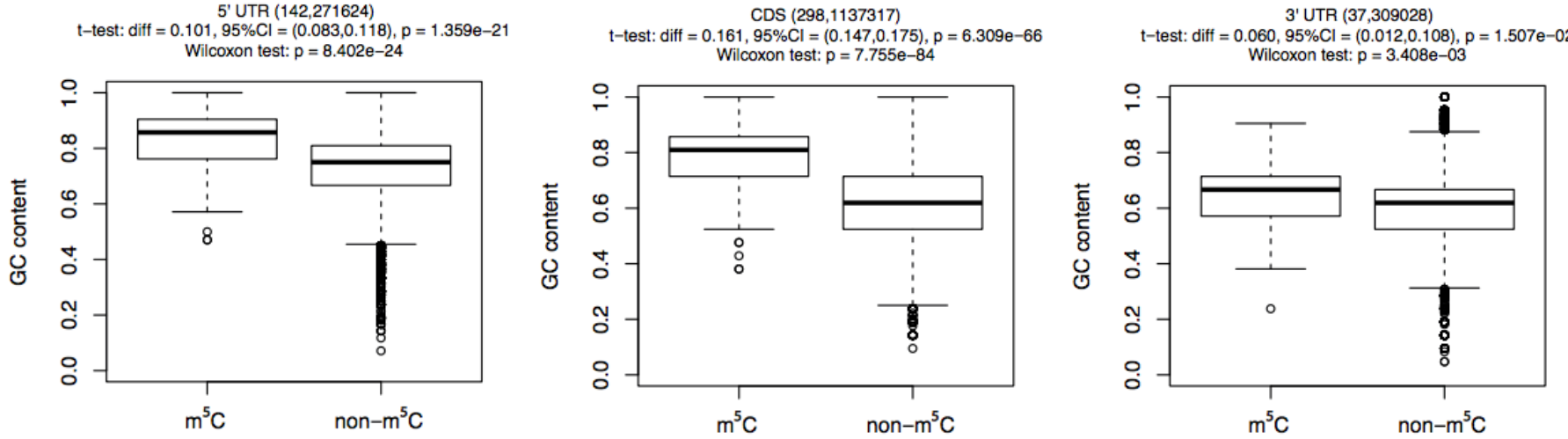

5' UTR (142,271624)
t−test: diff = 0.101, 95%CI = (0.083,0.118), p = 1.359e−21
Wilcoxon test: p = 8.402e−24
CDS (298,1137317)
t−test: diff = 0.161, 95%CI = (0.147,0.175), p = 6.309e−66
Wilcoxon test: p = 7.755e−84
3' UTR (37,309028)
t−test: diff = 0.060, 95%CI = (0.012,0.108), p = 1.507e−02
Wilcoxon test: p = 3.408e−03
GC content
0.0
0.2
0.4
0.6
0.8
1.0
m5C
non−m5C

Figure 5.7: GC content of mRNAs with and without the m^5C modification.

(A) Local GC content is calculated +/- 10 nt around methylated and unmethylated cytosines, normalised to the GC content for each mRNA region. GC content tends to be higher around m^5C sites than unmethylated cytosines, particularly in the CDS, where a large significant increase in local GC content is observed around m^5C sites compared to unmethylated cytosines in the same region. Significance from the null is determined through the t- test (p values are given).

(**B)** In absolute terms, m^5C sites are also associated with high GC content. The GC content surrounding CDS localised m^5C sites is significantly higher than the GC content around unmethylated cytosines in the 5′ UTR and CDS but not 3′ UTR. Significance from the null is determined through the t-test (p values are given).

Only genes that have a detectable m^5C site that are sufficiently expressed in the Illumina data are used in this

analysis. Box limits represent 25^{th} percentile, median and 75^{th} percentile. Whiskers represent 2.5 and 97.5 percentiles.

In this thesis, MEME identified a 21-nt GC rich sequence occurring in 18.4% of input sequences and is centered at the m^5C position. As well, DREME identified an 8-mer corresponding to a section of the 21-nt motif discovered by MEME. This motif is present in 37% of input sequences, and is centered at the m^5C position (Fig. 5.8). This lack of a clear motif, and the highly structured nature of the GC rich environment surrounding the m^5C site, suggest that a combination of sequence and structure features may be required to specify m^5C methylation positions. Interestingly this is similar to sequence features of m^1A sites in human mRNA. Dominissini *et al.* (2016) show that m^1A sites are present within GC rich regions, however no clear 'representative' motif was identified.

5.3.3 m^5C is associated with low minimum free energy structures

Next, it was investigated whether m^5C sites are present in highly structured regions. To this end, MFEs were calculated for mRNA regions containing an m^5C candidate site compared to the MFEs of mRNA regions lacking an m^5C site. MFEs were calculated using RNAfold (Gruber *et al.*, 2008). The means of the sample and null site MFE distributions were compared using two-sided two-sample t-tests to assess statistical significance.

As the MFE positively correlates with increasing mRNA length, the calculated MFEs were normalised to the length, referred to here as the adjusted MFE (aMFE). That is, the aMFE is calculated by dividing the MFE by the sequence length and the value is reported as the folding energy in kcal/mol per nt. After this adjustment, the MFEs of all nucleotide sequences are comparable, as these normalised MFEs have a reported weak relationship with sequence length (Catania & Lynch, 2010; Zhang *et al.*, 2006). This analysis revealed that m^5C-containing mRNA regions have a lower mean aMFE compared to mRNA regions lacking unmethylated cytosines. Messenger RNAs containing 5′ UTR, CDS and 3′ UTR localised m^5C sites showed a 14.1%, 5.0% and 0.3% decrease in aMFEs respectively when compared with their unmethylated counterparts (Fig. 5.9). This suggests that m^5C sites in the 5′ UTR and CDS preferentially reside within more ordered secondary structures. Also as GC content can give rise to stable secondary structure, this observation correlates with the finding that m^5C sites in the 5′ UTR and CDS are predominantly located within a GC rich region, whilst m^5C sites localised in the 3′ UTR are not.

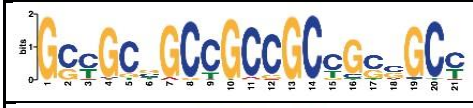
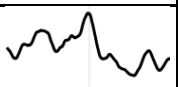
Motif	P-value	% of sites	Motif Distribution	Discovery Program used
	6.3e-260	18.4		MEME
	3.4e-026	37.0		DREME

Figure 5.8: Relatively enriched sequence motifs identified around the m^5C site using MEME and DREME.

Input sequences were +/- 50 nt around the m^5C position. The most enriched motif detected using MEME is a 21-nt motif, rich in G's and C's, centered on the m^5C, and was discovered in 18.4% of the input sequences. The most enriched motif detected using DREME is an 8-mer motif corresponding to a section of the 21-nt motif identified by MEME. Again this motif is centrally distributed around the m^5C site and is present in 37% of the input sequences.

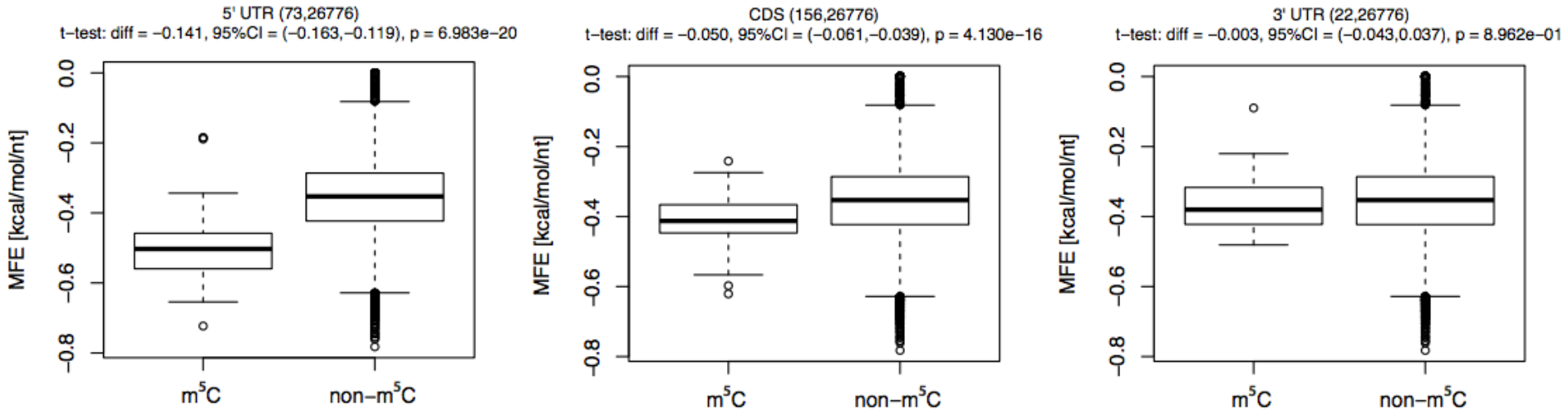


Figure 5.9: Minimum free energies (MFEs) of m^5C-containing transcripts compared to unmethylated transcripts by mRNA region.

MFEs are normalised to mRNA region length and are given as kcal/mol/nt. m^5C sites localised to the 5′ UTR or CDS are present within lower MFE structures compared to unmethylated cytosines in the same region i.e. methylated cytosines are located in more structured regions than unmethylated cytosines in the same mRNA region. Significance from the null is determined through the t-test and the Wilcoxon test (p values are given). The difference in means between MFEs of 5′ UTR or CDS m^5C localised sites compared to unmethylated cytosines in the 5′ UTR or CDS is a decrease of 14.1% and 5.1% respectively. Only genes that have a detectable m^5C site i.e. are sufficiently expressed in the BGI data are used in this analysis. Box limits represent 25^{th} percentile, median and 75^{th} percentile. Whiskers represent 2.5 and 97.5 percentiles.

5.3.4 m^5C is associated with poorly translated mRNAs

As secondary structure and GC content are known to affect translation efficiency, it was next asked whether m^5C affected the rate of mRNA translation. The translation efficiency score is a measure of mRNA translation into proteins. The score is calculated by dividing ribosome-profiling (Ribo-Seq) RPKM values with the total gene expression (mRNA-Seq) RPKM values of the same transcript. This analysis makes use of published Ribo-Seq and mRNA-Seq datasets obtained for the HeLa cell line (Wang, Lu, *et al.*, 2014). Ribosome profiling reads reflect the extent of mRNAs that are associated with ribosomes i.e. ribosome profiling is the deep sequencing of ribosome-protected mRNA fragments, essentially providing a snapshot of the proportion of mRNA molecules that are being actively translated. On the other hand, mRNA-Seq data is a measure of total mRNA expression. Thus the fraction of actively translating mRNAs can be determined by comparing the two datasets; effectively this is a proxy for the efficiency by which an mRNA is translated. It is important to note that only genes that showed a RPKM > 1 in both the ribosome profiling and gene expression datasets were used in the present analysis (only mRNAs that are sufficiently expressed are used in the analysis). Translation efficiencies were calculated for each mRNA containing an m^5C candidate site, and compared with translation efficiencies of mRNA transcripts completely lacking m^5C. Translation efficiencies scores were categorised based on the location of the m^5C site in the mRNA. The difference in means between the translation efficiency distributions for transcripts with and without m^5C sites was assessed using a two-sided two-sample t-test.

Investigations into translation efficiencies of 5′ UTR and CDS localised m^5C-containing transcripts revealed a statistically significant reduction as compared to transcripts that did not contain an m^5C site anywhere across the transcript (Fig. 5.10). A comparison of the means between the methylated and unmethylated sets, indicated that the translation efficiency of the 5′ UTR localised m^5C-containing mRNAs decreased by 8.8% compared to the unmethylated transcripts (p-value = 1.621E-02), whilst a similar trend was also observed for m^5C sites in the CDS compared to the unmethylated set, where a 7.8% decrease in translation efficiencies was observed as compared to the non-m^5C distribution (p-value = 1.914E-03). The small number of 3′ UTR m^5C sites meant that no definitive statement could be made regarding the affect of 3′ UTR sites on mRNA translation efficiency (Fig. 5.10). Even though the observed decrease in translation

efficiency of transcripts containing a 5′ UTR or CDS m^5C site is statistically significant (at the 5% significance level), the effect size is small; this observed reduction in translation efficiency may be primarily driven by the low MFEs structures associated with m^5C sites in the 5′ UTR and CDS. Indeed, as alluded to previously, highly structured 5′ UTRs may interfere with 40S scanning resulting in lower translation efficiencies. Thus to determine whether secondary structure was the main driving force behind the observed reduction in translation efficiencies, a comparison between the translation efficiencies of the most structured transcripts with and without m^5C sites was performed. Genes were selected from the m^5C and non-m^5C lists that show the lowest MFE/nt values; genes that lie within the 1^{st} quartile of the MFE/nt distribution are chosen. This comparison revealed a significant but small decrease between the average translation efficiencies of the most structured methylated transcripts compared to the most structured unmethylated transcripts (Fig. 5.11). Again, as the effect size is quite small, a definitive statement regarding the function of m^5C in the regulation of mRNA translation efficiency cannot be made. Furthermore, considering that the 5′ UTR and CDS median methylation rate is low (Fig. 5.12), the actual effect of m^5C could be even more pronounced than what is being calculated/observed here. For example, it is possible that the m^5C modification may function to enhance translation of these poorly translated transcripts, however, because only a small number of transcripts are producing the effect, such a signal may be diluted out and therefore not observable in the present analysis.

Ultimately, these analyses suggest that m^5C sites are associated with weakly translated mRNAs, however whether the m^5C mark functions to repress or enhance translation requires further investigation

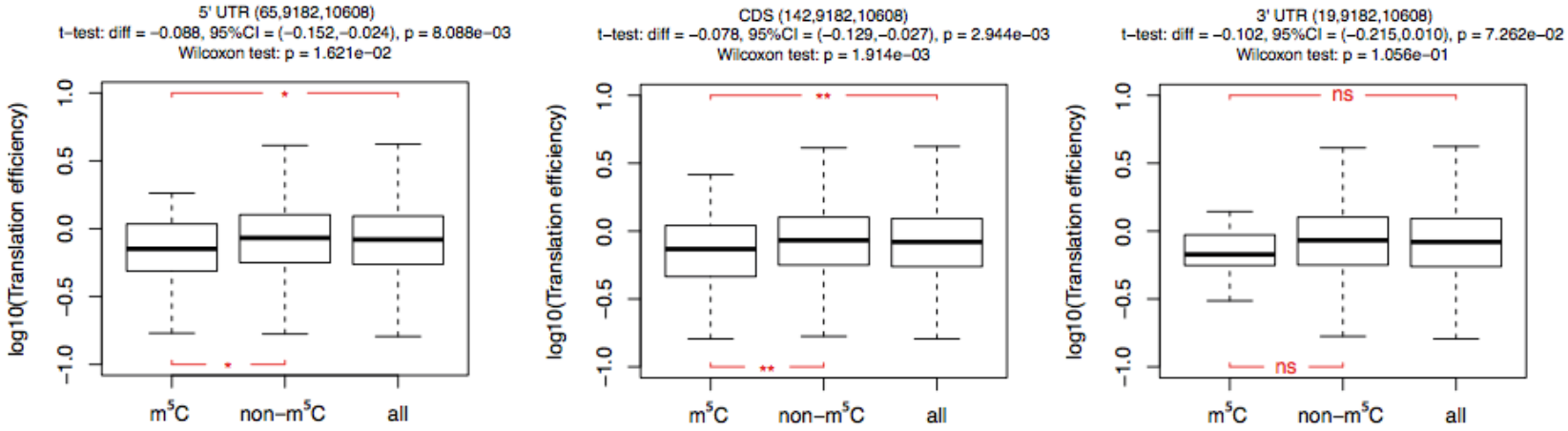


Figure 5.10: mRNA translation efficiency scores of m^5C containing mRNAs compared to mRNAs completely lacking m^5C across the transcript.

mRNAs harbouring 5′ UTR and CDS localised m^5C sites tend to show a significant reduction in translation efficiency compared to mRNAs completely devoid of any m^5C. Significance from the null is determined through the t-test and the Wilcoxon test (p values are given). The difference in means between translation efficiencies of mRNAs containing a 5′ UTR or CDS m^5C localised sites compared to mRNAs without m^5C is a decrease of 8.8% and 7.8% respectively. Only genes that are sufficiently expressed (RPKM > 1) in the ribosome profiling and gene expression datasets from Wang, Lu,

et al. (2014) are used. There are 9182 genes expressed in the BGI data that are completely void of any m^5C site anywhere across the transcript, this is the "non-m^5C" distribution. In total there are TE values for 10,608 genes (following RPKM filtering), this is the "all" distribution. t-test: Diff gives the actual difference (effect size), 95% CI is the 95% confidence interval of the difference, and p is the associated p-value for the two-tailed t-test of H(null): diff = 0 versus H(alternative): diff ≠ 0. Statistical significance is represented by the usual star notation; one star meaning $p < 0.05$, two stars meaning $p < 0.01$. Box limits represent 25th percentile, median and 75th percentile. Whiskers represent 2.5 and 97.5 percentiles.

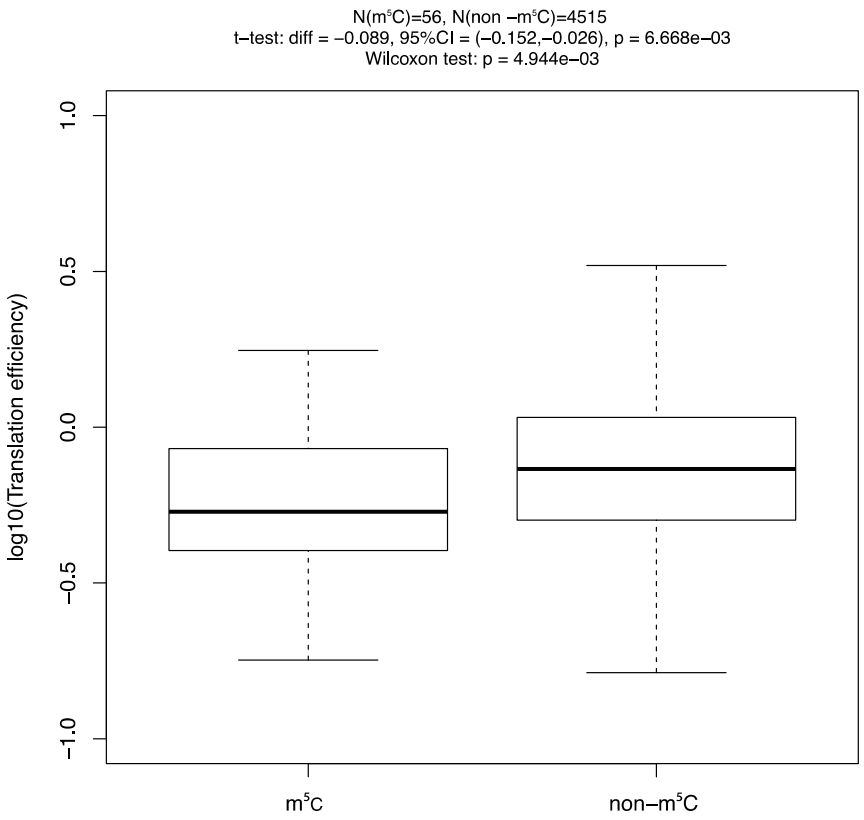


Figure 5.11: A comparison of translation efficiencies of the most structured m^5C containing mRNAs compared to the most structured mRNAs devoid of m^5C.

A statistically significant decrease of 8.9% is observed between the means of the two sets. Only genes that lie within the 1st quartile of the MFE/nt distribution are chosen. Box limits represent 25th percentile, median and 75th percentile. Whiskers represent 2.5 and 97.5 percentiles. p-values are indicated.

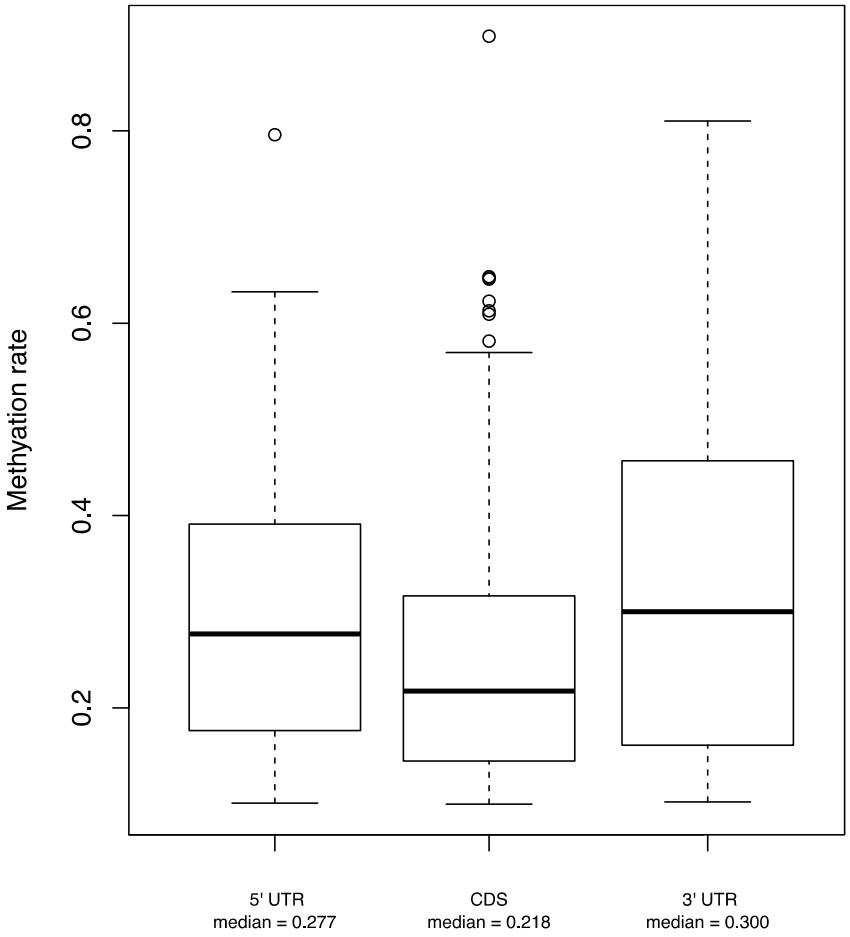


Figure 5.12: Methylation rate (i.e. number of transcripts containing m^5C compared to the number of transcripts without m^5C) per each mRNA region.

Median methylation rate for the 5′ UTR, CDS and 3′ UTR sites is 27.7%, 21.8%, and 30.0% respectively. A cutoff 10% methylation rate filter is imposed. Box limits represent 25th percentile, median and 75th percentile. Whiskers represent 2.5 and 97.5 percentiles.

5.4 Discussion

This Chapter has focused on understanding the features of m^5C methylation in mRNAs, in order to gain some insight into the functionality of the m^5C mark in mRNAs. As a post-transcriptional modification of mRNA, it is conceivable that m^5C may have a role in mRNA stability or translational regulation. Following from the previous Chapter, this Chapter has further focused on investigating the affects of m^5C (if any) on mRNA translational regulation.

One of the strengths of the bsRNA-seq approach is in identifying m^5C sites, at single nucleotide resolution throughout the transcriptome. This allows m^5C function to be studied through metagene analyses. That is, the affect of m^5C on mRNAs can be investigated across all candidates, drawing global trends, and specifically being able to examine the role of m^5C in a context-dependent manner (e.g. it is possible to study the affects of 5′ UTR, CDS, and 3′ UTR m^5C sites independently). More specifically, these analyses revealed an association between 5′ UTR, CDS localised m^5C candidate sites with high GC content, low MFE structures and low mRNA translation efficiencies.

5.4.1 Transcriptome-wide map of m^5C sites in HeLa cells

In the present study a total of 847 candidate m^5C sites were identified (mapping to the genome), using the HeLa cell line as a model. As expected, many sites mapped to tRNAs, however m^5C was also detected within other RNA biotypes including mRNAs.

A total of 226 candidate m^5C sites were detected in a wide variety of ncRNAs in this thesis. Squires *et al.* (2012) identified a total of 1,780 candidate m^5C sites in various ncRNAs. Presently, 27% of candidate m^5C sites reside in annotated ncRNAs, whereas Squires *et al.* (2012) identified 17% of candidate m^5C sites in ncRNAs. A total of 595 candidate m^5C sites were detected in mRNAs in the present study, in comparison to 8,495 candidate m^5C sites detected in the (Squires *et al.*, 2012) data. This corresponds to 70% and 81% of the candidate m^5C sites mapping to regions in mRNAs respectively. As the coverage of mRNAs in both studies was similar, this suggested the Squires *et al.* (2012) data contained a high proportion of false positives.

5.4.2 Methylation level

Overall, it is interesting to note that the methylation levels vary, with some sites showing near stoichiometric methylation (~100% methylation), whereas other sites are moderately methylated (60-70%) and still others are lowly methylated (10-30%), with the majority of sites showing low methylation (10-20%) (Fig. 5.4). Interestingly, candidate m^5C sites in tRNAs and miRNAs tend to be methylated at a >60% transcript stoichiometry. Whilst all the candidate m^5C sites in snoRNAs occur at a transcript stoichiometry of 10-20% (Fig. 5.4). Generally, m^5C methylation in mRNAs is present at a low transcript stoichiometry methylation rate (<30%), with only a small proportion of sites exhibiting a high transcript stoichiometry methylation rate (>60%). However, mRNAs are represented in each of the low, medium and high methylation groups. For example, Peptidylglycine Alpha-Amidating Monooxygenase mRNA is lowly methylated (11% methylation rate), Guanine Nucleotide Binding Protein (G Protein), Alpha Stimulating mRNA has a moderate methylation stoichiometry (35% methylation rate) and CTS Telomere Maintenance Complex Component 1 mRNA is highly methylated (80% methylation rate). These varying methylation stoichiometries will likely be of functional significance. A lowly methylated site is expected to indicate a regulatory function of m^5C, whereas a highly methylated site may suggest a mandatory function of the methylation. Also these varying methylation levels may indicate the involvement of the m^5C modification in various cellular processes.

However, it is also possible that the range of observed methylation levels in mRNAs might reflect differential abundance of available transcript variants for a given gene. For example, consider gene A that has two transcript variants A1 and A2. Variant A1 is the dominant form and is expressed at 80%, whilst A2 is expressed at 20%. An observed 20% methylation could therefore represent methylation in A2 in 100% of cases. Thus even though, overall a 20% methylation is observed, actually, 100% of the intended target may be methylated. This selective methylation of a particular transcript variant may contribute to function. For example, m^5C may only be present in the transcript variant which gives rise to a functional protein and could function to enhance mRNA translation. Alternatively, m^5C may be present in the less abundant form and mediate mRNA decay. Furthermore, m^5C may have a role in splicing and therefore is placed only in a particular transcript variant. Therefore, comparing methylation rate with transcript variant abundance may provide some insight into m^5C function.

5.4.3 m^5C in mRNAs

In this thesis, candidate m^5C sites were detected throughout all regions of mRNAs. Overall, a large number of candidate m^5C sites were located in the CDS, whilst m^5C localised 5′ UTR sites were significantly enriched relative to sites in the other two regions. The enrichment of candidate m^5C sites in the 5′ UTR is consistent with the previous study by Squires *et al.* (2012).

Presently, an enrichment of m^5C sites was also observed just downstream of the start codon. Clearly this non-random distribution of sites in the 5′ UTR, and just downstream of the AUG start codon, will have implications for the role of m^5C. Because of the locality of these enriched sites, it is conceivable that these sites may be involved in regulating translation initiation.

5.4.4 m^5C and mRNA translation rates

Investigations into GC content and MFEs of m^5C containing transcripts revealed m^5C sites tend to reside in GC rich areas, and in a structure associated with a lower (more negative) MFE compared to unmethylated cytosines. m^5C sites are strongly defined by these features as compared to unmethylated cytosines, to the extent that they could even act as determinants for predicting candidate m^5C sites. Further, analyses into m^5C function revealed 5′ UTR and CDS localised m^5C sites are present in weakly translated mRNAs. It is important to note that as the methylation rate is quite low, any change in translation rate resulting from the m^5C mark may be diluted out and therefore not observable in the performed metagene analysis in this thesis. Indeed this analysis does not directly indicate a function for m^5C, rather it shows m^5C may have a role in translation and moreover, that m^5C sites tend to be located in weakly translated mRNAs.

As m^5C sites are located in GC rich and low MFE regions, it is possible that the introduction of the m^5C mark may induce or stabilise the surrounding structured environment, which could impede ribosome scanning resulting in reduced protein levels. This reduced translation efficiency may directly result in the degradation of these translationally repressed mRNAs. On the other hand, it is also possible that the m^5C mark may relax the secondary structure resulting in translation of those previously translationally repressed mRNAs, possibly in response to certain cellular cues.

Also, the m^5C mark may function to recruit specific m^5C 'reader' protein/s to mediate its function. Such proteins may act either to repress or promote translation of a particular subset of m^5C-containing mRNAs. Interestingly, the current study provides evidence that particular proteins, involved with translation machinery, may function as 'readers' of the m^5C modification (see Chapter IV). For example, the m^5C mark may recruit 'reader' proteins ultimately resulting in the enhanced translation of weakly translated mRNAs. Indeed, translationally repressed mRNAs may be reactivated for translation following modification by m^5C, possibly in response to an environmental cue e.g. stress.

5.5 Future Directions – Outlook/open-ended questions

The gene ontology analysis of the total candidate m^5C sites in HeLa cells did not reveal any enrichment of complexes or pathways. Identified GO terms were diverse and included metabolic and cellular processes, suggesting that m^5C may regulate various processes involved in post-transcriptional gene regulation. Further work is required in order to determine which processes m^5C has a regulatory role, and the mechanism by which this occurs.

This thesis indicates that the majority of targets are lowly methylated, whilst only a small subset of mRNAs are modified with m^5C. Other work should be directed at understanding why certain mRNAs are targeted. How does the methyltransferase recognise its target? Are proteins involved in directing the methyltransferase to its target for methylation? Why and how are only a proportion of transcripts recognised for modification? What is the biological significance of this low methylation? Or does m^5C methylation occur in response to a particular cellular cue, resulting in increased methylation of targets? For example, NSUN2 is overexpressed in various cancers; does this result in stoichiometric methylation of mRNA targets? What would be the biological consequence of this? Are m^5C-containing mRNAs translated into proteins that function in the cancer pathway?

There is also a need to discern the cellular machinery involved in mediating m^5C function. For example, are there specific m^5C 'readers' that elicit m^5C function (this being the focus of Chapter IV)? Do different 'readers' recognise m^5C sites in a

particular RNA biotype or context? Does reading occur in response to particular environmental cue? When do demethylases remove the m^5C mark?

There is a strong motivation 1) to establish the identity of m^5C 'readers' and demethylases, 2) to determine the subset of m^5C-containing mRNAs they target, and 3) to elucidate any phenotypes associated with perturbed expression of 'writers', 'readers' or 'erasers'. Understanding such information is important as it may shed light on the role of m^5C in different pathways/processes. Clearly, m^5C is expected to affect multiple aspects of mRNA metabolism, as it is present in all mRNA regions. This thesis suggested that m^5C sites localised around the start codon may be involved in regulating mRNA translation.

Chapter VI

General Discussion and Future Directions

6.1 General Discussion

Understanding the function of post-transcriptional modifications, especially in the relatively lesser abundant RNAs, presents a challenging problem. With the advent of next-generation sequencing (NGS), advances are being made. However, NGS as an output requires the application of a specific protocol geared at detecting a particular modification. For example, methods to identify m^6A, m^6Am, m^5C, pseudouridine and m^1A have been developed – NGS is then coupled with those approaches to identify sites transcriptome-wide. Clearly, many other modifications exist which cannot yet be identified using NGS as an output. For all of the aforementioned modifications (except for m^1A), the use of NGS as an output allows for the identification of sites at single-nucleotide resolution and reveals information regarding stoichiometry. On the other hand, mass spectrometry identifies the full range of RNA modifications simultaneously with high accuracy in a high-throughput manner. Mass spectrometry is a very sensitive approach. RNA fragments are prepared by complete digestion with RNase T1, which generates a catalogue of fragments ending with a G residue at the 3' end. However, as RNA is a polyanion, ionisation of oligonucleotides is difficult and therefore its 'flight' through the vacuum, that is required for mass spectrometry, is hampered. This approach reveals sequence information, however, the amount of RNA required for an analysis of the mRNA transcriptome is unattainable using current experimental methods. Alternatively, the RNA sample may be subjected to total enzymatic degradation and separated by reverse-phase HPLC. This results in nucleosides that 'fly' more efficiently through the mass spectrometer enabling the detection and quantification of modified nucleosides in the femtomol to attomol range. Importantly, positional isomers of modified nucleosides i.e. different modified nucleosides with the same molecular mass e.g. m^1A, m^6A may be identified by differences in retention time (the amount of time a compound spends on the column after it has been injected) at the liquid chromatography stage. Thus, the power of this technique is the separation in two dimensions (separation by HPLC and by MS) that allows for the unequivocal identification of RNA modifications present in the sample. A major drawback is that sequence information is lost. Therefore NGS is the output of choice to identify post-transcriptional modifications, however liquid chromatography coupled with mass spectrometry can be used as a powerful validation tool.

Currently, the molecular function/s of the m^5C modification in RNA are poorly

understood. However, m^5C in both mitochondrial and nuclear-encoded tRNAs and rRNAs has been shown to generally affect stability (Schaefer *et al.*, 2010); Tuorto *et al*., 2012; Blanco *et al*., 2014; (Yamasaki *et al.*, 2009); Sharma *et al*., 2013). This thesis has primarily focused on understanding the role of this modification in RNA, particularly in mRNAs. Work presented in this thesis suggests m^5C has a complex, context-dependent role, but broadly affects mRNA stability and translation.

Three main techniques to identify m^5C sites transcriptome-wide have been advanced. The method used in this thesis was bisulfite treatment of total RNA, which allowed for the detection of m^5C sites at single nucleotide resolution. Using this approach, m^5C sites were detected in both mRNAs and various ncRNAs. However, RNA secondary structure is known to impede bisulfite conversion, which may result in false-positives. The other two approaches (5-Aza-IP and miCLIP) are based on immunoprecipitations, and therefore only provide peak width information as to the location of m^5C sites. As both these techniques effectively trap the methyltransferase to its RNA target, over-expression of the methyltransferase of interest is required. This may result in non-physiological targets being detected. Interestingly, though, the 5-Aza-IP method does not detect m^5C in mRNA, whilst the miCLIP approach identifies only a small number of sites in mRNA. Recently, Hussain *et al*. (2013) compare the number of sites identified using bsRNA-seq, 5-Aza-IP and miCLIP approaches. Even though it is anticipated that the majority of m^5C sites are controlled by NSUN2, overall there is little overlap between the three methods, especially when focusing on non-tRNA targets. However the non-overlap suggests that other enzymes are also responsible for m^5C. For example, the mitochondrial-localised methyltransferases, NSUN3 and NSUN4, are the prime candidates to control m^5C sites present within mt-RNA.

With the advent of NGS, identification of the m^5C modification transcriptome-wide has become relatively straightforward. This has revealed a number of m^5C sites in many different RNA biotypes. In this thesis, using a metagene analysis approach, features of m^5C in mRNAs were determined. m^5C was found to be significantly enriched in the 5≠ UTR of mRNAs, and particularly concentrated just downstream of the start codon. Further analyses showed that m^5C sites were significantly enriched in higher GC regions and lower MFE structures compared to unmethylated cytosines. Overall, m^5C sites are not randomly distributed; rather they are present in areas of regulatory importance, whilst also being associated with features known to affect the efficiency by

which mRNAs are converted into protein. Thus the possibility that m^5C could function in regulating mRNA translation, at least, in a subset of m^5C-containing RNAs was investigated. By calculating mRNA translation efficiencies, it became apparent that m^5C sites resided in weakly translated mRNAs compared to unmethylated cytosines. This suggested that m^5C may indeed have a role in regulating mRNA translation. As mean methylation rate is low, no information can be gained as to the mechanism by which m^5C regulates translation. Conceivably, m^5C could either function to repress or promote translation. This could be achieved by the recruitment of m^5C 'reader' proteins that could either physically block the translation process, or enhance translation by further recruiting initiation factors. The exploration into m^5C 'readers' yielded Ezrin and Radixin as candidates. It was recently shown that Ezrin interacts with poly(A)-binding protein 1 (PABP1) and enhances translation of mRNAs containing a structured 5´ UTR through a suspected closed loop model. Thus, as many 5´ UTR localised m^5C-contianing RNAs have been identified as oncogenic mRNAs, and generally are characterised by poor translation, it is possible that recognition of these modified mRNAs by Ezrin may result in increased translation and the progression of cancer. This could represent a model by which m^5C promotes translation in the event of cancer by recruiting specific 'reader' proteins.

Furthermore, NSUN2 was also identified as a possible 'reader' of m^5C, akin to METTL3, which was recently shown to function as both a 'writer' and 'reader' of m^6A. It could be expected that NSUN2 would specifically target m^5C-modified mRNAs which function in a pathway in which NSUN2 has a role.

Another example where a potential m^5C 'reader' may mediate m^5C function is highlighted in this thesis. NSUN4 was shown to target human 12S mt-rRNA for m^5C modification in this thesis. Sub-stoichiometric m^5C methylation (40%) at C841 in the small subunit (12S mt-rRNA) was observed suggesting a regulatory function for the methylation. (Metodiev *et al.*, 2014) show that mitoribosome assembly and mitochondrial translation are inhibited in the absence of NSUN4. Thus in this thesis it was postulated that NSUN4 'reads' the m^5C mark on modified small subunits which are destined for mitoribosme assembly. Mitoribosome assembly is hypothesised to occur by NSUN4 'reading' the m^5C mark in the small subunit and bringing it to the large subunit by interacting with MTERF4, which is known to target the large subunit. In this way,

unmodified small subunits are expected to be degraded, therefore, the concentration of mitoribosomes, and protein, is controlled. This example reveals the potential for the involvement of m^5C in mRNA stability/half-life, which may be mediated by an m^5C 'reader' that also functions to 'write' the modification.

This thesis also showed another context where m^5C is expected to have a potential role in mRNA stability. The presence of m^5C in the 3≠UTR of mRNAs suggested a possible link between m^5C and post-transcriptional gene control. To investigate this further, AGO2 footprints were overlapped with candidate m^5C sites. This revealed an enrichment of AGO2 footprints just upstream of candidate m^5C sites in both the CDS and 3≠UTR. Furthermore, a depletion of m^5C sites was observed at conserved 7-mer seeds. These complementary analyses indicate that m^5C sites are abundant downstream, but not in, AGO2 binding sites. Therefore, the possibility that AGO2/miRNAs may be involved in guiding m^5C formation was explored. This idea was explored by performing co-IP experiments of AGO2 and NSUN2. This result suggested that AGO2 and NSUN2 do not likely interact, and therefore AGO2 is not expected to be involved in m^5C formation. It is, then, more probable that regulatory interaction exists between AGO2 and m^5C. For example, one of the factors may either block or promote binding of the other. To explore this further the association between AGO2 footprints and m^5C sites, categorised by increasing stoichiometry, was performed. This revealed a negative correlation between highly methylated sites and AGO2 footprints, implicating m^5C as an antagonist of AGO2 binding. In this way, at least a subset of 3≠UTR and CDS localised m^5C sites are expected to promote mRNA stability/half-life.

It may be argued that as most sites are sub-stoichiometrically modified, at least a proportion of them are non-functional. For example, if the modification destabilises the mRNA then a very low level of methylation will not have an effect. Indeed, it has been suggested by Graur *et al*. (2013) that some modifications may be functionally neutral. This is possible as NSUN2 is primarily a tRNA methyltransferase and therefore mRNA structures resembling tRNA structures may be methylated as an off-target effect. Generally, though, the large number of m^5C sites identified, suggests that at least the majority of sites are functional. Furthermore, the functions of m^5C in mRNAs are beginning to emerge suggesting that m^5C methylation in the context of mRNAs is not an off-target effect.

It is interesting to note here that levels of methylation may be dynamic in response to a particular cellular/environmental cue. For example, an mRNA may be methylated at 20%, which results in 'normal' levels of translation. However, in response to a specific cellular/environmental cue, m^5C methylation may be increased. This could then result in a burst of translation of that particular mRNA. For example, m^5C levels may increase/decrease in specific disease contexts or following stress. This suggests that even low-level modified sites serve a purpose, and are not necessarily an off-target effect.

It has been argued in this thesis that there is exquisite selection for sites, and the proportion of sites that are modified is of critical importance. Overall m^5C is suggested to have a complex role in RNA, which is dependent on its context.

6.2 Future Directions

To definitively elucidate the role of m^5C in mRNAs, a comprehensive high confidence set of m^5C sites needs to be identified. This could be facilitated by the development of a RNA m^5C antibody, which would enable the easier detection of RNA m^5C sites transcriptome-wide. Moreover, identified m^5C sites should be assigned to a particular MTase. Determining the substrate specificity for each of the NSUN MTases will provide a greater overall insight into m^5C function. For example, is NSUN2 the only m^5C methyltransferase to target mRNA? By what mechanism does NSUN2 recognise its targets? Do NSUN2 targets elicit a different function to m^5C synthesised by other MTases?

The dynamics between 'writing', 'reading' and 'erasing' of the modification should be understood. For example, in which compartment does each function take place and by what mechanism? How does the MTase recognise its target? How is only a proportion of a given RNA selected for methylation? Does this relate to transcript variant abundances? How does a 'reader' protein specifically recognise a subset of m^5C-containing mRNAs? Will a 'reader' always affect the same outcome upon mRNA binding? What cues, if any, are required for demethylation to occur? What is the specificity of the 'eraser' enzymes? When/where does 'writing', 'reading', 'erasing' take place?

NSUN MTases are known to have other roles in addition to their methyltransferase activity. For example, NSUN2 is involved in spindle assembly and chromosome segregation. An interesting question is whether these other functions define, at least a subset, of the MTase's methylation targets. Or is methyltransferase activity independent of other MTase function? In this case, what dictates when the enzyme functions as an MTase?

Also does m^5C exert a function in a physiological setting, or is m^5C function 'switched on' following a stress response? For example, the most comprehensive list of Ψ-containing RNAs was achieved during stress conditions (Li *et al.*, 2015). N6-methyladenosine was also shown to enhance translation of stress response genes following heat shock (Zhou *et al.*, 2015).

Finally, there should also be an emphasis on understanding the interplay, if any, between different post-transcriptional modifications in mRNAs and affect on mRNA fate. For example, is more than one modification required for function?

Chapter VII

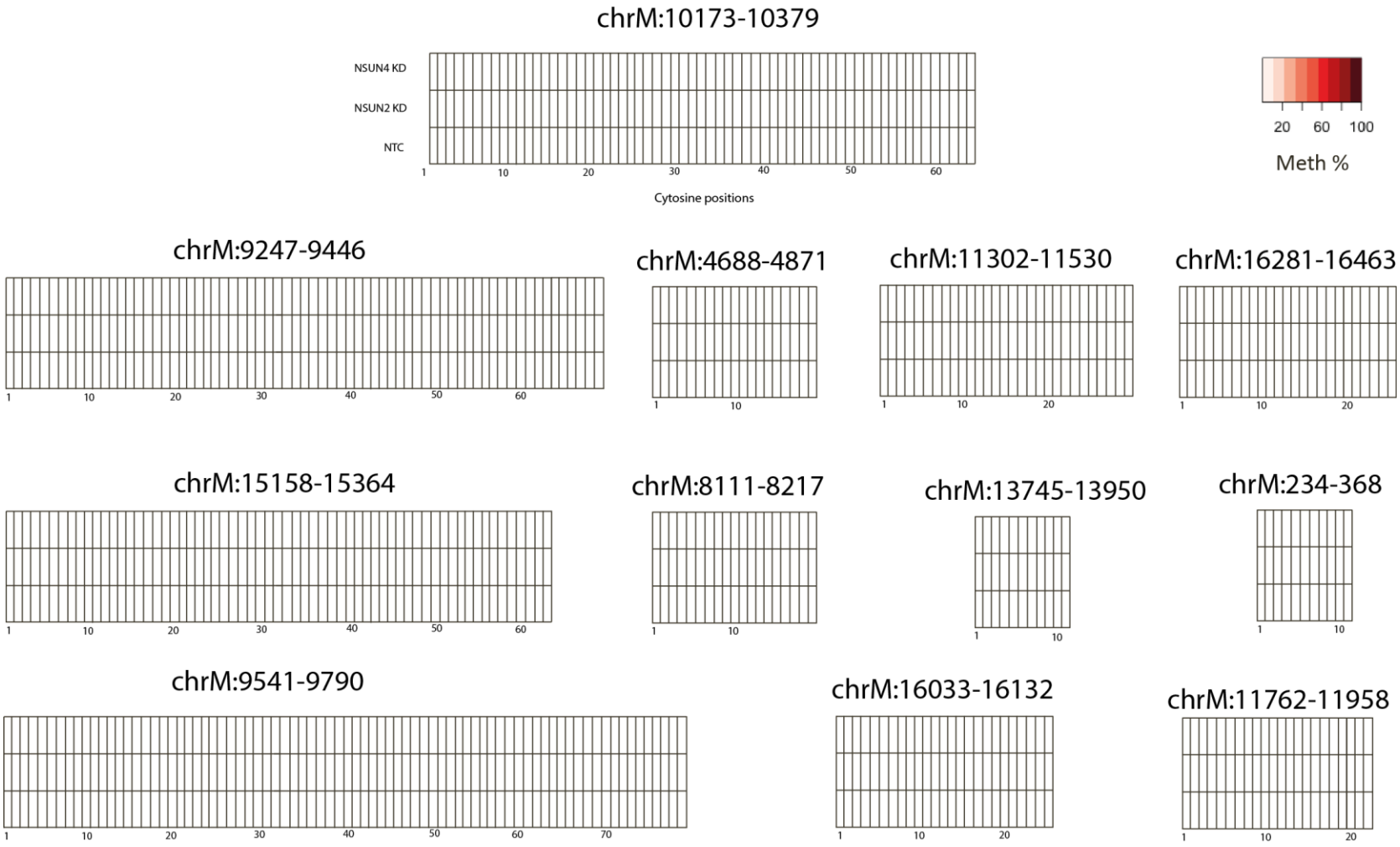
Appendix

Some of the following Appendix files have been included digitally. Please refer to the attached DVD in the pocket at the back.

7.1 Sequence for non-humanised *R-Luc in vitro* transcript.

ATGATCCAGAACAAAGGAAACGGATGATAACTGGTCCGCAGTGGTGG
GCCAGATGTAAACAAATGAATGTTCTTGATTCATTTATTAATTATTA
TGATTCAGAAAAACATGCAGAAAATGCTGTTATTTTTTTACATGGTA
ACGCGGCCTCTTCTTATTTATGGCGACATGTTGTGCCACATATTGAG
CCAGTAGCGCGGTGTATTATACCAGACCTTATTGGTATGGGCAAATC
AGGCAAATCTGGTAATGGTTCTTATAGGTTACTTGATCATTACAAAT
ATCTTACTGCATGGTTTGAACTTCTTAATTTACCAAAGAAGATCATT
TTTGTCGGCCATGATTGGGGTGCTTGTTTGGCATTTCATTATAGCTA
TGAGCATCAAGATAAGATCAAAGCAATAGTTCACGCTGAAAGTGTAG
TAGATGTGATTGAATCATGGGATGAATGGCCTGATATTGAAGAAGAT
ATTGCGTTGATCAAATCTGAAGAAGGAGAAAAAATGGTTTTGGAGAA
TAACTTCTTCGTGGAAACCATGTTGCCATCAAAAATCATGAGAAAGT
TAGAACCAGAAGAATTTGCAGCATATCTTGAACCATTCAAAGAGAAA
GGTGAAGTTCGTCGTCCAACATTATCATGGCCTCGTGAAATCCCGTT
AGTAAAAGGTGGTAAACCTGACGTTGTACAAATTGTTAGGAATTATA
ATGCTTATCTACGTGCAAGTGATGATTTACCAAAAATGTTTATTGAA
TCGGACCCAGGATTCTTTTCCAATGCTATTGTTGAAGGTGCCAAGAA
GTTTCCTAATACTGAATTTGTCAAAGTAAAAGGTCTTCATTTTTCGC
AAGAAGATGCACCTGATGAAATGGGAAAATATATCAAATCGTTCGTT
GAGCGAGTTCTCAAAAATGAACAATAATTCTAGAGCGGCCGCTTCGA
GCAGACATGATAAGATACATTGATGAGTTTGGACAAACCACAACTAG
AATGCAGTGAAAAAAATGCTTTATTTGTGAAATTTGTGATGCTATTG
CTTTATTTGTAACCATTATAAGCTGCAATAAACAAGTTAACAACAAC
AATTGCATTCATTTTATGTTTCAGGTTCAGGGGGAGGTGTGGGAGGT
TTTTTAAAGCAAGTAAAACCTCTACAAATGTGGTAAAATCGATAAGG
ATCCAAAAAAAAAAAAAAAAAAAAAA

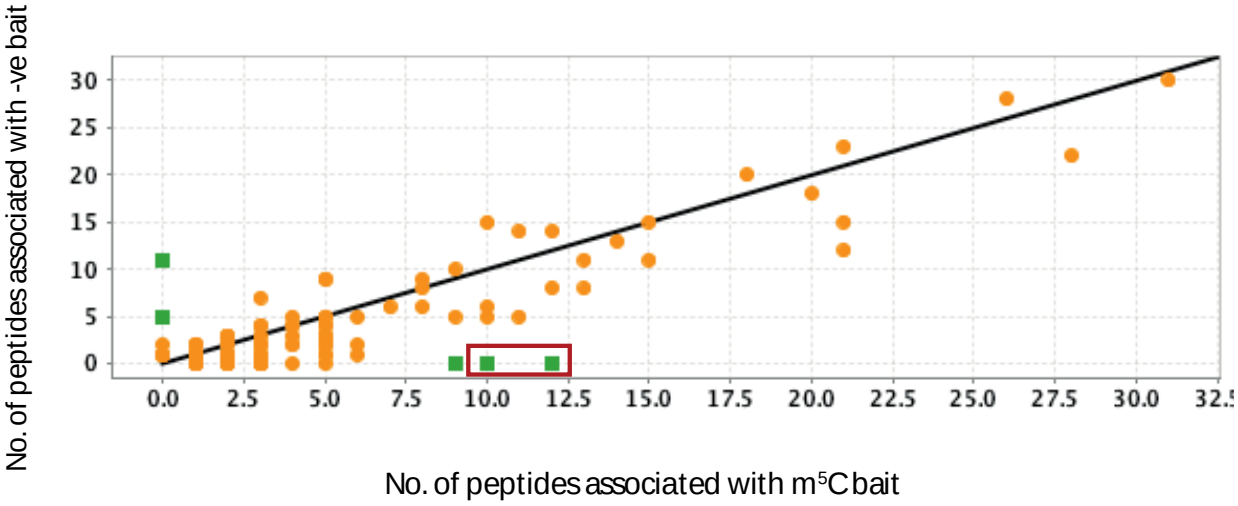
7.2 Heatmaps showing candidate mitochondrial m^5C sites not to be methylated.



7.3 List of RNA binding proteins identified following mass spectrometry of methylated and unmethylated RNA baits.

Attached is an excel file showing the unique peptide counts associated with the baits when incubated with NSUN2 knockdown lysate. Sample 1 and 3 represent two biological replicates for the methylated bait; samples 2 and 4 represent two biological replicates for the unmethylated bait.

7.4 Scatter plot showing number of peptides associated with the methylated and unmethylated baits when incubating with wild type lysate.



Numbers of peptides associated with each bait from the mass spectrometry analysis is plotted. The x-axis shows the number of peptides associated with the m^5C bait, and the y-axis shows the number of peptides associated with the negative control bait. Proteins specific to the m^5C bait fall on the x-axis and protein specific to the negative control bait fall on the y-axis. Proteins showing an insignificant odds ratio are coloured orange. Proteins with a significant odds ratio are coloured green. Fisher's exact test was applied to determine any significant changes between the odds ratio for a given peptide, compared to the odds ratios of all the other detected peptides. Significance was determined at a p-value ≤ 0.01. The statistical test was performed using Scaffold4 mass spectrometry analysis software. This plot is representative of n = 1 experiment. The red rectangle indicates potential m^5C 'readers'; these are the ezrin and radixin proteins.

7.5 RNAModR example code.

The following lines of R code will load the RNAModR library, and plot the distribution of m^6A sites Linder *et al.* (2015) across the 5' UTR, CDS and 3' UTR of the human hg38-based transcriptome.

```
# Load the library.
library(RNAModR);

# Build reference transcriptome.
# This might take a few minutes.
BuildTx("hg38");

# Load and map m6A sites to reference transcriptome.
posSites <- ReadBED(system.file("extdata",
"miCLIP_m6A_Linder2015_hg38.bed", package =
"RNAModR"));
posSites <- SmartMap(posSites, "m6A_Linder");

# Keep sites located in the 5'UTR, CDS and 3'UTR
posSites <- FilterTxLoc(posSites, filter = c("5'UTR",
"CDS", "3'UTR"));

# Plot distribution across transcript sections
PlotSectionDistribution(posSites);
```

7.6 The bash script for the alignment of Illumina bsRNA- seq reads.

```
#!/bin/bash

# Prepare and index reference genome
bismark_genome_preparation --verbose --bowtie2 ../refGenome/

# HeLa 1
bismark --phred64-quals --bowtie2 -N 1 --non_bs_mm ../refGenome/
HeLa1_trimmed.fastq
samtools view -h -o HeLa1_trimmed.fastq_bismark_bt2.sam
HeLa1_trimmed.fastq_bismark_bt2.bam
bismark_methylation_extractor -s --CX --cytosine_report --
genome_folder ../refGenome/ HeLa1_trimmed.fastq_bismark_bt2.sam

# HeLa 2
bismark --phred64-quals --bowtie2 -N 1 --non_bs_mm ../refGenome/
HeLa2_trimmed.fastq
samtools view -h -o HeLa2_trimmed.fastq_bismark_bt2.sam
HeLa2_trimmed.fastq_bismark_bt2.bam
bismark_methylation_extractor -s --CX --cytosine_report --
genome_folder ../refGenome/ HeLa2_trimmed.fastq_bismark_bt2.sam

# Merge bismark bam files and extract methylation sites
samtools merge HeLa_merged.bam
HeLa1_trimmed.fastq_bismark_bt2.bam
HeLa2_trimmed.fastq_bismark_bt2.bam
samtools view -h -o HeLa_merged.sam HeLa_merged.bam
bismark_methylation_extractor -s --CX --cytosine_report --
genome_folder ../refGenome/ HeLa_merged.sam
```

Alignment is performed using bowtie.

-N1: only one mismatch allowed in the first L bases (-l is 20).

'--non_bs_mm': prints an extra column at the end of SAM files showing the number of non-bisulfite mismatches of a read.

Input qualities are ASCII chars equal to the Phred quality plus 64.

Samtools coverts the SAM file to a BAM file.

Bismark provides the methylation report. By default, this mode will only consider cytosines in CpG context, but it was extended to cytosines in any sequence context by using the option '--CX'.

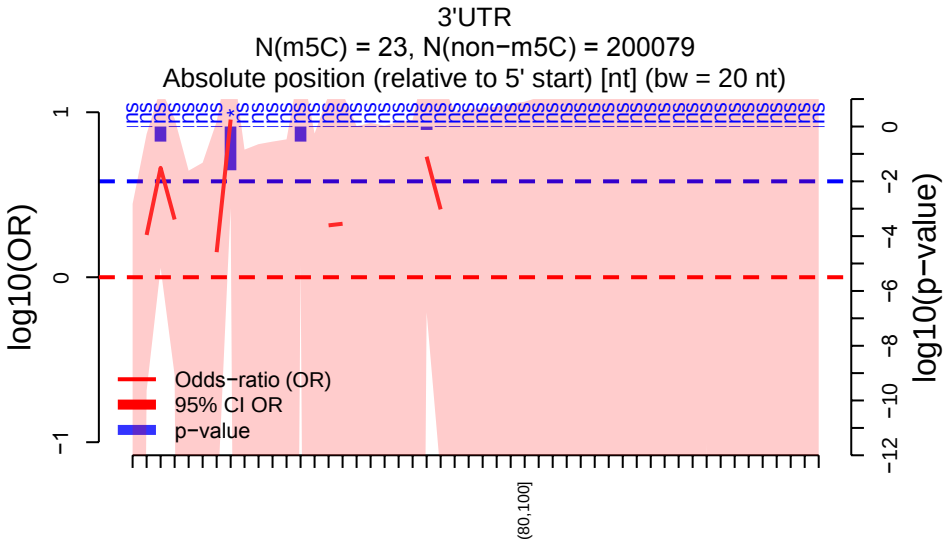
6. 7.7 List of identified m^{5}C sites identified in the Illumina bsRNA-seq experiment.

This excel file shows the 1) identity of m^{5}C sites by chromosomal location, 2) strand information, 3) context of methylation, 4) total number reads (coverage), 5) number of non-methylated reads, 6) number of methylated reads 7) methylation rate, and 8) gene information including symbol, ID, start and stop coordinates, length and type.

7. 7.8 Full list of gene ontology terms obtained using the complete set of m^{5}C sites identified in mRNAs from the Illumina bsRNA-seq experiment.

The excel file shows the gene ontologies associated with the Illumina m^{5}C sites identified within mRNAs. Enriched terms classified by molecular function, biological process and cellular compartment are shown.

8. 7.9 Spatial distribution of m^{5}C sites in 3' UTR.



The plot is shown relative to the 5' end of the 3' UTR. The distribution of m^{5}C sites within the 3′ UTR is not significantly different from the distribution of unmethylated cytosines.

References

Abbasi-Moheb L, Mertel S, Gonsior M, Nouri-Vahid L, Kahrizi K, et al. 2012. Mutations in NSUN2 cause autosomal-recessive intellectual disability. *American Journal of Human Genetics* 90: 847-55

Alarcon CR, Lee H, Goodarzi H, Halberg N, Tavazoie SF. 2015. N6-methyladenosine marks primary microRNAs for processing. *Nature* 519: 482-5

Amort T, Souliere MF, Wille A, Jia XY, Fiegl H, et al. 2013. Long non-coding RNAs as targets for cytosine methylation. *RNA Biology* 10: 1003-8

Andersen NM, Douthwaite S. 2006. YebU is a m5C methyltransferase specific for 16 S rRNA nucleotide 1407. *Journal of Molecular Biology* 359: 777-86

Anderson KD, Morin MA, Beckel-Mitchener A, Mobarak CD, Neve RL, et al. 2000. Overexpression of HuD, but not of its truncated form HuD I+II, promotes GAP-43 gene expression and neurite outgrowth in PC12 cells in the absence of nerve growth factor. *Journal of Neurochemistry* 75: 1103-14

Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, et al. 2000. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nature Genetics* 25: 25-9

Auxilien S, Guerineau V, Szweykowska-Kulinska Z, Golinelli-Pimpaneau B. 2012. The human tRNA m (5) C methyltransferase Misu is multisite-specific. *RNA Biology* 9: 1331-8

Babendure JR, Babendure JL, Ding J, Tsien RY. 2006. Control of mammalian translation by mRNA structure near caps. *RNA* 5: 851-861.

Baer R, Dubin DT. 1980. The 3'-terminal sequence of the small subunit ribosomal RNA from hamster mitochondria. *Nucleic Acids Research* 8: 4927-41

Bailey TL. 2011. DREME: motif discovery in transcription factor ChIP-seq data. *Bioinformatics* 27: 1653-9

Bailey TL, Williams N, Misleh C, Li WW. 2006. MEME: discovering and analyzing DNA and protein sequence motifs. *Nucleic Acids Research* 34: W369-73

Baker B, Muckenthaler M, Vives E, Blanchard A, Braddock M, et al. 1994. Identification of a novel HIV-1 TAR RNA bulge binding protein. *Nucleic Acids Research* 22: 3365-72

Barrett LW, Fletcher S, Wilton SD. 2012. Regulation of eukaryotic gene expression by the untranslated gene regions and other non-coding elements. *Cellular and Molecular Life Sciences : CMLS* 69: 3613-34

Batista PJ, Molinie B, Wang J, Qu K, Zhang J, et al. m^{6}A RNA Modification Controls Cell Fate Transition in Mammalian Embryonic Stem Cells. *Cell Stem Cell* 15: 707-19

Beilharz TH, Humphreys DT, Clancy JL, Thermann R, Martin DI, et al. 2009. microRNA-mediated messenger RNA deadenylation contributes to translational repression in mammalian cells. *PloS One* 4: e6783

Bhan A, Mandal SS. 2015. LncRNA HOTAIR: A master regulator of chromatin dynamics and cancer. *Biochimica et Biophysica Acta* 1856: 151-64

Blanco S, Dietmann S, Flores JV, Hussain S, Kutter C, et al. 2014. Aberrant methylation of tRNAs links cellular stress to neuro-developmental disorders. *The EMBO Journal* 33: 2020-39

Blanco S, Frye M. 2014. Role of RNA methyltransferases in tissue renewal and pathology. *Current Opinion in Cell Biology* 31: 1-7

Blanco S, Kurowski A, Nichols J, Watt FM, Benitah SA, Frye M. 2011. The RNA-methyltransferase Misu (NSun2) poises epidermal stem cells to differentiate. *PLoS Genetics* 7: e1002403

Bodi Z, Bottley A, Archer N, May ST, Fray RG. 2015. Yeast m6A Methylated mRNAs Are Enriched on Translating Ribosomes during Meiosis, and under Rapamycin Treatment. *Plos One* 7: e0132090.

Bolger AM, Lohse M, Usadel B. 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* 30: 2114-20

Briggs JW, Ren L, Nguyen R, Chakrabarti K, Cassavaugh J, et al. 2012. The ezrin metastatic phenotype is associated with the initiation of protein translation. *Neoplasia* 14: 297-310

Bringmann P, Luhrmann R. 1987. Antibodies specific for N6-methyladenosine react with intact snRNPs U2 and U4/U6. *FEBS letters* 213: 309-15

Brzezicha B, Schmidt M, Makalowska I, Jarmolowski A, Pienkowska J, Szweykowska-Kulinska Z. 2006. 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 Research* 34: 6034-43

Bujnicki JM, Feder M, Ayres C, Redman KL. 2004. Sequence-structure-function studies of tRNA:m5C methyltransferase Trm4p and its relationship to DNA:m5C and RNA:m5C methyltransferases. *Nucleic Acids Research*: 2453- 63

Camara Y, Asin-Cayuela J, Park CB, Metodiev MD, Shi Y, et al. 2011. MTERF4 regulates translation by targeting the methyltransferase NSUN4 to the mammalian mitochondrial ribosome. *Cell Metabolism* 13: 527-39

Cantara WA, Crain PF, Rozenski J, McCloskey JA, Harris KA, et al. 2011. The RNA Modification Database, RNAMDB: 2011 update. *Nucleic Acids Research* 39: D195-201

Castello A, Fischer B, Eichelbaum K, Horos R, Beckmann BM, et al. 2012. Insights into RNA biology from an atlas of mammalian mRNA-binding proteins. *Cell* 149: 1393-406

Catania F, Lynch M. 2010. Evolutionary dynamics of a conserved sequence motif in the ribosomal genes of the ciliate Paramecium. *BMC Evolutionary Biology* 10: 129

Chan PP, Lowe TM. 2009. GtRNAdb: a database of transfer RNA genes detected in genomic sequence. *Nucleic Acids Research* 37: D93-7

Chatr-Aryamontri A, Breitkreutz BJ, Oughtred R, Boucher L, Heinicke S, et al. 2015. The BioGRID interaction database: 2015 update. *Nucleic Acids Research* 43: D470-8

Chen T, Hao YJ, Zhang Y, Li M, Wang M, et al. 2015. m6A RNA methylation Is Regulated by MicroRNAs and Promotes Reprogramming to Pluripotency. *Cell Stem Cell*: 289-301

Cui X, Meng J, Rao MK, Chen Y, Huang Y. 2015. HEPeak: an HMM-based exome peak-finding package for RNA epigenome sequencing data. *BMC Genomics* 16 Suppl 4: S2

Dandeker T and Bengert P. 2002. RNA Motifs and Regulatory Elements. *Springer* 2:1-233.

Decatur WA, Fournier MJ. 2002. rRNA modifications and ribosome function. *Trends in Biochemical Sciences* 27: 344-51

Delatte B, Wang F, Ngoc LV, Collignon E, Bonvin E, et al. 2016. RNA Biochemistry. Transcriptome-wide distribution and function of RNA hydroxymethylcytosine. *Science* 351: 282-5

Desai PM, Culver GM, Rife JP. 2011. Site-directed mutants of 16S rRNA reveal important RNA domains for KsgA function and 30S subunit assembly. *Biochemistry* 50: 854-63

Doll A, Grzeschik KH. 2001. Characterization of two novel genes, WBSCR20 and WBSCR22, deleted in Williams-Beuren syndrome. *Cytogenetics and Cell Genetics* 95: 20-7

Dominissini D, Moshitch-Moshkovitz S, Schwartz S, Salmon-Divon M, Ungar L, et al. 2012. Topology of the human and mouse m6A RNA methylomes revealed by m6A-seq. *Nature* 485: 201-6

Dominissini D, Nachtergaele S, Moshitch-Moshkovitz S, Peer E, Kol N, et al. 2016. The dynamic N(1)-methyladenosine methylome in eukaryotic messenger RNA. *Nature* 530: 441-6

Dubin DT. 1974. Methylated nucleotide content of mitochondrial ribosomal RNA from hamster cells. *Journal of Molecular Biology* 84: 257-73

Dunn DB. 1961. The occurrence of 1-methyladenine in ribonucleic acid. *Biochimica et Biophysica Acta* 46: 198-200

Edelheit S, Schwartz S, Mumbach MR, Wurtzel O, Sorek R. 2013. Transcriptome-wide mapping of 5-methylcytidine RNA modifications in bacteria, archaea, and yeast reveals m5C within archaeal mRNAs. *PLoS Genetics* 9: e1003602

El Yacoubi B, Bailly M, de Crecy-Lagard V. 2012. Biosynthesis and function of posttranscriptional modifications of transfer RNAs. *Annual Review of Genetics* 46: 69-95

Eulalio A, Huntzinger E, Izaurralde E. 2008. GW182 interaction with Argonaute is essential for miRNA-mediated translational repression and mRNA decay. *Nature Structural & Molecular Biology* 15: 346-53

Fehon RG, McClatchey AI, Bretscher A. 2010. Organizing the cell cortex: the roel of ERM proteins. *Nature Reviews Molecular Cell Biology* 11:276-287.

Frye M, Jaffrey SR, Pan T, Rechavi G, Suzuki T. 2016. RNA modifications: what have we learned and where are we headed? *Nature reviews. Genetics* 17: 365-72

Frye M, Watt FM. 2006. The RNA methyltransferase Misu (NSun2) mediates Myc-induced proliferation and is upregulated in tumors. *Current Biology : CB* 16: 971-81

Fu Y, Dominissini D, Rechavi G, He C. 2014. Gene expression regulation mediated through reversible m(6)A RNA methylation. *Nature reviews. Genetics* 15: 293-306

Goll MG, Kirpekar F, Maggert KA, Yoder JA, Hsieh CL, et al. 2006. Methylation of tRNAAsp by the DNA methyltransferase homolog Dnmt2. *Science* 311: 395-8

Graff JR, Zimmer SG. 2003. Translational control and metastatic progression: enhanced activity of the mRNA cap-binding protein eIF-4E selectively enhances translation of metastasis-related mRNAs. *Clinical & Experimental Metastasis* 20: 265-73

Gruber AR, Lorenz R, Bernhart SH, Neubock R, Hofacker IL. 2008. The Vienna RNA websuite. *Nucleic Acids Research* 36: W70-4

Gustilo EM, Vendeix FA, Agris PF. 2008. tRNA's modifications bring order to gene expression. *Current Opinion in Microbiology* 11: 134-40

Gutschner T, Hammerle M, Eissmann M, Hsu J, Kim Y, et al. 2013. The noncoding RNA MALAT1 is a critical regulator of the metastasis phenotype of lung cancer cells. *Cancer Research* 73: 1180-9

Ha M, Kim VN. 2014. Regulation of microRNA biogenesis. *Nature reviews. Molecular Cell Biology* 15: 509-24

Haag S, Warda AS, Kretschmer J, Gunnigmann MA, Hobartner C, Bohnsack MT. 2015. NSUN6 is a human RNA methyltransferase that catalyzes formation of m5C72 in specific tRNAs. *RNA* 21: 1532-43

Haag S, Sloan KE, Ranian N, Warda AS, Kretschmer J, et al. 2016. NSUN3 and ABH1 modify the wobble position of mt-tRNAMet to expand codon recognition in mitochondrial translation. *EMBO Journal* 4: 2104-2119

Hafner M, Landthaler M, Burger L, Khorshid M, Hausser J, et al. 2010. Transcriptome-wide identification of RNA-binding protein and microRNA target sites by PAR-CLIP. *Cell* 141: 129-41

Harris T, Marquez B, Suarez S, Schimenti J. 2007. 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. *Biology of Reproduction* 77: 376-82

Hoernes TP, Clementi N, Faserl K, Glasner H, Breuker K, et al. 2016. Nucleotide modifications within bacterial messenger RNAs regulate their translation and are able to rewire the genetic code. *Nucleic Acids Research* 44: 852-62

Hong B, Brockenbrough JS, Wu P, Aris JP. 1997. Nop2p is required for pre-rRNA processing and 60S ribosome subunit synthesis in yeast. *Molecular and Cellular Biology* 17: 378-88

Hsieh AC, Ruggero D. 2010. Targeting eukaryotic translation initiation factor 4E (eIF4E) in cancer. *Clinical cancer research : an official journal of the American Association for Cancer Research* 16: 4914-20

Huang DW, Sherman BT, Lempicki RA. 2009. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nature Protocols* 4: 44-57

Huang DW, Sherman BT, Tan Q, Collins JR, Alvord WG, et al. 2007. The DAVID Gene Functional Classification Tool: a novel biological module-centric algorithm to functionally analyze large gene lists. *Genome Biology* 8: R183

Huber SM, van Delft P, Mendil L, Bachman M, Smollett K, et al. 2015. Formation and abundance of 5-hydroxymethylcytosine in RNA. *Chembiochem : a European Journal of Chemical Biology* 16: 752-5

Huntzinger E, Izaurralde E. 2011. Gene silencing by microRNAs: contributions of translational repression and mRNA decay. *Nature Reviews. Genetics* 12: 99- 110

Hussain S, Benavente S, Nascimento E, Dragoni I, Kurowski A, et al. 2009. The nucleolar RNA methyltransferase Misu (NSun2) is required for mitotic spindle stability. *Journal of Cell Biology.*: 27-40

Hussain S, Sajini AA, Blanco S, Dietmann S, Lombard P, et al. 2013. NSun2-mediated cytosine-5 methylation of vault noncoding RNA determines its processing into regulatory small RNAs. *Cell Reports* 4: 255-61

Jaksik R, Rzeszowska-Wolny J. 2012. The distribution of GC nucleotides and regulatory sequence motifs in genes and their adjacent sequences. *Gene* 492: 375-81

Jansen RP. 2001. mRNA localization: message on the move. *Nature Reviews Molecular Cell Biology* 2:247-256.

Jia G, Fu Y, Zhao X, Dai Q, Zheng G, et al. 2011. N6-methyladenosine in nuclear RNA is a major substrate of the obesity-associated FTO. *Nature Chemical Biology* 7: 885-7

Jurkowski TP, Meusburger M, Phalke S, Helm M, Nellen W, et al. 2008. Human DNMT2 methylates tRNA(Asp) molecules using a DNA methyltransferase-like catalytic mechanism. *RNA* 14: 1663-70

Kaushal PS, Sharma MR, Booth TM, Haque EM, Tung CS, et al. 2014. Cryo-EM structure of the small subunit of the mammalian mitochondrial ribosome. *Proceedings of the National Academy of Sciences of the United States of America* 111: 7284-9

Khoddami, V. (2013). 'Transcriptome-wide RNA Modification Profiling', PhD thesis, University of Utah.

Khoddami V, Cairns BR. 2013. Identification of direct targets and modified bases of RNA cytosine methyltransferases. *Nature Biotechnology* 31: 458-64
Kim S, Coulombe PA. 2010. Emerging role for the cytoskeleton as an organizer and regulator of translation. *Nature Reviews. Molecular cell biology* 11: 75-81
Kiran A, Baranov PV. 2010. DARNED: a DAtabase of RNa EDiting in humans. *Bioinformatics* 26: 1772-6
Kirchner S, Ignatova Z. 2015. Emerging roles of tRNA in adaptive translation, signalling dynamics and disease. *Nature Reviews. Genetics* 16: 98-112
Kirino Y, Suzuki T. 2005. Human mitochondrial diseases associated with tRNA wobble modification deficiency. *RNA Biology* 2: 41-4
Konicek BW, Dumstorf CA, Graff JR. 2008. Targeting the eIF4F translation initiation complex for cancer therapy. *Cell Cycle* 7: 2466-71
Krol J, Loedige I, Filipowicz W. 2010. The widespread regulation of microRNA biogenesis, function and decay. *Nature Reviews. Genetics* 11: 597-610
Krueger F, Andrews SR. 2011. Bismark: a flexible aligner and methylation caller for Bisulfite-Seq applications. *Bioinformatics* 27: 1571-2
Kullmann M, Gopfert U, Siewe B, Hengst L. 2002. ELAV/Hu proteins inhibit p27 translation via an IRES element in the p27 5'UTR. *Genes & Development* 16: 3087-99
Lascaro D, Castellana S, Gasparre G, Romeo G, Saccone C, Attimonelli M. 2008. The RHNumtS compilation: features and bioinformatics approaches to locate and quantify Human NumtS. *BMC Genomics* 9: 267
Lebedeva S, Jens M, Theil K, Schwanhausser B, Selbach M, et al. 2011. Transcriptome-wide analysis of regulatory interactions of the RNA-binding protein HuR. *Molecular Cell* 43: 340-52
Lee S, Rose S, Metodiev MD, Becker L, Vernaleken A, et al. 2015. Overexpression of the mitochondrial methyltransferase TFB1M in the mouse does not impact mitoribosomal methylation status or hearing. *Human Molecular Genetics* 24: 7286-94
Levanon EY, Eisenberg E, Yelin R, Nemzer S, Hallegger M, et al. 2004. Systematic identification of abundant A-to-I editing sites in the human transcriptome. *Nature Biotechnology* 22: 1001-5
Li X, Xiong X, Wang K, Wang L, Shu X, et al. 2016. Transcriptome-wide mapping reveals reversible and dynamic N(1)-methyladenosine methylome. *Nature chemical biology* 12: 311-6
Li X, Zhu P, Ma S, Song J, Bai J, et al. 2015. Chemical pulldown reveals dynamic pseudouridylation of the mammalian transcriptome. *Nature Chemical Biology* 11: 592-7
Lin S, Choe J, Du P, Triboulet R, Gregory RI. 2016. The m(6)A Methyltransferase METTL3 Promotes Translation in Human Cancer Cells. *Molecular Cell* 62: 335-45
Linder B, Grozhik AV, Olarerin-George AO, Meydan C, Mason CE, Jaffrey SR. 2015. Single-nucleotide-resolution mapping of m6A and m6Am throughout the transcriptome. *Nature Methods* 12: 767-72
Liu J, Yue Y, Han D, Wang X, Fu Y, et al. 2014. A METTL3-METTL14 complex mediates mammalian nuclear RNA N6-adenosine methylation. *Nature Chemical Biology* 10: 93-5
Liu N, Dai Q, Zheng G, He C, Parisien M, Pan T. 2015. N(6)-methyladenosine-dependent RNA structural switches regulate RNA-protein interactions. *Nature* 518: 560-4

Liu N, Parisien M, Dai Q, Zheng G, He C, Pan T. 2013. Probing N6-methyladenosine RNA modification status at single nucleotide resolution in mRNA and long noncoding RNA. *RNA* 19: 1848-56

Livingstone M, Atas E, Meller A, Sonenberg N. 2010. Mechanisms governing the control of mRNA translation. *Physical Biology* 7: 021001

Louie E, Ott J, Majewski J. 2003. Nucleotide frequency variation across human genes. *Genome Research* 13: 2594-601

Luo G, Macqueen A, Zheng G, Duan H, Dore LC, Lu Z, Liu J, Chen K, Jia G, Bergelson J, He C. 2014. Unique features of the m6A methylome in Arabidopsis thaliana. *Nature Communications* 5:5630.

Machnicka MA, Milanowska K, Osman Oglou O, Purta E, Kurkowska M, et al. 2013. MODOMICS: a database of RNA modification pathways--2013 update. *Nucleic Acids Research* 41: D262-7

Mercer TR, Neph S, Dinger ME, Crawford J, Smith MA, et al. 2011. The human mitochondrial transcriptome. *Cell* 146: 645-58

Metodiev MD, Lesko N, Park CB, Camara Y, Shi Y, et al. 2009. Methylation of 12S rRNA is necessary for in vivo stability of the small subunit of the mammalian mitochondrial ribosome. *Cell Metabolism* 9: 386-97

Metodiev MD, Spahr H, Loguercio Polosa P, Meharg C, Becker C, et al. 2014. NSUN4 is a dual function mitochondrial protein required for both methylation of 12S rRNA and coordination of mitoribosomal assembly. *PLoS Genetics* 10: e1004110

Meyer KD, Jaffrey SR. 2014. The dynamic epitranscriptome: N6-methyladenosine and gene expression control. *Nature reviews. Molecular Cell Biology* 15: 313- 26

Meyer KD, Patil DP, Zhou J, Zinoviev A, Skabkin MA, et al. 2015. 5' UTR m(6)A Promotes Cap-Independent Translation. *Cell* 163: 999-1010

Meyer KD, Saletore Y, Zumbo P, Elemento O, Mason CE, Jaffrey SR. 2012. Comprehensive analysis of mRNA methylation reveals enrichment in 3' UTRs and near stop codons. *Cell* 149: 1635-46

Mignone F, Gissi C, Liuni S, Pesole G. 2002. Untranslated regions of mRNAs. *Genome Biology* 3: REVIEWS0004

Mobarak CD, Anderson KD, Morin M, Beckel-Mitchener A, Rogers SL, et al. 2000. The RNA-binding protein HuD is required for GAP-43 mRNA stability, GAP-43 gene expression, and PKC-dependent neurite outgrowth in PC12 cells. *Molecular Biology of the Cell* 11: 3191-203

Moon HJ, Redman KL. 2014. Trm4 and Nsun2 RNA:m5C methyltransferases form metabolite-dependent, covalent adducts with previously methylated RNA. *Biochemistry* 53: 7132-44

Motorin Y, Grosjean H. 1999. Multisite-specific tRNA:m5C-methyltransferase (Trm4) in yeast Saccharomyces cerevisiae: identification of the gene and substrate specificity of the enzyme. *RNA* 5: 1105-18

Motorin Y, Helm M. 2010. tRNA stabilization by modified nucleotides. *Biochemistry* 49: 4934-44

Motorin Y, Lyko F, Helm M. 2010. 5-methylcytosine in RNA: detection, enzymatic formation and biological functions. *Nucleic Acids Research* 38: 1415-30

Mourier T, Hansen AJ, Willerslev E, Arctander P. 2001. The Human Genome Project reveals a continuous transfer of large mitochondrial fragments to the nucleus. *Molecular Biology and Evolution* 18: 1833-7

Nagaike T, Suzuki T, Ueda T. 2008. Polyadenylation in mammalian mitochondria: insights from recent studies. *Biochimica et Biophysica Acta* 1779: 266-9

Nakano S, Suzuki T, Kawarada L, Iwata H, Asano K, Suzuki T. 2016. NSUN3 methylase initiates 5-formylcytidine biogenesis in human mitochondrial tRNA(Met). *Nature Chemical Biology* 12: 546-51

Noon KR, Bruenger E, McCloskey JA. 1998. Posttranscriptional modifications in 16S and 23S rRNAs of the archaeal hyperthermophile Sulfolobus solfataricus. *Journal of Bacteriology* 180: 2883-8

Ondov BD, Varadarajan A, Passalacqua KD, Bergman NH. 2008. Efficient mapping of Applied Biosystems SOLiD sequence data to a reference genome for functional genomic applications. *Bioinformatics* 24: 2776-7

Parr RL, Maki J, Reguly B, Dakubo GD, Aguirre A, et al. 2006. The pseudo-mitochondrial genome influences mistakes in heteroplasmy interpretation. *BMC Genomics* 7: 185

Pasquinelli AE. 2012. MicroRNAs and their targets: recognition, regulation and an emerging reciprocal relationship. *Nature Reviews. Genetics* 13: 271-82

Perlman S, Abelson HT, Penman S. 1973. Mitochondrial protein synthesis: RNA with the properties of Eukaryotic messenger RNA. *Proceedings of the National Academy of Sciences of the United States of America* 70: 350-3

Persson H, Kvist A, Vallon-Christersson J, Medstrand P, Borg A, Rovira C. 2009. The non-coding RNA of the multidrug resistance-linked vault particle encodes multiple regulatory small RNAs. *Nature Cell Biology* 11: 1268-71

Piechota J, Tomecki R, Gewartowski K, Szczesny R, Dmochowska A, et al. 2006. Differential stability of mitochondrial mRNA in HeLa cells. *Acta Biochimica Polonica* 53: 157-68

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

Piekna-Przybylska D, Przybylski P, Baudin-Baillieu A, Rousset JP, Fournier MJ. 2008b. Ribosome performance is enhanced by a rich cluster of pseudouridines in the A-site finger region of the large subunit. *The Journal of Biological Chemistry* 283: 26026-36

Ping XL, Sun BF, Wang L, Xiao W, Yang X, et al. 2014. Mammalian WTAP is a regulatory subunit of the RNA N6-methyladenosine methyltransferase. *Cell Research* 24: 177-89

Polak P, Oren A, Ben-Dror I, Steinberg D, Sapoznik S, et al. 2006. The cytoskeletal network controls c-Jun translation in a UTR-dependent manner. *Oncogene* 25: 665-76

Reid R, Greene PJ, Santi DV. 1999. Exposition of a family of RNA m(5)C methyltransferases from searching genomic and proteomic sequences. *Nucleic Acids Research* 27: 3138-45

Rhee HW, Zou P, Udeshi ND, Martell JD, Mootha VK, et al. 2013. Proteomic mapping of mitochondria in living cells via spatially restricted enzymatic tagging. *Science* 339: 1328-31

Rodriguez JM, Maietta P, Ezkurdia I, Pietrelli A, Wesselink JJ, et al. 2013. APPRIS: annotation of principal and alternative splice isoforms. *Nucleic Acids Research* 41: D110-7

Rorbach J, Minczuk M. 2012. The post-transcriptional life of mammalian mitochondrial RNA. *The Biochemical Journal* 444: 357-73

Rossmanith. 2012. Of P and Z: Mitochondrial tRNA processing enzymes. *Biochimica et Biophysica Acta* 1819: 1017-26.

Rottman F, Shatkin AJ, Perry RP. 1974. Sequences containing methylated nucleotides at the 5' termini of messenger RNAs: possible implications for processing. *Cell* 3: 197-9

Sakita-Suto S, Kanda A, Suzuki F, Sato S, Takata T, Tatsuka M. 2007. Aurora-B regulates RNA methyltransferase NSUN2. *Molecular Biology of the Cell* 18: 1107-17

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

Saneyoshi M, Harada F, Nishimura S. 1969. Isolation and characterization of N6-methyladenosine from Escherichia coli valine transfer RNA. *Biochimica et Biophysica Acta* 190: 264-73

Santos MC, Goldfeder MB, Zanchin NI, Oliveira CC. 2011. The essential nucleolar yeast protein Nop8p controls the exosome function during 60S ribosomal subunit maturation. *PloS One* 6: e21686

Schaefer M, Hagemann S, Hanna K, Lyko F. 2009a. Azacytidine inhibits RNA methylation at DNMT2 target sites in human cancer cell lines. *Cancer Research* 69: 8127-32

Schaefer M, Pollex T, Hanna K, Lyko F. 2009b. RNA cytosine methylation analysis by bisulfite sequencing. *Nucleic Acids Research* 37: e12

Schaefer M, Pollex T, Hanna K, Tuorto F, Meusburger M, et al. 2010. RNA methylation by Dnmt2 protects transfer RNAs against stress-induced cleavage. *Genes & Development* 24: 1590-5

Schneider CA, Rasband WS, Eliceiri KW. 2012. NIH Image to ImageJ: 25 years of image analysis. *Nature Methods* 9: 671-5

Schosserer M, Minois N, Angerer TB, Amring M, Dellago H, et al. 2016. Methylation of ribosomal RNA by NSUN5 is a conserved mechanism modulating organismal lifespan. *Nature Communications* 6: 1-17

Schumann U, Shafik A, Preiss T. 2016. METTL3 Gains R/W Access to the Epitranscriptome. *Molecular Cell* 62: 323-4

Shafik A, Schumann U, Evers M, Sibbritt T, Preiss T. 2016. The emerging epitranscriptomics of long noncoding RNAs. *Biochimica et Biophysica Acta* 1859: 59-70

Shanmugam R, Fierer J, Kaiser S, Helm M, Jurkowski TP, Jeltsch A. 2015. Cytosine methylation of tRNA-Asp by DNMT2 has a role in translation of proteins containing poly-Asp sequences. *Cell Discovery* 1: 15010

Sharma S, Yang J, Watzinger P, Kotter P, Entian KD. 2013. Yeast Nop2 and Rcm1 methylate C2870 and C2278 of the 25S rRNA, respectively. *Nucleic Acids Research* 41: 9062-76

Shen L, Liang Z, Gu X, Dedon PC, Liu L et al. 2016. N6-methyladenosine RNA modification regulates Shoot Stem Cell Fate in *Arabidopsis*. *Developmental Cell* 38: 1-15

Sherry ST, Ward MH, Kholodov M, Baker J, Phan L, et al. 2001. dbSNP: the NCBI database of genetic variation. *Nucleic Acids Research* 29: 308-11

Sibbritt, T. (2015). 'Investigating the Enzymology and Molecular Function of 5-methylcytosine in RNA', PhD thesis, Australian National University

Sibbritt T, Shafik A, Clark SJ, Preiss T. 2016. Nucleotide-Level Profiling of m(5)C RNA Methylation. *Methods in Molecular Biology* 1358: 269-84

Silvera D, Formenti SC, Schneider RJ. 2010. Translational control in cancer. *Nature reviews. Cancer* 10: 254-66

Slomovic S, Laufer D, Geiger D, Schuster G. 2006. Polyadenylation of ribosomal RNA in human cells. *Nucleic Acids Research* 34: 2966-75

Squires JE, Patel HR, Nousch M, Sibbritt T, Humphreys DT, et al. 2012. Widespread

occurrence of 5-methylcytosine in human coding and non-coding RNA. *Nucleic Acids Research* 40: 5023-33
Srivastava R, Gopinathan KP. 1987. Ribosomal RNA methylation in Mycobacterium smegmatis SN2. *Biochemistry International* 15: 1179-88

Suzuki T, Nagao A, Suzuki T. 2011. Human mitochondrial tRNAs: biogenesis, function, structural aspects, and diseases. *Annual Review of Genetics* 45: 299- 329
Suzuki T, Suzuki T. 2014. A complete landscape of post-transcriptional modifications in mammalian mitochondrial tRNAs. *Nucleic Acids Research* 42: 7346-57
Tang H, Fan X, Xing J, Liu Z, Jiang B, et al. 2015. NSun2 delays replicative senescence by repressing p27 (KIP1) translation and elevating CDK1 translation. *Aging* 7: 1143-58
Tress ML, Wesselink JJ, Frankish A, Lopez G, Goldman N, et al. 2008. Determination and validation of principal gene products. *Bioinformatics* 24: 11-17
Tuorto F, Liebers R, Musch T, Schaefer M, Hofmann S, et al. 2012. RNA cytosine methylation by Dnmt2 and NSun2 promotes tRNA stability and protein synthesis. *Nature Structural & Molecular Biology* 19: 900-5
van der Velden AW, Thomas AA. 1999. The role of the 5' untranslated region of an mRNA in translation regulation during development. *International Journal of Biochemistry and Cell Biology* 31:87-106.
Van Haute L, Dietmann S, Kremer L, Hussain S, Pearce SF, et al. 2016. Deficient methylation and formylation of mt-tRNA(Met) wobble cytosine in a patient carrying mutations in NSUN3. *Nature Communications* 7:1-10
Wang X, Lu Z, Gomez A, Hon GC, Yue Y, et al. 2014a. N6-methyladenosine-dependent regulation of messenger RNA stability. *Nature* 505: 117-20
Wang X, Zhao BS, Roundtree IA, Lu Z, Han D, et al. 2015. N(6)-methyladenosine Modulates Messenger RNA Translation Efficiency. *Cell* 161: 1388-99
Wang Y, Li Y, Toth JI, Petroski MD, Zhang Z, Zhao JC. 2014b. N6-methyladenosine modification destabilizes developmental regulators in embryonic stem cells. *Nature Cell Biology* 16: 191-8
Wei CM, Gershowitz A, Moss B. 1976. 5'-Terminal and internal methylated nucleotide sequences in HeLa cell mRNA. *Biochemistry* 15: 397-401
Wutz A. 2011. Gene silencing in X-chromosome inactivation: advances in understanding facultative heterochromatin formation. *Nature Reviews. Genetics* 12: 542-53
Xiao W, Adhikari S, Dahal U, Chen YS, Hao YJ, et al. 2016. Nuclear m(6)A Reader YTHDC1 Regulates mRNA Splicing. *Molecular cell* 61: 507-19
Xing J, Yi J, Cai X, Tang H, Liu Z, et al. 2015. NSun2 Promotes Cell Growth via Elevating Cyclin-Dependent Kinase 1 Translation. *Molecular and Cellular Biology* 35: 4043-52
Xu C, Wang X, Liu K, Roundtree IA, Tempel W, et al. 2014. Structural basis for selective binding of m6A RNA by the YTHDC1 YTH domain. *Nature Chemical Biology* 10: 927-9
Yakubovskaya E, Guja KE, Mejia E, Castano S, Hambardjieva E, et al. 2012. Structure of the essential MTERF4:NSUN4 protein complex reveals how an MTERF protein collaborates to facilitate rRNA modification. *Structure* 20: 1940-7
Yamasaki S, Ivanov P, Hu G, Anderson P. 2009. Angiogenin cleaves tRNA and promotes stress-induced translational repression. *Journal of Cell Biology* 1: 35-42
Ye J, Coulouris G, Zaretskaya I, Cutcutache I, Rozen S, Madden TL. 2012. Primer-

BLAST: a tool to design target-specific primers for polymerase chain reaction. *BMC Bioinformatics* 13: 134

Zhang BH, Pan XP, Cox SB, Cobb GP, Anderson TA. 2006. Evidence that miRNAs are different from other RNAs. *Cellular and Molecular Life Sciences* 63: 246-54

Zhang X, Liu Z, Yi J, Tang H, Xing J, et al. 2012. The tRNA methyltransferase NSun2 stabilizes p16INK(4) mRNA by methylating the 3'-untranslated region of p16. *Nature Communications* 3: 712

Zhao J, Sun BK, Erwin JA, Song JJ, Lee JT. 2008. Polycomb proteins targeted by a short repeat RNA to the mouse X chromosome. *Science* 322: 750-6

Zheng G, Dahl JA, Niu Y, Fedorcsak P, Huang CM, et al. 2013. ALKBH5 is a mammalian RNA demethylase that impacts RNA metabolism and mouse fertility. *Molecular Cell* 49: 18-29

Zheng Y, Miskimins WK. 2011. CUG-binding protein represses translation of p27Kip1 mRNA through its internal ribosomal entry site. *RNA Biology* 8: 365-71

Zhou J, Wan J, Gao X, Zhang X, Jaffrey SR, Qian SB. 2015. Dynamic m(6)A mRNA methylation directs translational control of heat shock response. *Nature* 526: 591-4