INVESTIGATING THE ENZYMOLOGY AND MOLECULAR FUNCTION OF 5-METHYLCYTOSINE IN RNA

Tennille Sibbritt

February 2015

A thesis submitted for the degree of

Doctor of Philosophy

of 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

Wow, I cannot believe I have finished! It feels so surreal! It has been a tough journey but it was definitely worth it in the end. I mentioned to several people that there have been only a few moments in my life where I have genuinely been proud of myself, and right now is one of those moments. Not only because I managed to persist with this and finish, but because I also managed to move interstate during the beginning of my PhD, organise a wedding and the purchase a block of land in Sydney, and continuously travel back and forth between Sydney and Canberra to see my loved ones. I certainly learnt how to manage my time brilliantly during this journey. I also had so many people who helped me out throughout the course of my PhD to make all of this possible.

I wanted to start off by thanking my supervisor, Thomas Preiss, for your guidance, support and patience. I have really enjoyed our discussions and I'm so thankful for the guidance you have provided me. A big thank you for reading my thesis over and over again! I also wanted to thank my co-supervisor, Susan Clark, for your support, guidance and for having me in your lab for four months.

To Brian Parker, thank you for all your help with the bioinformatics in this thesis and for putting up with my persistence. Without you I wouldn't have been able to complete this. I also wanted to thank Jiayu Wen for helping develop the statistical model used in this thesis. Thank you to Jeffrey Squires and David Humphreys for your help during the beginning of my PhD. You both taught me the skills required for this project and I genuinely appreciate it. Thank you to both Andrew Shafik and Natalie Beveridge for taking some of the weight off me and helping me out with a few of the things in this thesis. To Terry Neeman, thank you so much for all your help and advice with the statistical analyses.

Thank you to everyone from the Clark lab, in particular Clare Stirzaker (for calming me down when things went wrong), as well as Jenny Song, Aaron Statham and Wenjia Qu for your help with the MiSeq. Thank you also to Brian Gloss for helping me out with R.

To the former and current members of the Preiss lab who I have not yet mentioned, I'd like to thank you: Stuart Archer (for your advice, patience and your sense of humour),

Hardip Patel (for the bioinformatics earlier in my PhD), Grace Wei, Jennifer Clancy, Rina Soetanto, Tony Duan and Hao Yang. To Eloisa Pagler, Yalin Liao and Renae Domaschenz, thank you for listening to me, for your help, advice and coffee/lunch breaks.

To Carly Hynes, thank you for everything, particularly your friendship, patience and advice. You have been there for me throughout the course of my PhD including moving down with me to Canberra, living with me, working with me, going to the gym/bootcamp with me, holidaying with me- we're practically married! I'm so lucky to have a friend like you and I don't think I could have finished this without you. To Nicola Smith, thank you so much for being there for me throughout the course of my PhD, including reading several of my thesis chapters. Your advice and friendship has been absolutely invaluable and I don't know how to repay you. To my demonstrating buddy, Piyankarie Jayatilaka, thank you for your friendship and the coffee breaks.

Finally, to my family and friends who I have not yet mentioned and have been there for me for the past four years and longer. Thank you so much for being so understanding, particularly when I moved interstate and I've been too busy to be social. Thank you to my Dad for his support. A really big thank you to my Mum and sister, Sheena, for everything! You ladies are amazing and have been there for me every step of the way, including when I have broken down and cried over the last few months. I can't wait to be back up in Sydney to spend more time with all of you. To my amazing fiancé, Chris, thank you so much for reading my thesis and for being there for me every step of the way. I'm the luckiest woman in the world to have such a selfless, compassionate and understanding person in my life. You have been my absolute rock and backbone, and I know I couldn't have done this PhD without you. Your positive attitude is inspirational and I can't wait to spend the rest of my life with you. I love you. Bring on 2015 and the future! ☺

ABSTRACT

Internal nucleoside methylation of eukaryotic RNA in the form of N6-methyladenosine (m^6A) and 5-methylcytosine (m^5C) have been known to exist for decades, however absence of facile methods to map modified sites have limited the understanding of their role. With the availability of next-generation sequencing, these drawbacks have been overcome, revealing non-random distribution of internal methylation across different transcript biotypes. Recently, we implemented a bisulfite sequencing-based technique for transcriptome-wide detection of m^5C (bsRNA-seq) and mapped thousands of m^5C sites in the human transcriptome including in ncRNA and mRNA. Biased distribution of m^5C within mRNAs was evident, consistent with roles in post-transcriptional gene regulation. Two RNA methyltransferases are known to be active in humans, NSUN2, which is overexpressed in various tumours, and TRDMT1. The aims of this thesis were to determine the RNA targets of NSUN2 and TRDMT1, explore the molecular functions of m^5C in mRNA and compare RNA m^5C profiles of normal prostate cells (PrEC), prostate cancer cells (LNCaP) and cervical cancer cells (HeLa) using improved high-throughput methods.

RNAi-mediated knockdown of *NSUN2* and *TRDMT1* in HeLa cells coupled with bsRNA-seq revealed 4,241 candidate m^5C sites across tRNA, ncRNA and mRNA; 393 of these sites were called as NSUN2 targets and 120 as TRDMT1 targets. NSUN2 mediated the bulk of tRNA methylation, targeting several structural positions and a broad range of isotypes and isoacceptors, while TRDMT1 was specific to one structural position across four isoacceptors, consistent with previously published work. Interestingly, m^5C was evident in tRNAs prior to end processing and splicing, indicating that methylation occurs early during tRNA biogenesis. Five tRNAs exhibiting m^5C sites were independently validated by locus-specific bsRNA-sequencing. Importantly, five m^5C sites in ncRNAs and nine in mRNAs were independently validated as NSUN2 targets with NSUN2 exhibiting multisite specificity, indicating its importance as a ncRNA and mRNA modifying enzyme.

Enrichment of Ago2 footprints upstream of m^5C sites and depletion of miRNA target regions at m^5C sites was evident in the CDS and 3’ UTR, suggestive of regulatory interactions between m^5C and RISC. A subtle decrease in average steady-state levels of

Abstract

NSUN2-targeted mRNAs was evident in the NSUN2 knockdown sample, suggesting a role for m^5C in mRNA stability. Investigation of the correlation of m^5C with translation state of individual mRNAs revealed two examples with opposing trends, suggesting context-dependent effects of m^5C on translation.

Comparison of the RNA m^5C profiles of PrEC and LNCaP cells revealed 5,653 candidate m^5C sites in common across tRNA, ncRNA and mRNA, with few discernible differential m^5C sites. Inspection of differential sites suggested that many might have resulted from technical imperfections of bsRNA-seq, which was independently confirmed. A comparison of prostate cells with HeLa cells showed high concordance of m^5C patterns across all three cell lines, with independent examples validated, indicating that m^5C is conserved across multiple cellular contexts.

The discovery that m^5C is abundant in mRNA across different cellular contexts and correlative links between the presence of m^5C in mRNA and several aspects of post-transcriptional gene regulation highlight a functional importance for m^5C in mRNA.

CONFERENCES, PUBLICATIONS AND AWARDS

CONFERENCES

Oral Presentations:

Sibbritt T., Parker B. J., Humphreys D. T., Squires J. E., Beveridge N. J., Shafik A., Statham A. L., Patel H. R., Clark S. J. and Preiss T., Exploring the role of RNA cytosine methylation in post-transcriptional gene regulation and cancer, *ComBio, Canberra, ACT, Australia*, 2014

Sibbritt T., The association of RNA cytosine methylation with post-transcriptional regulation and cancer, *RNA Network of Australasia: RNA Society Seminar, Canberra, ACT, Australia*, 2013

Sibbritt T., Patel H. R, Humphreys D. T., Squires J. E., Clark S. J., Preiss T., Investigating the role of RNA methylation in post-transcriptional regulation using a prostate cancer cell model, *34th Annual Lorne Genome Conference, Lorne, Australia*, 2013

Sibbritt T., Patel H. R, Humphreys D. T., Squires J. E., Clark S. J., Preiss T., RNA methylation: A hidden mechanism of post-transcriptional control deregulated in cancer?, *ComBio, Adelaide, Australia*, 2012

Sibbritt T., Humphreys D. T., Squires, J. E., Patel H. R., Nousch M., Clark S. J., Preiss T., Investigating the function of 5-methylcytosine in RNA by next-generation sequencing, *Annual Epigenetics Highlights from NSW/ACT AEpiA node, Sydney, Australia*, 2011

Poster Presentations:

Sibbritt T., Parker B. J., Humphreys D. T, Squires J. E., Beveridge N. J., Shafik A., Statham A. L., Patel H. R., Clark S. J. and Preiss T., Investigating the enzymology and molecular function of 5-methylcytosine in RNA, *Joint Australia and Japan RNA Meeting, Sydney, NSW, Australia*, 2014

Sibbritt T., Parker B. J., Humphreys D. T, Squires J. E., Beveridge N. J., Shafik A., Statham A. L., Patel H. R., Clark S. J. and Preiss T., The mysteries of RNA methylation: exploring the enzymology and molecular function of 5-methylcytosine in RNA, *Post-transcriptional Gene Regulation Gordon Research Conference, Newport, Rhode Island, USA*, 2014

Sibbritt T., Parker B. J., Humphreys D. T., Patel H. R., Statham A. L., Clark S. J., Preiss T., Investigating the association of RNA cytosine methylation with post-transcriptional regulation and cancer, *ASMR New Investigator Forum, Canberra, ACT, Australia*, 2014

Sibbritt T., Parker B. J., Humphreys D. T., Patel H. R., Statham A. L., Clark S. J., Preiss T., Investigating the association of RNA cytosine methylation with post-transcriptional regulation and cancer, *35th Annual Lorne Genome Conference, Lorne, VIC, Australia*, 2014

Sibbritt T., Parker B. J., Patel H. R., Humphreys D. T., Squires J. E., Beveridge N. J., Clark S. J., Preiss T., RNA methylation: a mechanism for post-transcriptional control that is deregulated in cancer?, *18th Annual meeting of the RNA Society, Davos, Switzerland*, 2013

Sibbritt T., Humphreys D. T., Squires, J. E., Patel H. R., Nousch M., Clark S. J., Preiss T., Investigating the function of 5-methylcytosine in RNA by next-generation sequencing, *ASMR New Investigator Forum, Canberra, Australia*, 2012

Sibbritt T., Humphreys D. T., Squires, J. E., Patel H. R., Nousch M., Clark S. J., Preiss T., Investigating the function of 5-methylcytosine in RNA by next-generation sequencing, *33rd Annual Lorne Genome Conference, Lorne, Australia*, 2012

Other presentations:

Guest speaker, Year 12 Career Development Session, Canberra Girls' Grammar School, Canberra, ACT, Australia, 2013

PUBLICATIONS

Sibbritt T., Patel H. R., Preiss T., Mapping and significance of the mRNA methylome, *WIREs RNA*, 2013, **4**: 397-422.

Squires J. E., Patel H. R., Nousch M., **Sibbritt T.**, Humphreys D. T., Parker B. J., Suter C. M., Preiss T., Widespread occurrence of 5-methylcytosine in human coding and non-coding RNA, *Nucleic Acids Res*., 2012, **40**: 5023-5033.

AWARDS

I was grateful to receive the following throughout my PhD studies:

Scholarships:

- The Australian Postgraduate Award, 2011-2014
- The Australian National University Supplementary Scholarship, 2011-2014

Conference Awards:

- Student Poster Prize, 35th Annual Lorne Genome Conference, Lorne, VIC, Australia, 2014
- Promega Student Award, 34th Annual Lorne Genome Conference, Lorne, VIC, Australia, 2013
- University of Canberra Best Poster Presentation Award, ASMR ACT New Investigator Forum, Canberra, ACT, Australia, 2012
- Student Poster Prize, 33rd Annual Lorne Genome Conference, Lorne, VIC, Australia, 2012

Travel Awards:

- Vice-Chancellor's Higher Degree Research Travel Grant to attend the 18th Annual Meeting of the RNA Society, Davos, Switzerland, 2013
- RNA 2013 Travel Fellowship to attend the 18th Annual Meeting of the RNA Society, Davos, Switzerland, 2013
- ASBMB Student Travel Bursary to attend ComBio, Adelaide, SA, Australia, 2012
- Travel Bursary Award to attend the Winter School in Mathematical and Computational Biology, Brisbane, QLD, Australia, 2012

COMMONLY USED ABBREVIATIONS

Ψ	Pseudouridine
5-aza-C	5-azacytidine
aa	Amino acid
AdoMet	S-adenosyl L-methionine
Ago	Argonaute
AlkB	Alkylated DNA repair protein B
ALKBH5	Alkylation repair homolog 5
Aza-IP	5-azacytidine-mediated RNA immunoprecipitation
bp	Base pair
BSA	Bovine serum albumin
bsRNA-seq	Sequencing of whole bisulfite-converted transcriptomes
cDNA	Complementary DNA
CDS	Coding sequence
CLIP	Crosslinking and immunoprecipitation
CPM	Counts per million
CTD	Carboxy-terminal domain
dH_2O	Distilled H_2O
DMSO	Dimethyl sulfoxide
DNA	Deoxyribonucleic acid
DNase	Deoxyribonuclease
DNMT	DNA methyltransferase
ds	Double stranded
DTT	Dithiothreitol
ECL	Enhanced chemiluminescence
EDTA	Ethylenediamine-tetraacetic acid
EtBr	Ethidium bromide
EtOH	Ethanol
FDR	False discovery rate
FTO	Fat mass and obesity associated
GO	Gene ontology
GSEA	Gene set enrichment analysis
GWAS	Genome-wide association studies
h	Hour
HAc	Acetic acid
HeLa	Human cervical cancer cells
HITS-CLIP	High-throughput sequencing of RNA isolated by CLIP
hm^5C	5-hydroxymethylcytosine
HuR	Human antigen 1
iCLIP	Individual-nucleotide resolution CLIP
IGV	Integrative Genomics Viewer
IP	Immunoprecipitation
KAc	Potassium acetate
kDa	Kilodalton
m^5C	5-methylcytosine
m^6A	N6-methyladenosine
KD	knockdown
lincRNA	Long intergenic non-coding RNA
LNCaP	lymph node metastatic prostate cancer cells
MeDIP	Methylated DNA immunoprecipitation
MeRIP	Methyl-RNA immunoprecipitation

METTL3	Methyltransferase like 3
METTL14	Methyltransferase like 14
miCLIP	Methylated iCLIP
min	Minute
miRNA	MicroRNA
mRNA	Messenger RNA
Myc	Avian Myelocytomatosis Viral Oncogene Homolog
NaAc	Sodium acetate
ncRNA	Non-coding RNA
Nop	Nucleolar protein
NSUN2	NOP2/Sun domain family, member 2
nt	Nucleotide
NTC	Non-targeting control
NTD	Amino-terminal domain
P body	Processing body
PAR-CLIP	Photoactivatable ribonucleoside-enhanced CLIP
PBS	Phosphate buffered saline
PCR	Polymerase chain reaction
Poly(A)	Polyadenylation
Pre-tRNA	Precursor tRNA
PrEC	Prostate epithelial cells
qPCR	Quantitative PCR
Q	Quality score
R-Luc	Renilla Luciferase
RBP	RNA binding protein
RCMT	RNA (cytosine-5)-methyltransferases
RFM	Rossmann-fold methyltransferase
RISC	RNA-induced silencing complex
RNA	Ribonucleic acid
RNA-seq	RNA sequencing
RNAi	RNA interference
RNase	Ribonuclease
rpm	Revolutions per minute
RPM	Reads per million
rRNA	Ribosomal RNA
RT	Reverse Transcriptase
RT-qPCR	Real time quantitative PCR
scaRNA	Small Cajal body-specific RNA
SD	Standard deviation
sec	Seconds
SEM	Standard error of the mean
siRNA	Small interfering RNA
SNP	Single nucleotide polymorphism
snRNA	Small nuclear RNA
snoRNA	Small nucleolar RNA
ss	Single stranded
svRNA	Small vault RNA
TET	ten-eleven translocation
TLR	Toll-like receptor
TRDMT1	tRNA aspartic acid methyltransferase 1
tRNA	Transfer RNA
uRNA	Ubiquitous RNA
UTR	Untranslated region
vtRNA	Vault RNA
WTAP	Wilms tumor 1 associated protein
X-gal	5-bromo-4-chloro-indolyl-β-D-galactopyranoside
YTHDF2	YTH domain family 2

LIST OF FIGURES

Figure 1.1. Schematic representation of known eukaryotic mRNA modifications. 3
Figure 1.2. Transcriptome-wide approaches to map m^6A and m^5C in RNA. 13
Figure 1.3. m^6A and m^5C writers, erasers and readers known to act on mRNA. 27
Figure 1.4. Biased distribution of m^6A and m^5C within mRNAs. 44
Figure 1.5. Reported roles for m^6A in mRNA and speculative roles of m^5C in mRNA. 46
Figure 2.1. Method for preparation of RNA for bsRNA-seq. 77
Figure 2.2. Whole-transcriptome library preparation method for bsRNA-seq. 79
Figure 2.3. TruSeq DNA LT sample preparation method. 88
Figure 3.1. Confirmation of methyltransferase knockdown in HeLa cells by RT-qPCR. 94
Figure 3.2. Assessment of rRNA depletion by the RNA 6000 Nano Chip and RT-qPCR. 97
Figure 3.3. Schematic demonstrating the bisulfite conversion of RNA. 98
Figure 3.4. Schematic demonstrating the bisulfite conversion of RNA, reverse transcription, primer design and PCR amplification. 100
Figure 3.5. Confirmation of efficient bisulfite conversion by Sanger bisulfite sequencing. 103
Figure 3.6. Confirmation of NSUN2 and TRDMT1 knockdown and bisulfite conversion efficiency by Sanger bisulfite sequencing. 104
Figure 3.7. Confirmation of fragmentation of RNA from bisulfite conversion. 106
Figure 3.8. Representative cDNA size selection of the bsRNA-seq libraries. 107
Figure 3.9. Size selection of amplified libraries by polyacrylamide gel electrophoresis. 109
Figure 3.10. Confirmation of the bsRNA-seq purity and library size. 110
Figure 3.11. Representative satay plot of the bsRNA-seq libraries for primer B cycle one. 112
Figure 3.12. SOLiD™ sequencing uses di-base encoding, where one colour represents two bases. 113
Figure 3.13. Schematic of the global and local mapping of 50 nt SOLiD™ sequencing reads by B-SOLANA. 116
Figure 3.14. Distribution of sequencing reads of the HeLa methyltransferase knockdown bsRNA-seq libraries. 119
Figure 3.15. Distribution of sequencing reads of the PrEC and LNCaP bsRNA-seq libraries. 120
Figure 3.16. Schematic demonstrating the detection of candidate m^5C sites. 124
Figure 3.17. Preparation of DNA and RNA dot blots for the detection of m^5C. 128
Figure 3.18. Dot blot assay for the detection of m^5C in DNA and RNA. 129
Figure 3.19. Dot blot assay for the detection of m^5C in DNA and RNA after denaturation in 2x RNA loading buffer. 130

Figure 3.20. Representative snapshot from IGV depicting a subset of sequencing reads mapping to $tRNA^{Leu(CAA)}$. 132
Figure 3.21. Preparation of PrEC and LNCaP cells for bsRNA-seq and MiSeq amplicon sequencing. 133
Figure 3.22. Preparation of RNAi-mediated knockdown of *NSUN2*, *TRDMT1* and *DNMT1* in HeLa cells for bsRNA-seq and MiSeq amplicon sequencing. 135
Figure 3.23. Confirmation of knockdown of *NSUN2*, *TRDMT1* and *DNMT1* in HeLa cells as biological replicates. 136
Figure 3.24. Validation of MiSeq amplicon libraries by agarose gel electrophoresis. 138
Figure 3.25. Schematic of the alignment of 150 bp paired-end reads from Illumina® MiSeq amplicon sequencing using Bismark. 141
Figure 3.26. Bisulfite conversion efficiency of the R-Luc *in vitro* transcripts assessed by bsRNA-seq and MiSeq amplicon sequencing of the HeLa methyltransferase knockdown samples. 145
Figure 3.27. Bisulfite conversion efficiency of the R-Luc *in vitro* transcripts assessed by bsRNA-seq and MiSeq amplicon sequencing of PrEC and LNCaP cells. 146
Figure 3.28. bsRNA-seq and MiSeq amplicon sequencing of $tRNA^{Asp(GUC)}$ and $tRNA^{Leu(CAA)}$ in the HeLa methyltransferase knockdown samples. 148
Figure 3.29. bsRNA-seq and MiSeq amplicon sequencing of $tRNA^{Asp(GUC)}$ and $tRNA^{Leu(CAA)}$ in PrEC and LNCaP cells. 149
Figure 3.30. Schematic displaying the issues associated with multiple m^5C sites within a single 50 nt SOLiD™ colour-space sequencing read. 154
Figure 4.1. Distribution of candidate m^5C sites detected in HeLa cells across different transcript biotypes. 168
Figure 4.2. The distribution of candidate m^5C sites across different transcript biotypes called as NSUN2 or TRDMT1 targets. 172
Figure 4.3. Contribution of NSUN2 and TRDMT1 to m^5C across transcript biotypes based on bsRNA-seq. 174
Figure 4.4. tRNA structural positions and isotypes exhibiting candidate m^5C sites called as NSUN2 and TRDMT1 targets. 178
Figure 4.5. tRNA isoacceptors exhibiting candidate m^5C sites called as NSUN2 and TRDMT1 targets. 181
Figure 4.6. Validation of candidate m^5C sites in tRNAs by MiSeq amplicon sequencing. 182
Figure 4.7. Representative snapshots from IGV depicting a subset of sequencing reads from the global tRNA mapping of $tRNA^{Leu(CAA)}$ and $tRNA^{Asp(GUC)}$ in the NTC sample. 186
Figure 4.8. Mean read coverage of all tRNAs containing candidate m^5C sites anchored at the 5' end of the mature tRNA in the NTC sample. 187

Figure 4.9. Mean read coverage of all tRNAs containing candidate m^{5}C sites anchored at the 3' end of the mature tRNA in the NTC sample. 188
Figure 4.10. The coverage and proportion of m^{5}C at known sites in 28 rRNA and 12S mitochondrial rRNA. 191
Figure 4.11. Validation of candidate m^{5}C sites in other ncRNA by MiSeq amplicon sequencing. 194
Figure 4.12. Secondary structures of ncRNAs with validated m^{5}C sites obtained either from the RNase P database, Rfam database or RNA STRAND. 197
Figure 4.13. The distribution and location bias of candidate m^{5}C sites in mRNA. 198
Figure 4.14. Validation of candidate m^{5}C sites in mRNA by MiSeq amplicon sequencing. 203
Figure 4.15. Significant structural motifs identified by CMFinder for candidate m^{5}C sites detected in mRNA. 206
Figure 4.16. Sequence logos surrounding candidate m^{5}C sites in mRNA and tRNA. 207
Figure 4.17. Comparison of the number of tRNAs exhibiting candidate m^{5}C sites called as NSUN2 targets by bsRNA-seq, miCLIP and Aza-IP. 213
Figure 4.18. Comparison of candidate m^{5}C sites in tRNAs and ncRNAs called as NSUN2 or TRDMT1 targets by bsRNA-seq and Aza-IP. 215
Figure 4.19. Comparison of candidate m^{5}C sites in tRNA structural positions called as NSUN2 or TRDMT1 targets by bsRNA-seq and Aza-IP. 219
Figure 4.20. Comparison of tRNA isotypes called as NSUN2 and TRDMT1 targets by bsRNA-seq, Aza-IP and miCLIP. 222
Figure 4.21. Comparison of tRNA isoacceptors called as NSUN2 or TRDMT1 targets by bsRNA-seq, miCLIP and Aza-IP. 226
Figure 5.1. HeLa Ago2 and HuR footprint density surrounding candidate m^{5}C sites in mRNA. 248
Figure 5.2. Density of conserved 7-mer motifs surrounding candidate m^{5}C sites in mRNA. 250
Figure 5.3. Global analysis of the average steady-state level of mRNAs containing candidate m^{5}C sites called as NSUN2 targets based on bsRNA-seq. 252
Figure 5.4. Confirmation of depletion of m^{5}C at known target sites of NSUN2 and TRDMT1 in the replicate NSUN2 and TRDMT1 knockdown samples. 255
Figure 5.5. Schematic demonstrating the samples used for the validation of candidate m^{5}C sites and those used for the analysis steady-state levels. 256
Figure 5.6. Validation of m^{5}C sites in six mRNAs in which steady-state levels were investigated by RT-qPCR in the replicate NSUN2 and TRDMT1 knockdown samples. 258
Figure 5.7. Steady-state level of mRNAs validated to contain NSUN2-mediated m^{5}C sites in the HeLa methyltransferase knockdown cells. 260

Figure 5.8. Batch variation evident after RT-qPCR of mRNAs validated to contain NSUN2-mediated m^5C sites. 262
Figure 5.9. Experimental design to investigate the reproducibility of bisulfite conversion and PCR amplification. 265
Figure 5.10. Confirmation of bisulfite conversion in each HeLa technical replicate by MiSeq amplicon sequencing. 266
Figure 5.11. The proportion of methylation at m^5C sites in three HeLa cell technical replicates as determined by MiSeq amplicon sequencing. 269
Figure 5.12. Schematic of sucrose density gradient centrifugation. 271
Figure 5.13. Absorbance trace at 254 nm of HeLa cells and analysis of the RNA integrity after polysome gradient fractionation. 274
Figure 5.14. Confirmation of bisulfite conversion in polysome fractions for each replicate by MiSeq amplicon sequencing. 276
Figure 5.15. Confirmation of known m^5C sites in 28S rRNA by MiSeq amplicon sequencing. 278
Figure 5.16. Gradient distribution and methylation profiles of mRNAs validated to contain m^5C sites mediated by NSUN2 at different translation states. 281
Figure 6.1. NSUN2 levels in PrEC and LNCaP cells based on RT-qPCR and western blotting. 291
Figure 6.2. Confirmation of RNAi-mediated knockdown of *NSUN2* in LNCaP cells and verification of bisulfite conversion efficiency 293
Figure 6.3. The distribution of candidate differential m^5C sites across different transcript biotypes in PrEC and LNCaP cells 295
Figure 6.4. Lack of validation of candidate differential m^5C sites in mRNA by MiSeq amplicon sequencing in PrEC and LNCaP cells. 299
Figure 6.5. Representative snapshots from IGV depicting a subset of sequencing reads containing candidate m^5C sites that could not be validated in PrEC cells. 301
Figure 6.6. Mean methylation detected in each sample for all candidate m^5C sites in common between PrEC and LNCaP cells based on bsRNA-seq. 304
Figure 6.7. Representative snapshot from IGV depicting a subset of sequencing reads mapping to vtRNA1-3 surrounding the candidate m^5C site. 304
Figure 6.8. The distribution of candidate m^5C sites across different transcript biotypes in PrEC and LNCaP cells. 306
Figure 6.9. Schematic demonstrating the steps taken to make the HeLa methyltransferase knockdown and prostate datasets comparable 307
Figure 6.10. Comparison of candidate m^5C sites in different transcript biotypes in prostate cells and HeLa cells. 309

Figure 6.11. tRNA structural positions and isotypes containing candidate m^5C sites in PrEC and LNCaP cells. 313
Figure 6.12. Comparison of tRNA structural positions and isotypes exhibiting candidate m^5C sites detected by bsRNA-seq in PrEC and LNCaP cells with HeLa cells. 314
Figure 6.13. tRNA isoacceptors exhibiting candidate m^5C sites in PrEC and LNCaP cells. 315
Figure 6.14. Comparison of tRNA isoacceptors exhibiting candidate m^5C sites in PrEC and LNCaP cells with HeLa cells. 316
Figure 6.15. Validation of candidate m^5C sites in tRNAs in PrEC and LNCaP cells by MiSeq amplicon sequencing. 318
Figure 6.16. Validation of candidate m^5C sites in other ncRNA in PrEC and LNCaP cells by MiSeq amplicon sequencing. 320
Figure 6.17. Validation of candidate m^5C sites in mRNA in PrEC and LNCaP cells by MiSeq amplicon sequencing. 325
Figure 6.18. Comparison of tRNA structural positions and isotypes exhibiting candidate m^5C sites in PrEC and LNCaP cells with the Aza-IP study [77]. 333
Figure 6.19. Comparison of tRNA isoacceptors exhibiting candidate m^5C sites in PrEC and LNCaP cells with the Aza-IP study [77]. 333
Figure 7.1. Schematic demonstrating m^5C in tRNA prior to end processing and splicing. 343
Figure 7.2. Confirmed and speculative molecular functions of m^5C in RNA. 347

LIST OF TABLES

Table 1.1. Comparison of MeRIP-seq, m^{6}A-seq and bsRNA-seq techniques, analyses and findings. 12

Table 1.2. Examples of human RNA:m^{5}C and m^{6}A methyltransferases (MTases) and their orthologues in *S. cerevisiae*. 16

Table 1.3. Biological functions and links to human health of RNA:m^{5}C and m^{6}A methyltransferases, readers and demethylases. 37

Table 2.1. Oligo sequences for the random primer mix 53

Table 2.2. siRNAs and their corresponding sequences used for RNAi-mediated knockdown. 56

Table 2.3. Bisulfite PCR primers and experimental conditions. 65

Table 2.4. RT-qPCR primers and experimental conditions. 68

Table 3.1. The distribution of colours of standard RNA and bisulfite-converted RNA using SOLiD™ sequencing. 113

Table 3.2. The number of total mappable sequencing reads in each bsRNA-seq library. 118

Table 3.3. General features of the MiSeq amplicon libraries. 142

Table 3.4. Mean cytosine conversion of the R-Luc *in vitro* transcripts based on bsRNA-seq. 146

Table 4.1. The proportion of cytosines with ≥5 reads containing candidate m^{5}C sites across different transcript biotypes 169

Table 4.2. Proportion of m^{5}C at C34 in unspliced and spliced tRNA$^{Leu(CAA)}$ based on MiSeq amplicon sequencing. 184

Table 4.3. Sequence context of m^{5}C sites discovered and validated as NSUN2 targets in this thesis. 208

Table 5.1. GO term enrichment of mRNAs containing candidate m^{5}C sites in HeLa cells based on topGO. 244

Table 5.2. GO term enrichment of NSUN2-targeted mRNAs in HeLa cells based on topGO. 245

Table 5.3. GO term enrichment of TRDMT1-targeted mRNAs in HeLa cells based on topGO. 246

Table 5.4. Two-way ANOVA statistical analysis to determine differences in expression levels of NSUN2-targeted mRNAs. 263

Table 5.5. Two-way ANOVA statistical analysis to determine differences in overall steady-state levels of NSUN2-targeted mRNAs. 263

Table 5.6. Mean cytosine conversion of the non-humanised R-Luc *in vitro* transcript. 266

Table 5.7. Mean cytosine conversion of the non-humanised R-Luc *in vitro* transcript. 276

Table 6.1. Overlap of candidate m^{5}C sites detected in HeLa cells and sites in common between PrEC and LNCaP cells at different RPM expression thresholds 308

TABLE OF CONTENTS

Originality statement ..... **II**

Acknowledgements ..... **III**

Abstract ..... **V**

Conferences, publications and awards ..... **VII**

Conferences VII

Publications IX

Awards IX

Commonly used abbreviations ..... **X**

List of figures ..... **XII**

List of tables ..... **XVII**

CHAPTER ONE INTRODUCTION **1**

1.1 Introduction to RNA modifications **2**

1.2 Transcriptome-wide mapping of ψ, m^6A and m^5C **5**

1.2.1 Prevalence and global distribution of ψ 5

1.2.2 Prevalence and global distribution of m^6A 8

1.2.3 Prevalence and global distribution of m^5C 10

1.3 Enzymology **14**

1.3.1 RNA:m^6A methyltransferases METTL3 and METTL14 16

1.3.1.A Discovery 16

1.3.1.B Structure/catalysis 17

1.3.1.C Substrate specificity 17

1.3.2 RNA:m^6A reader YTHDF2 19

1.3.2.A Discovery 19

1.3.2.B Structure/Catalysis 19

1.3.2.C Substrate specificity 20

1.3.3 RNA:m^6A demethylases FTO and ALKBH5 21

1.3.3.A Discovery 21

1.3.3.B Structure/Catalysis 21

1.3.3.C Substrate specificity 22

1.3.4 RNA:m^5C methyltransferase NSUN2 22

1.3.4.A Discovery 22

1.3.4.B Structure/Catalysis 23

1.3.4.C Substrate specificity 23
1.3.5 RNA:m^{5}C methyltransferase TRDMT1 25
1.3.5.A Discovery 25
1.3.5.B Structure/Catalysis 25
1.3.5.C Substrate specificity 26
1.4 Role in biology and human health 28
1.4.1 RNA:m^{6}A methyltransferases METTL3 and METTL14 29
1.4.2 RNA:m^{6}A reader YTHDF2 31
1.4.3 RNA:m^{6}A demethylases FTO and ALKBH5 32
1.4.4 RNA:m^{5}C methyltransferase NSUN2 33
1.4.5 RNA:m^{5}C methyltransferase TRDMT1 35
1.5 Molecular function 39
1.5.1 Biased distribution of m^{6}A in mRNA 40
1.5.2 Molecular function of m^{6}A 42
1.5.3 Biased distribution of m^{5}C in mRNA 43
1.5.4 Molecular function of m^{5}C 45
1.6 Aims 46
1.6.1 Detailed aims 47
CHAPTER TWO MATERIALS & METHODS 48
2.1 Materials 49
2.1.1 Chemicals, reagents and consumables 49
2.1.2 Kits 51
2.1.3 Antibodies 52
2.1.4 Enzymes 52
2.1.5 Buffers and solutions 53
2.1.5.A Buffers and solutions for general molecular biology 53
2.1.5.B Buffers and solutions for dot blot assays and western blotting 53
2.1.5.C Buffers and solutions for bsRNA-seq and MiSeq amplicon sequencing 54
2.1.5.D Buffers and solutions for tissue culture 54
2.1.5.E Buffers and solutions for sucrose density gradient centrifugation 55
2.1.6 DNA Markers 55
2.1.7 Oligonucleotides 56
2.1.8 siRNAs 56
2.1.9 *In vitro* transcripts 56

2.1.10 Plasmids 57
2.1.11 Bacterial Strains 57
2.1.12 Bacterial Growth Media 57
2.2 Methods 58
2.2.1 Tissue culture techniques 58
2.2.1.A Maintenance of HeLa cells 58
2.2.1.B Maintenance of PrEC cells 58
2.2.1.C Maintenance of LNCaP cells 58
2.2.1.D Cryopreservation and revival of cell lines 59
2.2.1.E Reverse transfection of cells with siRNAs 59
2.2.2 Nucleic acid techniques 60
2.2.2.A RNA isolation and precipitation 60
2.2.2.B Removal of genomic DNA from RNA 60
2.2.2.C phenol/chloroform extraction 60
2.2.2.D Agarose gel electrophoresis 61
2.2.2.E Bisulfite conversion of RNA 61
2.2.2.F Reverse transcription 62
2.2.2.G *In vitro* transcription 62
2.2.2.H DNA/RNA dot blot assay 62
2.2.3 Nucleic acid quantification and integrity analysis 63
2.2.3.A NanoDrop spectrophotometric analysis for quantification 63
2.2.3.B Qubit fluorometric quantification 63
2.2.3.C Agarose gel electrophoresis for integrity analysis 64
2.2.3.D Agilent RNA 6000 Nano Kit for integrity analysis 64
2.2.4 Polymerase Chain Reaction (PCR) 64
2.2.4.A Bisulfite PCR 64
2.2.4.B Real time quantitative PCR (RT-qPCR) for expression analysis 67
2.2.4.C TaqMan® non-coding RNA assay 69
2.2.5 Protein techniques 69
2.2.5.A Protein isolation 69
2.2.5.B Protein quantification 69
2.2.5.C SDS PAGE and western blotting 69
2.2.5.D Densitometry of western blots 70
2.2.6 sucrose density gradient centrifugation for polysome analsysis 71
2.2.6.A Sucrose gradient construction 71
2.2.6.B Cell lysis 71

2.2.6.C Polysome profiling and fraction collection 71
2.2.6.D Preparation of polysome fractions for RNA extraction 72
2.2.7 Cloning and Sanger sequencing 72
2.2.7.A Preparation of *E. coli* DH5α for cloning 72
2.2.7.B Cloning of PCR products and bacterial transformation 73
2.2.7.C Plasmid DNA extraction 73
2.2.7.D Sanger sequencing of plasmids 73
2.2.8 Next-generation sequencing 74
2.2.8.A Transcriptome-wide bisulfite sequencing (bsRNA-seq) 74
2.2.8.B SOLiD™ bsRNA-seq analysis pipeline 79
2.2.8.C Functional genomic analyses for m^5C in RNA 82
2.2.8.D TruSeq amplicon sequencing using Illumina® MiSeq 85
2.2.8.E TruSeq amplicon sequencing analysis pipeline 89
2.2.8.F Generating coverage plots and heatmaps 89

CHAPTER THREE METHODS FOR DETECTION OF m^5C IN RNA 90

3.1 Introduction 91

3.2 Results 92

3.2.1 Preparation of cell lines and RNA for bsRNA-seq 93
3.2.1.A Preparation of cells and RNAi-mediated knockdown of *NSUN2*, *TRDMT1* and *DNMT1* 93
3.2.1.B Bisulfite conversion of RNA 94
3.2.1.C Assessment of bisulfite conversion and knockdown efficiencies 98
3.2.2 Preparation of libraries for bsRNA-seq 105
3.2.3 Mapping of bsRNA-seq data and detection of candidate m^5C sites 114
3.2.3.A Mapping and general features of the bsRNA-seq libraries 114
3.2.3.B Detection of candidate m^5C sites in RNA 121
3.2.4 Detection of m^5C in RNA using an antibody 125
3.2.5 Validation of m^5C sites using Illumina® MiSeq 131
3.2.5.A Selecting candidate m^5C sites for validation 131
3.2.5.B Preparation of libraries for MiSeq amplicon sequencing 132
3.2.5.C quantification and loading of MiSeq libraries 138
3.2.6 General features and mapping of the MiSeq libraries 140
3.2.6.A Mapping of MiSeq amplicon sequencing data 140
3.2.6.B General features of the MiSeq amplicon libraries 141

3.2.7 Bisulfite conversion and knockdown efficiency based on bsRNA-seq and MiSeq 143
3.2.7.A R-Luc *in vitro* transcripts 143
3.2.7.B Known m^5C sites in $tRNA^{Asp(GUC)}$ and $tRNA^{Leu(CAA)}$ 147
3.3 Discussion 150
3.3.1 knockdown of *NSUN2* and *TRDMT1* in HeLa cells 150
3.3.2 features of the bsRNA-seq libraries 151
3.3.2.A distribution of sequencing reads 151
3.3.2.B Alignment of sequencing reads and detection of candidate m^5C sites 152
3.3.2.C experimental design for the detection of high-confidence candidate m^5C sites 155
3.3.2.D Bisulfite conversion efficiency 157
3.3.2.E RNA modifications resistant to bisulfite conversion 157
3.3.3 Detection of m^5C using an antibody 159
3.3.4 Improvements to bisulfite sequencing validation techniques 160
3.3.5 Conclusions 162
3.4 Chapter highlights 163

CHAPTER FOUR RNA m^5C PROFILING OF HELA CELLS DEPLETED OF NSUN2 OR TRDMT1 164
4.1 Introduction 165
4.2 Results 166
4.2.1 m^5C sites detected in HeLa cells 166
4.2.1.A Total m^5C sites in HeLa cells 166
4.2.1.B m^5C sites mediated by NSUN2 and TRDMT1 170
4.2.2 m^5C in tRNA 174
4.2.2.A tRNA structural positions and isotypes targeted by NSUN2 and TRDMT1 175
4.2.2.B tRNA isoacceptors targeted by NSUN2 and TRDMT1 178
4.2.2.C Validation of candidate m^5C sites in tRNA 181
4.2.2.D Methylation of tRNAs prior to processing 183
4.2.3 m^5C in other ncRNA 189
4.2.3.A m^5C in rRNA 189
4.2.3.B m^5C in other ncRNA 191
4.2.3.C Sequence and structural context of m^5C in other ncRNA 194
4.2.4 m^5C in mRNA 197
4.2.4.A Distribution and location bias of m^5C in mRNA 197

4.2.4.B Validation of m^5C sites in mRNA 199
4.2.4.C Structural and sequence context of m^5C in mRNA 204
4.3 Discussion 209
4.3.1 m^5C sites in HeLa cells 209
4.3.1.A Total m^5C sites in HeLa cells 209
4.3.1.B m^5C sites mediated by NSUN2 and TRDMT1 211
4.3.1.C Global comparison of m^5C sites mediated by NSUN2 and TRDMT1 with the literature 212
4.3.2 m^5C sites in tRNA 216
4.3.2.A tRNA structural positions and isotypes targeted by NSUN2 and TRDMT1 ... 216
4.3.2.B tRNA isoacceptors targeted by NSUN2 and TRDMT1 223
4.3.2.C Methylation of tRNA prior to processing 226
4.3.3 m^5C in ncRNA 228
4.3.3.A m^5C in rRNA 229
4.3.3.B sequence and structural context of m^5C in other ncRNA 230
4.3.4 m^5C in mRNA 233
4.3.4.A Distribution of m^5C in mRNA 233
4.3.4.B Validation of m^5C sites in mRNA 234
4.3.4.C Structural and sequence context of m^5C in mRNA 235
4.3.5 Conclusions 237
4.4 Chapter highlights 237
4.4.1 tRNA 237
4.4.2 Other ncRNA 238
4.4.3 mRNA 239
CHAPTER FIVE EXPLORING THE MOLECULAR FUNCTIONS OF m^5C IN MRNA 240
5.1 Introduction 241
5.2 Results 242
5.2.1 Gene set enrichment and GO analyses of methylated mRNAs 242
5.2.2 Investigating the association of m^5C with Ago and miRNAs 247
5.2.3 Analysis of global steady-state levels of methylated mRNAs 251
5.2.4 Analysis of steady-state levels of methylated mRNAs using RT-qPCR 253
5.2.4.A Determining housekeeping genes for normalisation of RT-qPCR data 253
5.2.4.B Steady-state level analysis of candidate mRNAs targeted by NSUN2 254

5.2.4.C Analysis of overall steady-state level of candidate mRNAs targeted by NSUN2 263
5.2.5 Investigating the role of m^{5}C in translation 264
5.2.5.A Assessing the reproducibility of bisulfite conversion and PCR amplification 264
5.2.5.B Sucrose density gradient centrifugation and polysome fractionation 270
5.2.5.C Translation efficiency of candidate mRNAs 274
5.2.5.D preparation of fractions for MiSeq amplicon sequencing 275
5.2.5.E Investigating the relationship between m^{5}C in mRNA and translation 279
5.3 Discussion 281
5.3.1 Ago2 and miRNA analyses 281
5.3.2 mRNA steady-state level analyses 282
5.3.3 mRNA translation analysis 284
5.3.4 Conclusions 285
5.4 Chapter highlights 287
CHAPTER SIX COMPARISON OF THE RNA M^{5}C PROFILES OF PREC, LNCAP & HELA CELLS 288
6.1 Introduction 289
6.2 Results 290
6.2.1 Expression level of NSUN2 in PrEC and LNCaP cells 290
6.2.2 Knockdown of *NSUN2* in LNCaP cells 292
6.2.3 Differential m^{5}C sites 294
6.2.3.A Global distribution of differential m^{5}C sites 294
6.2.3.B differential m^{5}C sites in mRNA 296
6.2.4 Stoichiometry of m^{5}C sites in common between PrEC and LNCaP cells 302
6.2.5 m^{5}C sites in common between PrEC and LNCaP cells 305
6.2.6 Global comparison with HeLa cells 306
6.2.7 m^{5}C sites in common in tRNA 310
6.2.7.A tRNA structural positions and isotypes exhibiting m^{5}C sites 310
6.2.7.B tRNA isoacceptors exhibiting m^{5}C sites 315
6.2.7.C Validation of m^{5}C sites in tRNA 317
6.2.8 m^{5}C sites in common in other ncRNA 319
6.2.9 m^{5}C sites in common in mRNA 321
6.3 Discussion 326
6.3.1 Differential m^{5}C sites in PrEC and LNCaP cells 326

6.3.1.A ncRNAs 326
6.3.1.B mRNAs 327
6.3.2 m^5C sites in common between PrEC and LNCaP cells 329
6.3.2.A Global comparison with HeLa cells 329
6.3.2.B m^5C sites in tRNAs 330
6.3.2.C m^5C sites in other ncRNA 334
6.3.2.D m^5C sites in mRNA 335
6.3.3 Conclusions 336
6.4 Chapter highlights 337
6.4.1 Differential m^5C sites 337
6.4.2 m^5C sites in common 338

CHAPTER SEVEN GENERAL DISCUSSION & FUTURE DIRECTIONS 339
7.1 Discussion 340
7.1.1 Substrate range of NSUN2 and TRDMT1 341
7.1.2 m^5C prior to processing of transcripts 342
7.1.3 Differential RNA m^5C profiles between normal and cancer cells 343
7.1.4 Similar m^5C patterns are present across multiple cell lines 344
7.1.5 Molecular function of m^5C in mRNA 345
7.2 Future directions 348
7.3 Concluding remarks 350

CHAPTER EIGHT APPENDIX 352
8.1 Sequences for the R-Luc *in vitro* transcripts 353
8.2 R script for structural motif analyses 354
8.3 Candidate m^5C sites in HeLa cells, PrEC and LNCaP cells 354
8.4 Global and local tRNA alignments 355
8.5 Rfam structural alignment of tRNAs exhibiting candidate m^5C sites from the local mapping 355
8.6 MiSeq amplicon sequencing of transcripts exhibiting m^5C sites 356
8.7 GSEA and GO analysis for candidate m^5C sites in HeLa cells 356
8.8 Complete list of housekeeping genes for RT-qPCR 356
8.9 statistical analysis for RT-qPCR gene expression analysis 356
8.10 Overlap of candidate m^5C sites in PrEC, LNCaP and HeLa cells 356
References 357

CHAPTER ONE

INTRODUCTION

1.1 INTRODUCTION TO RNA MODIFICATIONS

The following introduction has been adapted from a recent literature review I published [1].

A wealth of chemically diverse nucleobase and ribose modifications have been characterised in RNA, amounting to over 100 distinct ribonucleosides that are known across all phyla and types of RNA [2, 3]. Much of this diversity occurs among tRNAs and the other non-coding (nc) RNAs of the translational apparatus. Messenger (m) RNA is comparatively devoid of modifications, both in terms of diversity and density. The RNA modification database [2] lists thirteen non-standard nucleosides as present in eukaryotic cellular mRNA (Figure 1.1). These include 7-methylguanosine (m^7G) and several other base and/or ribose-methylated nucleosides that are accounted for by the presence of the 5' cap structure m^7G(5')ppp(5')N (termed "cap 0") and its further methylated variants that are either common to many eukaryotes ("cap 1" & "cap 2") [4, 5] or selectively found in certain species such as the "cap 4" of Trypanosomes or the trimethylguanosine 'TMG cap' of *C. elegans* trans-spliced mRNAs [6]. This leaves four modifications, inosine (I), pseudouridine (ψ), N6-methyladenosine (m^6A) and 5-methylcytidine (m^5C) that were reported to occur internally within mRNAs. A-I as well as C-U editing by ADAR and APOBEC enzymes changes base-pairing thus having clearly predictable effects on RNA function such as codon alteration or changes to RNA structure, as reviewed elsewhere [7, 8]. ψ is abundant in transfer (t) RNAs and ribosomal (r) RNAs and enhances their function by stabilising their structure [9-15]. Natural occurrence of ψ in mRNA remained unknown until recently, with early studies suggesting that artificial ψ in mRNAs facilitated non-canonical base pairing in the ribosome decoding centre [16, 17]. By contrast, m^6A and m^5C do not change base identity and thus would be expected to have more subtle or hidden effects. For example, m^6A has been shown to block the non-standard A:G base pairing, which can play important structural roles [18].

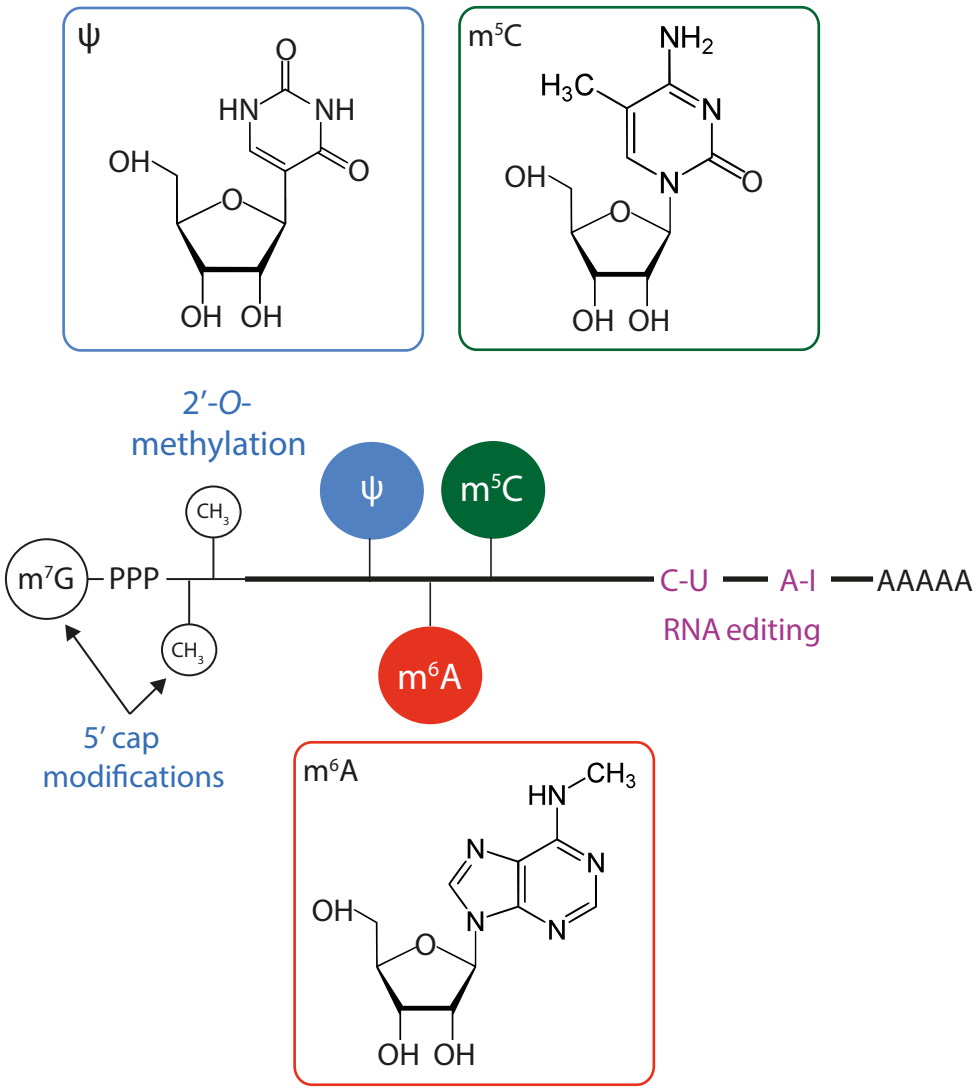


Figure 1.1. Schematic representation of known eukaryotic mRNA modifications.

The internal base modifications ψ, m^6A and m^5C are highlighted as blue, red and green lollipops, respectively.

A flurry of research conducted during the mid-seventies in the United States, Japan and Australia focused on the methylation state of eukaryotic mRNA and rapidly pieced together the puzzle of the cap structure present on eukaryotic cellular and viral mRNA [19]. This spawned a whole field of research characterising the enzymology of the cap structure and its functions in multiple aspects of RNA metabolism, particularly during processing, translation and decay of mRNA [6]. Generally, these pioneering studies attributed about half of the detectable methyl groups to modifications within the extended cap structure while the other half was seen to occur internally within mRNA,

or more accurately polyadenylated RNA, which might also have included some ncRNA species. With this caveat, m^6A was established as the most abundant internal mRNA modification with estimates typically ranging from 1-5 internal residues per molecule of either cellular (mammalian, insect, plant) or viral mRNA [20-39], though not all mRNAs contain m^6A residues. Characterisations of the m^6A methyltransferase in human (MT-A70) [40], plant (MTA) [41] and yeast (Ime4) [42], also known as METTL3 (for methyltransferase like 3), have since advanced this research as has the identification of a consensus sequence context, originally proposed to be RRm^6ACH (R = A, G; H = A, C, U) in its most extended form [43]. Sites for m^6A have also been mapped in three mRNAs, one each of viral, mammalian and yeast origin [44-46] and over the years some evidence was gathered for cellular and molecular functions of this modification in mRNA.

Some of the early studies also detected m^5C as an internal modification in mammalian cellular or viral mRNA [26, 30, 47, 48]. Specifically, methylation in hamster cell mRNA was reported to be split into approximately 50% deriving from the cap structure, while 40% and 10% came from internal m^6A and m^5C, respectively [26]. This represents relatively sparse modification levels (e.g. m^6A: on average 0.72 sites/1000 nt, 1.8/mRNA, 0.3% of all internal adenosines; m^5C: 0.18 sites/1000 nt, 0.45/mRNA, 0.07% of all internal cytidines), although it still compares favourably to A-I editing frequency (highest levels in brain: 0.059 sites/1000 nt, 0.15/mRNA, 0.02% of all internal adenosines; levels in other tissues are up to ~10 fold lower) [49]. Adenovirus and Sindbus virus mRNA were also found to contain detectable levels of m^5C; for the latter it represented a substantial proportion of internal methylation [30, 47, 48]. While there is considerable information on the occurrence and roles of m^5C in tRNA and rRNA (reviewed in refs. [50-52]), no sites had been mapped in any mRNA and no further research into the role of m^5C as a modification of mRNA had been published until recently. Similarly, despite ψ being the most abundant RNA modification, its presence in mRNA remained elusive up until recently.

An overriding impediment to further progress in understanding the role of internal mRNA methylation and pseudouridylation has been the lack of suitable methods to map multiple modified sites with ease and sensitivity. In this chapter, I will review the

current state of this field with a particular emphasis on recent studies that have overcome these hurdles and provided the first transcriptome-wide maps of ψ [53-55], m^6A [56, 57] and m^5C [58].

1.2 Transcriptome-wide mapping of ψ, m^6A and m^5C

Traditional techniques to detect RNA modifications often rely on combinations of metabolic (radioactive) labelling, purification of a particular transcript class or individual RNA from cellular material, followed by hydrolysis and identification of modified nucleosides through chromatographic separation [50, 51, 59]. For instance, both m^6A and m^5C were initially identified as internal mRNA modifications in this way, e.g. ref. [26], using ^{14}C-labelling of methyl groups and oligo(dT)-mediated purification of mRNAs via their then newly discovered poly(A) tail [60]. Mapping modifications to particular sequence positions required additional laborious and relatively insensitive approaches, and in combination this precluded progress with mapping internal mRNA base methylation for all but a few cases [44-46]. This changed recently with the development of transcriptome-wide, high-throughput sequencing-based approaches that have adapted methods developed for DNA in epigenetics research (Table 1.1) [56-58]. The association of m^5C with epigenetic regulation of genome activity is well established and many methods are available to map its presence in genomic DNA; often these have been combined with microarray or high-throughput sequencing as the ultimate readout. Typically, they are based on either recording patterns of DNA cleavage by methylation-sensitive restriction enzymes, enrichment of DNA fragments by binding to antibodies against m^5C, or the bisulfite sequencing chemistry [61, 62].

1.2.1 Prevalence and global distribution of ψ

Up until recently, ψ was only detected in tRNAs, rRNAs, small nuclear (sn) RNAs and small nucleolar (sno) RNAs, however three recent global approaches have detected ψ in mRNAs, thoroughly extending what was previously known about this modification. As these studies have been published very recently and the technique used for the detection of ψ is substantially different to those used to detect m^6A and m^5C in RNA, I will only

briefly summarise what has been found thus far, including the biological implications of this modification in mRNA.

The transcriptome-wide approaches used for the detection of ψ took advantage of its selective modification with *N*-cyclohexyl-*N'*-(2-morpholinoethyl)-carbodiimide metho-p-toluenesulphonate (CMC) as identified by previous primer-extension methods [63-65]. The modification of ψ by CMC creates a block for reverse transcription one nucleotide 3' to ψ. This was then coupled with next-generation sequencing, providing single nucleotide resolution of ψ, termed pseudo-seq [53], ψ-seq [54] or PSI-seq [55]. The preparation of samples for sequencing were similar between each study, with the pseudo-seq study performing poly(A)+ selection of *S. cerevisiae* and human cervical cancer cell (HeLa) RNA, fragmentation and denaturation of the RNA, CMC treatment, followed by library preparation and next-generation sequencing. The ψ-seq study performed poly(A)+ selection of *S. cerevisiae* and human embryonic kidney cell (HEK293) RNA, CMC treatment, fragmentation of the RNA followed by library preparation and next-generation sequencing. The PSI-seq study also performed poly(A)+ selection of *S. cerevisiae* RNA, fragmentation of the RNA, CMC treatment, followed by library preparation and next-generation sequencing.

Pseudo-seq confidently detected 260 novel candidate ψ sites in 238 mRNAs and 74 novel candidate ψ sites in 38 ncRNAs in yeast. In HeLa cells, pseudo-seq confidently detected 96 novel candidate ψ sites in 89 mRNAs and 12 ψ novel candidate sites in 9 ncRNAs [53]. In both yeast and humans, ψ was found to be distributed in untranslated regions (UTR) and coding regions (CDS) of mRNA. Examination of ψ synthases (*Pus*) yeast deletion strains revealed that the majority of ψ sites in mRNAs and ncRNAs were mediated by Pus1, with a smaller number of sites mediated by Pus2, Pus4 and Pus7. Interestingly, ψ was investigated under different growth phases of yeast, with differential ψ sites detected in mRNAs and ncRNAs, indicating that ψ in these transcripts is regulated in response to environmental signals. Similarly, some ψ sites in human mRNAs were reported to be regulated by different cellular growth states.

ψ-seq confidently detected 328 candidate ψ sites across mRNA and ncRNA in yeast, 108 of which were Pus-dependent sites [54]. In HEK293 cells, 353 candidate ψ sites in mRNAs and 43 candidate ψ sites in ncRNAs were detected confidently. As with pseudo-seq, candidate ψ sites were distributed throughout mRNAs and differential ψ patterns were observed under different yeast growth phases. Interestingly, heat shock induced the formation of Pus7-mediated ψ in mRNA and deletion of Pus7 was reported to decrease its target mRNAs, suggesting a stabilising role for this modification in mRNA.

PSI-seq was performed for two replicates of yeast log-phase samples and one heat shocked yeast sample. This technique confidently detected 103 candidate ψ sites in 56 mRNAs in log-phase 1, 335 candidate ψ sites in 150 mRNAs in log-phase 2 and 335 candidate ψ sites in 208 mRNAs in heat shocked yeast, in addition to the 43 known ψ sites in 18S and 25S rRNA [55]. There was a modest overlap in candidate sites detected, with 5 sites overlapping between the log-phase replicates, 9 sites overlapping between log-phase 1 and heat shock, and 14 sites overlapping between log-phase 2 and heat shock. However, two mRNAs, *RPL11A* and *TEF1*, exhibited ψ in all three datasets and were identified as Pus1- and Pus4-mediated sites, respectively. Interestingly, the candidate ψ site in *RPL11A* was conserved in *S. mikitae* and *S. pombe*, and the candidate ψ site in *TEF1* was conserved in *S. mikitae*, indicating that ψ is likely to have functional importance in mRNA. As with pseudo-seq and ψ-seq, ψ was distributed throughout the mRNA with no biases evident. However, gene ontology (GO) analysis of the PSI-seq data revealed a significant enrichment of cytoplasmic translation GO term, suggesting a potential role in post-transcriptional gene regulation.

Thus, all three studies have shown the presence of ψ in mRNA, albeit at low levels, as well as various ncRNA in yeast and humans. ψ is largely conserved across different species and appears to be sensitive to different growth phases, suggesting that this modification in mRNA may be important for development. Early investigation into the molecular function of this modification suggests potential roles in post-transcriptional gene regulation. Further explorations into the molecular and biological functions of this modification are warranted.

1.2.2 Prevalence and global distribution of m^{6}A

Recent work addressing the transcriptome-wide distribution of m^{6}A adapted approaches equivalent to MeDIP-seq (methylated DNA immunoprecipitation sequencing). Similar protocols were developed independently by two separate groups who termed it either MeRIP-seq (methyl-RNA immunoprecipitation sequencing) and applied it to total mouse brain and HEK293T cells [56], or m^{6}A-seq, which was performed on human hepatocellular carcinoma cells (HepG2) and mouse liver [57]; see Table 1.1 for a comparison of both methods. In each case, cellular RNA was enriched for mRNA, broken into fragments of ~100 nucleotide (nt) length and immunopurified using an anti-m^{6}A antibody immobilised on beads, followed by next-generation sequencing analysis of input as well as enriched RNA on Illumina® next-generation sequencing systems (Figure 1.2A). The similarities and differences in approach were expertly summarised in ref. [66]. Briefly, Meyer et al. [56] used depletion of rRNA, two rounds of immunopurification and elution by proteinase K digestion, whereas Dominissini et al. [57] used poly(A)+ selection, a single round of immunopurification and elution with N6-methyladenosine. In addition to these differences in experimental approach, the two studies also varied in the computational algorithms that were applied to the alignment of sequencing reads and the detection of candidate m^{6}A regions. As investigated in ref. [66], these bioinformatics methods are critical to the overall outcome as the presence and exact position of m^{6}A is not directly measured but instead inferred by threshold-based detection of peaks in read density across transcript regions.

MeRIP-seq confidently detected 13,471 candidate m^{6}A sites in RNA derived from 4,654 genes in mouse brain [56]. 94.5% of these were found within mRNAs, whereas the remainder mapped to several classes of (long) ncRNA. Analyses of HEK293T cells returned 18,756 m^{6}A peaks in RNA derived from 5,768 genes [56]. This demonstrated that, while many m^{6}A sites were conserved between mouse and human RNAs, over 2,000 transcripts were uniquely methylated in either mouse brain or HEK293T cells. Strengths of the MeRIP-seq study are that results were compared across different next-generation sequencing platforms using different anti-m^{6}A antibodies and biological replicate samples. Furthermore, three mRNA candidates were independently validated through sequence-specific RNA pull-down and anti-m^{6}A immunoblotting. m^{6}A-seq analysis of the HepG2 transcriptome yielded 12,769 candidate m^{6}A sites distributed

among 6,990 mRNAs and 250 ncRNAs [57]. On average, ~1.7 m^6A sites per gene were detected, broadly in agreement with earlier research as mentioned above [20-37]. Applying the same approach to mouse liver RNA revealed 4,513 m^6A peaks within 3,376 mRNAs and 66 ncRNAs [57]. Of 2,023 expressed genes conserved between human and mouse, 997 had candidate m^6A sites in orthologous positions.

Recent optimisation of the m^6A-seq protocol has enabled the detection of m^6A in *S. cerevisiae* undergoing meiosis to near single nucleotide resolution by decreasing the fragment size and implementing a ligation-based strand-specific library preparation protocol, allowing the capturing of m^6A within the sequenced fragment [67]. This yielded a total of 1,308 candidate m^6A sites in 1,183 mRNAs within 3 nt of the known m^6A consensus sequence. The candidate m^6A sites identified during meiosis were also conserved across *S. mikatae*, with 229 m^6A-containing transcripts in common with *S. cerevisiae*.

The discovery of the RNA:m^6A methyltransferases, METTL3, and more recently, METTL14, have further advanced research into this modification. Implementation of photoactivatable ribounucleoside crosslinking and immunoprecipitation (PAR-CLIP) to locate the binding sites of METTL3 and METTL14 yielded a total of 2,110 and 1,147 transcripts, respectively [68]. Binding sites of the methyltransferases sites were detected predominantly in intergenic regions and introns, although there were a larger proportion of binding sites in mRNAs compared to ncRNAs. There was also high overlap between the target transcripts, with ~56% in common between METTL3 and METTL14. Another recent study that performed PAR-CLIP of METTL3 in HEK293 cells revealed 922 binding sites distributed across mRNA (>90%) as well as microRNA (miRNA) and large intergenic ncRNAs (lincRNAs) [69]. Both of these studies showed wide distribution of m^6A sites in mRNA and to a lesser extent, ncRNA, however the presence of other RNA:m^6A methyltransferases that may mediate m^6A formation remain to be explored.

Thus, all studies indicated widespread occurrence of m^6A in human, mouse and yeast transcriptomes, particularly in mRNAs, with highly significant evolutionary conservation of some sites as well as pronounced variability in others.

1.2.3 Prevalence and global distribution of m^5C

Recent efforts to map m^5C sites in RNA capitalised on the bisulfite sequencing technique so popular in DNA studies. The method, originally developed in Australia, relies on the differential chemical reactivity of m^5C compared to unmethylated cytidine when treated with sodium bisulfite [70, 71]. Unmethylated cytidine is deaminated to uracil whilst m^5C remains unconverted and is still read as 'C' in subsequent sequencing. Two groups independently developed adaptations of bisulfite sequencing to RNA. Schaefer et al. adapted a commercial DNA bisulfite sequencing kit to RNA and published a series of papers focused on the role of m^5C in tRNA biology [59, 72, 73]. Squires et al. started with a bisulfite treatment protocol originally devised to detect tRNA modifications by primer extension. They re-optimised treatment parameters and combined it with a next-generation sequencing readout [74]. This approach, termed bsRNA-seq (sequencing of whole bisulfite-converted transcriptomes) in the following, was used to obtain a first transcriptome-wide map of m^5C sites in HeLa cells [58]. Cellular RNA was enriched for mRNA through a combination of rRNA depletion and poly(A)+ selection and then subjected to bisulfite treatment, which also fragmented the RNA. cDNA libraries of insert sizes ranging from 50-120 nt were then sequenced on the SOLiD™ next-generation sequencing platform (Figure 1.2B) and mapped to references of all known human transcripts using SOCS-B, an alignment tool designed to map bisulfite-converted SOLiD™ colour-space reads [75]. In principle, candidate m^5C sites are thus detectable with nucleotide precision through identification of mapped positions that exhibit unconverted Cs. Notwithstanding this methodological advantage, some caveats also apply to this approach. For instance, bisulfite treatment is sensitive to incomplete denaturation of RNA secondary structure and some cytosine modifications other than m^5C may also be resistant to bisulfite treatment [58]. Also, during computational analysis trade-offs again need to be made between sensitivity and accuracy in the final candidate m^5C site selection when choosing mapping parameters and thresholds for read coverage and proportion of cytidine conversion [58]. An interesting option might be to complement bsRNA-seq with MeRIP-seq using available antibodies against m^5C. Recently, a study applied MeRIP-seq for validation of candidate m^5C sites detected within archaeal RNAs by bsRNA-seq [76]. Thus, application of MeRIP-seq for m^5C detection in RNA is a promising alternative method.

bsRNA-seq detected 10,530 candidate m^5C sites, of which 8,495 were mapped in mRNAs whereas 255 were found in tRNAs and 1,780 sites in other (long) ncRNAs. Analysis of unmethylated spike-in RNA and endogenous tRNA mapping data suggested that bsRNA-seq performed well. The spike-in control displayed complete cytidine conversion, while 21 of the 28 previously known m^5C sites in tRNA were detected and approximately 80% of the 234 identified novel candidate sites in tRNA were situated in anticipated structural positions. 13 candidate m^5C sites were chosen for independent validation by Sanger bisulfite sequencing, which confirmed 12 sites, including 3 within mRNAs. Conservation of m^5C sites was not assessed as RNA from other cellular sources or species was not analysed.

In addition to bsRNA-seq and MeRIP-seq, newly developed techniques for transcriptome-wide detection of m^5C have recently come to the fore with the discovery of NSUN2 and TRDMT1 as human RNA:m^5C methyltransferases. The first technique, termed 5-azacytidine-mediated immunoprecipitation (Aza-IP), takes advantage of the DNA/RNA methyltransferase inhibitor, 5-azacytidine (5-aza-C), which is randomly incorporated into nucleic acids as a cytidine analogue [77]. Prior to methylation, NSUN2 and TRDMT1 covalently bind to the RNA and this bond is rendered irreversible in the presence of 5-aza-C. The complex is then immunoprecipitated and subjected to next-generation sequencing for identification of NSUN2 and TRDMT1 targets. The Aza-IP study, which was performed on HeLa cells, reported NSUN2 and TRDMT1-mediated m^5C sites predominantly in tRNAs, with a smaller proportion in ncRNAs, and 8 were validated by Sanger bisulfite sequencing [77]. One mRNA exhibited a candidate m^5C site in this study, found to be mediated by TRDMT1. The second technique, termed methylated individual-nucleotide resolution crosslinking and immunoprecipitation (miCLIP) involved the mutation of the second conserved cysteine of NSUN2, which normally aids in the release of the methylated RNA [78]. This resulted in the formation of an irreversible bond between NSUN2 and the RNA target, which was immunoprecipitated and subjected to next-generation sequencing. miCLIP identified 80% of NSUN2-mediated m^5C sites in tRNAs while the remainder in ncRNAs and mRNAs in HEK293 cells and human fibroblasts [78]. While both of these studies were performed transcriptome-wide, a caveat of these approaches was the inability to detect m^5C sites mediated by other methyltransferases, although this remains to be explored (see section 1.3).

In summary, bsRNA-seq, Aza-IP and miCLIP demonstrated the widespread presence of m^5C throughout coding [58, 78] and non-coding sequences [58, 77, 78] in a human transcriptome.

Table 1.1. Comparison of MeRIP-seq, m^6A-seq and bsRNA-seq techniques, analyses and findings.

The table is adapted from ref. [66].

Study	MeRIP-seq [56]		m^6A-seq [57]		bsRNA-seq [58]
Tissue/cell line	C57BL/6 brain	HEK293T	C57BL/6 liver	HepG2	HeLa
Organism	Mouse	Human	Mouse	Human	Human
Replicates	3		1		1
RNA preparation	Ribominus™		GenElute™ mRNA		RiboMinus™, Streptavidin Magnetic Beads
m^6A/m^5C enrichment method	IP-Synaptic Systems, NEB, 2 IP rounds		IP-Synaptic Systems, 1 IP round		Bisulfite conversion
RNA fragment size/cDNA size selection	100 nt		100 nt		Various fragment sizes, cDNA size selection 50-120 nt
RNA sequencing platform	Illumina® GAII & HiSeq2000	Illumina® HiSeq2000	Illumina® GAII		SOLiD™ 3
Sequenced control	Yes, input control				Yes, conversion and non-conversion controls
Aligners	BWA, TopHat		Novoalign, BowTie		SOCS-B
m^6A/m^5C site detection method	MeRIPPeR Fisher's exact test of IP read enrichment		Proprietary Computed Winscore >2 (4x enrichment) + filtering		Threshold parameters ≤80% cytosine conversion, coverage of ≥10, ≤2 mismatches
Peaks/sites reported	13,471	18,756	12,769	4,513	10,530
Genes/transcripts reported	4,654 genes	5,768 genes	7,240 transcripts	3,442 transcripts	5,359 genes
Enrichment of candidate site	Stop codon, internal exons		Stop codon, TSS, internal exons, AS exons		5' UTR, 3' UTR, Argonaute binding sites

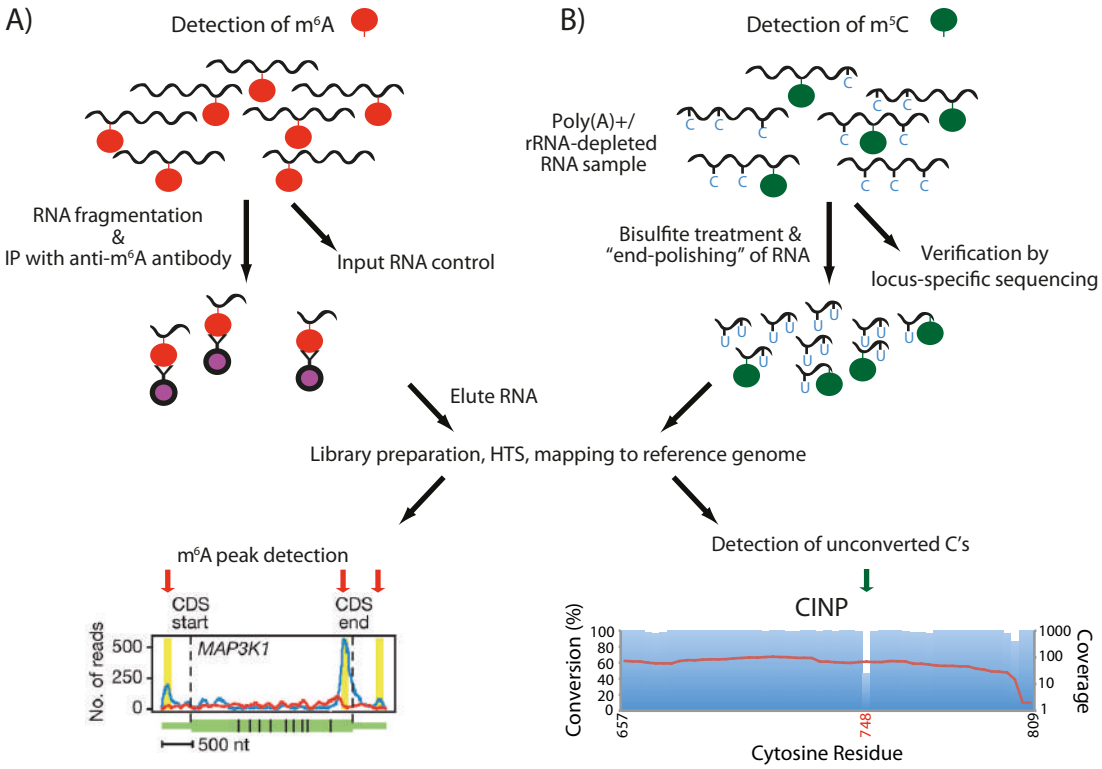


Figure 1.2. Transcriptome-wide approaches to map m^6A and m^5C in RNA.

A) MeRIP-seq or m^6A-seq employs immunoprecipitation of fragmented RNA using an anti-m^6A antibody prior to next-generation sequencing. Peaks representing m^6A sites are detected relative to RNA-seq read coverage generated from input RNA, B) bsRNA-seq uses bisulfite conversion chemistry and next-generation sequencing to detect m^5C sites. After mapping to the reference by allowing ambiguity of C or T positions with sufficiently high coverage, the proportion of unconverted cytosines is determined, representing candidate m^5C sites. Bottom panels show examples of m^6A sites detected in *MAP3K1* (left) [57] and a m^5C site detected in *CINP* (right) [58]. Sites called as m^6A or m^5C are indicated above the panels by red or green arrows, respectively. (Bottom left image reprinted with permission from ref. [57]. Copyright 2012 Nature Publishing Group; bottom right image reprinted with permission from ref. [58]. Copyright 2012 Oxford University Press.)

1.3 ENZYMOLOGY

Methyltransferases are an abundant class of cellular proteins (e.g. 208 are found in the human proteome, 86 in *S. cerevisiae*) [79, 80]. They act by transferring a methyl group to their substrate, usually from S-Adenosyl-L-methionine (AdoMet) as a co-factor. Based on structural criteria, methyltransferases are grouped into several superfamilies [81]. The largest among the three major superfamilies is the Rossmann-fold methyltransferase family (RFM), which targets a wide range of substrates including DNA, protein, RNA, lipids, polysaccharides and small molecules. RFM methyltransferases share a set of sequence/structural motifs commonly called motifs I-X after the linear order in which they appear in the DNA:m^5C methyltransferases. They correspond to the key protein features associated with AdoMet-binding (I-III), catalysis (X, IV, VI and VIII) and elements implicated in preservation of the common fold (V and VII). Of note, both DNA and RNA methyltransferases can exhibit permuted topologies with changes to the linear order of these conserved motifs [82]. The SPOUT methyltransferases (named after the founding members spoU and TrmD) form the second largest, though relatively poorly characterised, family expected to mainly target RNA but also protein (at arginine residues) [83]. Finally, the SET methyltransferases (short for (Su(var)3-9, Enhancer of Zeste, Trithorax) modify protein at lysine residues, mainly histones and ribosomal proteins [84]. Methylation is involved in approximately two-thirds of all known RNA modifications, thus over sixty distinct RNA methyltransferases have been described [52]; all eukaryotic cases belong to either the RFM or SPOUT superfamilies [85]. Three modes of RNA substrate recognition are known for RNA modifying enzymes. First, the catalytic and RNA-binding functions reside within the same protein domain. Second, separate catalytic and RNA-binding domains are fused together within the same polypeptide or are contributed by distinct subunits of a tight complex. Third, the methyltransferase, together with auxiliary proteins, employs a guide RNA to recognise the substrate [85].

All known RNA:m^5C and RNA:m^6A methyltransferases belong to the RFM superfamily [79] and show some homology to their DNA counterparts [86, 87]. For instance, sequence/structural alignments of 5-methylpyrimidine generating enzymes suggested that during evolution the RNA:m^5U-forming enzymes spawned the RNA:m^5C methyltransferases, which gave rise to DNA:m^5C methyltransferases [82,

86], although other models have also been advanced [88]. Phylogenetic analysis of eukaryotic RNA:m^5C methyltransferases identified five subfamilies termed RCMT1, -2, -7, -8 and -9 (for RNA (cytosine-5)-methyltransferases) (Table 1.2) [51, 86, 89]. RCMT2 includes the *S. cerevisiae* methyltransferase 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 [90, 91]. Nop2 also mediates m^5C at C2870 in 25S rRNA [92]. RCMT8 includes the *S. cerevisiae* methyltransferase Rcm1, and its mammalian orthologue, NSUN5A (NOP2/Sun domain family, member 5A), the former of which mediates m^5C at C2278 in 25S rRNA [86, 92]. The *S. cerevisiae* methyltransferase Nop8, and its mammalian orthologue NSUN4 (NOP2/Sun domain family, member 4), are important for ribosomal biogenesis [93-96]. NSUN4 has recently been reported to mediate m^5C at C911 in 12S rRNA in mice [97]. Prototypic members of the RCMT7 subfamily are the *S. cerevisiae* Ncl1/Trm4 (nuclear protein 1/tRNA methyltransferase 4) [98] and its mammalian orthologue NSUN2 (NOP2/Sun domain family, member 2, also known as MISU [Myc-induced SUN-domain-containing protein]) [99, 100]. Mammalian DNMT2/TRDMT1 (DNA methyltransferase 2/tRNA aspartic acid methyltransferase 1) is an outlier as it is structurally more similar to DNA methyltransferases [101]. Finally, the RNA:m^6A methyltransferase METTL3 (methyltransferase like 3) features a circularly permutated version of the RFM domain and belongs to a protein family that includes bacterial DNA:m^6A methyltransferases [87].

How cells deploy their repertoire of methyltransferases to methylate mRNAs internally is poorly understood. Below I summarise what is known about those m^6A and m^5C methyltransferases (“writers”) that have already been implicated in methylation of mRNA. I also refer to the emerging literature on demethylases (“erasers”) that may serve to dynamically remove methyl marks from mRNA and effector proteins (“readers”) that may interpret methylation marks to serve a molecular function.

Table 1.2. Examples of human RNA:m^{5}C and m^{6}A methyltransferases (MTases) and their orthologues in *S. cerevisiae*.

Modification	**m^{5}C**					**m^{6}A**
MTase-family	7β strands RFM					RFM 7β rearranged
Examples	NSUN1	NSUN2	NSUN4	NSUN5	TRDMT1	METTL3
Known substrates	tRNA? rRNA?	tRNA ncRNA mRNA	Mitochondrial rRNA	rRNA?	tRNA mRNA	mRNA
***S. cerevisiae* orthologue**	Nop2	Ncl1	Nop8	Rcm1	-	Ime4

1.3.1 RNA:m^{6}A METHYLTRANSFERASES METTL3 AND METTL14

1.3.1.A Discovery

A mRNA:m^{6}A methyltransferase activity was first characterised in an *in vitro* methylation system based on HeLa nuclear extract and model RNA substrates [38, 102, 103]. It was shown that the enzyme is composed of three dissociable components, MT-A1 (~30 kDa apparent molecular mass), -A2 (~200 kDa) and –B (~875 kDa), which were all necessary for full activity [102]. MT-A2 was shown to contain a ~70 kDa AdoMet-binding subunit. Microsequencing and cDNA cloning subsequently identified this as the methyltransferase, which was named MT-A70 [40] (later renamed as METTL3 for methyltransferase-like 3) [102].

Structure prediction and phylogenetic analysis revealed a METTL3 family of proteins comprising three paralogous eukaryotic lineages (with METTL3 representing lineage A) as well as one small subfamily of bacterial DNA:m^{6}A methyltransferases [87]. Methyltransferase activity has since been confirmed for METTL3 orthologues in *A. thaliana* (called MTA for mRNA adenosine methylase) [41] and *S. cerevisiae* (Ime4 for inducer of meiosis 4) [42]. Interestingly, efficient mRNA methylation in yeast requires formation of the MIS complex between Mum2, Ime4 and Slz1 [104]. The *A. thaliana* Mum2 orthologue, FKBP12 interacting protein 37 kDa (FIP37), also interacts with MTA [41]. Similarly, through GST-pulldown and co-immunoprecipitation experiments efficient mRNA methylation in mammals was reported to require the formation of a complex consisting of METTL3, the METTL3 homologue METTL14,

and the Mum2 orthologue, Wilms tumour 1 associated protein (WTAP) [68, 69, 87, 105].

1.3.1.B Structure/catalysis

METTL3 and its orthologues exhibit a permuted topology of the methyltransferase domain (amino acids (aa) 350-570 of the 580 aa long human enzyme) otherwise characteristic of b class DNA amino-methyltransferases (Figure 1.3A) [87]. Subdomains for AdoMet binding and catalysis can readily be discerned, however it lacks an intervening subdomain for nucleic acid recognition typically found in b class DNA amino-methyltransferases [106]. Thus, the WTAP component of the methyltransferase complex may provide RNA targeting. The yeast MIS complex component Mum2 was suggested to activate Ime4 catalytic activity and/or target it to mRNA substrates [104]. The latter option is given additional credence by the fact that FIP37 in plants and WTAP in humans are RNA-binding proteins implicated in mRNA splicing [41]. Perhaps WTAP constitutes the MT-B component of the purified human mRNA:m^{6}A methyltransferase, which was also suggested to bind RNA [102, 104]. WTAP lacks any obvious catalytic domains and has been suggested to be indispensible for regulating mRNA:m^{6}A methyltransferase activity and localisation of the complex into nuclear speckles [69, 107]. Thus, WTAP most likely functions as a regulatory subunit of the methyltransferase complex.

1.3.1.C Substrate specificity

Painstaking early work in yeast, plants and mammals established that m^{6}A is present only within the Gm^{6}AC and Am^{6}AC trinucleotide contexts, with a 75% preference for Gm^{6}AC [24, 41, 44, 108-110]. An extended consensus sequence context was proposed to be RRm^{6}ACH (R = A, G; H = A, C, U; Figure 1.3C) [43]. Until recently, mapping of specific sites or regions for m^{6}A had been achieved for only three native mRNAs [44-46]. Seven m^{6}A sites were mapped at nucleotides 6394, 6447, 6507, 6718, 7414, 7424, and 8014 in Rous sarcoma virus genomic RNA [46, 111, 112]. m^{6}A sites further were narrowed to regions near the 3’ end for bovine prolactin and yeast *ime2* mRNAs [44, 45]; for prolactin mRNA a single site in the context AGm^{6}ACU was pinpointed [103]. The thousands of m^{6}A-containing peaks mapped in human and mouse transcriptomes in

the recent MeRIP-seq [56] and m^{6}A-seq studies [57] also showed non-random distribution across mRNA regions (see below). Importantly, peaks from both studies were highly enriched in the motif RRm^{6}ACU, thus concurring with and refining the previous consensus [66]. Specific investigation into the sequence specificity of both METTL3 and METTL14 also revealed the same extended motif [68, 69, 105]. It has been reported that METTL3 and METTL14 function synergistically to catalyse the formation of m^{6}A. Incubation of METTL3 and METTL14 individually with a RNA probe containing the m^{6}A consensus motif revealed that each enzyme methylated RNA at relatively low levels [68]. However, co-incubation of METTL3 and METTL14 with the RNA probe resulted in stronger levels of methylation. This is highlighted by high overlap in METTL3 and METTL14 target transcripts [68, 105]. While a large proportion of m^{6}A sites were found to be mediated by METTL3 and METTL14, m^{6}A sites were found in mRNAs that were not mediated by either enzyme [68, 105].

A common observation in all this work was that not every cellular mRNA contained m^{6}A and that many more suitable sequence motifs were present in a mRNA than were actually methylated. It remains unclear what additional features of the mRNA substrate are recognised to achieve such selective methylation, however the structural context does not appear to be a major requirement for METTL3 and METTL14 activity. A C-to-U mutation in the RRm^{6}ACU consensus motif resulted in substantially decreased methylation efficiency compared to the standard consensus sequence [68]. RNA probes synthesised with the consensus motif located in the stem or a loop revealed slightly higher activity towards the motif in the stem, however very high activity of the methyltransferase complex was evident for a probe exhibiting a random structure [68].

In some eukaryotes, m^{6}A is also present in snRNA [110, 113] and rRNA [2, 3], although the enzymatic pathways of m^{6}A formation in these RNA types remain largely unknown [52]. The m^{6}A sites in human 18S, and probably 28S, rRNA match the context consensus found in mRNAs, while that in U6 snRNA does not. A snRNA:m^{6}A methyltransferase activity was partially characterised in HeLa nuclear extract and found to be biochemically distinct from the mRNA:m^{6}A methyltransferase [110]. Genetic depletion of MTA/Ime4 in *A. thaliana* or *S. cerevisiae* led to clear loss of m^{6}A in polyadenylated RNA [41, 42, 114]; a residual amount in plant tissues was likely due to incomplete shut-down of MTA expression [114], although it could in principle be due

to activity of the METTL3 homolog AT4G09980 (a member of lineage B in ref. [87]). Thus, it appears that METTL3/MTA/Ime4 and METTL14 are responsible for a large proportion of m^6A formation in cellular mRNA, with a small proportion of sites possibly being mediated by another methyltransferase.

1.3.2 RNA:m^6A READER YTHDF2

1.3.2.A DISCOVERY

YTHDF2 (YTH domain family member 2) is part of the YTH domain family of proteins that are widespread across eukaryotes and bind single-stranded (ss) RNA [115, 116]. The founding member of this family, YT521-B, is a nuclear protein that was identified in yeast two-hybrid screens along with splicing factors and was found to interact with several splicing factors through two-hybrid and coimmunoprecipitation assays [117]. Further investigation revealed that YT521-B modulates alternative splice site selection in a concentration-dependent manner [117]. YT521-B and the *S. cerevisiae* RNA binding protein meiotic mRNA interception (Mmi1), the latter of which eliminates meiotic-specific mRNAs [118, 119], are the only characterised YTH family members and the function of other members remain unexplored. Very recently, YTHDF2 was reported to bind a m^6A-containing RNA bait, suggesting that this protein may mediate various cellular processes including alternative splicing in response to binding m^6A [57, 120].

1.3.2.B STRUCTURE/CATALYSIS

The YTH domain family members were found to contain YTH domains through a YT521-B homology search using BLAST [115]. A conserved portion of the rat YT521-B was identified between residues 356 and 499, and further BLAST searches revealed multiple proteins from a number of organisms containing this domain. The YTH domain was predicted to bind RNA, with portions of this domain showing similarity to the RNA recognition motif [115]. This binding was confirmed by expression of YT521-B in *D. melanogaster*, revealing that this domain bound ssRNA, but not DNA or double-stranded (ds) RNA [116]. YTFDH2 exhibits the RNA-binding YTH domain in the carboxy-terminal and a P/Q/N-rich region at the amino-terminal (Figure 1.3A) [120]. Proteins containing P/Q/N-rich regions have been reported to aggregate within

processing bodies (P bodies), with many of these proteins involved in mRNA degradation [121]. *In vitro* pulldown assays have revealed that C-YTHDF2 enriches m^{6}A-containing mRNA from total RNA and N-YTHDF2 colocalised with the P body marker DCP1a [120]. This suggests that the YTH domain located in the C-terminus binds m^{6}A-containing mRNA and the P/Q/N-rich N-terminus is involved in localising the m^{6}A-containing mRNA to P bodies possibly for degradation.

Two studies have recently determined the crystal structure of YTHDF2 in complex with m^{6}A-containing RNA, providing information on the mechanism of recognition of m^{6}A [122, 123]. A basic patch of residues was identified to be important for binding to the backbone of the mRNA, including R411, K416 and R441 and R527 [123]. K416A and R527A mutations resulted in decreased binding affinity to both unmethylated mRNA and m^{6}A-containing mRNA, indicating that these residues are not involved in the recognition of m^{6}A. Comparison of the crystal structure of YTHDF2 in complex with m^{6}A-containing RNA to that without RNA revealed a conformational change such that m^{6}A was tightly locked in an aromatic cage consisting of W486, W432 and W491 [122, 123]. Mutations of several residues within the aromatic cage resulted in significantly reduced binding affinity of YTHDF2 to m^{6}A-containing RNA compared to unmethylated RNA, indicating that this aromatic ring is largely responsible for m^{6}A recognition [122, 123].

1.3.2.C Substrate specificity

As a newly characterised YTH domain family member, the substrate specificity of YTHDF2 still remains to be explored in further detail. Based on gel-shift assays, YTHDF2 preferentially bound a m^{6}A-containing RNA probe compared to an unmethylated probe, and there was a slight preference for the consensus sequence [120]. YTHDF2 PAR-CLIP revealed an enrichment of the Gm^{6}AC motif, in agreement with the previously identified m^{6}A consensus [120]. The majority of YTHDF2 binding sites were located in mRNAs, but 1% was also in ncRNAs. Interestingly, overlapping of m^{6}A peaks identified by m^{6}A-seq with YTFDH2 PAR-CLIP peaks in this study revealed a 50% overlap, indicating that not all m^{6}A peaks are bound by YTFDH2 [120]. It is unclear what additional features are required for recognition by YTHDF2, however there may be other m^{6}A readers. This has been highlighted by the ability of YTFDH1

and YTFDH3 to bind m^{6}A-containing RNA, suggesting that these enzymes may also function as m^{6}A readers [57, 120].

1.3.3 RNA:M^{6}A DEMETHYLASES FTO AND ALKBH5

1.3.3.A DISCOVERY

Two related proteins, fat mass and obesity associated (FTO) and ALKBH5 (alkylation repair homolog 5), were recently described as m^{6}A demethylases acting on RNA [124, 125]. Originally discovered in a study of the Fused toes (Ft) mutation in mice [126], the FTO gene came into sharp focus when genome-wide association studies (GWAS) strongly associated it with body mass index (BMI), risk of obesity and type 2 diabetes mellitus in multiple populations [127-131].

1.3.3.B STRUCTURE/CATALYSIS

FTO and ALKBH5 are both members of the AlkB (alkylated DNA repair protein B) family of nonheme Fe(II)/2-oxoglutarate (2-OG)- dependent dioxygenases [132-134]. *E. coli* AlkB is an enzyme that repairs cytotoxic 1-methyladenine and 3-methylcytosine in DNA, releasing the methyl groups as formaldehyde [135]. The crystal structure of FTO revealed an amino-terminal AlkB-like catalytic core domain with the characteristic double-stranded b-helix (aa 32-325) and a carboxy-terminal domain (CTD aa 326-505) with a novel, predominantly helical structure (Figure 1.3A). Tight interactions between the two domains are important for FTO activity. The insertion of an 11 amino acid-loop within the catalytic core, conserved between FTO orthologues but absent in other AlkB members, indicated a bias of FTO against double-stranded nucleic acids and thus a likely preference for RNA [136, 137]. Double FTO mutants, either H231A D233A- mutating two Fe (II) ligands- or R316Q R322Q -mutating two 2-OG ligands, each completely lost m^{6}A-demethylation activity [124]. Similarly, mutation of the Fe (II) ligand residues H204 or H266 in ALKBH5 also severely abrogated activity [125]. FTO possesses a bipartite nuclear localisation signal at the N-terminus [126] consistent with its nuclear localisation [132] and enrichment in nuclear splicing-related speckles [124]. ALKBH5 also co-localises with mRNA processing factors in nuclear speckles [125].

1.3.3.C Substrate specificity

Recombinant FTO was initially shown to preferentially demethylate 3-methylthymine or 3-methyluracil in synthetic, ssDNA and ssRNA, respectively [132, 138]. However, more recent work demonstrated a clear preference of FTO for m^6A in ssRNA *in vitro* [124]. The prevalence of m^6A in eukaryotic mRNA, but virtual absence in genomic DNA, thus suggested that the physiological substrate of FTO is RNA, not DNA. This was confirmed by FTO knockdown and/or overexpression in several human cell lines, and detection of commensurate increases or decreases in m^6A content of poly(A)+ RNA [124]. Analogous experiments with ALKBH5 essentially gave the same results; additionally it was found *in vitro* to have a mild preference for m^6A within the consensus motif and appeared not to act on rRNA *in vivo* [125]. Interestingly, ALKBH5 was also found as a mRNA-binding protein in recent quantitative proteomics screens [139, 140].

In summary, FTO and ALKBH5 together represent a major proportion of the m^6A demethylase activity in mammalian cells and tissues. Both enzymes are broadly, though not completely evenly, expressed and display very similar enzymatic/molecular characteristics. Beyond their preference for m^6A in ssRNA it is essentially unknown if and how they might select their targets for demethylation. Orthologues of both FTO and ALKBH5 are mostly confined to vertebrates [134, 141]. Given the much wider occurrence of m^6A in RNA, it seems likely that additional, as yet undiscovered m^6A demethylases exist in other organisms.

1.3.4 RNA:m^5C methyltransferase NSUN2

1.3.4.A Discovery

NSUN2 belongs to the RCMT7 subfamily of eukaryotic RNA:m^5C methyltransferases. The *S. cerevisiae* enzyme Ncl1/Trm4 was discovered first through searches for novel NOP2/Sun domain-containing proteins [98, 142], followed some years later by a characterisation of the orthologues in human (hTrm4; through homology with yeast Trm4) [100] and mouse (MISU; in a screen for genes activated by Myc) [99].

1.3.4.B Structure/Catalysis

NSUN2 and its mammalian orthologues have highly similar sequences [100, 143] and contain eight of the signature motifs for RFM methyltransferases in a characteristic permuted topology (Figure 1.3B) [86, 89]. The enzyme contains the conserved methyltransferase domain in its amino-terminal half, followed by a characteristic carboxy-terminal extension. Limited proteolysis of yeast Ncl1 showed that these two regions form separable domains. The amino-terminal domain (NTD, aa 1-437) is active as a methyltransferase but its catalytic efficiency is enhanced by the CTD (aa 438-684) [143]. The CTD on its own has low affinity for RNA and it does not markedly augment the RNA-binding properties of the whole enzyme, thus it might promote formation of a productive conformation of the enzyme/substrate complex. Since the NTD lacks a variable region for target recognition as typically seen in DNA:m^5C methyltransferases, it is unclear how the enzyme specifically recognises its RNA substrates. The conserved cysteine residues in motif IV and VI of Ncl1, Cys260 and Cys310, respectively, are important for catalysis, with Cys310 being essential for the initial nucleophilic attack at position C6 of the target cytosine and Cys260 assisting with the breakdown of the covalent adduct [144-146]. Mammalian NSUN2 was shown to be catalytically active using an *in vivo* complementation approach in *S. cerevisiae* and *in vitro* assays [99, 100]. The enzyme has no classical nuclear localisation signal; nevertheless, yeast Ncl1 localises to nucleoli and the nucleoplasm with some concentration at the nuclear periphery [142]. A similar pattern was described for mammalian NSUN2 [99, 100], although some fluctuation was seen during the cell cycle with a minority of cells in G2 phase displaying localisation to cytoplasmic vesicles. In M phase cells, NSUN2 was found along the spindle and at centrioles [99].

1.3.4.C Substrate specificity

Biochemical studies identified yeast Ncl1 as a multisite-specific tRNA:m^5C methyltransferase responsible for modifying yeast tRNAs at all known sites, i.e. structural positions 34, 40, 48 and 49 [98]. Positions 48 and 49 are commonly modified, whereas 34 and 40 are modified only in the intron-containing precursors of $tRNA^{Leu(CAA)}$ and $tRNA^{Phe(GAA)}$, respectively. By contrast, human NSUN2, when expressed in yeast cells appeared unable to modify positions 48 and 49 of tRNA and restricted to modifying C34 in $tRNA^{Leu(CAA)}$, acting at the level of the intron-containing

tRNA precursor. NSUN2 specificity was found to be influenced by both the structural (intron retention results in an elongated anticodon stem) and sequence context, though no consensus was derived [100]. The view that mammalian NSUN2 has a much narrower substrate specificity was nevertheless called in question by several more recent findings. First, when validating the bsRNA-seq data set [58], m^5C modification levels at structural positions 48 and 49 of $tRNA^{Asp(GUC)}$ were substantially reduced after siRNA-mediated *NSUN2* knockdown. This was also seen in human *NSUN2*-mutant cells [147]. Second, recombinant NSUN2 was shown to methylate position 48 in $tRNA^{Leu(CAA)}$ with or without an intron and positions 48, 49 and 50 in $tRNA^{Gly2(GCC)}$ *in vitro* [148] (Figure 1.3D). Third, tRNA methylation analysis of *Trdmt1* and *NSun2* single-knockout mice demonstrated complementary target-site specificities for the two enzymes (TRDMT1: C38 and NSUN2: C34, 40, 48, 49, 50; across four diagnostic tRNAs) and double-knockout mice showed a complete loss of m^5C in the tRNA fraction [149]. Fourth, examination of the substrate specificity of NSUN2 by Aza-IP and miCLIP in human and mouse cells revealed a number of additional m^5C sites in non-canonical tRNA structural positions [77] as well as tRNA isotypes and isoacceptors [77, 150]. Fifth, NSUN2 was shown to methylate sites in the *RPPH1* subunit of RNase P, vault (vt) RNAs and various ncRNAs [58, 77, 78], as well as *CINP*, *NAPRT1* and *NDUFB7* mRNAs [58]. Sixth, NSUN2 was further found to bind the 3' UTR of cyclin-dependent kinase inhibitor 2A, (*CDKN2A*, $p16^{Ink4A}$) mRNA, though curiously implicated in m^6A modification in this case [151]. Finally, NSUN2 was identified as a mRNA-binding protein by immunoprecipitation and RNA-seq [152] and quantitative proteomics screens [139, 140].

Collectively, these results indicate that both mammalian NSUN2 and yeast Ncl1 are multisite-specific tRNA:m^5C methyltransferases responsible for all (yeast) or most (mammalian) tRNA m^5C sites. At least the mammalian enzyme extends its activity to other ncRNA and mRNA. Whether NSUN2 is largely or solely responsible for the widespread occurrence of m^5C in these RNA biotypes as well remains to be tested. Some scope exists here for roles of eight other NSUN2/NOP2-related proteins in humans, NSUN1-7, with NSUN5 encoded by three genes NSUN5A-C [51]. Apart from NSUN2, enzyme activity and substrate specificity have not been characterised for any of these homologues in humans, although most of them retain a potentially active catalytic domain.

1.3.5 RNA:m^{5}C methyltransferase TRDMT1

1.3.5.A Discovery

A first TRDMT1-like protein was discovered in *S. pombe* by homology to bacterial DNA methyltransferases [153], followed by identification of mammalian TRDMT1 [154, 155]. It is the most widely conserved 'DNMT-type' enzyme with orthologues in protists, plants, fungi, and animals. Phylogenetic evidence suggests that the evolutionary ancestor of TRDMT1 was a DNA methyltransferase and an early TRDMT1 enzyme changed preference to RNA. DNMT1 and 3 have evolved as independent lineages [156]. Despite its prevalence, progress in understanding the physiological role of TRDMT1 initially lagged behind, due to difficulties in demonstrating activity with dsDNA as a substrate [157, 158]. It was then discovered that human TRDMT1 has robust RNA methyltransferase activity modifying tRNA$^{Asp(GUC)}$ at position C38 [159]. This was confirmed by others for TRDMT1 in human cells [160], *D. melanogaster* [59, 73] and the parasite *E. histolytica* [161]. Whether its (weak) DNA methyltransferase activity still has a biological role is in dispute [157, 158].

1.3.5.B Structure/Catalysis

TRDMT1 contains all the sequence motifs in the linear order characteristic of DNA:m^{5}C methyltransferases [154-156] (Figure 1.3B) and the TRDMT1 crystal structure strongly resembles those of prokaryotic DNA methyltransferases [162]. TRDMT1 employs a DNA methyltransferase-like catalytic mechanism of RNA methylation [160]. It is initiated by a nucleophilic attack of a cysteine residue (C79) located in sequence motif IV, supported by a glutamate (E119) from motif VI and two arginine residues (R160 and R162) from motif VIII [160]. Unlike all other mammalian DNA methyltransferases, TRDMT1 does not possess a large N-terminal extension. In DNMT1, 3A and 3B, this extension is important for specific targeting of the DNA substrate [157]. The tRNA binding site on human TRDMT1 has been mapped through a combination of directed mutagenesis and structural modelling. It consists of a cleft next to the AdoMet binding pocket that mainly interacts with the anticodon stem and loop of the tRNA substrate [163].

TRDMT1 does not have a canonical nuclear localisation signal and different findings have been reported on its intracellular localisation. For instance, expression of tagged versions showed it to be primarily cytoplasmic in NIH3T3 cells [159] while in *Dictyostelium* it appeared to be primarily nuclear [164]. In *D. melanogaster*, tagged TRDMT1 resides in both the nucleus and cytoplasm and its localisation is highly dynamic during the cell cycle. It enters nuclei during prophase and shows a spindle-like localisation pattern during mitotic divisions [165]. Studies with endogenous and tagged proteins in HeLa cells indicate that TRDMT1 is normally present in the nucleus, but under conditions of stress, it relocalises to the cytoplasm and concentrates in stress granules and P bodies [166], consistent with findings in *D. melanogaster* [73].

1.3.5.C Substrate specificity

The initial report was that human TRDMT1 modifies solely $tRNA^{Asp(GUC)}$ at position C38. This was shown with tRNA substrates from mouse, *D. melanogaster* and *A. thaliana* [159]. More recently it was demonstrated that *D. melanogaster* TRDMT1 can also modify $tRNA^{Gly(GCC)}$ and $tRNA^{Val(AAC)}$ at the equivalent position [73] (Figure 1.3D). Aza-IP performed on HeLa cells also revealed these tRNAs as TRDMT1 targets, as well as $tRNA^{Val(CAC)}$ in the equivalent position [77]. This exquisite specificity is not well understood but it probably involves in part the recognition of tRNA secondary structure and it might be noted that in each substrate the modified cytosine is in a CpG context. The bsRNA-seq study did not reveal any non-tRNA substrates for TRDMT1, although only a limited panel of five exemplary RNAs were screened [58]. What warrants the listing of TRDMT1 in this context is that Aza-IP identified one mRNA target for this enzyme [77].

No m^5C demethylase or reader has thus far been described.

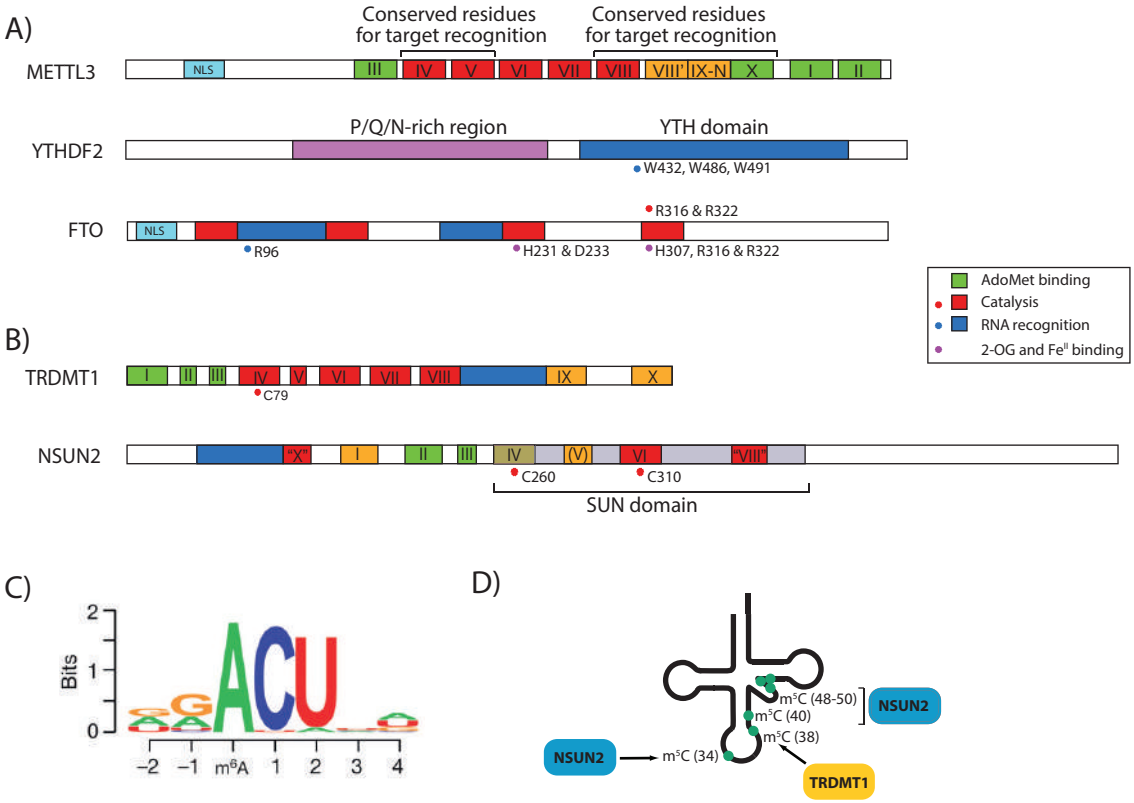


Figure 1.3. m^{6}A and m^{5}C writers, erasers and readers known to act on mRNA.

A) Schematic representations of the domain structures of the m^{6}A methyltransferase METTL3, reader YTHDF2 and demethylase FTO and B) the m^{5}C methyltransferases TRDMT1 and NSUN2. Roman numerals refer to secondary structure elements constituting the RFM methyltransferase fold. Elements making up particular enzyme domains or specific amino acid residues are colour-coded as follows: green: AdoMet binding; red: catalytic domain; blue RNA recognition; purple 2-oxoglutarate (2-OG) and iron (Fe) binding; see text for details and references. C) Sequence logo representing the preferred sequence context for m^{6}A in human mRNAs as determined by m^{6}A-seq. (Reprinted with permission from ref. [57]. Copyright 2012 Nature Publishing Group). D) Elements recognised by NSUN2 and TRDMT1 in mRNA are not known. Shown are positions in the tRNA cloverleaf structure that are methylated by the two enzymes in animal tRNAs.

1.4 ROLE IN BIOLOGY AND HUMAN HEALTH

To my knowledge there are no reports thus far demonstrating a direct causative link between internal mRNA methylation patterns and a specific physiological process or disease. Nevertheless, it has been commonly observed that the relative abundance of m^6A within mRNA populations can vary with cellular context and between species, e.g. refs. [20, 22, 24, 25, 32-37, 41, 44, 114]. Little is known about changes to the proportion of m^5C in cellular mRNAs under different conditions. The recent transcriptome-wide mapping studies added some new information. Anti-m^6A immunoblotting demonstrated presence of the modification in poly(A)+ RNA across a panel of murine tissues with highest levels in liver, kidney and brain. m^6A was detectable in poly(A)+ RNA throughout murine neural development with a dramatic increase by adulthood [56]. This resonates with dynamic changes in m^6A levels during plant development and during meiosis in *S. cerevisiae* (see below) [42, 114, 167]. Together with the reversible nature of m^6A through demethylation, this might warrant usage of the term 'epitranscriptomics' [66] or 'RNA epigenetics' [168, 169]. m^6A-seq analyses of responses by HepG2 cells to hepatocyte growth factor or interferon g addition, UV irradiation and heat shock showed some changes, although the majority of m^6A sites were shared between conditions. Comparisons of RNAs from HepG2 cells, mouse liver and human brain revealed that the principal characteristics of the m^6A methylome are not unique to cancer [57]. GO analyses of both sets of m^5C or m^6A-containing mRNAs did not reveal strong enrichments. Rather, it showed that both types of methylation affect mRNAs involved in a variety of cellular functions [56, 58].

Use of pharmacological methylation inhibitors also provided some insight. For instance, the antiviral adenosine analogue Neplanocin A (NPC) and the amino acid analogue cycloleucine have both been used to trigger accumulation of undermethylated cellular RNA and study the consequences for RNA processing, transport, translation and steady-state levels, e.g. refs. [38, 170-178]. However, both drugs generally decrease AdoMet levels and thus have pleiotropic effects at the cellular level. The ribonucleoside analogue, 5-aza-C, is an anti-cancer drug that promotes DNA hypomethylation and tumour-suppressor gene reactivation. 5-aza-C is incorporated into DNA, leading to covalent trapping of DNA methyltransferases. However, 5-aza-C is preferentially incorporated into RNA and also known to inhibit tRNA methylation [179-181], acting

at least in part through inhibition of TRDMT1 by a mechanism similar to that described for DNA [72]. The Aza-IP study also demonstrated that NSUN2 is inhibited by 5-aza-C [77]. The anti-cancer properties of 5-aza-C will thus likely also include effects on tRNA function; whether it also inhibits m^5C formation in mRNA remains to be shown.

In addition to the findings described above, considerable and highly specific information on knockdown/knockout phenotypes of several RNA:m^5C as well as m^6A methyltransferases, readers and demethylases is available. Aspects of these phenotypes will very likely be attributable to defective mRNA methylation and thus I provide a summary of these observations in the following.

1.4.1 RNA:m^6A METHYLTRANSFERASES METTL3 AND METTL14

There are no known disease-causing mutations for METTL3 or METTL14 and the consequences of a knockout of mammalian *METTL3* or *METTL14* have not yet been explored in great detail. Silencing of METTL3 in HepG2 cells resulted in apoptosis and pronounced changes in gene expression [57]. ~2,000 genes were differentially expressed and a further ~500 genes, whose overall level was unchanged, exhibited alternative splicing (see next section). Methylated mRNAs were enriched in both sets. GO analysis of differentially expressed genes showed significant enrichment of the p53-signalling pathway. Isoform switching was further noted for many mRNAs encoding components of this pathway, suggesting that modulation of p53 signalling through splicing might be relevant to induction of apoptosis by silencing of METTL3 [57]. Several recent *METTL3* knockdown studies have also highlighted its role in cell survival and development; knockdown of *METTL3* in mouse embryonic stem cells (ESCs) resulted in flatter and compact cells with subsequent loss of their self-renewal capacity [105], while in HeLa cells led to decreased cell viability [68]. In agreement with Dominissini et al., genes targeted by METTL3 were enriched for functional pathways related to cell death and cellular response to stress [68]. Finally, silencing of METTL3 in human and mouse cells resulted in elongation of the circadian period through changes in the transcription of RNA-processing machinery [182].

The biological role of METTL14 still remains to be explored in further detail, however its knockdown in HeLa cells resulted in cell death, consistent with the phenotype associated with *METTL3* knockdown [68].

A panel of mammalian cell lines, including tumorigenic, non-tumorigenic and virally-transformed cells, were all found to express a short form and a long form of *METTL3* mRNA [183]. The total amount of *METTL3* mRNA was similar between lines, but up to nine times more of the long than the short splice variant was detected in tumorigenic cell lines, compared to a 1.5-fold difference in the non-tumorigenic cell lines. Since the short mRNA lacks exons 2 and 4 and is likely non-functional, these findings are consistent with earlier reports of an 8-15-fold increase of METTL3 enzyme activity as a correlate of cellular transformation. Simple increases in cellular proliferation did not account for this effect [184].

Both METTL3 and WTAP appear to function in a number of developmental processes in zebrafish. Morpholino knockdown of *WTAP* in embryos resulted in multiple developmental defects while knockdown of *METTL3* resulted in more modest defects [69]. Interestingly, simultaneous knockdown of both enzymes led to severe developmental defects and increased apoptosis, consistent with observations in mammalian cells. This suggests that at least two components of the methyltransferase complex are important for development.

Knockout of MTA in *A. thaliana* arrests embryos at the globular stage of development, resulting in embryonic lethality. Expression of MTA in wild-type plants is strongly associated with dividing tissues, particularly reproductive organs, shoot meristem and emerging lateral roots [41]. Consistent with this, severe reduction of MTA and m^6A levels during later growth stages led to altered trichome branching and changes in leaf as well as flower structures indicative of reduced apical dominance and problems with organ definition [114]. *D. melanogaster* embryos with homozygous inactivation of *Dm ime4* died at larval and pupal stages; rescue of this phenotype was dependent on the enzyme's catalytic activity. *Dm ime4* is highly expressed in ovaries and testes. Hypomorphic alleles of *Dm ime4* revealed reduced fertility of both sexes and a critical function in Notch signalling during oogenesis, although no specific RNA target was identified [185]. *S. cerevisiae ime4* is a non-essential gene that is not appreciably

expressed during normal growth but strongly induced in starved diploid cells, concurrent with an increase in mRNA m^{6}A content [42, 167]. Increases in m^{6}A as well as the ensuing entry into meiosis and spore formation are all dependent on Ime4's catalytic activity [42, 44, 104].

The picture that emerges from these observations is that m^{6}A formation by the METTL3, METTL14 and WTAP complex and their orthologues is typically an essential function in multicellular organisms and it has important roles in cell differentiation and developmental decisions.

1.4.2 RNA:m^{6}A READER YTHDF2

Several phenotypes were found resulting from translocation events and microsatellite polymorphisms of *YTHDF2*. Reciprocal translocations of *YTHDF2* with *RUNX1* along with two other translocation events were detected in patients exhibiting acute myeloid leukaemia (AML) [186]. *RUNX1* encodes a transcription factor involved in regulating differentiation of haemotopoietic stem cells into mature blood cell [187]. The predicted effect of this hybrid transcript is the production of a truncated N-terminus RUNX1 protein without translating *YTHDF2*. However, as several translocation events were reported, it is difficult to discern whether this phenotype can be attributed to the *YTHDF2* and *RUNX1* reciprocal translocation. Interestingly, a microsatellite polymorphism located in the fourth intron of the *YTHDF2* gene has been reported to be associated with human longevity [188]. This genotype was also associated with increased *YTHDF2* mRNA in immortalised lymphocytes.

Functional clustering of YTHDF2 target transcripts against non-target transcripts revealed functions in gene expression and cell death and survival, consistent with the phenotypes associated with knockdown of components of the methyltransferase complex [120]. As with the m^{6}A methyltransferase complex, knockdown of *YTHDF2* in HeLa cells resulted in drastically reduced cell viability [120].

Thus, YTHDF2 may function in binding m^{6}A-containing RNA to regulate cell differentiation and survival.

1.4.3 RNA:m^{6}A DEMETHYLASES FTO AND ALKBH5

Interest in FTO skyrocketed when GWAS in multiple populations identified common genetic variants in the *FTO* gene as risk alleles for increased body mass, obesity and type 2 diabetes mellitus, reviewed in refs. [128, 137, 169, 189-192]. This was further extended to potential roles of FTO in neurological diseases including depression [193], dementia and Alzheimer's disease [194, 195]. An autosomal-recessive lethal syndrome in a consanguineous family was found to be due to an inactivating FTO mutation (R316Q in the catalytic core domain). All affected individuals suffered from severe intrauterine and postnatal growth retardation, microcephaly, severe psychomotor delay, functional brain deficits, as well as characteristic facial dysmorphism, and died before the age of three. Homozygous mutant skin fibroblast cultures showed impaired proliferation and accelerated senescence [196].

FTO is expressed from embryonic day 8.5 in the mouse [126, 197] and is widely expressed in human and murine adult tissues, with particularly high expression in many brain regions [129, 132, 198, 199]. Several mouse models with impaired or increased FTO function have been generated [200], with a number of these exhibiting impaired regulation of fat mass and/or energy expenditure [198, 201-203]. However, the direct role of FTO in regulating fat mass has been recently called into question. One study has suggested that the obesity associated non-coding sequences within *FTO* directly interact with the promoter region of *IRX3* and *FTO* in humans, mice and zebrafish [204]. A direct link was reported between *Irx3* expression and regulation of body mass and composition, with *Irx3*-deficient mice exhibiting 25-30% reduction in body weight. This indicates that the association of *FTO* with *IRX3* may be causative of obesity rather than FTO itself, and that m^{6}A does not function in regulating fat mass and obesity. However, a number of developmental defects have also been reported in mice exhibiting loss of FTO. The original Fused toes (Ft) homozygous mutant mice exhibited severe developmental defects with death during gestation [126, 197]. However, the 1.6 Mb deletion of 6 genes including *Fto* in these mice is not specific enough to assign this phenotype to a lack of FTO. Whole-body homozygous *Fto* knockout mice were viable, although they exhibited high perinatal lethality and postnatal growth retardation [198, 203]. Recently, a series of mice with distinct temporal and spatial loss of *Fto* expression were reported [205]. Global germline loss of *Fto* resulted in high perinatal lethality and

a reduction in body length in addition to a reduction in fat and lean mass. Adult onset global knockout mice circumvented the lethality. Ablation of FTO in mediobasal hypothalamic regions of the brain had only mild effects [205].

Finally, a role for FTO in cellular nutrient sensing through signalling by the mammalian target of rapamycin complex 1 (mTORC1) has been described. Cells lacking FTO catalytic activity display decreases in activation of the mTORC1 pathway and mRNA translation, as well as increased autophagy, all of which are likely to contribute to the FTO phenotype of stunted growth seen in humans and mouse models [206].

No association of ALKBH5 with human disease has thus far been reported. Murine *Alkbh5* mRNA was detected in all organs studied, with the highest expression found in testes. *Alkbh5*-deleted male mice displayed impaired fertility resulting from apoptosis that affects meiotic metaphase-stage spermatocytes [125]. Transcriptome profiling of wild-type and *Alkbh5*-deleted mouse testes revealed a total of 1,551 genes with significant differences in expression. 646 genes were differentially expressed at the gene level and 1,752 showed differentially expressed isoforms. These genes belonged to a range of ontologies, including spermatogenesis-related functions and members of the p53 functional interaction network, consistent with the induction of apoptosis and differentiation defects.

Although a number of studies have suggested that FTO is involved in regulating obesity, the recent discovery that IRX3 is causative of obesity indicates that m^6A patterns are not involved in regulating this phenotype. Rather, FTO is suggested to be important for the developmental process. Similarly, ALKBH5 appears to be important for growth and differentiation, consistent with the roles identified for the m^6A methyltransferase complex and YTHDF2.

1.4.4 RNA:m^5C methyltransferase NSUN2

The earliest link of NSUN2 to human disease stems from its identification as a direct target of Myc, a well-known proto-oncogene [99]. NSUN2 showed low expression in normal tissues but was induced in a variety of tumours, including oral and colorectal cancers [99, 207, 208]. Myc-induced proliferation required NSUN2, and growth of

human squamous-cell-carcinoma xenografts in nude mice was decreased by *NSUN2* RNAi [99]. In 2012, it was shown that mutations in the *NSUN2* gene cause autosomal recessive intellectual disability syndromes. Three separate studies presented one missense, two nonsense and two splicing mutations found in five independent consanguineous families [147, 209, 210]. Common features of homozygous patients were moderate to severe intellectual disability, facial dysmorphism, short stature, microcephaly and muscular hypertonia. Consistent with the neurological defects in patients, murine NSUN2 was enriched in Purkinje cells of the cerebellum [210] and *D. melanogaster* lacking the NSUN2 orthologue CG6133 had severe short-term memory loss [209]. The missense mutation Gly679Arg was shown to cause mislocalisation from a nuclear/nucleolar distribution to a nucleoplasmic and cytoplasmic pattern [210]. Consistent with a role in neurological development, *Nsun2* deficient mice exhibited an accumulation of 5' tRNA fragments, activating stress pathways and leading to increased apoptosis in the cortical, hippocampal and striatal neurons [150].

Levels and localisation of wild-type NSUN2 can vary during the cell cycle [99, 208]. One study found that NSUN2 was concentrated in the nucleolus during G1/S/G2 phases but showed perichromosomal and cytoplasmic enrichment during mitosis. NSUN2 was further phosphorylated by Aurora-B (targeting Ser139) during mitosis, when nucleoli disassemble. Aurora-B phosphorylation inhibited interaction of NSUN2 methyltransferase activity as well as its interaction with the nucleolar proteins NPM1/nucleophosmin/B23 and nucleolin/C23 [208]. In the other study, NSUN2 was also predominantly nucleolar in G1 phase. Levels were highest in S phase cells where NSUN2 distributed more evenly throughout the nucleus. In G2 phase cells, NSUN2 was partially detected in cytoplasmic vesicles and in M phase cells NSUN2 decorated the spindle. The nucleolar localisation of NSUN2 was dependent on RNA polymerase III activity; its knockdown decreased nucleolar size and inhibited cell-cycle progression [99]. Interestingly, cultured cells depleted of NSUN2 exhibit increased apoptosis, reduced proliferation and a variety of spindle defects leading to abnormal mitosis. NSUN2's effects on mitotic spindle stability were independent of its methyltransferase activity. Given the above it has been hypothesised that NSUN2 plays a dual role during the cell cycle, one requiring its methyltransferase activity and another in mitotic spindle stability [211].

Homozygous *Nsun2* deleted mice were viable, but were ~30% smaller than wild-type controls at three months. Homozygous males were sterile due to a requirement of NSUN2 for testis differentiation. Interestingly, NSUN2 was found to localise to the chromatoid body, a cytoplasmic RNA processing organelle essential for post-transcriptional regulation of gene expression in haploid germ cells [212]. Both genders showed cyclic alopecia at around ten months of age. Reporter tagging showed that NSUN2 was dynamically expressed from E3.5 during embryogenesis and high in adult skin. Detailed analysis of the hair loss phenotype showed that NSUN2 was required to balance stem cell self-renewal and differentiation in the epidermis [213]. Interestingly, mice lacking both NSUN2 and TRDMT1 showed a striking synthetic lethal interaction (see below) [149].

In summary, NSUN2 displays strong links to cell cycle control as well as (stem) cell differentiation and proliferation. Of note, the mitotic spindle-related role of NSUN2 does not require its methyltransferase activity. Excess activity of NSUN2 is involved in cancer, whereas loss of function causes complex developmental phenotypes including stunted growth and intellectual deficits. In addition to NSUN2, two other RNA methyltransferases, FTSJ1 (TRM7 in *S. cerevisiae*; 2'-*O*-methylation in tRNA) and TRMT1 (TRM1 in *S. cerevisiae*; guanine dimethylation in tRNA), have also been shown to cause intellectual disability [214-216].

1.4.5 RNA:m^5C methyltransferase TRDMT1

Mutations of TRDMT1 have thus far not been found in human disease. Mice, *D. melanogaster* or *A. thaliana* homozygous for inactivating mutations in TRDMT1 failed to show any overt phenotypes under laboratory conditions [159]. Closer inspection of TRDMT1-deficient flies revealed an increased sensitivity to oxidative and thermal stress, perhaps related to TRDMT1's demonstrated function in preventing stress-induced cleavage of tRNAs [73]. TRDMT1-overexpressing flies and amoebas had increased stress tolerance [217, 218]. By contrast, morpholino knockdown of TRDMT1 in zebrafish embryos led to lethal differentiation defects in the retina, liver, and brain. This effect was dependent on the presence of TRDMT1 in the cytoplasm but not the nucleus [219]. TRDMT1 is detectable in a range of human and mouse tissues, with

highest expression in testes, ovary and thymus [155]. Taken together, information remains sparse on the relevance of TRDMT1 for normal cell and organismic function.

Interestingly, a recent study reported severe synthetic genetic interactions between TRDMT1 and NSUN2 in mice [149]. Efforts to breed double knockout mice revealed severe intrauterine and perinatal lethality. Pups that were born alive were small and underdeveloped in several tissues, including malformation of the cerebral cortex, an immature skeleton and reduction in brown adipose tissue. Mutant fibroblasts isolated from embryos had a reduced ability to differentiate into adipocytes. While the m^5C content of mRNAs was not assessed in these cells, they exhibited complete loss of m^5C in tRNA, consistent with the complementary tRNA site specificities of the two enzymes. As a consequence, levels of affected tRNAs and global protein synthesis rates were also reduced. These findings, together with a lack of appreciable changes to mRNA levels, suggested the severe *Nsun2*/*Trdmt1* double knockout phenotype was in large part due to defects in tRNA function [149].

In summary, the enzymes discussed are typically found to be of importance for proper embryonic and neural development and function (Table 1.3). Roles in germ cells, and links to cell differentiation, growth/transformation as well as cell cycle regulation are also frequently observed. A major impediment to understanding the molecular (patho)physiology of mRNA methylation is that the full extent of each enzyme's specificity for particular RNA biotypes, or lack thereof, is not known. For METTL3, METTL14, YTHDF2, FTO and ALKBH5 the available evidence is firmly in favour of mRNAs as their substrate, although activity towards other transcript classes cannot be excluded. A major role for NSUN2 and TRDMT1 clearly is to modify tRNAs. Nevertheless, there is clear evidence that NSUN2 further acts on mRNAs and other ncRNAs. A big gap in our knowledge is how transcriptome-wide methylation patterns differ between health and disease states or how they respond to experimental intervention. The newly developed global approaches to map m^6A and m^5C can now be applied to close this gap.

Table 1.3. Biological functions and links to human health of RNA:m^5C and m^6A methyltransferases, readers and demethylases.

Modification	**m^6A**				**m^5C**	
MTase/ reader/ demethylase	**METTL3**	**YTHDF2**	**FTO**	**ALKBH5**	**NSUN2**	**TRDMT1**
Biological Functions	- Required for embryonic development in plants and flies [41, 185] - Required for meiosis and sporulation in yeast [42] - Hypomorph yields fertility defects in flies [185] - Knockdown in mESCs causes loss of self-renewal capacity [105] - Knockdown in HeLa cells resulted in decreased cell viability [68] - Knockdown causes elongation of circadian period [182]	- Knockdown in HeLa cells causes reduced cell viability [120]	- Germline inactivation causes severe growth retardation and multiple malformations in mice [196, 198, 203] - Overexpression or conditional loss in adult mice causes body mass phenotypes [202, 205] - Role in cellular nutrient sensing [206]	- Germline deletion causes male infertility in mice [125]	- Germline deletion in mice causes growth retardation, cyclic alopecia and male infertility [213] - Required for mitotic spindle stability [208]	- No phenotype in mice - Stress tolerance phenotypes in flies and amoebae [73, 217, 218] - Knockdown in zebrafish is embryonic lethal [219]

Links to human health	- Overexpression in cancer [183]	- Reciprocal translocation with *RUNX1* detected in AML patients [186] - Microsatellite polymorphism in intron associated with longevity in humans [188]	- Common variant associated with risk for obesity [127, 129, 131, 220] depression [193], dementia and Alzheimer's disease [194, 195] Mutations cause lethal brain and growth deficits [196]	- No links reported	- Overexpressed in cancer and tumours [99, 207, 208] Mutations cause dysmorphogenesis and intellectual disability [147, 209, 210]	- Inhibited by 5-aza-C (anti-cancer drug) [72]

1.5 MOLECULAR FUNCTION

Recent studies investigating m^6A have reported several molecular functions for this modification in mRNA, whereas little definitive evidence exists for specific molecular consequences of the presence of m^5C in mRNA. One lead for the function of m^5C in mRNA might come from the study of innate immunity. Cells can detect dsRNA derived from invading viruses and trigger subsequent activation of adaptive immune responses through RIG-1-like (retinoic acid-inducible protein 1) and TLR (Toll-like) receptors. Additionally, RIG-1 and MDA5 (melanoma differentiation-associated gene 5) recognise foreign RNAs exhibiting a 5' triphosphate [221] or a cap lacking 2'-*O*-methylation, [222], respectively. Of note here, it was found that cellular responses to introduced 5' triphosphate carrying RNAs can be attenuated by incorporating a range of modified nucleosides at internal positions. RIG-1 activation was lessened by ψ and 2-thiouridine substitution [221] and TLR activation was prevented by substitution with 2-thiouridine, 5-methyluridine, ψ, m^6A, and m^5C in a dose-dependent manner [223]. Expression of introduced mRNA in cells was also typically affected by the presence of modified nucleosides. For instance, expression of reporter mRNAs was increased when substituted at high level with ψ or m^5C, while other modified nucleosides either had little effect or, in the case of 2-thiouridine and m^6A, abolished expression [224]. A high profile application of these findings was the efficient generation of induced pluripotent stem cells (iPSCs) using transfection of ψ- and m^5C-substituted mRNAs encoding the Yamanaka transcription factors KLF4, c-MYC, OCT4 and SOX2 to bypass antiviral responses [225]. Another study demonstrated that nonsense codons could be converted into sense codons by targeted pseudouridylation [17]. These results illustrate how mRNA modifications can trigger marked biological responses. Nevertheless, given the high substitution level required for these effects it is unclear whether they resemble the typical function of m^6A or m^5C when present in mRNA at physiological and therefore typically much lower density.

Transcriptome-wide mapping of m^6A and m^5C in different cellular contexts is expected to pave the way for a broad-based elucidation of the function of mRNA methylation and the existing datasets have already yielded interesting, if largely correlative data [56-58, 66]. More recent studies of m^6A have extensively investigated the molecular function of

this modification in mRNA, yielding substantial evidence for a role in post-transcriptional gene regulation.

1.5.1 Biased distribution of m^{6}A in mRNA

The MeRIP-seq study [56] noted that the vast majority of detected peaks were intragenic, with 50.9% residing in the CDS, 41.9% in UTRs and 2% in introns, although the latter figure could be an underestimate since sequencing libraries were not enriched for pre-mRNA. 93% of the UTR peaks fell into 3' UTRs, with only 7% in 5' UTRs. Although about half of all methylated mRNAs had only a single peak, those that had several displayed a strong tendency for contiguous or clustered peaks. Superimposing peak distribution along all methylated mRNAs revealed a prevailing pattern. m^{6}A density was low in the 5' UTR and in the 5' region of the CDS, with a steady increase towards the 3' end. m^{6}A within the 3' UTR was enriched near the stop codon and density fell again towards the 3' end. With respect to other mRNA landmarks, no association between polyadenylation signals or splice junctions and sites of m^{6}A was found. Instead, an interesting link to miRNA biology was uncovered. 67% of 3' UTRs containing m^{6}A also exhibited a predicted miRNA target site, higher than the ~30% expected by chance. There was an anti-correlation between m^{6}A sites, which tended to be near the stop codon, and miRNA target sites, which tended to be near the 3' end of the 3' UTR. In the brain, targets of highly expressed miRNAs were furthermore enriched among the m^{6}A-containing mRNAs. Overall, this suggested that miRNA levels might control mRNA methylation and/or that a spatial separation within the 3' UTR might be needed for the two to somehow affect each other [56].

The m^{6}A-seq study [57] reported qualitatively very similar findings, with peaks concentrated in intragenic regions, although only 1/3 to 1/2 of peaks were in common between both studies [66]. On average ~1.7 m^{6}A sites were detected per mRNA. Analysing m^{6}A site distribution along mRNAs, an enrichment near transcription start sites was noted, together with a moderate enrichment elsewhere in the 5' UTR. The enrichment near the mRNA 5' end was likely due to the common occurrence of (N^6-,2'-*O*-)-dimethyladenosine (m^{6}Am) at the first transcribed position adjacent to the cap structure. Interestingly, the enzyme responsible for formation of m^{6}Am appears to be distinct from METTL3 [40] and m^{6}Am/Am ratios vary in a tissue- and mRNA-specific

manner [226]. m^{6}A sites in the CDS were frequent (37% of all peaks), although less dense than expected at random, whereas a strong enrichment in the region ±200 nt from the stop codon was seen (28% of all peaks) with a depletion in the remainder of the 3' UTR (Figure 1.4A). A relative enrichment of m^{6}A sites in long internal exons and analyses of gene expression in METTL3 knockdown cells revealed multiple patterns suggestive of a link to (alternative) splicing. First, down-regulated genes in the METTL3 depleted cells were enriched for examples of m^{6}A sites in intronic regions. Second, mRNAs with differentially spliced exons were enriched for m^{6}A sites. Third, the density of m^{6}A sites was higher in mRNAs derived from multi-isoform genes than from single isoform genes. The authors further noted overlap of m^{6}A sites with multiple *cis*-acting RNA elements, among them conserved sequence motifs, internal ribosome elements, zip-code and destabilising elements. Together with the concentration around stop codons this was taken as highly suggestive of potentially diverse roles of m^{6}A in the regulation of mRNA translation or stability [57].

Prior studies are broadly consistent with the new global mapping data in that they showed m^{6}A to be present at 1-5 internal residues per molecule sited throughout the length of the mRNA, often at sub-stoichiometric levels per site [20-39]. Some sites were found to be clustered together as in the case of Rous sarcoma virus RNA [46]. Mature bovine prolactin mRNA exhibited a cluster of m^{6}A residues within the 3' region of the molecule, about 2/3 of them concentrated within a 108 nt region in the 3' UTR [45, 227]. The prolactin pre-mRNA furthermore had a predominant m^{6}A site in its final intron [228], indicating that mRNA methylation can take place in the nucleus before splicing. Intronic sites have also been detected in Simian Virus 40 RNA [35]. Enrichment in mRNA 3' regions was typically, though not exclusively, seen. Bulk analysis of *A. thaliana* cellular mRNAs showed that m^{6}A is strongly enriched (>90%) near the 3' end in a region 100-150 nt before the poly(A) tail [114]. In meiotic *S. cerevisiae* cells, m^{6}A was distributed across all mRNA size ranges and the *ime1, -2,* and *-4* mRNAs were specifically shown to contain m^{6}A. Again, m^{6}A sites in *ime2* mRNA were narrowed down to a region in its 3' half [44]. By contrast, older reports found that m^{6}A was not present in the 3' third of HeLa cellular mRNA [28]; m^{6}A and m^{5}C sites were seen only distant from the poly(A) tail of HeLa or Adenovirus RNA [30], and

Simian Virus 40 RNAs contained on average three m^6A sites in the 5' portion of the CDS [229].

1.5.2 Molecular function of m^6A

Treatment of cells with methylation inhibitors such as cycloleucine, NPC or S-tubericidinylhomocysteine has been used in different contexts to trigger accumulation of undermethylated cellular RNA and study the consequences for mRNA processing, transport, translation and steady-state levels, reviewed in refs. [38, 114]. An accumulation of viral or cellular pre-mRNA in the nucleus was often seen, consistent with a role of mRNA methylation in splicing and/or nuclear mRNA export [172, 175, 176, 228, 230]. Several of the aforementioned studies saw no change in mRNA translation efficiency; however, others did observe reduced translation of hypomethylated mRNA [174, 177]. Since all these studies were performed by generally depleting cellular AdoMet levels it is not certain whether these effects were due to a depletion of m^6A or other methylated nucleosides in mRNAs, or indeed due to hypomethylation of other cellular components. Nevertheless, they are consistent with functions of m^6A in mRNA splicing, stability and translation as suggested by MeRIP-seq and m^6A-seq [56, 57]. Further support for such functions comes from the aforementioned formation of complexes between Ime4/MTA/METTL3-METTL14 with Mum2/FIP37/WTAP, an RNA-binding protein involved in mRNA splicing [41, 68, 69, 104] and the recent discovery of the alternative splicing-related protein YTHDF2 as a m^6A reader [57, 120].

More recent efforts to understand the molecular function of m^6A have implemented transcriptome-wide techniques and have provided almost definitive roles for this modification in mRNA. Several studies have reported that depletion of METTL3, METTL14 or WTAP resulted in increased stability of mRNA targets [105, 107] and variations in alternative splicing [69]. Consistent with this finding, depletion of YTHDF2 resulted in increased levels of target mRNAs with prolonged lifetime and an increase in the m^6A/A ratio in total mRNA, mRNA associated with 40S-80S and mRNA from polysomes, indicating that YTHDF2 functions in destabilising m^6A-containing mRNA [120]. YTHDF2 was also found to mediate relocalisation of m^6A-containing mRNA to P bodies where mRNA decay occurs, consistent with the role of

m^6A in mRNA destabilisation [120]. In contrast, the mRNA stabiliser human antigen 1 (HuR) was reported to preferentially bind unmethylated mRNA in *METTL3* and *METTL14* knockdown cells compared to control cells, suggesting that m^6A may hinder HuR binding and mediate binding of YTHDF2 for degradation of m^6A-containing mRNA [105]. A recent study has also supported the role of m^6A in alternative splicing regulation. This study reported that FTO depletion promoted the RNA-binding affinity of serine/arginine-rich splicing factor 2 (SRSF2), leading to increased inclusion of exons [231].

The roles of m^6A determined in these studies are largely in concordance and strongly support the molecular function of m^6A in regulating mRNA stability and alternative splicing.

1.5.3 Biased distribution of m^5C in mRNA

The bsRNA-seq study [58] also noted an uneven distribution of m^5C in the HeLa cell transcriptome. m^5C sites were enriched in several ncRNA types including tRNAs while mRNAs had a relatively low density of sites. Nevertheless, 83% of all detected m^5C sites were found in mature mRNAs. Within mRNAs, sites were depleted relative to random distribution in the CDS but significantly enriched in the 5' and 3' UTRs, respectively (Figure 1.4B). Intron sequences were not surveyed. Enrichment of human regulatory RNA motifs [232] and, notably, Argonaute (Ago) 1-4 footprints [233] near m^5C sites in mRNA was further observed. This suggested a role of m^5C miRNA biology and more broadly in the post-transcriptional control of gene expression.

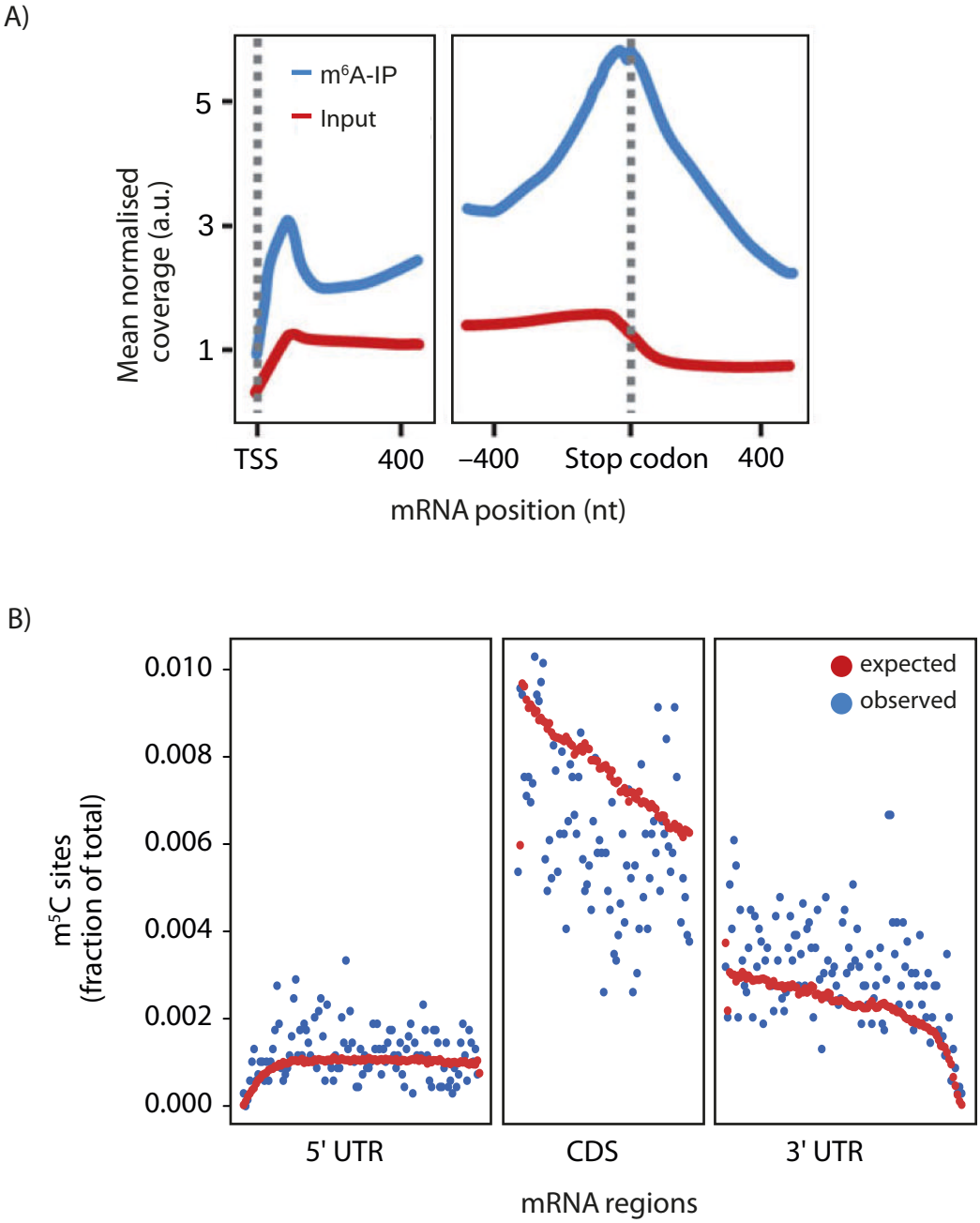


Figure 1.4. Biased distribution of m^{6}A and m^{5}C within mRNAs.

A) Metagene profiles of murine mRNAs displaying m^{6}A distribution in areas near the mRNA 5' end (transcription start site (TSS); left) and around the stop codon (right). Normalised next-generation sequencing read coverage from anti-m^{6}A immunoprecipitation (blue) and input RNA libraries (red) is shown. (Reprinted with permission from ref. [57]. Copyright 2012 Nature Publishing Group). B) Metagene profiles based on human bsRNA-seq data [58] showing relative enrichment of m^{5}C in the mRNA 5' and 3' UTRs and relative depletion within the CDS. UTRs and CDS of each mRNA were divided into an equal number of intervals and the fraction of m^{5}C candidate sites falling into each interval is plotted in blue. The fraction of all cytosines with ≥10 read coverage is plotted as background in red.

1.5.4 Molecular function of m^5C

No published prior attempts have been made to elucidate a molecular function of m^5C in mRNA. Nevertheless, the bsRNA-seq study [58] implicated NSUN2 as a methyltransferase that interacts with and modifies mRNA, and, as detailed above, there is evidence to corroborate this from three independent studies [139, 140, 152]. Interestingly, NSUN2 was further shown to promote the stability of the mRNA encoding the cell cycle regulator CDKN2A/$p16^{INK4}$ [151]. A NSUN2 binding site was mapped in the 3' UTR of $p16^{INK4}$ mRNA. *In vitro* methylation assays demonstrated specific methylation of this region by recombinant NSUN2, which was reduced by mutation of a local AAC motif centred on position 988. While the presence of this motif strongly suggests involvement of m^6A methylation, this is curiously at odds with the known specificity of NSUN2 for m^5C. $p16^{INK4}$ mRNA is further known to be destabilised by interactions of AU-rich element RNA binding protein 1 (AUF1), ELAV1 and Ago2 with its 3' UTR. A series of chimeric reporter studies and pull-down experiments then indicated that NSUN2-mediated methylation prevented the localisation of $p16^{INK4}$ 3′ UTR containing transcripts to processing bodies as well as antagonising the interaction and destabilising effects of HuR/AUF1/Ago2 on $p16^{INK4}$ mRNA [151]. With the caveat that this study suggested an involvement of the m^6A modification, these results indicate that mRNA methylation by NSUN2 can promote mRNA stability directly as well as through antagonising other regulatory RNA binding proteins (RBPs).

Altogether the evidence is clear that internal methylation is more than mere 'decoration' of the mRNA. A regulatory role is already strongly indicated for m^6A by the reversible and context-specific nature of its occurrence. Patterns of specific depletion or enrichment near landmarks along mRNAs suggest that both, m^5C and m^6A interact in some way with mechanisms of post-transcriptional gene control, particularly through miRNAs. Finally, multiple circumstantial pieces of evidence link m^6A to regulation of RBP-mRNA interactions, and as a consequence mRNA splicing, translation or stability. Additional transcriptomic profiles of m^5C and m^6A as well as wider availability of transcriptome-wide RBP footprint datasets will no doubt expand, refine and perhaps in some cases revise current working hypotheses of m^5C and m^6A function. Figure 1.5 summarises the reported and putative molecular steps that have been proposed thus far

to underpin such functions of either m^6A or m^5C, as well as some additional plausible scenarios.

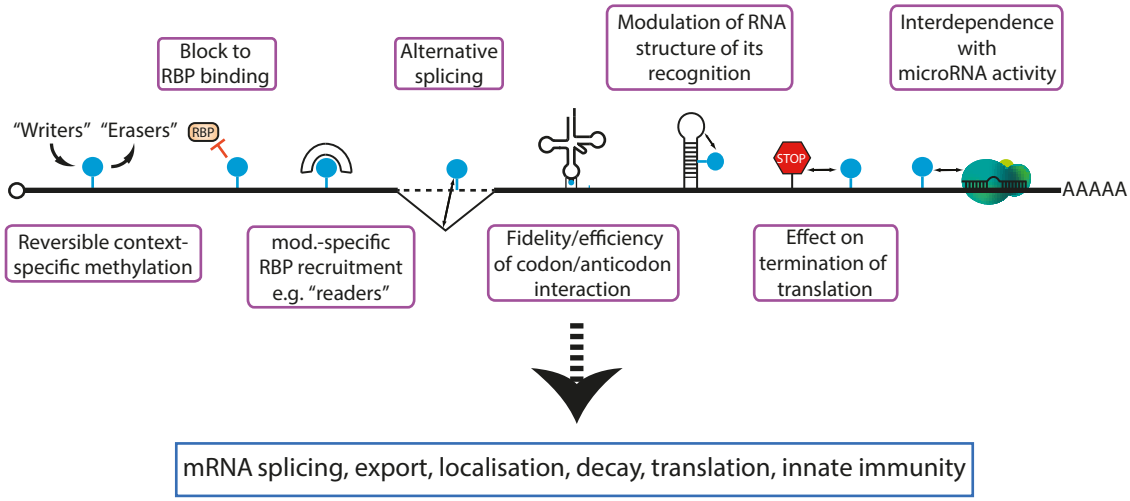


Figure 1.5. Reported roles for m^6A in mRNA and speculative roles of m^5C in mRNA.
Pictograms of reported molecular functions of m^6A and several plausible molecular consequences of the presence of m^5C (represented by blue lollipops) are shown along a stylised mRNA. These in turn could affect different aspects of mRNA utilisation. Evidence consistent with some of the depicted scenarios has been published and others have been proposed. See main text for details and references.

1.6 Aims

It is clear that RNA modifications such as m^6A and m^5C are non-randomly distributed throughout the transcriptome, suggesting that they are functionally important. A number of studies have provided evidence for the molecular function of m^6A in regulating mRNA stability and alternative splicing, although the function of m^5C is still largely speculative. Understanding the molecular and biological functions of these modifications would provide further insight into their role in human diseases and could be a therapeutic target in the future.

The overall aim of this thesis is to gain insight into the RNA targets of NSUN2 and TRDMT1, as well investigate the molecular functions of m^5C in mRNA. In order to understand the biological role of m^5C in RNA and the different cellular contexts that m^5C exists in, the RNA m^5C profiles between normal and cancer cells were compared.

Transcriptome-wide bisulfite sequencing, termed bsRNA-seq, will be used in addition to a bioinformatics pipeline to investigate the RNA m^5C profiles.

1.6.1 Detailed aims

To achieve the overall aims of determining the substrate range of NSUN2 and TRDMT1, exploring the molecular functions of m^5C in mRNA and investigating the presence of m^5C across different cellular contexts, the project was divided into the following four specific aims:

Aim 1: To prepare prostate epithelial cells (PrEC), lymph node metastatic prostate cancer cells (LNCaP) and RNAi-mediated knockdown of *NSUN2* and *TRDMT1* in HeLa cells for the detection of m^5C in RNA, and to improve and establish techniques for the detection of m^5C in RNA (Chapter 3)

Aim 2: To determine the substrate range of NSUN2 and TRDMT1 in RNA (Chapter 4)

Aim 3: To explore the molecular functions of m^5C in mRNA (Chapter 5)

Aim 4: To explore the role of RNA m^5C in cancer by examining the RNA m^5C profiles of PrEC and LNCaP cells, and to compare this with HeLa cells to confirm patterns of m^5C across multiple cell lines (Chapter 6)

CHAPTER TWO

MATERIALS & METHODS

2.1 MATERIALS

2.1.1 CHEMICALS, REAGENTS AND CONSUMABLES

Common chemicals and reagents used in this study and their suppliers are listed below.

Manufacturer	Chemicals/reagents/consumables
5 Prime (Hilden, Germany)	Phase Lock Gel Heavy (1.5 mL)
Amresco (Solon, OH, USA)	EZ-Vision® Three DNA Dye, DTT, glycine
BD Biosciences (Franklin Lanes, NJ, USA)	Bacto™ agar, Bacto™ tryptone peptone digest, Bacto™ yeast extract
Beckman Coulter (Brea, CA, USA)	Agencourt AMPure XP-PCR Purification, 14 x 89 mm polyallomer centrifuge tubes
Bio-Rad (Hercules, CA, USA)	Micro Bio-Spin™ Chromatography Columns
Coles (Hawthorn East, VIC, Australia)	Skim milk powder
Corning (Tewksbury, MA, USA)	T-75 flasks, 6-well plates, Costar® Spin-X® Plastic Centrifuge Tube Filters, Axygen® 8-strip thin wall PCR tube strips and caps
Diagenode (Liège, Belgium)	DNA methylation control pack
Eppendorf (Hamburg, Germany)	1.5 mL tubes, 2.0 mL tubes, 1.5 mL safe-lock tubes
Fisher Biotec (Wembley, WA, Australia)	Sets of dNTPs (100 mM each)
GE Healthcare (Little Chalfort, BKM, UK)	Amersham Hybond™-P PVDF transfer membrane, Amersham Hybond™-N^+ membrane, Whatman™ chromatography paper
Life Technologies (Carlsbad, CA, USA)	Glycogen (5 mg/mL), Lipofectamine®2000 Reagent, Lipofectamine RNAiMax, Opti-MEM® Reduced Serum Media, NuPAGE® Novex® 4-12% Bis-Tris

	Gels, NuPAGE® LDS Sample Buffer, NuPAGE® MES SDS Running Buffer, Novex® TBE-Urea Gels 10%, Novex® TBE-Urea Sample Buffer, Fetal Bovine Serum, DMEM, Dulbecco's phosphate buffered saline (PBS), T-Medium, 0.25% Trypsin-EDTA, L-glutamine, UltraPure™ Phenol:Water (3.75:1 v/v), Fast SYBR® Green Master Mix, TURBO™ DNase, RNaseOUT™, SYBR® Gold Nucleic Acid Stain, UltraPure™ DNase/RNase-Free Distilled Water (dH_2O), ampicillin, 3M sodium acetate (NaAc) (pH 5.5), Tris-HCl pH 8.0, TaqMan® Universal Master Mix, Eukaryotic 18S rRNA Endogenous Control (VIC®), yeast tRNA
Lonza (Basel, Switzerland)	PrEBM Basal Medium, PrEGM SingleQuot Kit Suppl. & Growth Factors, Hanks balanced salt solution, Trypsin neutralising solution, Trypsin-EDTA, SeaKem® LE Agarose
Merck Millipore (Billerica, MA, USA)	Cytobuster™ Protein Extraction Reagent, EMSURE® ethanol absolute for analysis (EtOH), EMSURE® methanol for analysis, EMSURE® isopropanol for analysis, ReBlot Plus Strong Antibody Stripping Solution 10x, ReBlot Plus Mild Antibody Stripping Solution 10x, Novagen® RedAlert™ 10x Western Blot Stain
National Diagnostics (Atlanta, GA, USA)	SequaGel - UreaGel System
New England Biolabs (Ipswich, MA, USA)	BSA 100x, Buffer EcoRI
Promega (Madison, WI, USA)	X-Gal, Tris base
Roche Applied Science (Indianapolis, IN, USA)	Complete protease inhibitor cocktail tablets (EDTA-free)
Sartorius AG	Minisart® syringe filters

(Gottingen, Germany)	
Sigma-Aldrich (St Louis, MO, USA)	Ethidium bromide, Tri Reagent®, BSA, sodium metabisulfite, hydroquinone, Penicillin-Streptomycin, Tween® 20, DMSO, Trypsin-EDTA, formamide, EDTA, PBS, NaOH, chloroform, luminol sodium salt, cycloheximide, HEPES, PMSF, bromophenol blue, xylene cyanol, phenol red, IGEPAL® CA-630, KCl, coumaric acid, NaN_3, $MnCl_4$, RbCl, NaMOPS, hydrogen peroxide (35% by weight), boric acid, Acetic acid (HAc), glycerol, SDS, methylene blue
Terumo (Tokyo, Japan)	20 CC/mL syringe
Thermo Scientific (Waltham, MA, USA)	Nunc® CryoTube® vials (1.8 mL), Nunc® cell culture dish (15 cm), IPTG solution
VWR International (Radnor, PA, USA)	Ammonium acetate, $CaCl_2$, Potassium acetate (KAc), $MgCl_2$, NaCl, sucrose

2.1.2 KITS

Manufacturer	**Kit**
Agilent Technologies (Santa Clara, CA, USA)	RNA 6000 Nano Kit, High Sensitivity DNA Kit
Illumina (San Diego, CA, USA)	TruSeq DNA LT Sample Prep Kit v2, MiSeq Reagent Kit v2 (300 cycles)
Kapa Biosystems (Woburn, MA, USA)	Library Quantification Kit- Illumina/Universal
Life Technologies	Qubit® dsDNA BR Assay Kit, RiboMinus™ Eukaryote Kit for RNA-Seq, SOLiD™ Total RNA-Seq Kit, MEGAscript® T7 Kit
Promega	PureYield™ Plasmid Miniprep System, pGEM®-T Easy Vector System, Wizard® SV Gel and PCR Clean-Up System

Qiagen (Hilden, Germany)	QiaQuick Gel Extraction Kit, MinElute® PCR Purification Kit
Thermo Scientific (Wilmington, DE, USA)	Pierce™ BCA Protein Assay Kit

2.1.3 ANTIBODIES

Manufacturer	Antibody	Dilution used
Abcam (Cambridge, UK)	Rabbit polyclonal anti-DNMT1 antibody	1:800
Agilent Technologies	Polyclonal rabbit anti-mouse immunoglobulin HRP	1:1000
Diagenode	5-mC monoclonal antibody 3D33	1:500
Millipore	Goat anti-rabbit IgG antibody, peroxidase conjugated	1:1000
Proteintech (Chicago, IL, USA)	Rabbit anti-NSUN2 polyclonal antibody, rabbit anti-TRDMT1 polyclonal antibody	1:1000
Sigma-Aldrich	Monoclonal anti-α-tubulin antibody produced in mouse	1:1000

2.1.4 ENZYMES

Manufacturer	Enzyme
Epicentre Biotechnologies (Madison, WI, USA)	Tobacco Acid Pyrophosphatase
Life Technologies	RNaseOUT™ Recombinant Ribonuclease Inhibitor, Superscript® III Reverse Transcriptase, TURBO™ DNase, Platinum® Taq DNA Polymerase
New England Biolabs	EcoRI, T4 Polynucleotide Kinase
Roche Applied Science (Penzberg, Upper Bavaria, Germany)	FastStart Taq DNA Polymerase

2.1.5 Buffers and solutions

2.1.5.A Buffers and solutions for general molecular biology

1x TAE:

40 mM Tris base (pH 8.2), 20 mM HAc and 1 mM EDTA

1x TBE:

89 mM Tris base (pH 7.6), 89 mM boric acid, 2 mM EDTA

1x TE:

10 mM Tris-HCl (pH 7.5), 1 mM EDTA

2x RNA loading dye:

95% formamide, 0.02% SDS, 0.2% bromophenol blue, 0.01% xylene cyanol
1 mM EDTA

20x random primer mix:

35 μM hexamers, 25 μM T12VN

Table 2.1. Oligo sequences for the random primer mix.

Oligo	Sequence
Hexamers	NNNNNN
T12VN	NVTTTTTTTTTTTT

2.1.5.B Buffers and solutions for dot blot assays and western blotting

Antibody storage solution:

5% BSA, 0.02% NaN_3, 1x PBS, phenol red

1x transfer buffer:

25 mM Tris, 192 mM glycine, 10% methanol

1x PBS-T:

1x PBS, 0.1% Tween® 20

Blocking buffer:

5% skim milk powder, 1x PBS, 0.1% Tween® 20

Methylene blue solution:

0.02% methylene blue, 0.3 M NaAc (pH 5.5)

ECL (enhanced chemiluminescence) solution:

Solution A: 0.4 mM coumaric acid, 2.5 mM luminol, 100 mM Tris-HCl, pH 8.0

Solution B: 21% hydrogen peroxide, 100 mM Tris-HCl, pH 8.0

Mix solutions together

2.1.5.C Buffers and solutions for bsRNA-seq and MiSeq amplicon sequencing

1x PAGE elution buffer

2.5 mM ammonium acetate in TE buffer

1x EBT:

10 mM Tris-HCl (pH 8.5), 0.1% Tween® 20

Library dilution buffer:

10 mM Tris-HCl (pH 8.0), 0.05% Tween® 20

2.1.5.D Buffers and solutions for tissue culture

Supplemented PrEBM media (PrEGM):

PrEBM basal medium was supplemented with PrEGM SingleQuot Kit Suppl. & Growth Factors to make PrEGM. PrEGM was stored at 4°C for a maximum of one month.

Supplemented T-media:

T-medium was supplemented with 10% FBS and 2 mM L-glutamine. Supplemented T-medium was stored at 4°C for a maximum of one month.

Supplemented DMEM media:

DMEM medium was supplemented with 5% FBS and 2 mM L-glutamine. Supplemented DMEM medium was stored at 4°C for a maximum of one month.

PrEC freezing media:

90% PrEGM, 10% FBS, 10% DMSO

LNCaP freezing media:

90% supplemented T-medium, 10% DMSO

2.1.5.E Buffers and solutions for sucrose density gradient centrifugation

Sucrose gradient buffer:

20 mM HEPES (pH 7.6), 125 mM KCl, 5 mM $MgCl_2$, 0.1 mg/mL cycloheximide, 1 mM DTT, 0.5 M PMSF

Lysis buffer:

20 mM HEPES (pH 7.6), 125 mM KCl, 5 mM $MgCl_2$, 0.1 mg/mL cycloheximide, 2 mM DTT and 0.5 M PMSF, 1x complete protease inhibitor, 0.5% IGEPAL® CA-630, 100 U/mL RNaseOUT™

2.1.6 DNA Markers

Manufacturer	**Marker**
Bioline (London, UK)	HyperLadder™-V
Life Technologies	1 Kb Plus DNA ladder, 10 bp DNA ladder

2.1.7 OLIGONUCLEOTIDES

Synthetic DNA oligonucleotides (purified by standard desalting) for RT-qPCR, PCR and Sanger sequencing were synthesised by Integrated DNA Technologies (Coralville, IA, USA). Primers that were designed for RT-qPCR and PCR are listed in the relevant sections of this chapter.

2.1.8 SIRNAS

SiRNAs were purchased from Dharmacon (Lafayette, CO, USA).

Table 2.2. siRNAs and their corresponding sequences used for RNAi-mediated knockdown.

siRNA	Catalogue number	Targeted sequences	Gene
siGENOME Human NSUN2 siRNA SMARTpool	M-018217-01-0005	5'-UAAGAAAGAUGGCGUGUGU-3' 5'-GAGUUGGUAUCCUGAAGAA-3' 5'-GAGCGAUGCCUUAGGAUAU-3' 5'-GGAGACGGCACUAUGAGAA-3'	NSUN2
siGENOME Human TRDMT1 siRNA SMARTpool	M-006671-01-0005	5'-GCGAUAUGCUCUUCUGUUA-3' 5'-CAGAAGAAAUUCACAGGAA-3' 5'-GAACGUUGAACCAAAUAUU-3' 5'-CAAAUUCAAGGCUACGAUA-3'	TRDMT1
siGENOME Human DNMT1 siRNA SMARTpool	M-004605-01-0005	5'-GGAAGAAGAGUUACUAUAA-3' 5'-GAGCGGAGGUGUCCCAAUA-3' 5'-GGACGACCCUGACCUCAAA-3' 5'-GAACGGCGCUCAUGCUUAC-3'	DNMT1
siGENOME non-targeting siRNA Pool #1	D-001206-13-05	5'- UAGCGACUAAACACAUCAA-3' 5'- UAAGGCUAUGAAGAGAUAC-3' 5'- AUGUAUUGGCCUGUAUUAG-3' 5'- AUGAACGUGAAUUGCUCAA-3'	N/A

2.1.9 *IN VITRO* TRANSCRIPTS

The *in vitro* transcripts synthesised and used for this study included humanised Renilla Luciferase (R-Luc) and non-humanised R-Luc. The humanised R-Luc *in vitro* transcript was transcribed from the humanised R-Luc insert located in pCl-Neo. The Filipowicz Lab cloned the humanised R-Luc insert into pCl-neo and provided it for this study [234]. The pCl-neo plasmid containing the R-Luc insert is known as humanised R-Luc or hRL 3xB pA in this study. The non-humanised R-Luc *in vitro* transcript was transcribed from the R-Luc insert in the pRL-TK vector. The *in vitro* transcription method is detailed in section 2.2.2.G. The sequences for these *in vitro* transcripts are provided in Appendix 8.1.

2.1.10 PLASMIDS

Manufacturer	Plasmid
Promega	pGEM®-T Easy Vector, pCl-neo, pRL-TK

2.1.11 BACTERIAL STRAINS

Escherichia coli (E. coli) DH5α

2.1.12 BACTERIAL GROWTH MEDIA

Transformation buffer 1 (TFB1):

30 mM KAc, 79 mM $MnCl_2$, 13 mM $CaCl_2$, 100 mM RbCl, 15% glycerol

0.294 g KAc, 0.989 g $MnCl_2$ and 0.147 g $CaCl_2$ and 15 mL 100% glycerol were added to 75 mL Milli-Q H_2O and the pH brought to 5.8 with HAc. 1.2 g of RbCl was then added and the final volume brought to 100 mL with Milli-Q H_2O. The buffer was then filter-sterilised using a 0.22 µm vacuum filter and stored at 4°C.

Transformation buffer 2 (TFB2):

13 mM NaMOPS, pH 7.0, 100 mM $CaCl_2$, 10 mM RbCl, 15% glycerol

0.294 g NaMOPS, pH 7.0 was added to 1.1 g $CaCl_2$, 0.12 g RbCl and 15 ml 100% glycerol. The final volume was brought to 100 mL with Milli-Q H_2O. The buffer was then filter-sterilised using a 0.22 µm vacuum filter and stored at 4°C.

Luria Broth (LB):

10 g/L Bacto™ tryptone peptone digest, 5 g/L Bacto™ yeast extract, 10 g/L NaCl made up to 1 L with Milli-Q H_2O before autoclaving.

Luria broth (LB) agar:

6.5 g Bacto™ agar was added to 400 mL LB and autoclaved. Once cooled to 55°C, ampicillin (100 µg/mL) was added and mixed. Plates were then poured under aseptic conditions and stored at 4°C.

2.2 METHODS

2.2.1 TISSUE CULTURE TECHNIQUES

2.2.1.A MAINTENANCE OF HELA CELLS

HeLa cells were cultured in supplemented DMEM medium and were routinely passaged at 90-100% confluency in either a 1:5 or 1:10 split ratio in a T-75 flask (flask with a 75 cm^2 surface area). For passaging, cells were washed with 5 mL pre-warmed PBS and 2 mL pre-warmed 0.25% Trypsin-EDTA was added to the flask and incubated at 37°C for 5 min. The trypsin was inactivated by addition of 8 mL supplemented DMEM. HeLa cells were maintained at 37°C with 5% CO_2.

2.2.1.B MAINTENANCE OF PREC CELLS

PrEC cells were cultured in PrEGM according to the manufacturer's protocol in a T-75 flask. 2,500 PrEC cells were seeded per cm^2 and media replaced every two days to ensure optimal growth. Upon reaching 80-90% confluency, cells were passaged. Cells were washed with 6 mL pre-warmed Hanks balanced salt solution and 2 mL pre-warmed Trypsin-EDTA was added to the flask and incubated at room temperature for 5 min. The trypsin was inactivated by addition of 12 mL trypsin neutralising solution to the flask and the cells were centrifuged at 1,200 rpm for 5 min before resuspending in PrEGM. PrEC cells were maintained at 37°C with 5% CO_2 until the 9th passage, at which the cells were discarded.

2.2.1.C MAINTENANCE OF LNCAP CELLS

LNCaP cells were cultured in supplemented T-medium and were routinely passaged at 90-100% confluency in a 1:3 split ratio in a T-75 flask. For passaging, cells were washed with 5 mL pre-warmed PBS and 2 mL pre-warmed 0.25% Trypsin-EDTA was added to the flask and incubated at 37°C for 5 min. Trypsin was inactivated by addition of 5 mL supplemented T-medium and the cells centrifuged at 1,200 rpm for 5 min before resuspending in supplemented T-medium. LNCaP cells were maintained at 37°C with 5% CO_2.

2.2.1.D Cryopreservation and revival of cell lines

Mammalian cell lines were preserved by cryopreservation in liquid nitrogen. Once confluent, cells were trypsinised and resuspended in ~1.5 mL freezing solution. 5-6 x 10^5 PrEC cells were stored in each cryovial and one confluent T-75 flask of LNCaP cells were stored in each cryovial. Cells were frozen to -80°C at a rate of ~1°C/min using a Nalgene freezing container for 24 h. For long-term storage, cryovials were transferred to liquid nitrogen. Revival of cells from liquid nitrogen storage was performed by thawing the cells quickly at 37°C and transferring them to ~9 mL of pre-warmed supplemented growth medium. Cells were then transferred to a T-75 flask containing an additional 5 mL supplemented growth medium. The medium was replaced 24 h post-thawing to remove traces of DMSO. Once confluent, cells were passaged as previously described.

2.2.1.E Reverse transfection of cells with siRNAs

For reverse transfection of siRNAs, siGENOME Human siRNA SMARTpools (Table 2.2) were reverse transfected using Lipofectamine® RNAiMAX. All transfections were performed in triplicate and in a 6-well plate format. 5 μL Lipofectamine® RNAiMAX was incubated with 245 μL Opti-MEM® Reduced Serum Medium for 5 min and added to 250 μL Opti-MEM® Reduced Serum Medium containing 150 pmoles siRNA. This was incubated at room temperature for 20 min to allow siRNA/Lipofectamine® RNAiMAX complexes to form. 500 μL of the complex was added to each well and cells were seeded at a density of 1.5 x 10^5 for HeLa cells or 4 x 10^5 for LNCaP cells. The volume was made up to 2.5 mL with supplemented growth medium without antibiotics that was specific to the cells and were maintained in an incubator at 37°C with 5% CO_2. Cells were split if necessary and reverse transfected again 72 h post-initial transfection and harvested 144 h post-initial transfection using 1 mL Tri Reagent® for RNA isolation or 300 μL Cytobuster™ Protein Extraction Agent for protein isolation.

2.2.2 NUCLEIC ACID TECHNIQUES

2.2.2.A RNA ISOLATION AND PRECIPITATION

Total RNA from adherent cells was extracted using Tri Reagent® according to the manufacturer's protocol. Briefly, medium was aspirated from the cells and 1 mL of Tri Reagent® was added to the cells per 10 cm^2, vortexed and placed in an Eppendorf tube. The Tri Reagent® suspension was incubated for 5 min at room temperature and 0.2 mL of chloroform per 1 mL Tri Reagent® was added to the suspension, mixed and incubated for 2 min at room temperature followed by centrifugation at 12,000 x g for 15 min at 4°C. The aqueous phase was transferred to another 1.5 mL Eppendorf tube containing 1 µL glycogen and the RNA precipitated with 0.5 mL isopropanol per 1 mL Tri Reagent®. The precipitated RNA was mixed, incubated for 10 min at room temperature and centrifuged at 12,000 x g for 10 min at 4°C. The RNA was washed with 75% EtOH, mixed, and centrifuged at 7,500 x g for 5 min at 4°C. Following the removal of the supernatant, the RNA pellet was air dried briefly in a biohazard cabinet and resuspended in dH_2O. The isolated RNA was precipitated again overnight at -80°C with 0.1 volumes of 3 M NaAc, 3 volumes of 100% EtOH and 1 µL glycogen. The suspension was centrifuged at 17,000 x g for 30 min at 4°C and the RNA washed with 75% EtOH as previously described and resuspended in dH_2O. RNA was stored at -80°C.

2.2.2.B REMOVAL OF GENOMIC DNA FROM RNA

For all downstream applications, RNA was subjected to genomic DNA removal using TURBO™ DNase according to the manufacturer's protocol. Briefly, up to 10 µg total RNA was treated with 2 U TURBO™ DNase in 1x TURBO™ DNase Buffer and incubated at 37°C for 30 min. TURBO™ DNase was removed by phenol/chloroform extraction (see next section).

2.2.2.C PHENOL/CHLOROFORM EXTRACTION

The reaction was made up to 100 µL in dH_2O and transferred to a Phase Lock Gel Heavy tube that had been centrifuged previously at 16,000 x g for 30 sec to settle the

gel bed. An equal volume of UltraPure™ Phenol:Water (3.75:1 v/v) and chloroform were added, and the solution mixed vigorously by hand for 15 sec. The mixture was centrifuged at 16,000 x g for 5 min at room temperature. An equal volume of chloroform was added again and the process repeated. The RNA-containing solution was transferred to a new Eppendorf tube and precipitated overnight in EtOH as previously described.

2.2.2.D Agarose gel electrophoresis

1-2% agarose gels were used for the analysis of PCR products or plasmid constructs. Agarose gels were prepared by mixing agarose with 1x TAE and pouring it into a horizontal gel box and allowing to set for 30 min-1 h. 1x EZ-Vision® Three DNA Dye was added to the DNA samples and size markers prior to loading. Agarose gels were run in 1x TAE at 100 V for 30-40 min and visualised by short wavelength UV light and photographed using the Syngene Gene Genius Bio Imaging System (Synoptics Ltd, Cambridge, UK).

2.2.2.E Bisulfite conversion of RNA

Sodium bisulfite solution was prepared using 40% (w/v) sodium metabisulfite and 600 μM hydroquinone. The pH was adjusted to 5.1 by addition of 10 M NaOH and the final solution filtered through a 0.22 μm filter. As a control for bisulfite conversion, 1/1000 of humanised and non-humanised R-Luc *in vitro* transcript was spiked into 2-4 μg DNase-treated RNA prior to bisulfite treatment. The RNA mixture was denatured by heating to 75°C for 5 min then 100 μL of pre-heated sodium bisulfite solution was added to each RNA mixture, mixed and incubated at 75°C for 4 h in the dark with 100 μL mineral oil layered on top to prevent evaporation. The reaction mixture was desalted by two passages through Micro Bio-spin 6 chromatography columns according to the manufacturer's protocol. RNA was desulfonated by addition of 1 volume of 1 M Tris-HCl (pH 9.0) and incubated for 1 h at 75°C in the dark with 150 μL mineral oil, followed by overnight EtOH precipitation as previously described. RNA was resuspended in dH_2O.

2.2.2.F Reverse transcription

Reverse transcription of isolated RNA was performed using Superscript® III Reverse Transcriptase (RT) according to the manufacturer's protocol. 50 ng-2 µg DNase treated RNA, 1 µL 20x random primer mix, 1 µL 1 mM dNTP mix and dH_2O to 13 µL were incubated at 65°C for 5 min, and then placed on ice for 1 min. For reverse transcription, a mixture of 1x First-strand Buffer, 1 µL 0.1 M DTT, 40 U RNaseOUT™ and 200 U Superscript® III (or dH_2O for -RT controls) were added to the denatured RNA sample and incubated at 25°C for 5 min, 50°C for 60 min and then at 70°C for 15 min. 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.

2.2.2.G *In vitro* transcription

For *in vitro* transcription, the MEGAscript® T7 Kit (Life Technologies, Carlsbad, CA, USA) was used according to the manufacturer's protocol. Briefly, linearised template DNA (~1 µg) was incubated with 1x Reaction Buffer, 7.5 mM each of ATP, CTP, GTP and UTP and 2 µL enzyme mix at 37°C for 2-4 h. Template DNA was removed by adding 2 U TURBO™ DNase and incubating at 37°C for 30 min. TURBO™ DNase was removed by phenol/chloroform extraction and RNA precipitated overnight in EtOH as previously described. After precipitation and resuspension in dH_2O, 10 µg of the *in vitro* transcript was treated with 2 U TURBO™ DNase again to remove any residual template DNA. The size of the *in vitro* transcript was assessed on a RNA 6000 Nano Chip on the Agilent® 2100 Bioanalyzer (Agilent Technologies, Santa Clara, CA, USA) according to the manufacturer's instructions.

2.2.2.H DNA/RNA dot blot assay

Methylated and unmethylated DNA controls were serially diluted two-fold from 100 ng-12.5 ng, denatured either in 2 µL 0.1 M NaOH at 99°C for 5 min and snap cooled on ice followed by neutralisation with 0.1x 6.6 M ammonium acetate, or 2x RNA loading dye at 75°C for 5 min and snap cooled on ice. Yeast tRNA and HeLa RNA were serially diluted two-fold from 5 µg-625 ng in dH_2O, denatured with or without 2x RNA loading buffer at 75°C and snap cooled on ice. ~10 µL of each sample was dotted onto an

Amersham Hybond™-N^+ membrane using a pipette and allowed to dry. The membrane was then UV crosslinked twice using the "auto-crosslink" function (120 mJ/cm^2) on the Stratalinker® 1800 UV Crosslinker (Stratagene, La Jolla, CA, USA). The membrane was blocked in blocking buffer for 1 h at room temperature while rocking, then incubated with the anti-m^5C antibody overnight at 4°C while rocking. The membrane was washed with 1x PBS-T three times over 30 min, incubated with a anti-mouse secondary antibody in blocking buffer for 1 h at room temperature while rocking, then washed again in 1x PBS-T three times for 30 min. Enough ECL solution to cover the membrane was added and incubated in the dark for 5 min. Membranes were visualised using the ImageQuant LAS 4000 (GE Healthcare, Little Chalfort, BKM, UK). To check hybridisation of nucleic acids to the membrane, the membrane was incubated with methylene blue solution for 5 min at room temperature while rocking, followed by rinsing with dH_2O. The membrane was photographed using the upper white light setting on the Syngene Gene Genius Bio Imaging System.

2.2.3 Nucleic acid quantification and integrity analysis

2.2.3.A NanoDrop spectrophotometric analysis for quantification

The NanoDrop spectrophotometer (Thermo Scientific, Wilmington, DE, USA) provides nucleic acid quantity and purity by assessing the 220-320 nm light absorbance profile using an inbuilt algorithm. Quantification of nucleic acids by this method was performed using 1.5 µL of purified and undiluted nucleic acid according to the manufacturer's protocol.

2.2.3.B Qubit fluorometric quantification

MiSeq DNA quantification was performed using the Qubit dsDNA BR Assay Kit and assayed using the Qubit Fluorometer (Life Technologies, Carlsbad, CA, USA) according to the manufacturer's protocol. In total, 2 µL of DNA per sample was assayed.

2.2.3.C Agarose gel electrophoresis for integrity analysis

For some experiments, 1% agarose gel electrophoresis was performed with 300 ng RNA. Prior to pouring the agarose into the horizontal gel box, ethidium bromide (EtBr) was added to a final concentration of 0.5 mg/mL. The samples were electrophoresed at 100 V for 30 min. Bands corresponding to 28S and 18S rRNA were visually assessed for RNA integrity. The 28S band should be ~2 times brighter than the 18S band, and minimal degradation should be observed on the gel.

2.2.3.D Agilent RNA 6000 Nano Kit for integrity analysis

For whole-transcriptome sequencing, RNA integrity was determined using the Agilent® 2100 Bioanalyzer according to the RNA 6000 Nano Kit protocol. High quality RNA should have an RNA Integrity Number (RIN) > 8.

2.2.4 Polymerase Chain Reaction (PCR)

2.2.4.A Bisulfite PCR

PCR was performed using diluted cDNA and indicated primers (Table 2.3). Fragments were amplified in a 25 μL reaction volume with Platinum® Taq DNA Polymerase using standard PCR conditions (1x PCR Buffer, 0.5 μL 10 mM dNTP mix, 0.5 μL 10 μM of each primer, 0.75 μL 50 mM $MgCl_2$ and 0.1 μL Platinum® Taq DNA Polymerase). A "touchdown" PCR program was used, which occurred in two phases. The first phase used an initial annealing temperature that was ~5°C above the highest primer melting temperature and decreased by 1°C for each successive cycle to 5°C below the lowest primer melting temperature: 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 implemented standard PCR conditions: 25-40 cycles at 95°C for 30 sec, the lowest annealing temperature for 30 sec, 72°C for 15-30 sec, followed by a final extension at 72°C for 5 min. All PCR products were validated by agarose gel electrophoresis and purified using either the QiaQuick Gel Extraction Kit or the MinElute PCR Purification Kit according to the manufacturer's instructions.

Table 2.3. Bisulfite PCR primers and experimental conditions.

Name	Sequence	Gene	Primer Type
A.Luc1 FWD	CTTTAAAAAACCTCCCACACCT	Non-humanised R-Luc *in vitro* transcript	Forward
A.Luc2 REV	TGATGAAATGGGAAAATATAT		Reverse
A.Luc4 FWD	CATTATAATTCCTAACAATTT		Forward
A.Luc4 REV	GAAGAAGGAGAAAAAATGGTT		Reverse
CAGE1PG_BS_F	CCACAATACTCACATAATC	*CAGE1 (intron)*	Forward
CAGE1PG_BS_R	GAAATTTTGAGGATAGAGAG		Reverse
CINP endogenous FWD	CTCTAACCAAAAAACCCCAA	*CINP*	Forward
CINP endogenous REV	TATGTGGAGAGTGATAGTAGG		Reverse
FTSJ3_BS_F	CACCAAAATTTTATTCTCCA	*FTSJ3*	Forward
FTSJ3_BS_R	AAAGGTTGAAGGTTATGTTG		Reverse
MCFD2_BS_F3	CCACCAACAACTTACACCAT	*MCFD2*	Forward
MCFD2_BS_R3	GGGGTTATAGGGTTTATGTAAG		Reverse
NAPRT1 endo FWD	ACCACTCCTCCACACACTATA	*NAPRT1*	Forward
NAPRT1 endo REV	TGGTGAGGGTTTTGGGGTGGA		Reverse
NDUFB7 WT FWD	CTCCCACTCAAACTCCTTCAT	*NDUFB7*	Forward
NDUFB7 WT REV	AAGGAGTGTGAGATGGTGG		Reverse
NEK7_BS_F2	CATCAAAAAACACACTAAAC	*NEK7*	Forward
NEK7_BS_R2	GTTGGGATGTTTATATTTTTG		Reverse
NSUN5P2_BS_F2	CTACTACCTACAATAAAAACC	*NSUN5P2*	Forward
NSUN5P2_BS_F2	GTAAAAAGTTGTAGTTGGTG		Reverse
OSBPL8_BS_F	CCCATACAAAACTCCCAAAA	*OSBPL8*	Forward
OSBPL8_BS_R	GTGGATTTTGGGGTTGAAAG		Reverse
PLD3_BS_F	CCAAACTACTATACACACCAA	*PLD3*	Forward
PLD3_BS_R	TTGTGGGTTTTGGAGTTTTG		Reverse
PWP2_BS_F2	CCAAATAACAAAACACTTC	*PWP2*	Forward
PWP2_BS_R2	GTTTGAGTTGGATATTAGTG		Reverse
RPPH1 endogenous FWD	CCATTAAACTCACTTCACT	*RPPH1*	Forward
RPPH1 endogenous REV	TGAGTTTTGGGGAGGTGAGTT		Reverse
RPS3_BS_F	CCACAATACTCACATAATC	*RPS3*	Forward
RPS3_BS_R	GAAATTTTGAGGATAGAGGG		Reverse
RTN3_BS_F2	CAACAATACACAACAATTACAA	*RTN3*	Forward
RTN3_BS_R2	GGGAGTTAGTAGGTATGTTG		Reverse
SCARNA2_BS_F	CAAAATCACCTCAATAATCA	*SCARNA2*	Forward
SCARNA2_BS_R	TTTTATTTGATTGGATTGTG		Reverse
SCO1_BS_FWD	CTAAAAACTTCCAAAAAACAAAC	*SCO1*	Forward
SCO1_BS_REV	GAGTTTTGGGGTTTAGTTG		Reverse
SNORD62A_BS_F	CTCTCAATCACTATCCTCTT	*SNORD62A*	Forward

SNORD62A_BS_R2	GATTTTTTTGATTTTTTATTGTGG		Reverse
SOX4_BS_F	CCACACCTTATACAACAAAT	*SOX4*	Forward
SOX4_BS_R	TTGGTGTAGAAAAAGAGTTG		Reverse
SRRT_BS_F2	CCAACATAAAACATCTCATAAC	*SRRT*	Forward
SRRT_BS_R2	GGATGATTGGAATGATAGAG		Reverse
SZRD1_BS_F	CTCATCTATCCTAACAACTA	*SZRD1*	Forward
SZRD1_BS_R	TGTAAGAAAAAGTGTATGGG		Reverse
SZRD1CDS_BS_F	CTACCATTCCCACAAAAAAC	*SZRD1*	Forward
SZRD1CDS_BS_R	GAGTAGTTGTGTGTATTTGG		Reverse
Asp tRNA for	CTCCCCATCAAAAAATCAA	$tRNA^{Asp(GUC)}$	Forward
Asp tRNA rev	TTTTTGTTAGTATAGTGGTG		Reverse
tRNAGluCTC_BS_F	CCCTAACCAAAAATCAAAC	$tRNA^{Glu(CUC)}$	Forward
tRNAGluCTC_BS_R	GTGGTTTAGTGGTTAGGATT		Reverse
tRNAGlyGCC_BS_F	CATAAACCAAAAATCAAACCC	$tRNA^{Gly(GCC)}$	Forward
tRNAGlyGCC_BS_R	GTATGGGTGGTTTAGTGGTA		Reverse
Leu CAA endogenous F	TATCAAAAATAAAATTCAAACCC	$tRNA^{Leu(CAA)}$	Forward
Leu CAA endogenous R	TTAGGATGGTTGAGTGGTT		Reverse
tRNAThrTGT_BS_F	CCCCAACAAAAATTAAACTC	$tRNA^{Thr(UGU)}$	Forward
tRNAThrTGT_BS_R	GTTTTATAGTTTAGGGGTTA		Reverse
tRNAValAAC_BS_F	CACCCAATTTCAAACCAC	$tRNA^{Val(AAC)}$	Forward
tRNAValAAC_BS_R	GTTTTTGTAGTGTAGTGGTTA		Reverse
tRNAValCAC_BS_F	TATTTCCACCCAAAATCAAA	$tRNA^{Val(CAC)}$	Forward
tRNAValCAC_BS_R	GTTTTTGTAGTGTAGTGGTTA		Reverse

2.2.4.B Real time quantitative PCR (RT-qPCR) for expression analysis

Specific forward and reverse primers were designed using Primer3 [235, 236] and were used with Fast SYBR® Green Master Mix for RT-qPCR (Table 2.4). Primers were designed such that they spanned exon-exon junctions to avoid amplification of genomic DNA. Each reaction was performed in triplicate using 2 µL diluted cDNA mixed with 8 µL master mix (5 µL Fast SYBR® Green Master Mix, 0.5 µL of 5 µM of each primer and 2 µL dH_2O). PCR amplification was performed in a 384-well plate format on the QuantStudio 12K Flex system (Life Technologies, Carlsbad, CA, USA). The reaction conditions were: 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.

Table 2.4. RT-qPCR primers and experimental conditions.

Name	Sequence	Gene	Annealing Temperature (°C)	Primer Type
ACTB_qPCR_F	CTGGAACGGTGAAGGTGACA	*ACTB*	60	Forward
ACTB_qPCR_R	CTGGAACGGTGAAGGTGACA			Reverse
B2M_qPCR FWD	ACTGAATTCACCCCCACTGA	*B2M*	60	Forward
B2M_qPCR REV	CCTCCATGATGCTGCTTACA			Reverse
NSUN2 real time FWD	TGGAATAAACGTCAGCCAAA	*NSUN2*	60	Forward
NSUN2 real time REV	GGGCTCAGCTGTGTGCTT			Reverse
DNMT2 real time FWD	TCCTAAAGAAATAGCAAATCTCCTTG	*TRDMT1*	60	Forward
DNMT2 real time REV	CGCTGTTTCACTGTTATCTTCTCA			Reverse
DNMT1 real time FWD	CCCAAGTAACTGGGATTAGAGC	*DNMT1*	60	Forward
DNMT1 real time REV	GGTTTGCCTGGTGCTTTTC			Reverse
Fil Rluc QPCR FOR	GGCGAGAAAATGGTGCTTGAG	*hRL 3xB*	60	Forward
Fil Rluc QPCR REV	TCCTTGAATGGCTCCAGGTAGG			Reverse
H2AFV_qPCR_F	GAAGCATTACCTGCCAGCTC	*H2AFV*	60	Forward
H2AFV_qPCR_R	TTGAAGACTCGCACCACAAG			Reverse
HPRT real time FWD	TGACACTGGCAAAACAATGCA	*HPRT*	60	Forward
HPRT real time REV	GGTCCTTTTCACCAGCAAGCT			Reverse
MCFD2_qPCR_F	TCCATCATTGTTCTTGTCATCA	*MCFD2*	60	Forward
MCFD2_qPCR_R	ACAGCCATCACTCATGTCCA			Reverse
MRPL9_qPCR_F	CACCTCACACCAATACTCGC	*MRPL9*	60	Forward
MRPL9_qPCR_R	AAATAGTTGCCCGCCACTTC			Reverse
NAPRT1_qPCR_F	AACTGCAGCATGTCCATGAG	*NAPRT1*	60	Forward
NAPRT1_qPCR_R	GAATGAAGCTGACCGAGGAC			Reverse
NDUFB7_qPCR_F	CCTTCATGCGCATCACATAG	*NDUFB7*	60	Forward
NDUFB7_qPCR_R	GTGACAGCTTCCCCAACTTC			Reverse
OSBPL8_qPCR_F2	ACTCCAAAGCATCCATCCAG	*OSBPL8*	60	Forward
OSBPL8_qPCR_R2	GTGAAGGGTCCAAAAGGTGA			Reverse
RTN3_qPCR_F2	CCAGGTAGGCTTTGAATGGA	*RTN3*	60	Forward
RTN3_qPCR_R2	TGGCAGCTTTCAGTGTCATC			Reverse
SDHA_qPCR_F	CAGTGCCAACGTCCACATAG	*SDHA*	60	Forward
SDHA_qPCR_R	GCCAGGGAAGACTACAAGGT			Reverse
SZRD1_qPCR_F	TCTTAGGGGTCTGCAGTGCT	*SZRD1*	60	Forward
SZRD1_qPCR_R	AGCCTAGGCCTCTGAGGAAC			Reverse
TBP_qPCR_F	AGCATCTCCAGCACACTCTT	*TBP*	60	Forward
TBP_qPCR_R	CCGGCTGTTTAACTTCGCTT			Reverse
TCF25_qPCR_F	CAAACGCAAAGAAGGAGAGG	*TCF25*	60	Forward
TCF25_qPCR_R	CAACGTGTGTACCCCAAGTG			Reverse

2.2.4.C TaqMan® non-coding RNA assay

For quantification of 18S rRNA using the TaqMan® non-coding RNA assay, each reaction was performed in triplicate using 2 µL cDNA with 8 µL master mix (5 µL TaqMan® Universal Master Mix, 0.5 µL Eukaryotic 18S rRNA Endogenous Control (VIC®) and 2 µL dH_2O). PCR amplification was performed in a 384-well plate format on the ABI 7900HT Real-Time PCR System (Life Technologies, Carlsbad, CA, USA). The reaction conditions were: 50°C for 2 min, 95°C for 10 min, 40 cycles of 95°C for 3 sec and 60°C for 1 min.

2.2.5 Protein techniques

2.2.5.A Protein isolation

Cells grown in 6-well plates were washed with ice cold PBS. 300 µL Cytobuster™ Protein Extraction Agent was added to each well and incubated at room temperature for 5 min. The cells were transferred to Eppendorf tubes and centrifugation performed at 16,000 x g for 5 min at 4°C. The supernatant was transferred to a new tube and stored at -80°C.

2.2.5.B Protein quantification

Protein was quantified using the Pierce™ BCA Protein Assay Kit according to the manufacturer's protocol. Briefly, BSA protein standards (0-2000 µg/mL) were prepared in duplicate in a 96-well plate and 25 µL of protein sample of interest was added to the 96-well plate in duplicate. 200 µL working reagent (50:1, Reagent A:B) was added to each standard and unknown sample, mixed and incubated with cover foil at 37°C for 30 min. After cooling the plate down to room temperature, the absorbance was measured at 562 nm on the FluoSTAR OPTIMA plate reader (BMG Labtech, Ortenberg, Germany).

2.2.5.C SDS PAGE and western blotting

A mixture of 1x LDS Sample Buffer and 2.8 µL 0.1 M DTT were added to 10 µg protein to a final volume of 28 µL. The protein sample was boiled for 2 min and chilled

on ice for another 2 min. The protein sample was loaded into NuPAGE® Novex® Bis-Tris 4-12% gels and run at 180 V for 50 min. Proteins were transferred onto a PVDF membrane in 1x transfer buffer at 100 V for 1 h at 4°C. Protein transfer efficiency was observed using RedAlert™ western blot stain. The membrane was blocked in blocking buffer for 1 h at room temperature while rocking. The membrane was incubated with the primary antibody for either 1 h at room temperature or overnight at 4°C while rocking, then washed with 1x PBS-T three times over 30 min. The membrane was incubated with the secondary antibody in blocking buffer for 1 h at room temperature while rocking, then washed again in 1x PBS-T three times for 30 min. Enough ECL solution to cover the membrane was added and incubated in the dark for 5 min. Membranes were visualised using the ImageQuant LAS 4000 (GE Healthcare, Little Chalfort, BKM, UK).

2.2.5.D Densitometry of western blots

In order to determine the intensity of protein bands on western blots and quantify the relative amounts of protein, ImageJ software was used. Effort was made to ensure that protein bands were not saturated. TIFF image files from the ImageQuant LAS 4000 were converted to greyscale by selecting the 8-bit image option. Equal sized rectangular selections were made for each band on the gel and the intensity of each band plotted using the “Plot lanes” function under the “Analyze” menu. To remove background noise, a straight line was drawn across the base of each peak, enclosing it. The wand tool was used to select each peak, providing intensity measurements. The intensity of each band was normalised to the intensity of the loading control.

2.2.6 SUCROSE DENSITY GRADIENT CENTRIFUGATION FOR POLYSOME ANALSYSIS

2.2.6.A SUCROSE GRADIENT CONSTRUCTION

17.5%, 25.6%, 33.8%, 41.9% and 50% sucrose solutions were prepared by dissolving the required amount of sucrose in sucrose gradient buffer. The sucrose was dissolved by placing the tubes on a rotator for 15 min followed by filtering through a 0.2 μm filter. Starting with the highest density sucrose, 2 mL was carefully laid at the bottom of a 14 x 89 mm polyallomer centrifuge tube, and placed in a -80°C freezer for a minimum of 15 min before applying the next layer. This process was repeated until all sucrose densities were layered. Sucrose gradients were stored at -80°C for several months. Prior to polysome analysis, sucrose gradients were thawed overnight at 4°C.

2.2.6.B CELL LYSIS

HeLa cells were plated in a 15 cm dish at a density of 7 x 10^6 cells. 24 h post-plating, the cells were treated with 0.1 mg/mL cycloheximide for 3 min at 37°C, placed on ice, and the media removed. The cells were washed twice in ice-cold PBS containing 0.1 mg/mL cycloheximide and drained for 2 min. The cells were lysed in 800 μL lysis buffer, scraped into an Eppendorf tube and vortexed for 2 min at 4°C. The lysate was incubated on ice for 10 min followed by centrifugation at 16,000 x g at 4°C for 10 min. The cleared lysate was carefully layered on top of the thawed sucrose gradient and centrifuged at 35,000 rpm for 2 h and 15 min at 4°C using SW40Ti rotor (Beckman Coulter).

2.2.6.C POLYSOME PROFILING AND FRACTION COLLECTION

The fractionation was conducted bottom up while the absorbance profile at 260 nm was recorded, using the BR-188 Density Gradient Fractionation System (Brandel, Gaithersburg, MD, USA). Prior to fraction collection, the baseline was set using Milli-Q H_2O. A 14 x 89 mm polyallomer centrifuge tube containing Milli-Q H_2O was attached to a syringe and H_2O was pushed through the system, with the baseline adjusted according to the manufacturer's protocol. Once set, the centrifuge tube containing the

sample was gently attached to the apparatus and the needle pierced through the bottom of the tube. The chase solution used was 60% sucrose in Milli-Q H_2O. The speed of the apparatus was set to 0.75 mL/min and fractions were collected every 37 seconds, resulting in 24 fractions.

2.2.6.D Preparation of polysome fractions for RNA extraction

Prior to RNA extraction, 1 ng of hRL 3xB plasmid, 2 µL glycogen and 3 volumes of EtOH were added to each of the fractions, followed by precipitating overnight at -80°C. The fractions were centrifuged at 16,000 x g for 20 min at 4°C. The supernatant was discarded and the pellet was carefully washed in 70% EtOH before spinning again at 16,000 x g for 5 min at 4°C. After discarding the supernatant and air-drying the pellet, 1 mL Tri Reagent® was added to the pellet and vortexed until dissolved. The sample was incubated for 5 min at room temperature to allow the nucleoprotein complexes to dissociate. Following this, the standard Tri Reagent® protocol was implemented for RNA extraction.

2.2.7 Cloning and Sanger sequencing

2.2.7.A Preparation of *E. coli* DH5α for cloning

To prepare *E. coli* DH5α cell stock for bacterial transformation, a frozen scraping was added to 1 mL LB and incubated for 3 h at 37°C with shaking at 220 rpm. The 1 mL was then inoculated into 100 mL LB and incubated at 37°C with shaking at 220 rpm until bacteria reached log phase (OD600 ~0.4-0.5). Bacteria were chilled on ice for 10 min and centrifuged at 3,000 rpm for 5 min at 4°C. The supernatant was removed and the pellet was gently resuspended in 30 mL ice cold TFB1 and chilled on ice overnight, followed by centrifugation at 3,000 rpm for 5 min at 4°C. The supernatant was removed and the cells were gently resuspended in 4 mL of ice cold TFB2, aliquoted into Eppendorf tubes (100 µL) and snap frozen immediately on dry ice. DH5α cell stocks were stored at -80°C until use.

2.2.7.B Cloning of PCR products and bacterial transformation

Approximately 50-200 ng purified PCR products were ligated to pGEM®-T Easy Vector according to the manufacturer's protocol, followed by bacterial transformation. The ligation product from the pGEM®-T Easy Vector Cloning System was incubated with 45 µL competent *E. coli* DH5α cells on ice for 30 min. Cells were heat-shocked at 42°C for 30 sec and placed on ice for 2 min. The cell/DNA solution was diluted with 350 µL LB and incubated at 37°C at 220 rpm for 1 h. The entire solution was then spread onto LB agar plates containing 50 µg/mL X-Gal, 40 µg/mL IPTG and 100 µg/mL ampicillin, and incubated overnight at 37°C. White colonies (X-Gal selection) were selected for further analysis.

2.2.7.C Plasmid DNA extraction

A single colony was picked and incubated in 5 mL LB containing 100 µg/mL ampicillin for 16 h at 37°C with shaking at 220 rpm. Plasmid DNA was isolated using the PureYield™ Plasmid Miniprep System according to the manufacturer's protocol. Purified plasmid DNA was quantified using the NanoDrop and digested with EcoRI for 2 h at 37°C to confirm ligation of insert. The digest was run on a 1% agarose gel at 100 V for 30 min.

2.2.7.D Sanger sequencing of plasmids

The purified plasmid DNA was submitted to either the Australian Cancer Research Foundation (ACRF) at the Garvan Institute of Medical Research in Sydney, or the Biomolecular Resource Facility (BRF) at the John Curtin School of Medical Research in Canberra for BigDye® Terminator Sequencing. Approximately 40 ng of plasmid DNA and 3.2 pmoles M13 reverse primer were submitted to ACRF. Approximately 100 ng/µL plasmid DNA and 1.6 µM M13 reverse primer were submitted to BRF. Analysis of Sanger sequencing data was performed using the EMBL-EBI CLUSTALW2 Multiple Sequence Alignment Program (EMBL-EBI, Cambridgeshire, UK).

2.2.8 NEXT-GENERATION SEQUENCING

2.2.8.A TRANSCRIPTOME-WIDE BISULFITE SEQUENCING (BSRNA-SEQ)

Preparation of RNA for bsRNA-seq

For bsRNA-seq, RNA was extracted from cells using Tri Reagent® and genomic DNA was removed using TURBO™ DNase as previously described. RNA quality was then assessed on the Agilent® 2100 Bioanalyzer RNA 6000 Nano Chip. Following this, 40 µg total RNA was depleted of rRNA using the RiboMinus™ Eukaryote Kit for RNA-seq according to the manufacturer's protocol and depletion assessed using the Agilent® 2100 Bioanalyzer RNA 6000 Nano Chip. 2-2.5 µg of rRNA-depleted RNA was then subjected to bisulfite conversion as described above. To ensure efficient bisulfite conversion (>98.5%), an 50-300 ng bisulfite-converted RNA was reverse-transcribed followed by PCR amplification of sections of the R-Luc *in vitro* transcript and $tRNA^{Asp(GUC)}$, which were subjected to Sanger sequencing. To confirm fragmentation of RNA required for bsRNA-seq, RNA was analysed using the Agilent® 2100 Bioanalyzer RNA 6000 Nano Chip.

End polishing of RNA

To ensure ligation of 5' SOLiD™ adaptors during the sequencing library preparation, the mRNA was de-capped; 500 ng of bisulfite-converted RNA was treated with 5 U Tobacco Acid Pyrophosphatase at 37°C for 1.5 h followed by phenol/chloroform extraction and overnight EtOH precipitation. The RNA was then resuspended in dH_2O. To ensure ligation of 3' SOLiD™ adaptors, 3' phosphate groups that may exist after bisulfite conversion were removed by treating the RNA with 10 U T4 Polynucleotide Kinase and 1 mM dATP in 1x T4 Polynucleotide Kinase Buffer at 37°C for 1 h. The RNA was phenol/chloroform extracted as previously described, followed by overnight EtOH precipitation. The RNA was then resuspended in dH_2O.

Ligation of adapters to RNA

Whole transcriptome libraries were prepared using the SOLiD™ Total RNA-Seq Kit. To prepare for the ligation of adaptors, 2 µL SOLiD™ Adaptor Mix and 3 µL Hybridisation Solution were incubated with 100 ng RNA at 65°C for 10 min and 16°C

for 5 min. Following this, 10 µL 2x Ligation Buffer and 2 µL Ligation Enzyme Mix were added to the hybridisation reaction, mixed and incubated in a thermal cycler with the lid off at 16°C for 16 h.

Reverse transcription

To reverse transcribe the adaptor bound RNA, 11 µL dH_2O, 4 µL 10x RT buffer, 2 µL dNTP Mix and 2 µL SOLiD™ RT Primer were added to the ligation mix followed by mixing and incubating at 70°C for 5 min and snap cooling on ice for 1 min. 1 µL ArrayScript™ Reverse Transcriptase was added to the reaction, gently mixed and incubated in a thermal cycler at 42°C for 30 min.

Size selection of cDNA

After reverse transcription, 10 µL cDNA was run on a 10% Novex® TBE-urea gel in 1x TBE at 180 V for 45 min along with the 10 bp DNA ladder. The gel was stained with 1x SYBR® Gold Nucleic Acid Stain for 5 min, and visualised using medium wavelength UV light and photographed using the Typhoon™ FLA 9000 (GE Healthcare, Little Chalfort, BKM, UK). cDNA of size between 100-160 nt, corresponding to an insert size of 50-120 nt, were excised from the gel. The size-selected area for each sample was divided into four vertical strips and placed in individual 0.2 mL tubes.

PCR amplification and purification of libraries

For PCR amplification of the libraries, two vertical strips from the size selection gel per sample were placed into separate tubes and 76.8 µL dH_2O, 10 µL 10x PCR Buffer, 8 µL dNTP Mix, 2 µL SOLiD™ 5' PCR Primer and 1.2 µL AmpliTaq® DNA Polymerase were added to each tube. 2 µL of barcoded 3' PCR Primers were added to each tube, with each tube having a separate index number. The reactions were then subjected to PCR using the following conditions: 95°C for 5 min, 18 cycles of 95°C for 30 sec, 62°C for 30 sec and 72°C for 30 sec, followed by a final extension at 72°C for 7 min.

For PCR product purification, up to 20 µL of each sample was combined with 1x Novex® TBE-Urea Sample Buffer and separated on a polyacrylamide gel (5% stacking gel, 8% resolving gel, SequaGel- UreaGel System) in 1x TBE at 300 V for 37 min. The

gel was stained with 1x SYBR® Gold Nucleic Acid Stain for 5 min and visualised using medium wavelength UV light and photographed using the Typhoon™ FLA 9000. PCR products corresponding to ~200 nt were excised from the gel and placed in an 0.5 mL tube that had been pierced with a 19-G needle at the base. The 0.5 mL tube was placed in a 1.5 mL Eppendorf tube and subjected to centrifugation for 1 min at 13,000 rpm in order to fragment the gel. To elute the DNA from the gel, 700 µL 1x PAGE Elution Buffer was added to the fragmented gel and placed on a rotator at room temperature overnight. The gel/buffer mixture was centrifuged briefly and the liquid transferred to a tube containing a Costar® Spin-X® Plastic Centrifuge Tube Filter prior to centrifuging at 13,000 rpm for 30 sec to remove any residual gel. The DNA was EtOH precipitated on dry ice for 1 h and centrifuged at 20,000 x g for 1 h at 4°C. The DNA was washed twice with 70% EtOH and after briefly air-drying, the sample was resuspended in 10 µL dH_2O. To confirm size, concentration and purity, 1 µL purified DNA was run on a polyacrylamide gel as well as an Agilent High Sensitivity DNA Chip. The products were diluted to similar concentrations (600 pM) and prepared for sequencing using emulsion PCR (ePCR) on the SOLiD™ EZ Bead™ Emulsifier according to manufacturer's instructions. The libraries were then sequenced using the SOLiD™ 5500 XL platform (Life Technologies, Carlsbad, CA, USA).

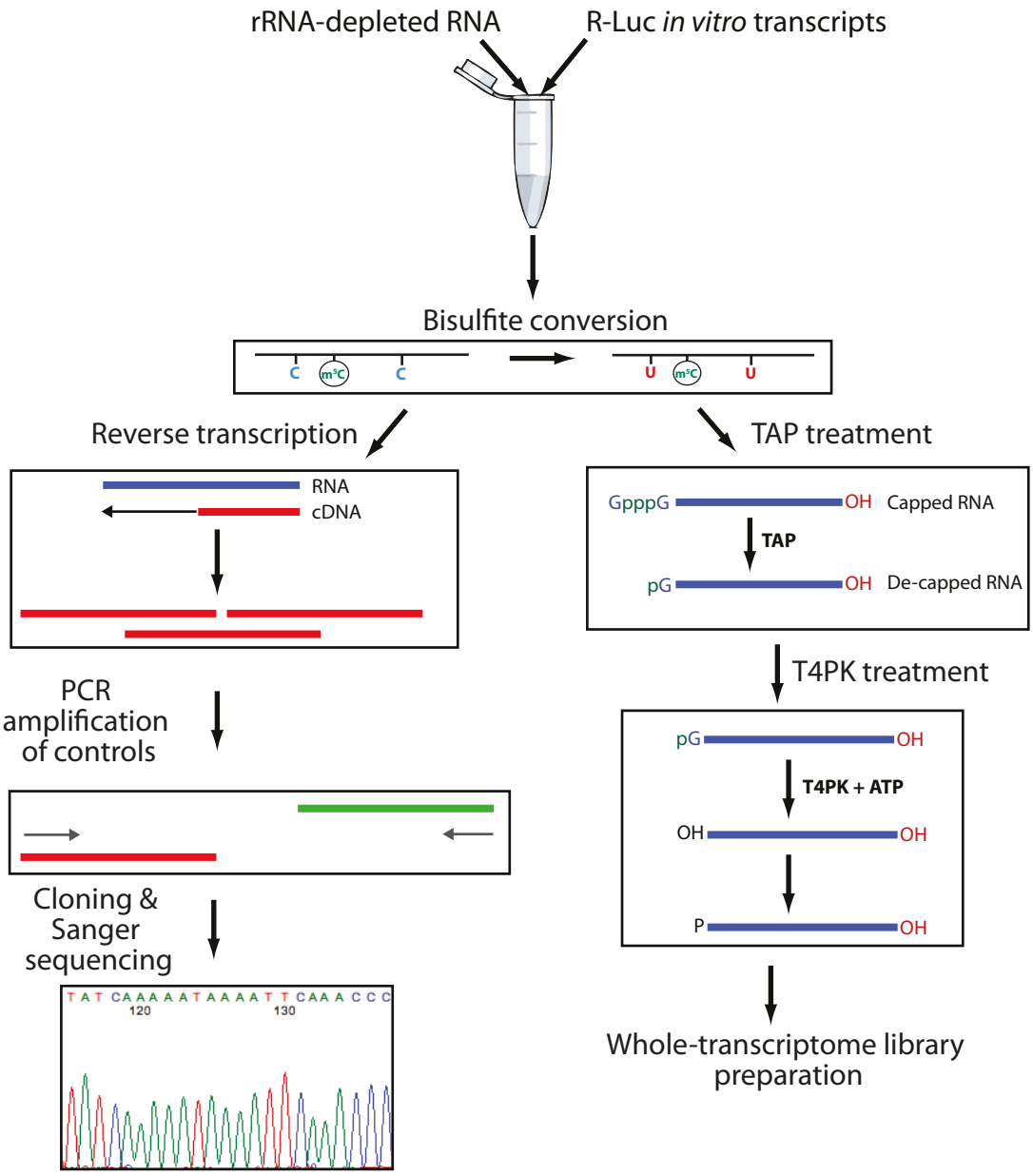
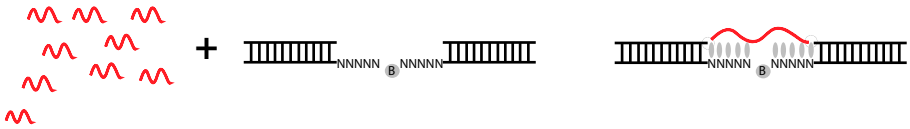


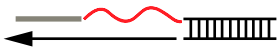
Figure 2.1. Method for preparation of RNA for bsRNA-seq.

Total RNA was depleted of rRNA using the RiboMinus™ Eukaryote Kit for RNA-Seq and the R-Luc *in vitro* transcripts were spiked in. The RNA was bisulfite converted and an aliquot reverse transcribed to inspect bisulfite conversion efficiency by Sanger sequencing. The bisulfite-converted RNA was end-polished using Tobacco Acid Pyrophosphatase and T4 Polynucleotide Kinase before commencement of library preparation.

1) Hybridise and ligate the RNA

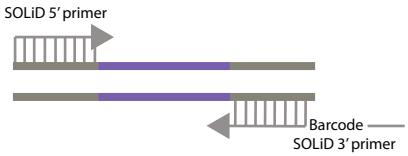


2) Perform reverse transcription

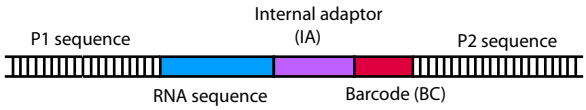


3) Purify and size-select the cDNA

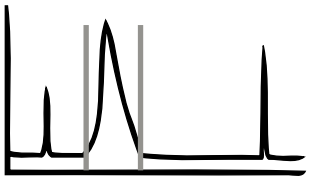
4) Amplify the cDNA



5) Purify the amplified DNA



6) Assess yield and size distribution of amplified DNA



7) Proceed with SOLiD™ System templated bead preparation

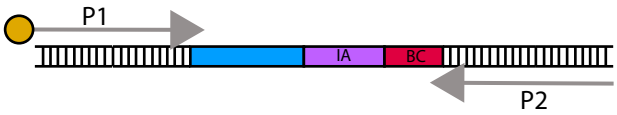


Figure 2.2. Whole-transcriptome library preparation method for bsRNA-seq.
The SOLiD™ Total RNA-Seq Kit was used for preparation of bsRNA-seq libraries. Step 1: The RNA was hybridised and ligated to SOLiD™ adaptors. Step 2: The RNA was reverse transcribed using universal primers. Step 3: The cDNA was run on a 10% TBE-Urea gel and fragments between 100 and 160 nt were size selected and purified. Step 4: The purified cDNA library was enriched by PCR using universal 5' primers and 3' barcoded primers. Step 5: The enriched library was purified by running on polyacrylamide gel. Step 6: The yield and size distribution was assessed by running a polyacrylamide gel and an Agilent High Sensitivity DNA chip. Step 7: Libraries were pooled and ePCR was performed on the SOLiD™ EZ Bead™. Picture modified from the SOLiD™ Total RNA-Seq manual.

2.2.8.B SOLiD™ bsRNA-seq analysis pipeline

Global mapping of sequencing data

Sequencing data was initially mapped to complete and non-CpG *in silico* bisulfite converted human Refseq transcriptomes, and any unmapped reads from this were mapped to complete and non-CpG *in silico* bisulfite converted hg19 reference genomes. The mapping was based on the method and implementation of the DNA methylation detection algorithm, B-SOLANA [237], but modified to be applicable at the RNA level. This was done by using Tophat [238], which enabled mapping of splice junction reads to the Refseq transcriptome in order to detect known transcript isoforms, followed by mapping to the reference genome. To ensure maximal specificity of candidate m^5C sites, only uniquely mapped reads were retained. Uniquely mapped reads were defined by the Bowtie parameters `--best` and `--strata`, where only reads with a single best mapping are used (i.e. least number of mismatches). Additional parameters included `-n 3 -l 35 -e 200`, where `-n` sets the limit for the number of mismatches in colour-space in the first `-l` bases of the 50 nt SOLiD™ sequencing reads, and `-e` sets the limit for the sum of the Phred quality values at all mismatch positions in the 50 nt sequencing read. In general, bad quality reads have higher numbers of mismatches and as such were excluded from further analysis.

Local mapping of sequencing data

For local mapping of specific highly repetitive sequences, B-SOLANA was implemented as above but sequencing data was mapped to specific complete *in silico*

bisulfite converted reference sequences. To map mature tRNA sequences, sequencing reads were mapped to the full set of Rfam tRNA annotations in hg19 that matched the tRNA database (tRNAdb) (i.e. excluded mitochondrial tRNA sequences) with the *in silico* addition of non-templated CCA to the 3' end. For tRNA sequences that are identical, only one single sequence was selected that was representative of the entire population. rRNA sequences were mapped to the rRNA tandem repeat sequences as follows: Human ribosomal DNA complete repeating unit (18S rRNA, 5.8S rRNA, 28S rRNA, 5' external transcribed spacer and internal transcribed spacers) and Human 5S DNA. The R-Luc *in vitro* transcript was mapped to the R-Luc reference sequences derived from the plasmid sequence.

Calling of candidate m^{5}C sites

Proportion statistic: The primary statistic used for detection of candidate m^{5}C sites was a proportion statistic $p = \frac{C+\varphi}{C+T}$, where C is the number of unconverted cytosines and T is the number of converted cytosines. φ was used to add pseudocounts ($\frac{1}{16}$ counts in prostate and $\frac{1}{8}$ counts in HeLa) to prevent issues associated by division by zero. This was transformed to adapt standard RNA-seq expression analysis approaches to such proportion data. Firstly, the data was transformed into an approximate Gaussian distribution by an arc sin of square root transformation ($p' = \frac{360}{2\pi}.arcsin\sqrt{p}$). [239], followed by a log transformation, as the arc sin transformed data was closer to a log-normal distribution.

Non-specific filtering: Non-specific (i.e. blinded to experimental/control labels) filtering to initially exclude candidate m^{5}C sites that are lowly expressed was performed by requiring ≥5 reads across all samples used in the model. The HeLa methyltransferase knockdown dataset and prostate dataset were processed separately. For differential methylation, filtering of low methylation sites across controls whilst allowing low methylation in the experimental samples (knockdown for HeLa methyltransferase knockdown dataset or differential methylation for prostate dataset) was performed. At least three samples in the HeLa dataset required ≥1% methylation ($p < 0.05$; binomial test) and at least two samples in the prostate dataset required ≥5% methylation. For

calling of methylation sites in HeLa cells or commonly methylated sites in PrEC and LNCaP cells, all samples required >0% methylation.

Statistical model: A weighted moderated linear model based on *p'* was used for differential methylation estimation using Limma [240]. The moderated t-test uses information across methylation sites to improve the variance estimate of the datasets. The weights incorporated count noise estimated by quadratic fit of variance versus total read counts over each site, using the Voom method [240]. A two-sample t-test was implemented to detect candidate differential m^5C sites (i.e. knockdown vs. control and cancer vs. normal) and a one-sample t-test against a 0% methylation level for calling of candidate m^5C sites.

Differential m^5C site calling: A balanced design with NSUN2 and TRDMT1 knockdown sets against the non-targeting control (NTC) and DNMT1 knockdown samples was initially used to detect individual sites controlled by both NSUN2 and TRDMT1. As no such sites were detected, it was assumed that most sites were controlled by either NSUN2 or TRDMT1 exclusively, and to increase power, an unbalanced design was implemented; for example, NSUN2 versus TRDMT1, DNMT1 and NTC as controls. Therefore, the experimental design was unbalanced with a single "experimental" sample and effectively three control "biological replicates". Although the NTC sample is the only true unmanipulated control, knockdown of DNMT1 should have minimal effect on RNA methylation being a DNA methyltransferase and as mentioned, TRDMT1 is responsible for a different subset of sites to NSUN2, so could be used as a control relative to NSUN2. This was also the case for the opposite, where NSUN2, DNMT1 and NTC behaved as controls for detecting TRDMT1-regulated sites. For the prostate dataset, a balanced design of two LNCaP and two PrEC samples were used. The false discovery rate (FDR) was estimated by the Benjamini-Hochberg method, using a cut-off of 30% for the HeLa cell dataset, and 40% for the prostate cell datasets.

Overall m^5C site calling: To call candidate m^5C sites across all samples, a one-sample moderated t-test was used, comparing a "control" set against a constant zero methylation. The control set used for the HeLa methyltransferase knockdown dataset

included TRDMT1, DNMT1 and NTC samples for sites detected as NSUN2-regulated or potentially other RNA methyltransferases, and NSUN2, DNMT1 and NTC for sites detected as TRDMT1-regulated or potentially other RNA methyltransferases. For the prostate dataset, all samples were compared against a 0% methylation level. The FDR was estimated by Adaptive Benjamini-Hochberg using a cut-off of 1%.

Post-filtering: Post-filtering was performed to exclude potential candidate m^5C sites overlapping known single nucleotide polymorphisms (SNPs) and editing sites. Candidate m^5C sites were required to be homogeneous ($C + T \geq 95\%$) to exclude calls in bases of high read error rate or unannotated SNPs. Candidate m^5C sites overlapping dbSNP 137 (all SNPs) or the DARNED editing database were excluded. Differential somatic mutations, non-canonical editing sites and unannotated SNPs were detected using RNA-seq data. For the prostate dataset, RNA-seq data was provided by Professor Susan Clark's lab at the Garvan Institute of Medical Research. For the HeLa methyltransferase knockdown dataset, HeLa ENCODE RNA-seq data (Caltech HeLa S3 1x75) was used. Bases that did not match the reference (<90% of reads) were excluded. Finally, candidate m^5C sites that exhibited <10% mean methylation across all control samples were removed.

2.2.8.C Functional genomic analyses for m^5C in RNA

Analysis of m^5C prior to end processing of tRNAs

Sequencing reads from globally mapped tRNAs that had ≥1 candidate m^5C site from the HeLa NTC library were separated into two subsets; those showing ≥1 T-to-C mismatches and those with no mismatches. All other sequencing reads were discarded. Mean coverage plots were generated across the 5' and 3' ends of tRNAs that contained candidate m^5C sites based on bsRNA-seq.

Location bias analysis of m^5C in mRNA

Location bias analysis of candidate m^5C sites across the 5' UTR, CDS and 3' UTR was performed using binomial tests by determining the total number of candidate m^5C sites detected in each region relative to the proportion of total potential m^5C sites (for cytosines ≥5 reads) within each genomic region.

Motif enrichment analysis of m^{5}C sites

CMFinder [241] was used to find *de novo* enriched structural motifs over 100 nt flanking regions around subsets of candidate m^{5}C sites in mRNAs (3' UTR and CDS). CMFinder is a RNA motif prediction tool for unaligned sequences. It is an expectation maximisation algorithm that uses covariance models for motif description, integrating multiple techniques to effectively search motif space and a Bayesian framework to predict secondary structures. Structures were selected based on the following criteria: 1) significant enrichment, 2) substantial coverage of candidate m^{5}C sites (>30% of candidate m^{5}C sites showed the structure), 3) the distribution of candidate m^{5}C sites localised on or surrounding the structural motif. The R script for the generation of structural motifs is provided in Appendix 8.2.

Sequence logos were generated over subsets of candidate m^{5}C sites in mRNAs using the seqLogo R package [242] and using the genomic coordinates of the candidate m^{5}C sites. For the generation of sequencing logos, the nucleotides at each position were scaled according to their frequency and the entire stack of nucleotides at a given position is represented as the information content. As the nucleotides did not have equal probability of being in a certain position, a formula based on the uncertainty measure was used, developed by Claude Shannon [243]. The uncertainty measure is defined as $H_i = \sum_{b=a}^{t} f(b,i) log_2\ f(b,i)$, where H_i is the uncertainty at position i, b is one of the four bases, and $f(b,i)$ is the frequency of bases at position i. The total information at a certain position is represented by the decrease in uncertainty as the m^{5}C site is aligned using $R_{sequence}(i) = 2 - (H_i + e_n)$, where $R_{sequence}(i)$ is the amount of information present at position i, 2 is the maximum uncertainty at a given position and e_n is a correction factor that is required if there are a small number of samples. The size of each base within the logo was determined by multiplying the frequency of each base by the information content using $f(b,i)R_{sequence}(i)$.

Analysis of steady-state levels of NSUN2-targeted mRNAs

The expression levels in counts per million (CPM) of the set of mRNAs exhibiting candidate m^{5}C sites called as NSUN2 targets were compared with the set of all mRNAs showing no candidate m^{5}C sites. The log fold-change was determined between the

NSUN2 knockdown samples and control samples (NTC and TRDMT1) by differential expression analysis using Limma:

$$log_2CPM_{NSUN2} - mean(log_2CPM_{TRDMT1}, log_2CPM_{NTC})$$

The significance of the expression fold-change difference between the NSUN2-targeted mRNAs and unmethylated mRNAs was determined by the Wilcoxon test. The DNMT1 knockdown sample was removed from the analysis on the basis of its known effect on gene expression levels.

Analysis of the relationship of m^{5}C with Ago2

bsRNA-seq data was integrated with publicly available HeLa Ago2 HITS-CLIP dataset from [233] and HuR PAR-CLIP dataset from [244]. Spatial distribution summary plots were generated by anchoring at candidate m^{5}C sites in mRNAs within the CDS or 3' UTR with 500 nt flanking regions showing Ago2 HITS-CLIP or HuR PAR-CLIP footprint density. The introns were computationally removed for the generation of spatial distribution plots. For the random shuffled control, candidate m^{5}C sites were assigned to random positions within each genomic region. The significance of Ago2 footprints relative to the random shuffled control was determined using a permutation test with 10 iterations.

Analysis of the relationship of m^{5}C with conserved miRNA target sites

Genome-wide 7-mer motifs that showed a phyloP [245] conservation score > 1.3 (across 46-way vertebrate species) were divided into two sets based on whether they were the reverse complement of miRNA seed regions in miRBase v19. Spatial distribution summary plots were generated by anchoring at the candidate m^{5}C site within the CDS or 3' UTR, with a 200 nt flank, showing the density of conserved miRNA target sites. The introns were computationally removed for the generation of spatial distribution plots. As controls, the density of 7-mer motifs that were not the reverse complement of miRNA seed regions in miRBase v19, and density of conserved miRNA target sites for unmethylated cytosines were also plotted. For the random shuffled control, candidate m^{5}C sites were assigned to random positions within each genomic region. The significance of the density of miRNA target regions relative to the

random shuffled control was determined using a permutation test with 10 iterations.

Gene set enrichment and GO analyses

Gene set enrichment (GSEA) and GO analyses were performed using gage [246] and topGO [247] R packages, respectively. The analyses of the genes from the list of candidate m^5C sites in HeLa cells and candidate m^5C sites called as NSUN2 or TRDMT1 targets were performed relative to the genes showing expression in the bsRNA-seq data. Genes showing expression were defined as having ≥5 reads in at least one cytosine in the gene.

2.2.8.D TruSeq amplicon sequencing using Illumina® MiSeq

Bisulfite PCR and preparation of amplicons for MiSeq

For amplicon sequencing of bisulfite converted RNA using the MiSeq platform (Illumina, San Diego, CA, USA), touchdown PCR was performed using Platinum® Taq DNA polymerase and annealing temperatures specific to the primers (see section 2.2.4.A). PCR for each primer set was performed in triplicate and run on a 1-2% agarose gel to confirm amplification. Triplicates were pooled and purified using the Wizard SV Gel and PCR Clean-Up System according to the manufacturer's protocol. Amplicons were then quantified using the Qubit dsDNA BR Assay Kit according to the manufacturer's protocol. 30 ng of each purified amplicon for each sample was pooled and concentrated, then made up to 55 μL with dH_2O. The pooled amplicons will be referred to as DNA in the following methods. Amplicon sequencing library preparation was performed according to the TruSeq DNA LT Sample Prep Kit protocol with some modifications, as described below.

End Repair

10 μL End Repair Control and 40 μL End Repair Mix were added to 50 μL of DNA and the mixture and gently mixed. The reaction mix was incubated for 30°C for 30 min. Following end repair, the reaction mix was purified using either the MinElute PCR Purification Kit (amplicons <100 nt) according to the manufacturer's instructions, or AMPure XP beads (amplicons >100 nt). For purification using AMPure XP beads, 136 μL of beads was diluted in 24 μL dH_2O, and 160 μL added to the reaction mixture and

incubated for 15 min at room temperature. The tubes containing the mixture were placed on a magnetic rack for 5 min and the supernatant removed. The beads were washed twice with 200 µL 80% EtOH and allowed to air dry for 15 min at room temperature. The beads were resuspended in 17.5 µL Resuspension Buffer and incubated at room temperature for 2 min. The sample was then placed on a magnetic rack for 5 min and 15 µL of the liquid transferred to a new tube.

A-tailing and ligation of adaptors

2.5 µL A-tailing Control and 12.5 µL A-tailing Mix were mixed with 15 µL DNA and incubated at 37°C for 30 min. For ligation of the adaptors, 2.5 µL Ligation Control, 2.5 µL Ligation Mix and 2.5 µL DNA Adaptor Index were added to the DNA, mixed and centrifuged at 280 x g for 1 min followed by incubating at 30°C for 15 min. As multiple libraries were being sequenced simultaneously, indexed adaptors were used. 5 µL Stop Ligation Buffer was then added to the reaction mix. The reaction was purified using 42.5 µL undiluted AMPure XP beads as described previously, resuspended in 52.5 µL Resuspension Buffer and 50 µL of the reaction mix was transferred to a new tube. The purified reaction mix was purified again using 50 µL undiluted AMPure XP beads, resuspended in 22.5 µL Resuspension Buffer and 20 µL of the purified sample was transferred to a new tube.

PCR amplification to enrich libraries

25 µL PCR master mix and 5 µL PCR Primer Cocktail were added to 20 µL of DNA and mixed. The reaction mixture was then placed in a thermal cycler with the following reaction conditions: 98°C for 30 sec, 10 cycles of 98°C for 30 sec, 60°C for 30 sec and 72°C for 30 sec, followed by a final extension at 72°C for 5 min. The PCR reaction was purified using 50 µL undiluted AMPure XP beads, resuspended in 32.5 µL Resuspension Buffer and 30 µL of the purified PCR reaction was transferred to a new tube. The sample was stored at -20°C.

Quality control and quantification of libraries

To determine the average size of the libraries, 1 μL of each library was run on a 2% agarose gel at 100 V for 40 min. In addition, 2 μL of the pooled amplicons for each sample prior to library preparation were run on the same gel.

For accurate quantification of the amplicon libraries, samples were quantified using the Qubit dsDNA BR Assay Kit. Following dilution to ~50 nM after determining the average library size, the libraries were quantified either by qPCR using the Illumina® Library Quantification Kit or a DNA High Sensitivity DNA chip on the Agilent Bioanalyzer, according to the manufacturers' protocols.

Sample sheet preparation and loading of the libraries into the cartridge

The workflow selected for each MiSeq run on the sample sheet was "Fastq only". This generated fastq files only and also enabled the deselection of adaptor trimming, allowing trimming and mapping to be performed separately.

The libraries were diluted and pooled prior to loading into the cartridge according to the concentrations obtained from the qPCR or the DNA High Sensitivity DNA chip. Libraries were diluted to 10 nM in buffer EBT and 10 μL of each library was pooled together, followed by a further dilution to 2 nM in EBT. At this point, the pooled libraries were either denatured and diluted further, or submitted to the BRF for this process. In either case, the pooled libraries and PhiX control were denatured separately at 1 nM with 0.1 M NaOH for 5 min at room temperature. Following further dilutions to 20 pM and 7-9 pM in HT1, the PhiX control was spiked into the pooled libraries to a final concentration of 10%. Air bubbles were removed from the thawed cartridge, and 600 μL of the final sample was loaded into the cartridge. The sequencing run selected was 2 x 150 bp.

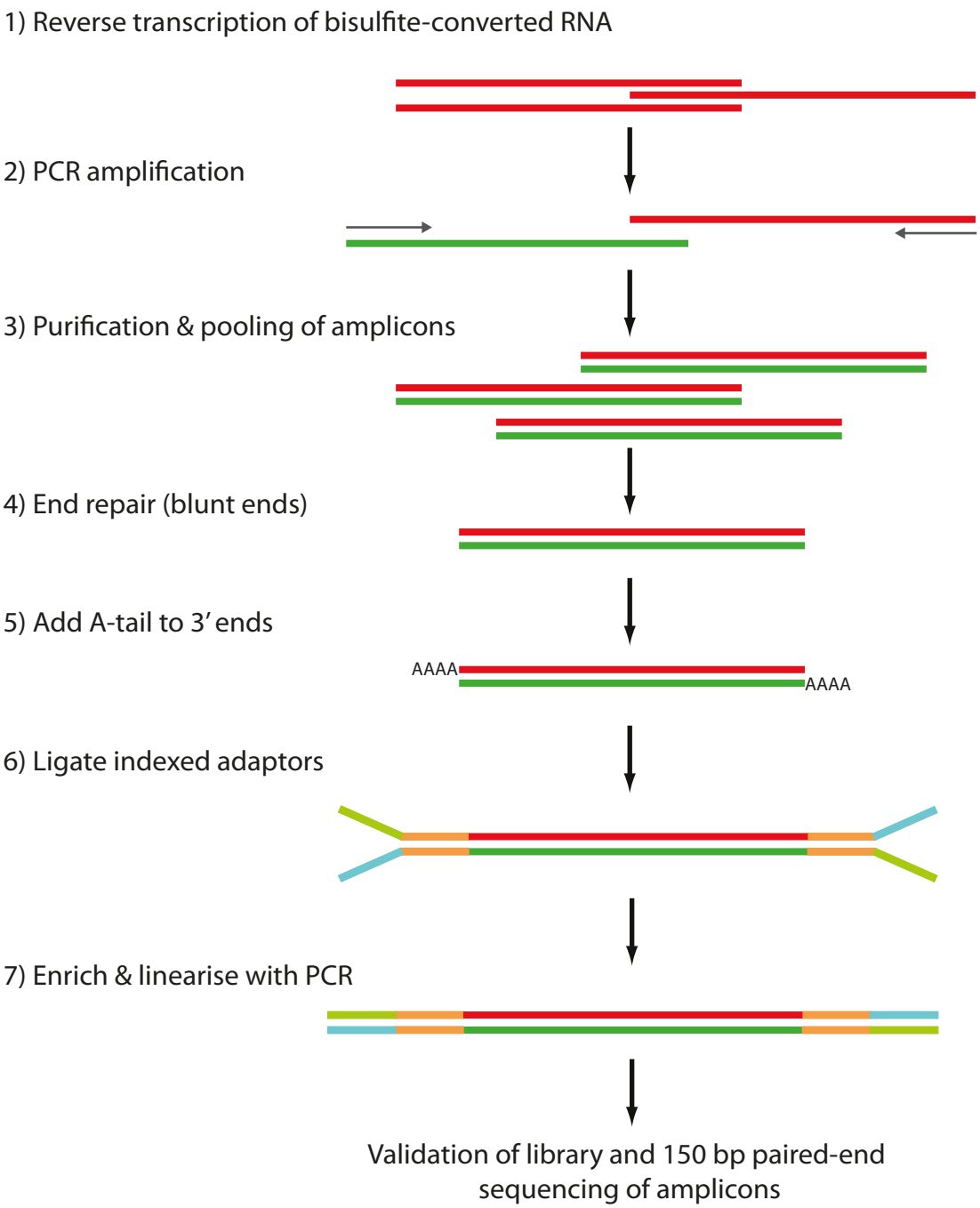


Figure 2.3. TruSeq DNA LT sample preparation method.

The TruSeq DNA LT Sample Prep Kit was used for preparing libraries. Step 1: Bisulfite-converted RNA was initially reverse transcribed using Superscript® III Reverse Transcriptase. Step 2: Specific amplicons were amplified in triplicate using Platinum® Taq DNA Polymerase. Step 3: Amplicons were purified and pooled for each sample. Step 4: The pooled amplicons were end repaired. Step 5: The pooled amplicons were A-tailed. Step 6: Indexed adaptors were ligated to the ends of the pooled amplicons. Step 7: The libraries were enriched by PCR and validated and quantified before sequencing on the Illumina® MiSeq platform.

2.2.8.E TruSeq amplicon sequencing analysis pipeline

Trimming of adaptors and poor quality ends of reads

Illumina® adaptor sequences that were ligated to the ends of amplicons to facilitate sequencing were first trimmed from the 150 bp paired-end reads by Trimmomatic [248] in palindromic mode (i.e. adaptor trimming 5' and 3' ends simultaneously), allowing single base precision. Bases at the 5' end that were below the Phred quality score of 3 and bases at the 3' end that were below the Phred quality score of 15 were removed. If any four base window within the read had a Phred quality score of <15, then the rest of the read was trimmed. The parameters for adaptor trimming were:

```
java -Xms16g -Xmx16g –classpath /home/bparker/bin/Trimmomatic-0.32/trimmomatic-0.32.jar org.usadellab.trimmomatic.TrimmomaticPE -phred33 ILLUMINACLIP:illuminaClipping.fa:4:30:10:1:true LEADING:3 TRAILING:15 SLIDINGWINDOW:4:15 MINLEN:36
```

Aligning of data

Sequencing reads were aligned to the reference sequence with Bismark, using Bowtie2 internally and implementing the parameters `bismark --non_directional --bowtie2` [249]. The reference sequence used to align the sequencing reads corresponded to the segment of RNA interrogated prior to bisulfite conversion. A single mismatch was allowed in the alignment, and candidate m^5C sites in overlapped paired-end reads were only counted once. The number of C and T calls at all C positions was extracted from the aligned sequencing reads in order to determine the proportion of m^5C.

2.2.8.F Generating coverage plots and heatmaps

Coverage plots were generated using the plot function in R Studio. The coverage for each transcript was normalised to the library size to convert to reads per million (RPM) before being plotted on a log scale in R Studio.

Heatmaps were generated using the heatmap function in the gplots package in R.

Chapter Three

Methods for detection of m^5C in RNA

3.1 INTRODUCTION

Prior to the availability of next-generation sequencing technologies, detection of nucleoside modifications in RNA relied on laborious techniques, either through metabolic or radioactive labelling, followed by hydrolysis and identification of modified nucleosides through chromatographic separation [50, 51, 59]. These techniques lacked sensitivity, and detection using these methods did not reveal the identity of transcripts or location of the modification. Now with the availability of next-generation sequencing platforms, detection of modified nucleosides and transcriptome-wide mapping of their positions is increasingly possible.

Several techniques that aid in the detection of modified nucleosides in RNA have come to the fore in the last six years. Recently, for the detection of m^6A in RNA, studies have coupled m^6A immunoprecipitation with next-generation sequencing, revealing its presence in a wide variety of transcript biotypes [56, 57]. Although this technique did not provide single nucleotide resolution, these studies have increased our understanding of the abundance of m^6A and follow-up studies are currently being conducted in order to understand the molecular function of this modification in more detail. For the detection of m^5C in RNA, the bisulfite sequencing technique that is the gold standard for the detection of DNA methylation has been adapted for use in RNA. Schaefer et al. adapted a commercial DNA bisulfite conversion kit for use on RNA, which was coupled with next-generation sequencing of PCR amplicons, focusing on m^5C in tRNAs and rRNAs [59]. Recently, we performed the first transcriptome-wide detection of m^5C in HeLa cells, coupling bisulfite conversion of RNA with transcriptome-wide sequencing, termed bsRNA-seq [58]. This revealed thousands of candidate m^5C sites in a wide variety of transcript biotypes at single nucleotide resolution, many of which were validated by Sanger bisulfite sequencing.

The overall aim of this chapter was to prepare cells for bsRNA-seq for the detection of m^5C in RNA and to establish and improve methods for the detection of m^5C in RNA. This involved RNAi-mediated knockdown of the *NSUN2* and *TRDMT1* methyltransferases (plus controls) in HeLa cells and the passaging of normal prostate epithelial cells (PrEC) and lymph node metastatic prostate cancer cells (LNCaP) cells. RNA extracted from these cells was then subjected to bisulfite conversion and library

preparation for bsRNA-seq. bsRNA-seq was performed on the SOLiD™ 5500XL platform and a bioinformatics pipeline modified from our previous study was used for the detection of candidate m^{5}C sites [58]. For subsequent validation of candidate m^{5}C sites, a high-throughput method was developed using the Illumina® MiSeq platform as a replacement for Sanger bisulfite sequencing.

3.2 Results

In addressing the aims of this thesis, the following biological samples were prepared.

To perform an initial screen to understand the substrate range of NSUN2 and TRDMT1 and to explore the molecular functions of m^{5}C in mRNA, RNAi-mediated knockdown of *NSUN2* and *TRDMT1* in HeLa cells was coupled with bsRNA-seq. The results of this approach will be addressed in Chapters 4 and 5. Non-targeting control (NTC) siRNAs were further transfected and the DNA methyltransferase *DNMT1* was also knocked down to serve as a control. Dr. Jeffrey E. Squires performed the RNAi-mediated knockdown in HeLa cells (one sample each), confirmation of knockdown and bisulfite conversion, and the preparation of the bsRNA-seq libraries for these samples. Dr. Natalie J. Beveridge prepared three additional biological replicates of the RNAi-mediated knockdown of *NSUN2*, *TRDMT1* and *DNMT1* and transfection of NTC siRNAs, which I used for subsequent validation of candidate m^{5}C sites.

To compare the RNA m^{5}C profiles of normal and cancer cells, I cultured and processed two biological replicate samples of each PrEC and LNCaP cells for bsRNA-seq, including confirmation of bisulfite conversion and bsRNA-seq library preparation for these samples. The same samples were also used for the validation of candidate m^{5}C sites. I also performed transfection of *NSUN2* siRNAs and NTC siRNAs into LNCaP cells for validation of candidate m^{5}C sites, which will be addressed in more detail in Chapter 6.

Dr. David T. Humphreys performed the emulsion PCR (ePCR) and prepared the sequencing runs for all bsRNA-seq runs and Dr. Brian J. Parker performed the mapping of all sequencing data and detection of candidate m^{5}C sites. All m^{5}C sites detected by bsRNA-seq in this study will be referred to as “candidate m^{5}C sites”.

3.2.1 Preparation of cell lines and RNA for bsRNA-seq

3.2.1.A Preparation of cells and RNAi-mediated knockdown of *NSUN2*, *TRDMT1* and *DNMT1*

For the preparation of PrEC and LNCaP cells, cells were thawed and passaged twice to ensure complete recovery from the thawing process prior to extracting RNA using Tri Reagent®. Cells were ~70% confluent prior to harvesting and two separate batches of PrEC and LNCaP cells were passaged at different times to serve as biological replicates.

Similarly, HeLa cells were thawed and passaged twice prior to transfection. RNAi-mediated knockdown of the RNA methyltransferases in HeLa cells was performed using siGENOME SMARTpool siRNAs, which contain pools of four siRNAs targeting various regions of the target transcript. The siRNAs used in this study included those targeted against *NSUN2*, *TRDMT1*, *DNMT1*, as well as a pool of four non-targeting control (NTC) siRNAs. For RNAi-mediated knockdown, a reverse-transfection protocol was employed, where the seeding of HeLa cells and transfection of cells were performed simultaneously. RNAi-mediated knockdown was performed over a 6-day period, which involved reverse-transfection of the cells, followed by the splitting of cells and an additional reverse-transfection 3 days after the initial transfection. Optimisation of the knockdown had been performed previously in our lab over 3, 4, 5 and 6 days and the depletion of methylation was investigated. Substantial depletion of m^5C in tRNAs was only observed after 6 days, due to the long half-lives of these transcripts (data not shown).

The cells were viable after 6 days of knockdown, although slower growth was evident in all knockdown samples. Knockdown of each methyltransferase was confirmed by RT-qPCR, which indicated more than 80% knockdown of *NSUN2* and *TRDMT1*, and ~60% knockdown of *DNMT1* (Figure 3.1). Depletion of the methyltransferases on the protein level could not be examined with these samples for bsRNA-seq, due to failure of the antibodies available at the time. Importantly, assessment of the depletion of methylation at the known m^5C sites targeted by NSUN2 and TRDMT1 in $tRNA^{Asp(GUC)}$ and $tRNA^{Leu(CAA)}$ was made after bisulfite conversion of the RNA. Depletion of

methylation at the NSUN2 and TRDMT1 target sites indicates effective functional depletion of the enzymes (see section 3.2.1.C).

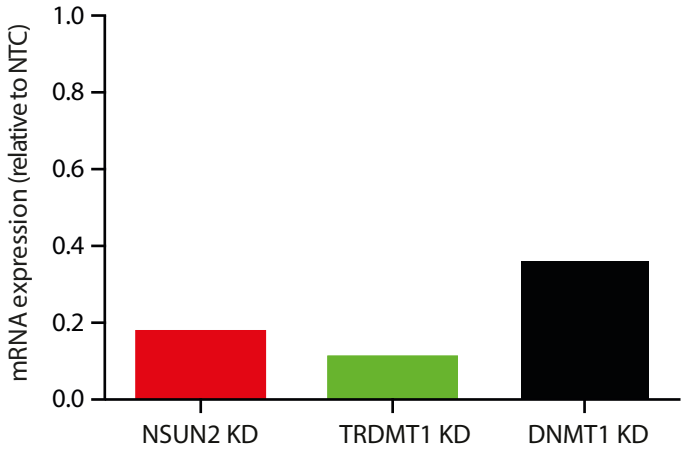
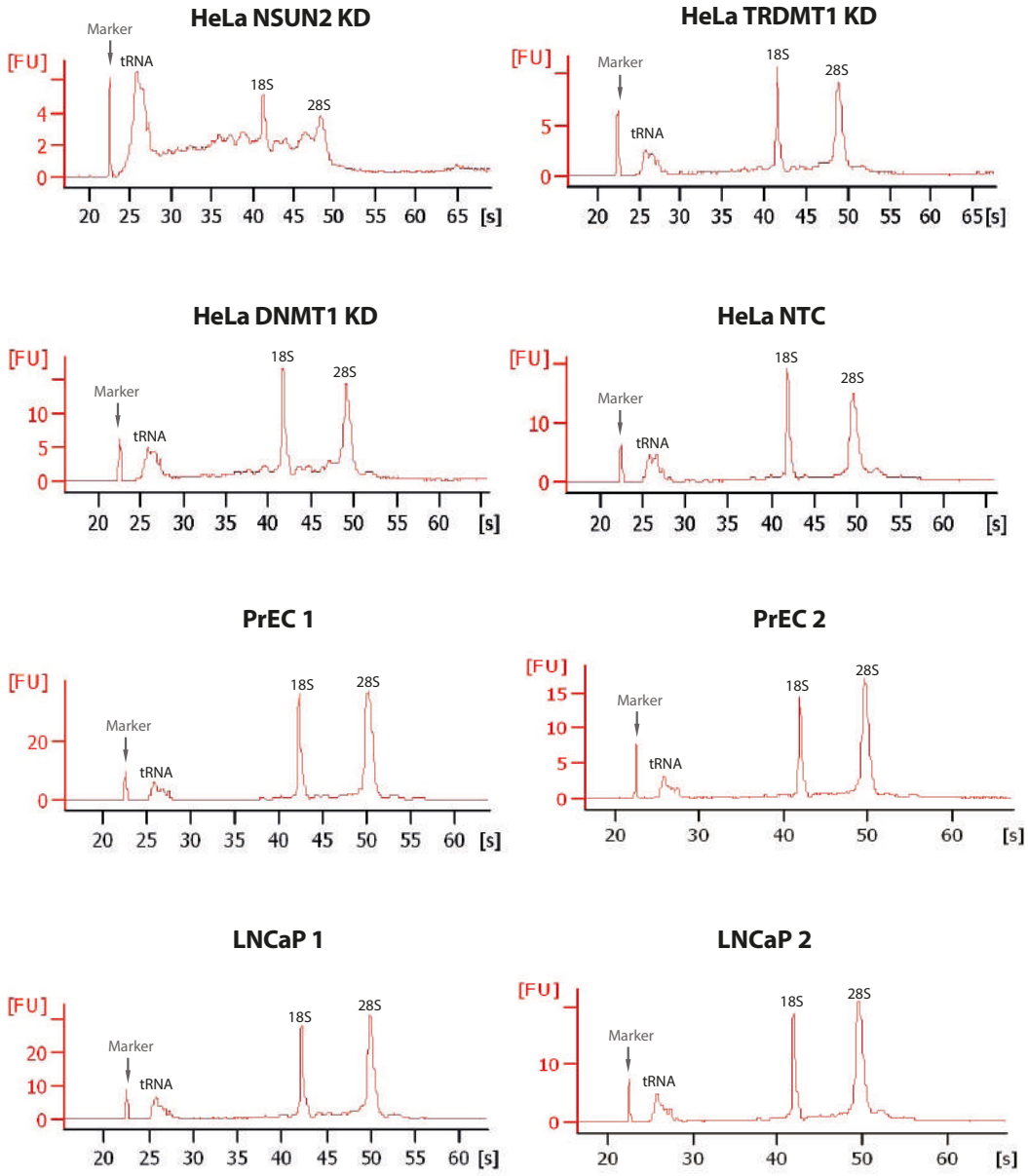


Figure 3.1. Confirmation of methyltransferase knockdown in HeLa cells by RT-qPCR.
Data has been normalised to *HPRT* and is relative to NTC (n=1).

3.2.1.B Bisulfite conversion of RNA

The RNA extracted from HeLa, PrEC and LNCaP cells was treated with TURBO™ DNase and depleted of rRNA using the RiboMinus™ Eukaryote Kit for RNA-Seq prior to bisulfite conversion. Depletion of rRNA prior to next-generation sequencing was necessary as its abundance results in biased sequencing reads towards rRNA, overshadowing sequencing reads towards more informative parts of the transcriptome such as mRNA. Depletion of rRNA was confirmed by a RNA 6000 Nano Chip, which showed almost complete depletion of 18S and 28S rRNA peaks in all samples, with only a peak at ~100 nt (migration time of 26-27 sec) evident, representing small transcripts including tRNAs (Figure 3.2A and B). However, this peak may also include small amounts of degraded rRNA that may not have been removed efficiently. mRNA is not visible on a RNA 6000 Nano Chip chromatogram of total RNA. Depletion of 18S rRNA in PrEC and LNCaP cells was confirmed by RT-qPCR, with >98% depletion observed (Figure 3.2 D).

A)


HeLa NSUN2 KD
HeLa TRDMT1 KD
HeLa DNMT1 KD
HeLa NTC
PrEC 1
PrEC 2
LNCaP 1
LNCaP 2
Marker
tRNA
18S
28S
[FU]
[s]

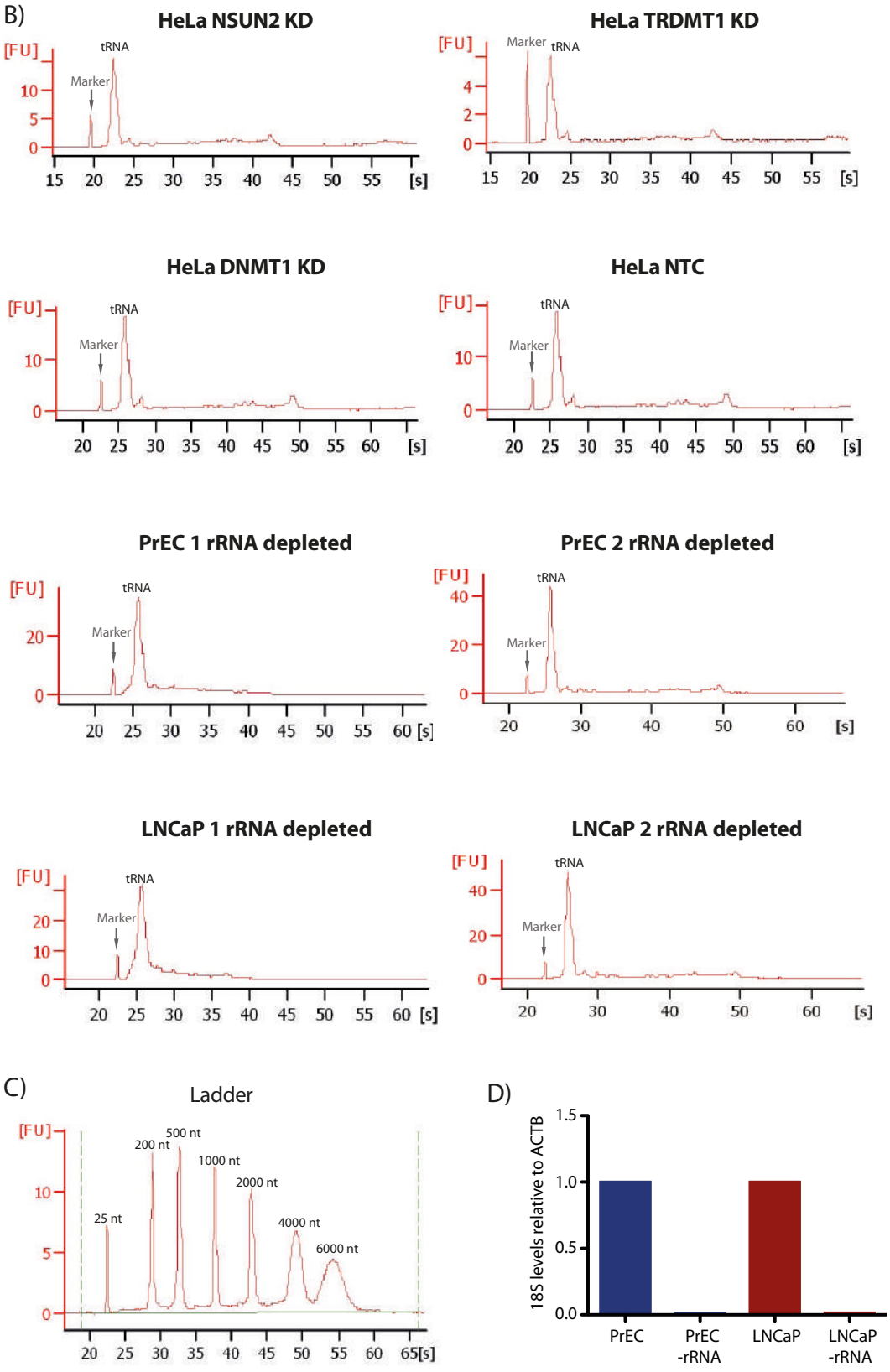

B)
HeLa NSUN2 KD
HeLa TRDMT1 KD
HeLa DNMT1 KD
HeLa NTC
PrEC 1 rRNA depleted
PrEC 2 rRNA depleted
LNCaP 1 rRNA depleted
LNCaP 2 rRNA depleted
[FU]
tRNA
Marker
[s]
C)
Ladder
25 nt
200 nt
500 nt
1000 nt
2000 nt
4000 nt
6000 nt
D)
18S levels relative to ACTB
PrEC
PrEC -rRNA
LNCaP
LNCaP -rRNA

Figure 3.2. Assessment of rRNA depletion by the RNA 6000 Nano Chip and RT-qPCR.
A) Agilent Bioanalyzer trace of DNase-treated RNA from HeLa, PrEC and LNCaP samples prior to rRNA depletion. The 18S and 28S bands are indicated. B) Agilent Bioanalyzer trace of HeLa, PrEC and LNCaP samples after rRNA depletion. The majority of 18S and 28S rRNA have been depleted. C) The sizes of the ladder corresponding to the same migration time as the samples are also indicated. D) RT-qPCR to assess depletion of 18S rRNA in the first replicate of PrEC and LNCaP cells after rRNA depletion. 18S rRNA levels are relative to *ACTB*. KD=knockdown.

Following confirmation of rRNA depletion, 2-2.5 μg RNA was denatured and bisulfite converted. Bisulfite conversion takes advantage of the differential chemical reactivity of m^5C compared to unmethylated cytosines; unmethylated cytosines are deaminated to uracil while m^5C remains as a cytosine (Figure 3.3). One of the challenges associated with bisulfite conversion is the propensity of cytosines located within highly structured regions of RNA to remain unconverted and therefore appear as clusters of non-converted cytosines. We have previously detected clusters of non-converted cytosines in both cytoplasmic and mitochondrial large subunit rRNAs, which are among the most highly structured RNAs [58]. To ensure efficient bisulfite conversion of cytosines, the RNA was denatured. However, too much denaturation and long incubation times with sodium bisulfite can lead to severe fragmentation of the RNA and therefore our treatment conditions represent a compromise between efficient RNA denaturation and limiting fragmentation [58]. Thus, it was expected that highly structured transcripts may still exhibit clusters of non-converted cytosines.

Humanised and non-humanised Renilla Luciferase (R-Luc) *in vitro* transcripts were spiked into the RNA samples prior to bisulfite treatment as controls for bisulfite conversion. Both *in vitro* transcripts contained no m^5C sites and therefore all cytosines should be converted to uracil during the process.

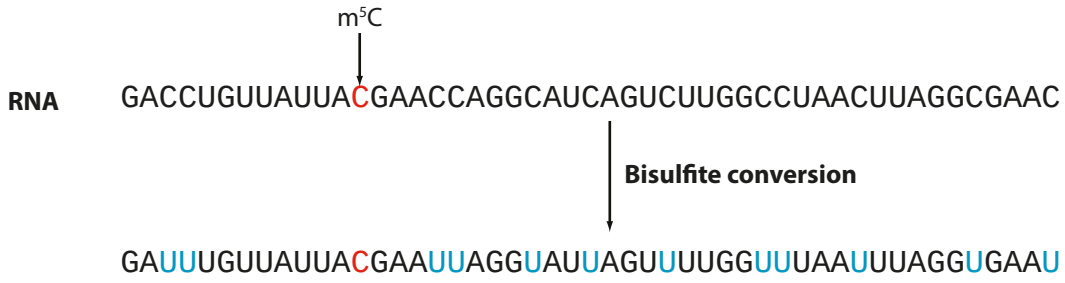


Figure 3.3. Schematic demonstrating the bisulfite conversion of RNA.
Unmethylated cytosines in RNA are converted to uracil after bisulfite conversion while m^{5}C remains as a cytosine. Red indicates the m^{5}C site and blue indicates unmethylated cytosines converted to uracil after bisulfite conversion.

3.2.1.C Assessment of bisulfite conversion and knockdown efficiencies

Bisulfite conversion and knockdown efficiencies were examined by Sanger bisulfite sequencing of several controls prior to bsRNA-seq library preparation. A 50-300 ng aliquot of bisulfite-converted RNA was reverse transcribed and the controls were PCR amplified. To confirm efficient bisulfite conversion, segments of the R-Luc *in vitro* transcript controls were PCR amplified and subjected to Sanger bisulfite sequencing. To confirm the knockdown efficiency of the RNA methyltransferases, the known m^{5}C sites in tRNA$^{Asp(GUC)}$ and tRNA$^{Leu(CAA)}$ were examined. NSUN2 is known to methylate C48 and C49 in tRNA$^{Asp(GUC)}$ and C34 in tRNA$^{Leu(CAA)}$, while TRDMT1 is known to methylated C38 in tRNA$^{Asp(GUC)}$, therefore these tRNAs served as appropriate controls. PCR amplification and Sanger bisulfite sequencing of the tRNAs were performed to confirm depletion of m^{5}C at the enzyme target sites and therefore knockdown of NSUN2 and TRDMT1. Also, while the reaction to convert m^{5}C to uracil is substantially slower than conversion of unmethylated cytosines to uracil, the m^{5}C sites in these tRNAs were examined in PrEC and LNCaP cells to ensure conversion of m^{5}C to uracil did not occur.

Primers for the PCR amplification of segments of the R-Luc *in vitro* transcripts were designed in such a way that they avoided areas of bisulfite-converted cytosines. Inefficient bisulfite conversion may result in unconverted cytosines, so it was necessary to ensure the primers detected this. (Figure 3.4A). Minimal amplification of sequences

containing unconverted cytosines in this case increased confidence in the bisulfite conversion efficiency. Primers for the amplification of the control tRNAs and validation of candidate m^5C sites were designed such that they spanned converted cytosines to avoid preferentially amplifying unconverted sequences, decreasing the likelihood of false positives (Figure 3.4B). To ensure that the primers were unique in amplifying the candidate of interest, the primer sequences were entered into BiSearch, a primer-design and search tool for bisulfite PCR [250]. Additionally, amplicons were designed to be between 70 bp-200 bp in length, depending on the surrounding sequences and the transcript biotype in which the candidate m^5C site is located. Smaller amplicons were preferable in order to avoid amplification of clustered regions of unconverted cytosines that may be attributed to the secondary structure of the RNA. We have previously found that amplification of longer amplicons from rRNA led to an increase in the detection of clusters of non-converted cytosines.

A “touchdown” PCR was used in all cases, preventing amplification of non-specific products by initially using a higher annealing temperature and decreasing by 1°C per cycle until ~5°C below the lowest primer melting temperature. All PCR products were run on an agarose gel, purified and subjected to ligation into the pGEM®-T Easy vector using T4 DNA ligase. Following bacterial transformation, culturing of colonies containing inserts and plasmid DNA purification, the DNA was submitted for Sanger bisulfite sequencing using the M13 reverse primer. For each amplicon, ~10 clones were selected for sequencing.

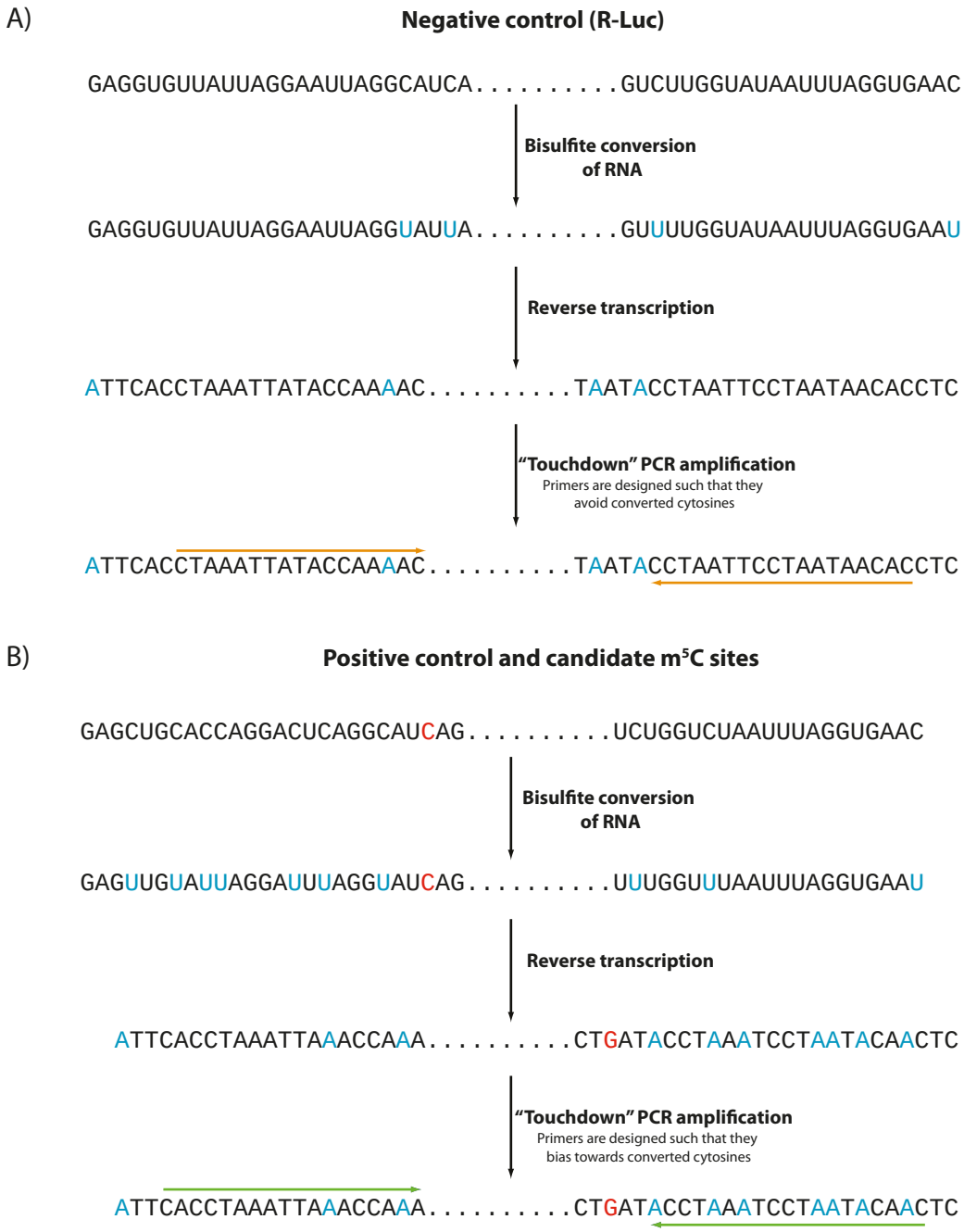


Figure 3.4. Schematic demonstrating the bisulfite conversion of RNA, reverse transcription, primer design and PCR amplification.

A) Primers designed for the R-Luc *in vitro* transcripts (orange arrows) avoided areas of converted cytosines to prevent preferential amplification of converted sequences, which may falsely indicate efficient bisulfite conversion. B) Primers designed for the tRNA controls and validation of candidate m^5C sites (green arrows) spanned areas containing converted cytosines to preferentially amplify converted sequences. Blue bases represent converted (unmethylated) cytosines and red bases represent m^5C. Primers were designed to amplify products that were 70 bp-200 bp.

Sanger bisulfite sequencing of two segments of the R-Luc *in vitro* transcript revealed 100% conversion of all cytosines in these segments for all samples except LNCaP 1, which exhibited 99.8% conversion. These findings indicate efficient bisulfite conversion of all samples (Figure 3.5).

For the HeLa methyltransferase knockdown samples, depletion of m^5C in $tRNA^{Asp(GUC)}$ and $tRNA^{Leu(CAA)}$ was examined (Figure 3.6A). As mentioned previously, NSUN2 is known to target C48 and C49 in $tRNA^{Asp(GUC)}$ and C34 in $tRNA^{Leu(CAA)}$, and TRDMT1 is known to target C38 in $tRNA^{Asp(GUC)}$. 75% depletion of m^5C at C48 and 50% depletion of m^5C C49 in $tRNA^{Asp(GUC)}$, as well as ~80% depletion of m^5C at C34 in $tRNA^{Leu(CAA)}$ was evident in the NSUN2 knockdown sample compared to the NTC sample. ~80% depletion of m^5C was evident at C38 in $tRNA^{Asp(GUC)}$ in the TRDMT1 knockdown sample compared to the NTC sample. As tRNAs are known to have exceptionally long half-lives of up to several days [251-253] and pronounced depletion of m^5C was still observed at tRNA sites targeted by NSUN2 and TRDMT1, this confirmed sufficient knockdown of NSUN2 and TRDMT1 in HeLa cells.

As expected, the DNMT1 knockdown and NTC samples showed high levels of m^5C at these positions, indicating that conversion of m^5C did not occur in these samples.

Methylation of both tRNAs was also examined in PrEC and LNCaP cells in order to determine if conversion of m^5C to uracil occurred. Sanger bisulfite sequencing of both tRNAs revealed high levels of m^5C at the known sites (Figure 3.6B). Thus, while almost complete conversion of cytosines was evident in segments of the R-Luc *in vitro* transcript, high levels of m^5C were evident at known m^5C sites indicating that conversion of m^5C did not occur.

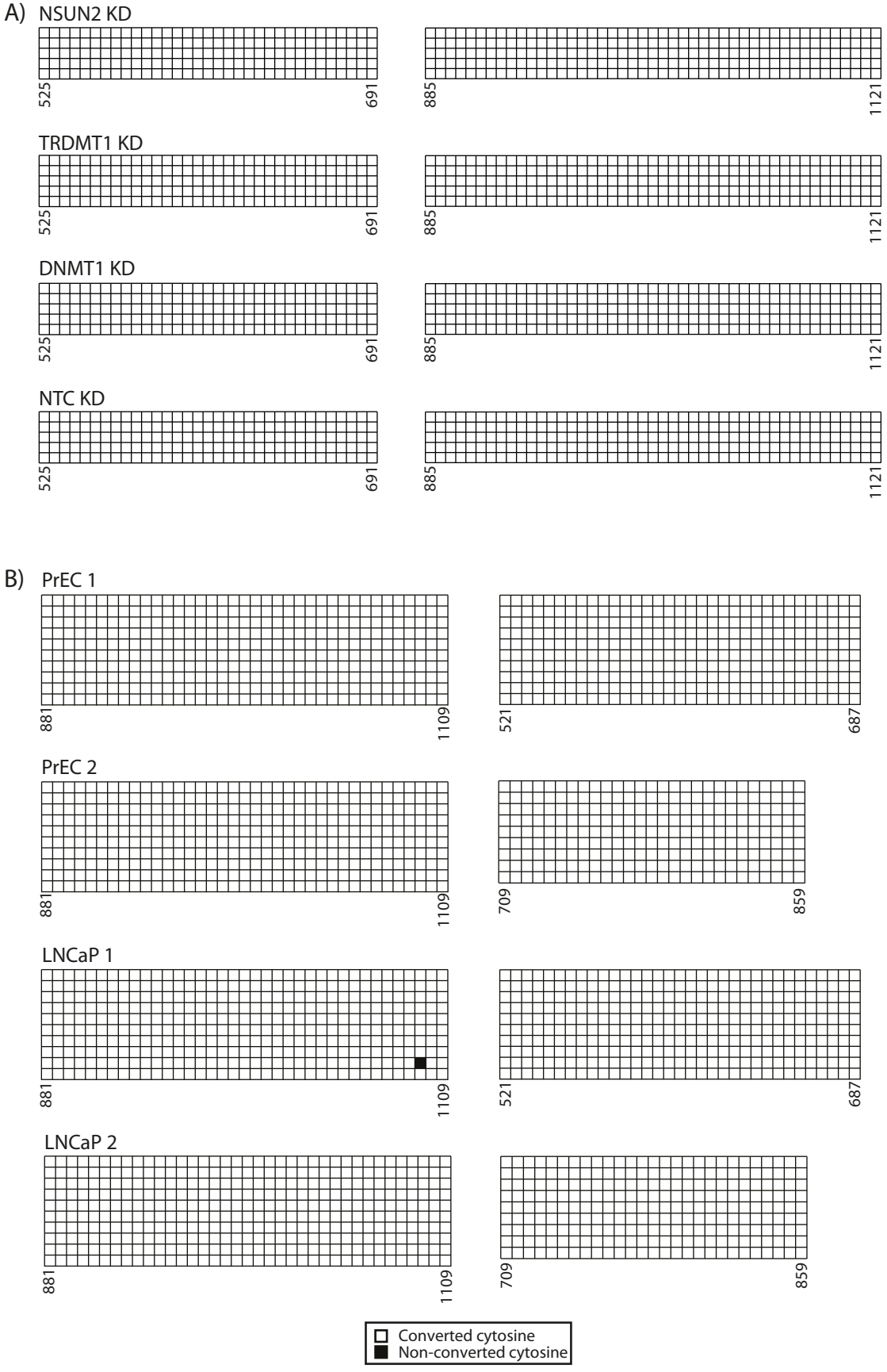

A)
NSUN2 KD
525
691
885
1121
TRDMT1 KD
525
691
885
1121
DNMT1 KD
525
691
885
1121
NTC KD
525
691
885
1121
B)
PrEC 1
881
1109
521
687
PrEC 2
881
1109
709
859
LNCaP 1
881
1109
521
687
LNCaP 2
881
1109
709
859
Converted cytosine
Non-converted cytosine

Figure 3.5. Confirmation of efficient bisulfite conversion by Sanger bisulfite sequencing. A) HeLa methyltransferase knockdown samples and B) PrEC and LNCaP samples. Two segments of a R-Luc *in vitro* transcript were amplified by PCR after bisulfite conversion and reverse transcription, and the bisulfite conversion efficiency assessed by Sanger bisulfite sequencing. Each column of squares represents a cytosine within the transcript and each row of squares represents a clone that was sequenced. Numbers represent the location of cytosines relative to the start of the transcript.

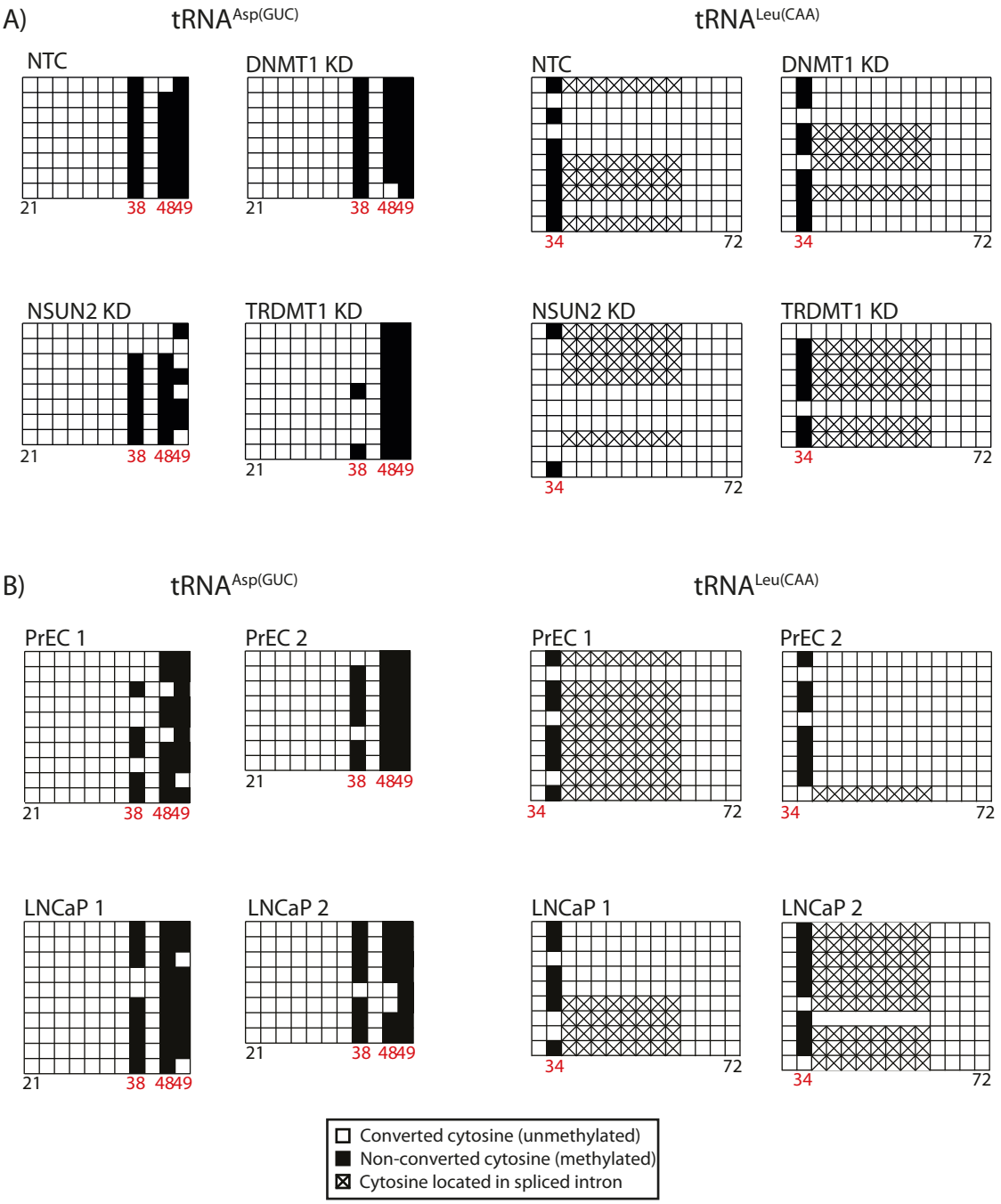


Figure 3.6. Confirmation of NSUN2 and TRDMT1 knockdown and bisulfite conversion efficiency by Sanger bisulfite sequencing.

Sanger bisulfite sequencing of endogenous tRNA$^{Asp(GUC)}$ containing known m^{5}C sites at C38, C48 and C49 and tRNA$^{Leu(CAA)}$ containing a known m^{5}C site at C34. A) Sanger bisulfite sequencing of the HeLa methyltransferase knockdown samples revealed depletion of m^{5}C at the known NSUN2 and TRDMT1 target sites. B) Sanger bisulfite sequencing of PrEC and LNCaP samples revealed high levels of m^{5}C at the known sites. Each column of squares represents a cytosine within the transcript and each row of squares represents a clone that was sequenced. The tRNA structural positions containing m^{5}C sites are indicated in red.

3.2.2 PREPARATION OF LIBRARIES FOR BSRNA-SEQ

Prior to transcriptome-wide sequencing, the RNA is normally fragmented by RNase III digestion or chemical hydrolysis. Sequencing of whole-transcriptome libraries on the SOLiD™ platform requires the RNA to be fragmented to a size between 35 nt to several thousand nucleotides depending on the sample, with an average size of 125-200 nt. Bisulfite conversion of RNA is a harsh process that results in its fragmentation and therefore fragmentation by RNase III or chemical hydrolysis was not necessary for the generation of bsRNA-seq libraries. Fragmentation of the bisulfite-converted RNA was assessed using a RNA 6000 Nano Chip, which showed sufficient fragmentation of the RNA within the desired size range (Figure 3.7). The bisulfite-converted RNA was then "end-polished" to aid in the ligation of adaptors during the library preparation. This involved de-capping of mRNA by Tobacco Acid Pyrophosphatase to ensure ligation of the 5' adaptors, and removal of any 3' phosphate groups that may exist after bisulfite conversion to ensure ligation of the 3' adaptors.

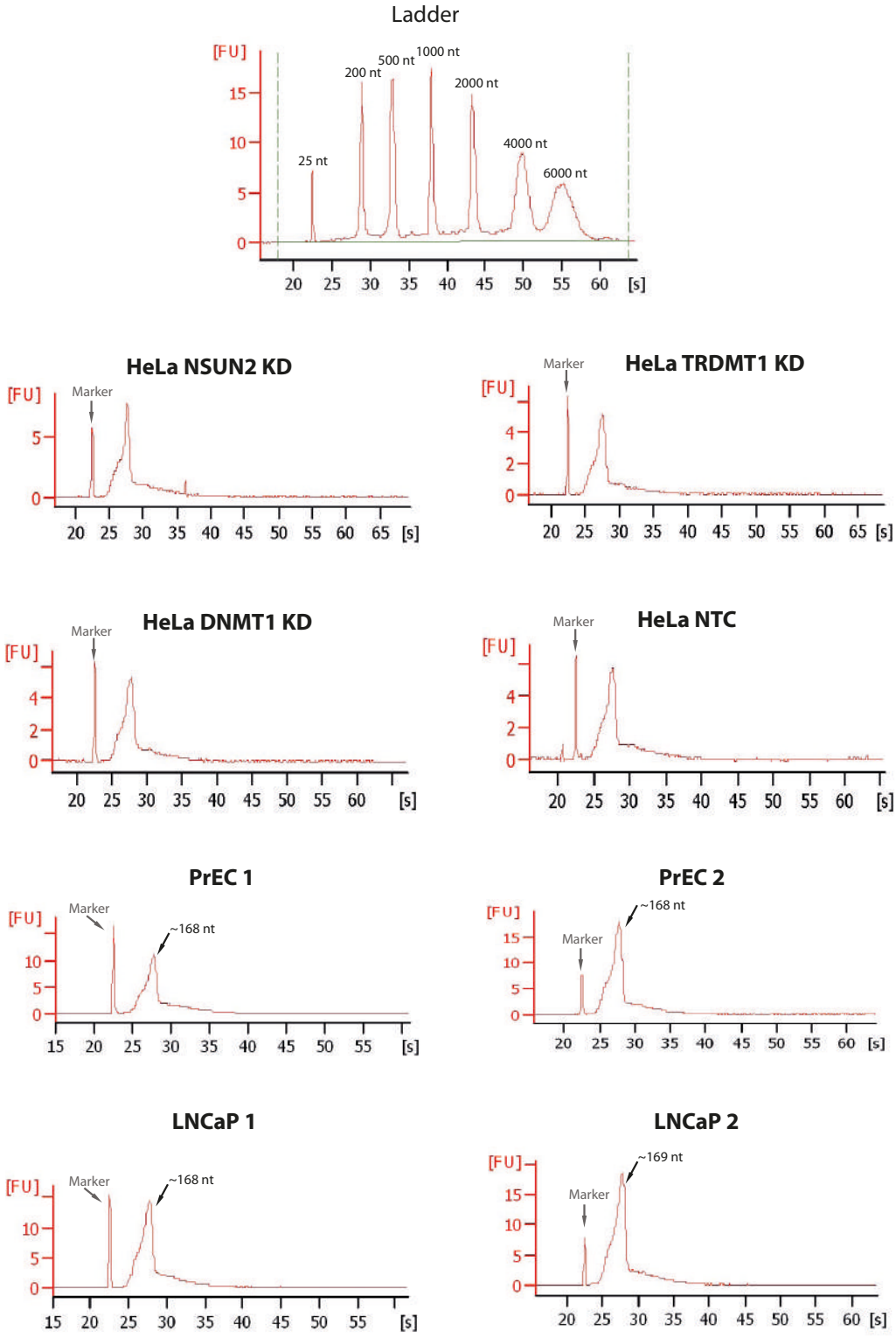


Figure 3.7. Confirmation of fragmentation of RNA from bisulfite conversion.

Bisulfite-converted RNA was subjected to a RNA 6000 Nano Chip on the Agilent Bioanalyzer to confirm fragmentation. The median fragment size obtained after bisulfite conversion and the marker are indicated. The sizes of the ladder corresponding to the same migration time as the samples are also indicated.

The libraries were prepared using the SOLiD™ Total RNA-Seq Kit, as described in Chapter 2. All HeLa methyltransferase knockdown samples were sequenced on a single run. PrEC 2 and LNCaP 2 were sequenced at a later date to PrEC 1 and LNCaP 1 to improve the accuracy of the initial screen.

First, adaptor sequences were ligated to the 5' and 3' ends of the RNA, which were used to reverse transcribe the RNA. Following sample clean-up using the MinElute® PCR Purification Kit, the cDNA was size-selected at a range of 100-160 nt, corresponding to an insert size of 50-120 nt (Figure 3.8). This ensured the selection of fragmented tRNAs and mRNAs.

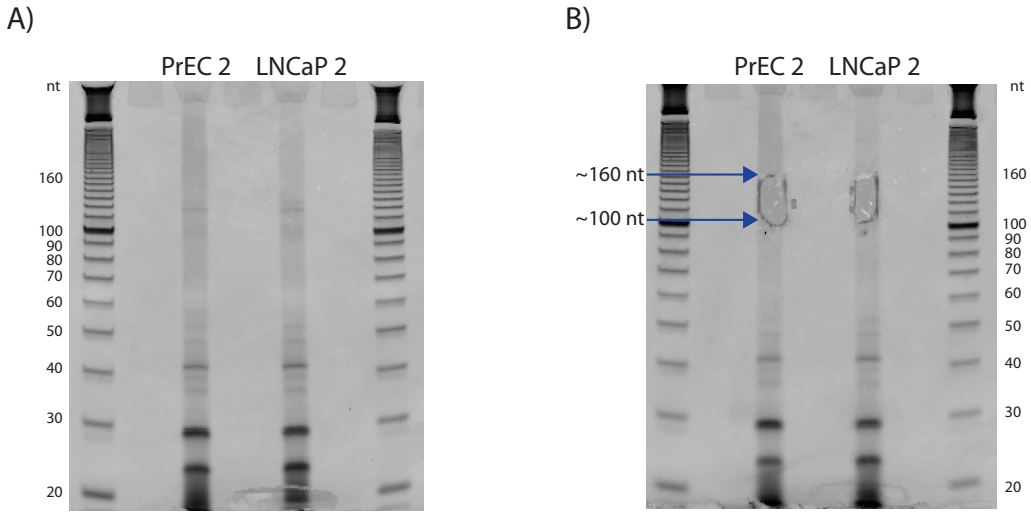


Figure 3.8. Representative cDNA size selection of the bsRNA-seq libraries.
A) 5 µL cDNA from each sample along with 1x Novex® TBE-Urea Sample Buffer was run on a 10% TBE-Urea gel at 180 V for ~45 min. The gel was stained with 1x SYBR® Gold Nucleic Acid Stain for 5 min and visualised, B) The cDNA was size-selected between ~100-160 nt, divided into four vertical strips, and the two inner strips were placed in separate 0.2 mL PCR tubes for PCR amplification. The sizes of the 10 bp DNA ladder are indicated.

The gel slice from each cDNA size selection was divided into four vertical strips, with the two inner strips being placed in separate 0.2 mL PCR tubes for PCR amplification. The samples were multiplexed allowing simultaneous sequencing of samples, which was achieved by the incorporation of 3' barcoded PCR primers. As each PrEC and LNCaP cell replicate was sequenced separately, two different 3' barcoded primers were used per sample. This was to avoid incorrectly assigning a sequencing read to the wrong barcoded sample as the SOLiD™ sequencing platform requires multiples of four barcodes. Only one 3' barcoded primer was required for each HeLa sample as there were four samples. 18 cycles of PCR were performed on all samples to enrich the libraries. After PCR amplification, the products were purified by size selection using polyacrylamide gel electrophoresis followed by EtOH precipitation (Figure 3.9). This ensured the removal of adaptors and primer dimers that may have formed during the PCR amplification step. The purification and size of the products were confirmed by a second polyacrylamide gel and a DNA High Sensitivity DNA Chip on the Agilent Bioanalyzer, respectively (Figure 3.10). As can be seen in Figure 3.10A, the concentration of each library varied, however this was adjusted for prior to ePCR to ensure equal loading of each library for sequencing. According to the polyacrylamide gel and the Agilent Bioanalyzer, all samples had the expected size range of ~150-170 nt. The increased length of the PCR amplified libraries was a result of the addition of the barcoded primer.

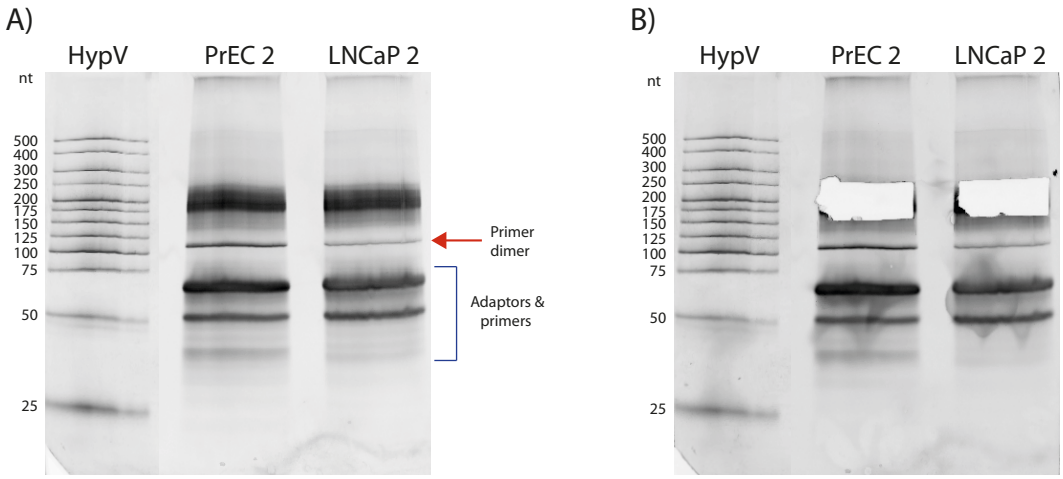


Figure 3.9. Size selection of amplified libraries by polyacrylamide gel electrophoresis.

A) Up to 20 µL amplified library along with 1x Novex® TBE-Urea Sample Buffer was separated on a polyacrylamide gel at 300 V for 37 min. The gel was stained with 1x SYBR® Gold Nucleic Acid Stain for 5 min and visualised. B) The band corresponding to ~200 nt was excised, removing all adaptors and primer dimers that could hinder the sequencing reaction. Expected sizes of adaptors and primer dimers are indicated. The sizes of HyperLadder V (HypV) are also indicated.

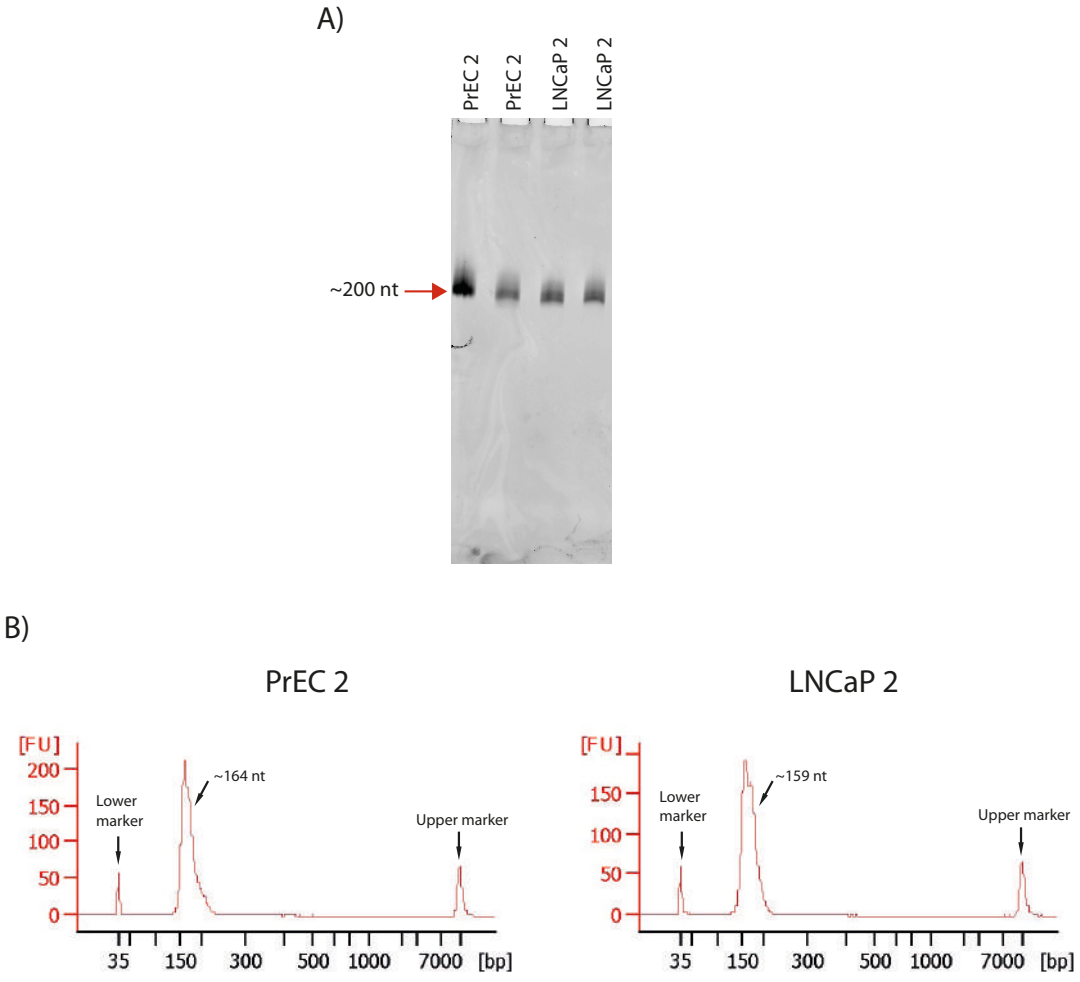


Figure 3.10. Confirmation of the bsRNA-seq purity and library size.

To confirm the purification process, A) 1 μL of the purified library was run on a polyacrylamide gel, confirming the presence of a single amplified product at ~200 nt. B) For further verification of the size and to determine the concentration of the libraries, the purified library was run on a High Sensitivity DNA Chip on the Agilent Bioanalyzer, confirming the expected size.

The libraries were pooled evenly to a total concentration of 600 pM and subjected to ePCR. In this reaction, the PCR products were mixed into a water-in-oil emulsion along with primer-coated beads, such that each water droplet contained a single PCR product. A polymerase then coated the beads with the amplified PCR product and the beads were then immobilised for sequencing on the SOLiD™ 5500XL platform.

For an indication of the bisulfite conversion efficiency on a global scale and the efficiency of the sequencing protocol, the beads were visualised on a satay plot for primer B cycle one, which provided information on the distribution of bases of the first nucleotide of the insert (Figure 3.11). The SOLiD™ platform uses a di-base encoding system, which relies on sequencing by ligation, rather than sequencing by synthesis as used in other platforms. With the use of a di-base probe that consists of a fluorescent dye at the 5' end, sequencing of every first and second base in each ligation reaction is performed with the fluorescent dye representing two nucleotides of the sequencing read (Figure 3.12). This allows each base to be interrogated twice and reduces the read error rate. For example, if only one of the two base interrogations leads to a mismatch, this most likely indicates a read error, whereas if both base interrogations lead to a mismatch it is most likely a SNP. This feature is quite advantageous for bisulfite sequencing, where bisulfite-induced mismatches occur. With the bisulfite conversion of either DNA or RNA, the distribution of colours changes (Table 3.1).

Sequencing of the last base of the adaptor, which is a T, and the first base of the insert revealed clear depletion of cytosines through lack of signal in the TXR channel (Figure 3.11). This provided an indication that the bisulfite conversion efficiency was high on a global level, for at least the first base of the insert.

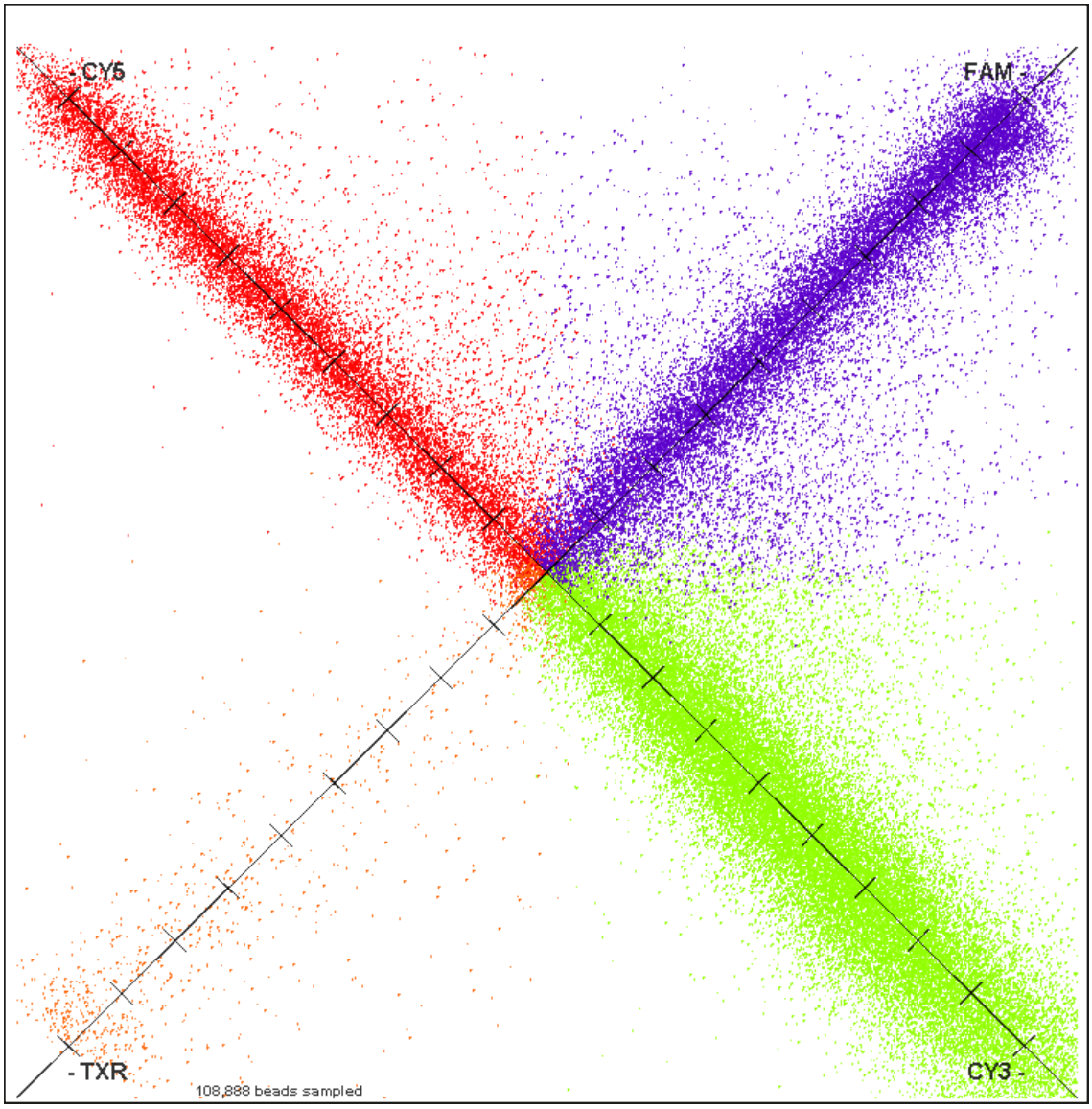


Figure 3.11. Representative satay plot of the bsRNA-seq libraries for primer B cycle one. On the SOLiD™ 5500XL platform, primer B sequences the last nucleotide of the adaptor, which is a T, as well as the first base of the insert. The TXR dye (orange) encodes TC, and there was a clear depletion of TC dinucleotides during cycle one of sequencing, indicating efficient bisulfite conversion of cytosines located in the first position of the insert.

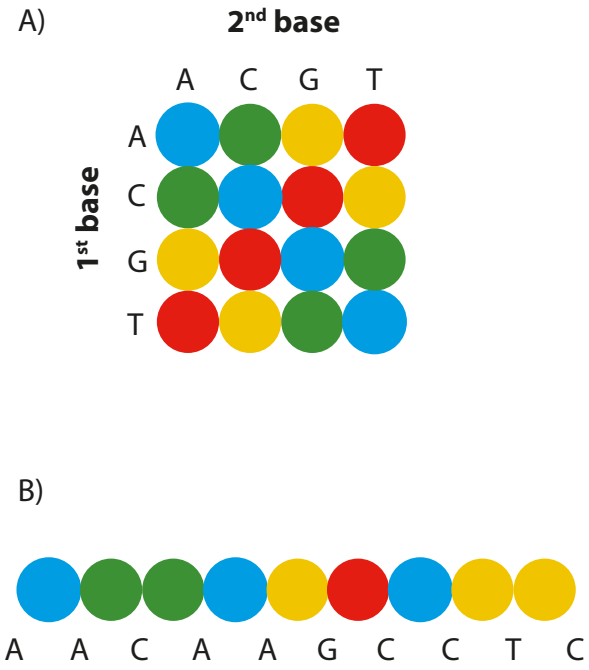


Figure 3.12. SOLiD™ sequencing uses di-base encoding, where one colour represents two bases.

A) The decoding matrix allows the conversion of colour-space to base-space, as long as one of the two bases is known. B) An example of how sequencing in colour-space allows double interrogation of each base. Adapted from Applied Biosystems.

Table 3.1. The distribution of colours of standard RNA and bisulfite-converted RNA using SOLiD™ sequencing.

	FAM	CY3	TXR	CY5
Normal RNA	AA CC GG TT	AC CA GT TG	AG GA CT TC	AT TA CG GC
Bisulfite-converted RNA	AA GG TT	 GT TG	AG GA 	AT TA

3.2.3 Mapping of bsRNA-seq data and detection of candidate m^{5}C sites

The following sections document details of the mapping of the bsRNA-seq libraries and the general features of the libraries. Dr. Brian J. Parker performed the mapping of the bsRNA-seq libraries and detection of candidate m^{5}C sites. The statistical model used for the detection of candidate m^{5}C sites was developed by Dr. Brian J. Parker and Dr. Jiayu Wen.

3.2.3.A Mapping and general features of the bsRNA-seq libraries

The alignment algorithm used for the mapping of the bsRNA-seq data was B-SOLANA, which was specifically designed to map bisulfite-converted DNA sequencing reads in colour-space [237]. While there are a number of alignment algorithms available that are specific to bisulfite sequencing of DNA, B-SOLANA is advantageous for bsRNA-seq as it is accurate and fast. SOCS-B, another bisulfite sequencing alignment algorithm we have previously used [58], becomes computationally intensive with large genomes such as the human genome, and is therefore quite slow. B-SOLANA improves speed and unique mapping rate by aligning against both fully *in silico* bisulfite converted and non-CpG *in silico* bisulfite converted references simultaneously with Bowtie. Following this, B-SOLANA determines the methylation levels.

The alignment algorithms available for the mapping of bisulfite-converted data are specifically designed for DNA. As such, it was necessary to adapt B-SOLANA for use with bisulfite-converted RNA. One of the major differences with m^{5}C in RNA is that it does not necessarily occur in a CpG context and clusters of m^{5}C sites in RNA have not yet been detected in a similar manner to m^{5}C in DNA [58]. Adjustments to the B-SOLANA alignment algorithm were applied for use with RNA using Tophat, while also keeping the features associated with mapping bisulfite-converted DNA. This enabled the mapping of splice junction reads to *in silico* bisulfite-converted Refseq transcriptomes in order to reliably detect known transcript isoforms, and then to *in silico* bisulfite-converted hg19 reference genomes (Figure 3.13A).

Several ncRNAs including rRNAs and tRNAs are either highly repetitive or identical sequences can be derived from multiple locations in the genome, which can cause multi-mapping issues. As such, specific “local” mapping was applied for rRNAs and tRNAs (Figure 3.13B). In addition, the R-Luc *in vitro* transcripts were mapped locally from the sequences derived from the plasmid they were generated from.

Although rRNA was depleted prior to bsRNA-seq, it was expected that a large proportion of rRNA would still be present in the library as no depletion method to date has been successful in completely removing rRNA. As such, efforts were made to map remaining rRNA sequencing reads to rRNA tandem repeat sequences as references. The non-CpG *in silico* bisulfite conversion of the references was not applied for the local mapping to limit detection of non-converted cytosines due to the strong secondary structure of rRNA.

tRNAs were mapped both genome-wide to all Rfam tRNA annotations and locally to Rfam tRNA annotations that matched tRNAdB. The genome-wide mapping of tRNAs was used for the detection of candidate m^5C sites in these transcripts. To focus on candidate m^5C sites in tRNAs in greater detail, local mapping was applied. The references used for the local mapping incorporated an *in silico* non-template 3’ CCA addition, which specifically allowed mapping of mature tRNA sequences. Additionally, as identical tRNA sequences may be derived from multiple locations in the genome, a single representative reference was used to map identical mature sequences in the local mapping. As the global mapping does not specifically map mature tRNA sequences, it was possible to distinguish between identical tRNA sequences in some cases when sequencing reads extended beyond the mature ends of the tRNA. For this study, the genome-wide mapping of tRNAs will be referred to as “global mapping”, and the mapping of mature tRNAs will be referred to as “local mapping”.

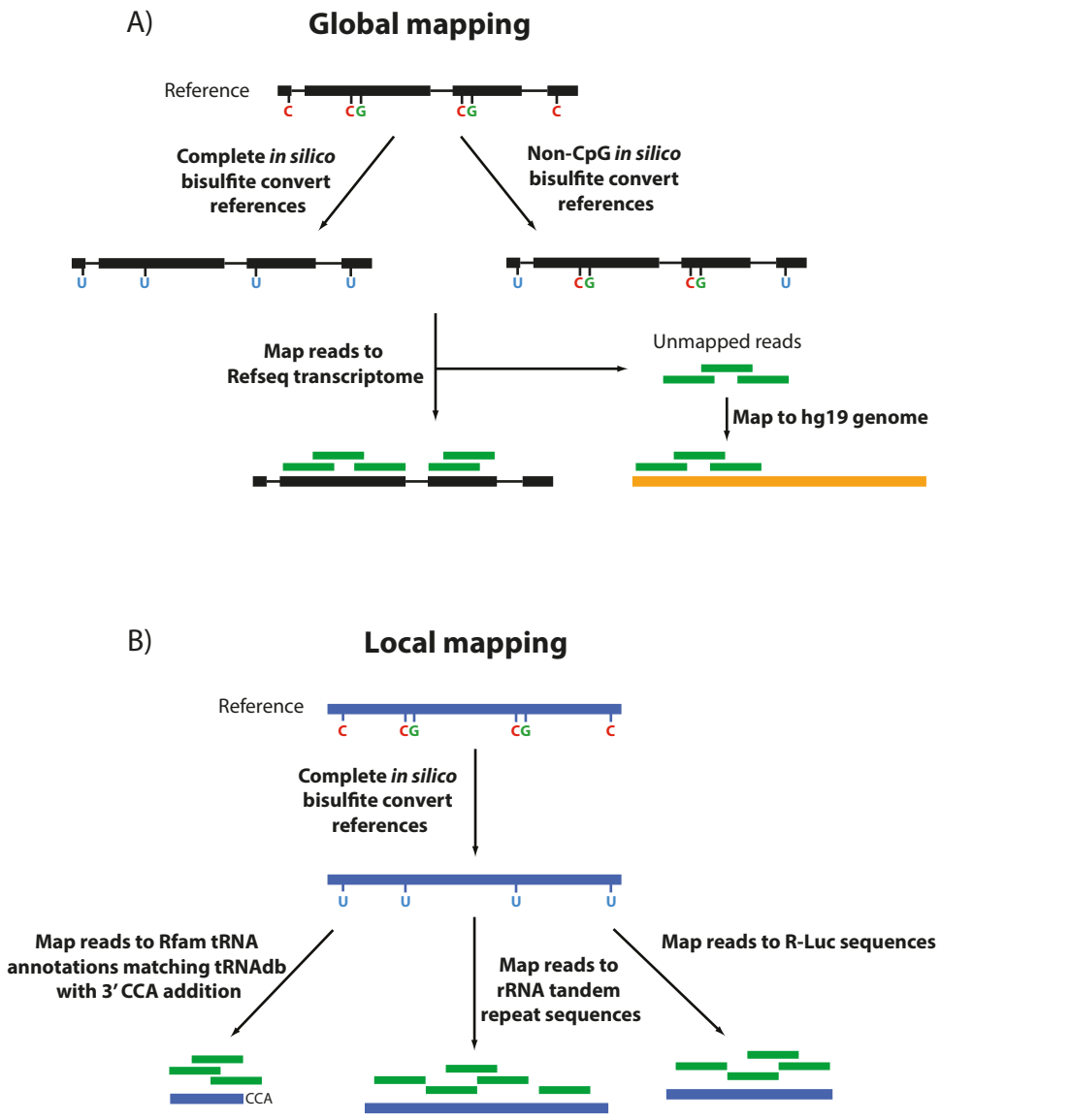


Figure 3.13. Schematic of the global and local mapping of 50 nt SOLiD™ sequencing reads by B-SOLANA.

A) Global mapping: the references were both completely *in silico* bisulfite converted and non-CpG *in silico* bisulfite converted prior to mapping. The 50 nt SOLiD™ sequencing reads were then mapped to both *in silico* bisulfite converted human Refseq transcriptomes. Sequencing reads that did not map to this reference were mapped to both *in silico* bisulfite converted human hg19 genomes. B) Local mapping: the references were completely *in silico* bisulfite converted prior to mapping. Sequencing reads were then mapped to complete *in silico* bisulfite converted Rfam tRNA annotations that matched tRNAdB with a 3' CCA addition, rRNA tandem repeat sequences and R-Luc *in vitro* transcript sequences.

In all cases of mapping, the Bowtie parameters `--best` and `--strata` were applied to ensure that only sequencing reads with a single best mapping were used and only mapped once to the best mapping transcript, minimising multi-mapping. These parameters were applied to ensure that those sequencing reads that mapped to multiple locations in the genome were removed from the analysis to prevent ambiguity. Additional parameters that were applied allowed a maximum of three colour-space mismatches in the first 35 nt of the 50 nt SOLiD™ sequencing reads, and a Phred score sum of 200 at all mismatch positions in the 50 nt sequencing read (not just the first 35 nt). While these are stringent criteria for the mapping of the sequencing reads, the high Phred score sum allowed for several base-space mismatches, particularly towards the 3' end of the sequencing read. This was because in our approach, non-converted cytosines and therefore candidate m^5C sites directly contribute to the mismatch score.

Between $5.0x10^7$-$12.1x10^7$ sequencing reads were generated for the HeLa methyltransferase knockdown libraries, of which ~38% were mappable (Table 3.2). For the PrEC and LNCaP libraries, between $1.32x10^8$-$1.83x10^8$ sequencing reads were generated of which ~44-76% were mappable. Despite each library being diluted to equivalent concentrations prior to sequencing, variations in the HeLa methyltransferase knockdown libraries were evident, with the NSUN2 knockdown library having less than half the amount of sequencing reads compared to the TRDMT1 knockdown and DNMT1 knockdown libraries. Similarly, the NTC library had a lower amount of sequencing reads compared to the TRDMT1 knockdown and DNMT1 knockdown libraries. This suggests that less of the NSUN2 knockdown and NTC libraries were loaded for sequencing. The PrEC and LNCaP libraries yielded relatively even numbers of sequencing reads even though the replicate libraries were prepared separately.

Table 3.2. The number of total mappable sequencing reads in each bsRNA-seq library.

Library	Total sequencing reads	Total mappable reads	% mappable reads
NSUN2	50,689,481	18,963,160	37.4%
TRDMT1	114,757,033	43,524,151	37.9%
DNMT1	120,221,774	45,944,290	38.2%
NTC	76,306,528	28,580,132	37.5%
PrEC 1	133,905,207	59,111,682	44.1%
PrEC 2	182,032,122	87,990,804	48.3%
LNCaP 1	132,529,838	60,361,771	45.5%
LNCaP 2	144,990,405	110,538,861	76.2%

As expected, ~40-75% of the sequencing reads mapped to rRNA (Figures 3.14 and 3.15). ~14-24% sequencing reads mapped to mRNAs, 2-7% to tRNAs and the remainder to other ncRNAs. It is clear that RNA extractions, library preparations and sequencing runs performed at different times, as is the case for each replicate of the PrEC and LNCaP samples, resulted in variations in the number of mappable sequencing reads (Table 3.2) and the distribution of the sequencing reads across different transcript biotypes (Figure 3.15). In the first library preparation for PrEC and LNCaP cells, the proportion of sequencing reads mapping to rRNA was significantly higher than the second library preparation. This may be attributed to batch variations in the RiboMinus™ Eukaryote Kit for RNA-Seq. The variability observed between the PrEC and LNCaP replicates may be prevented in the future by performing all experimental procedures simultaneously for all samples.

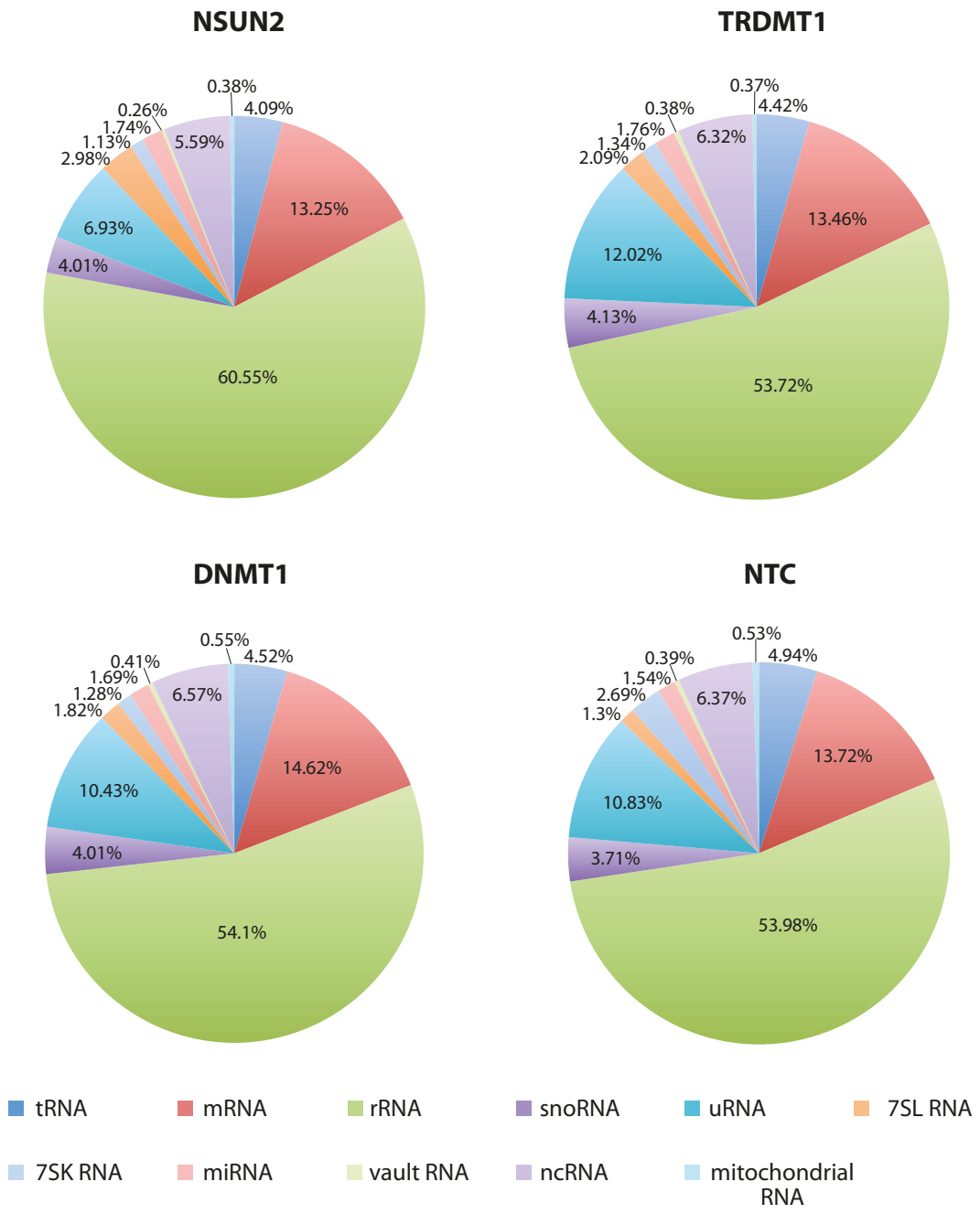


Figure 3.14. Distribution of sequencing reads of the HeLa methyltransferase knockdown bsRNA-seq libraries.

The sequencing reads were mapped using B-SOLANA to the appropriate *in silico* bisulfite-converted reference sequences. The distribution of sequencing reads in tRNAs are based on the global mapping.

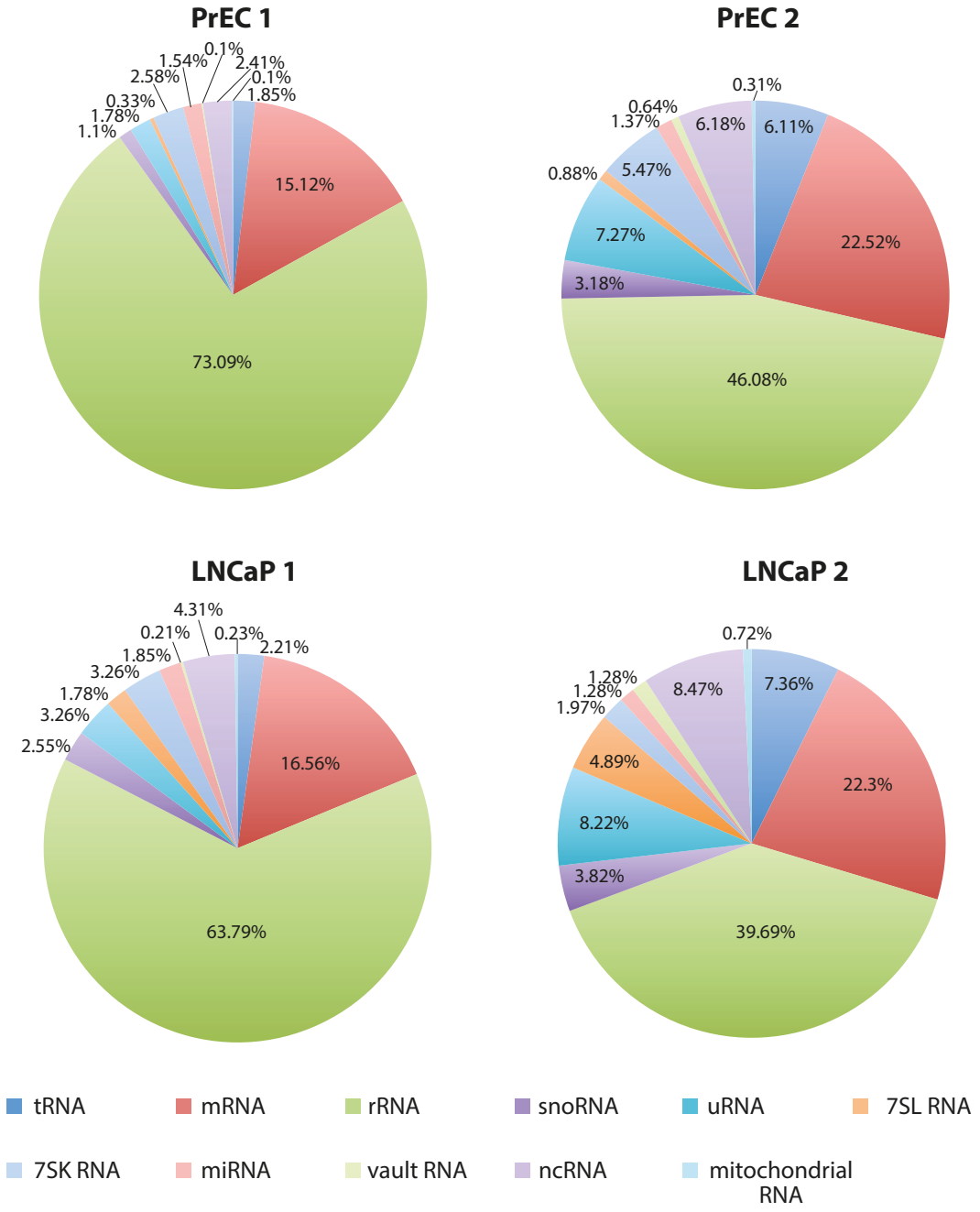


Figure 3.15. Distribution of sequencing reads of the PrEC and LNCaP bsRNA-seq libraries.

The sequencing reads were mapped using B-SOLANA to the appropriate *in silico* bisulfite-converted reference sequences. The distribution of sequencing reads in tRNAs are based on the global mapping.

3.2.3.B Detection of candidate m^{5}C sites in RNA

bsRNA-seq was performed on HeLa methyltransferase knockdown cells to determine the substrate range of NSUN2 and TRDMT1 and also to investigate the molecular functions of m^{5}C in mRNA. bsRNA-seq was performed on PrEC and LNCaP cells to compare their RNA m^{5}C profiles, and to compare this with HeLa cells to confirm patterns of m^{5}C across multiple cell lines. Thus, for the detection of candidate m^{5}C sites, the HeLa methyltransferase knockdown dataset and PrEC and LNCaP dataset were processed separately.

Prior to detection of candidate m^{5}C sites, non-specific filtering was applied to each dataset individually. Initially, all samples used in the model for each dataset required ≥5 reads at the candidate m^{5}C site, removing lowly expressed reads which may have led to incorrect detection of differential candidate m^{5}C sites. Second, non-specific filtering of low non-converted cytosines was applied in order to exclude noise due to spurious non-conversion. For differential methylation, 3 out of 4 samples from the HeLa methyltransferase knockdown dataset required >1% non-conversion, and at least 2 out of 4 samples required ≥5% non-conversion for the PrEC and LNCaP dataset. For detecting candidate m^{5}C sites in HeLa cells or candidate m^{5}C sites in common between PrEC and LNCaP cells, >0% non-conversion was required in all samples.

Detection of candidate m^{5}C sites mediated by NSUN2 and TRDMT1 in HeLa cells

Prior to identification of candidate m^{5}C sites selectively mediated by NSUN2 or TRDMT1, all candidate m^{5}C sites were first examined to determine whether any were targeted by both RNA methyltransferases by comparing both enzyme knockdown samples against the DNMT1 knockdown and NTC samples as controls. This revealed that no such sites were detectable in the bsRNA-seq data. To identify candidate m^{5}C sites that were mediated by either NSUN2 or TRDMT1, a two-sample moderated t-test was used. As no replicates were sequenced, the experimental design was set up in such a way that there was one “experimental” sample and three “controls”. Such an experimental design was required in order to implement statistical analyses for the detection of candidate m^{5}C sites. That is, if the NSUN2 knockdown sample was treated as the experimental sample, then the TRDMT1 knockdown, DNMT1 knockdown and NTC samples were treated as controls to detect candidate m^{5}C sites that were mediated

by NSUN2. Although the NTC sample was the only true control, DNMT1 should not target RNA for methylation and TRDMT1 should exclusively target sites other than those of NSUN2 for methylation, as described previously. The NSUN2 knockdown, DNMT1 knockdown and NTC samples were also treated as controls to detect sites that were mediated by TRDMT1. As RNAi-mediated knockdown was implemented in this study, it was expected that complete ablation of m^5C at NSUN2 or TRDMT1 target sites would not necessarily be observed. Due to the low sample replication, a FDR≤30% was set as the threshold for the detection of candidate m^5C sites mediated by NSUN2 and TRDMT1 to strike a balance between true positive enrichment while maintaining the proportion of false negatives at an acceptable level.

Detection of candidate differential m^5C sites in PrEC and LNCaP cells

To detect candidate differential m^5C sites in PrEC and LNCaP cells, a balanced design of two LNCaP and two PrEC samples was used, and a moderated two-sample t-test was implemented. Therefore, for detecting candidate m^5C sites that were hypermethylated in PrEC cells, the LNCaP samples were defined as controls, and vice-versa. Due to the low sample replication and an evident paucity of candidate differential m^5C sites between PrEC and LNCaP cells, a FDR≤40% was set as the threshold to obtain a set of such sites for further inspection.

Detection of all candidate m^5C sites in HeLa cells

While determining the substrate range and specificity of NSUN2 and TRDMT1 was one of the aims of this study, it was also of interest to detect all candidate m^5C sites in HeLa cells. The same unbalanced design was applied as described previously, using three control samples. A one-sample moderated t-test was used for the detection of candidate m^5C sites by comparing the cytosines in the control samples against a 0% methylation level to determine if they were statistically significantly increased. The TRDMT1 knockdown, DNMT1 knockdown and NTC samples were treated as controls to detect candidate m^5C sites that were mediated by NSUN2 or potentially other methyltransferases, and the NSUN2 knockdown, DNMT1 knockdown and NTC samples were treated as controls to detect candidate m^5C sites that were mediated by TRDMT1 or other methyltransferases. Combining the results from both sets of analyses provided the total number of candidate m^5C sites detected in HeLa cells, regardless of

whether they were mediated by NSUN2 or TRDMT1. As a result of the higher number of control samples used to detect all candidate m^5C sites in HeLa cells compared to candidate m^5C sites mediated by NSUN2 and TRDMT1, a stringent FDR≤1% threshold was applied.

Detection of candidate m^5C sites in common between PrEC and LNCaP cells

To detect candidate m^5C sites in common between PrEC and LNCaP cells, all four samples were used as controls. A one-sample moderated t-test was used for the detection of candidate m^5C sites by comparing the cytosines in the control samples against a 0% methylation level to determine if they were statistically significantly increased. As a result of the higher number of control samples used to detect candidate m^5C sites in common between PrEC and LNCaP cells compared to candidate differential m^5C sites, a FDR≤1% threshold was applied.

Post-filtering of candidate m^5C sites

Post-filtering of candidate m^5C sites was performed in order to remove any potential artefacts. Candidate m^5C sites had to be homogenous, where ≥95% of reads had to be a C or T, and any sites that overlapped a SNP or editing site were excluded. Finally, while many candidate m^5C sites would have been detected using the described proportion statistic, the amount of methylation for some of these may be too low for validation and further exploration. As such, for a site to be considered methylated, ≥10% mean methylation across all control samples was required.

Figure 3.16 shows a schematic for the detection of candidate m^5C sites in the different samples. Details into the number of candidate m^5C sites detected in HeLa cells and PrEC and LNCaP cells will be addressed in Chapters 4 and 6, respectively.

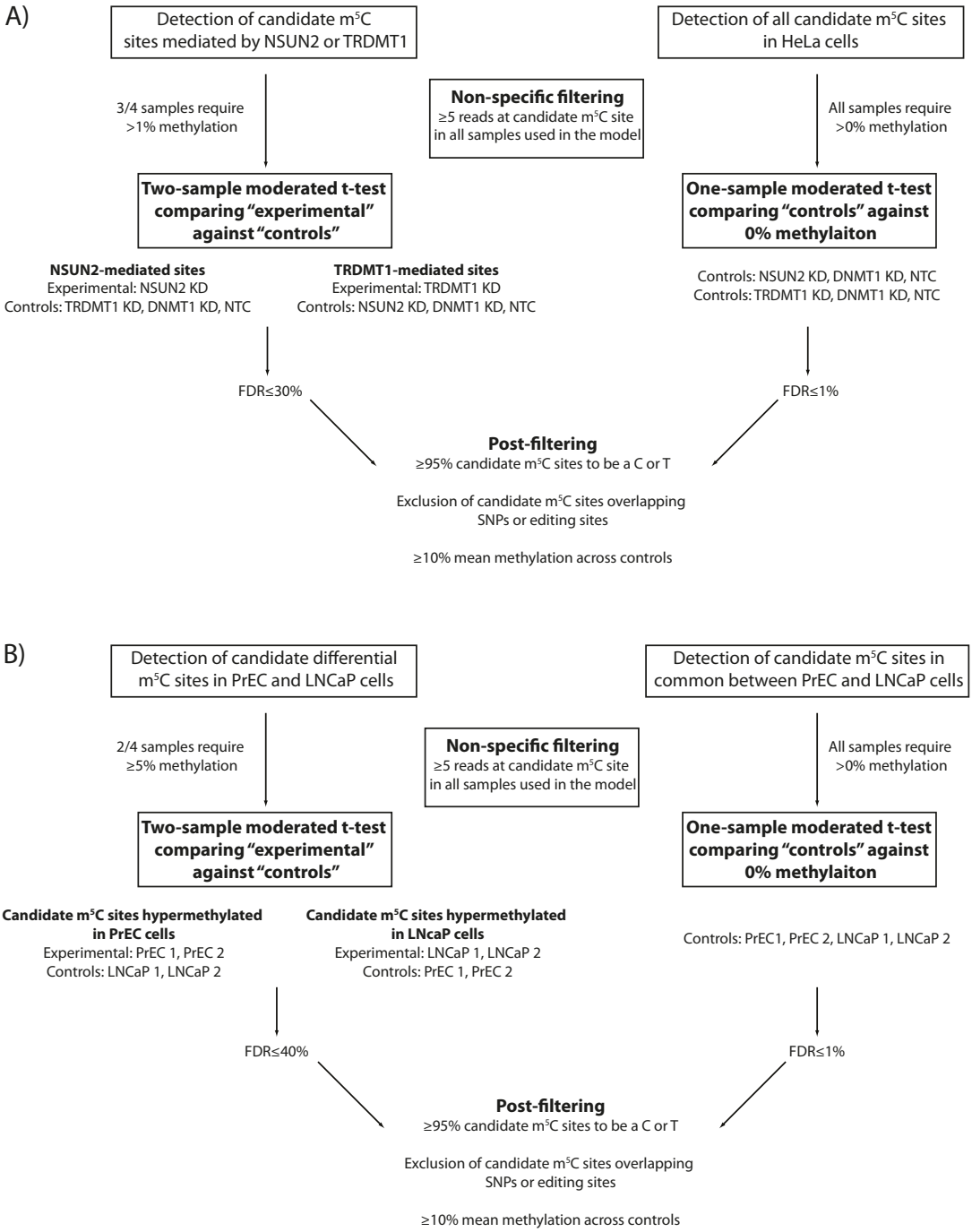


Figure 3.16. Schematic demonstrating the detection of candidate m^{5}C sites.

A) Left: detection of candidate m^{5}C sites mediated by NSUN2 or TRDMT1. Right: detection of all candidate m^{5}C sites in HeLa cells. KD=knockdown. B) Left: detection of candidate differential m^{5}C sites in PrEC and LNCaP cells. Right: detection of candidate m^{5}C sites in common between PrEC and LNCaP cells.

3.2.4 Detection of m^{5}C in RNA using an antibody

To ensure the accuracy of candidate m^{5}C sites detected using bsRNA-seq, it was desirable to validate these sites using different methods. Previously, this was performed by Sanger bisulfite sequencing, which is laborious, low-throughput and still relies on bisulfite conversion of RNA [58]. There is also the possibility that other modifications besides m^{5}C may be resistant to bisulfite conversion. As such, attempts were made to optimise an antibody for the detection of m^{5}C in RNA using MeRIP. As mentioned in the Chapter 1, recent studies have implemented MeRIP for the detection of RNA modifications and this is rapidly becoming a favourable technique in this field. It is versatile in that either RT-qPCR could be used to detect modification enrichments in candidate transcripts, or next-generation sequencing could be used to detect peaks transcriptome-wide corresponding to regions containing the modification (MeRIP-seq). Although this technique does not detect modifications at single nucleotide resolution, it would be a favourable validation tool, which could remove any ambiguity that bisulfite conversion may cause.

A single study thus far has implemented MeRIP-seq for the validation of m^{5}C sites detected by transcriptome-wide bisulfite sequencing in archaea [76]. As such, the same m^{5}C antibody was purchased for this study and a dot blot was initially performed to confirm that the antibody did detect m^{5}C in RNA. Yeast tRNA was used as a control due to its abundance in m^{5}C, and HeLa cell RNA was also used as the biological sample of interest. In addition to using RNA, methylated and unmethylated DNA controls were used and were purchased from the manufacturer of the antibody. These controls were double-stranded DNA of ~300 bp in length and were produced by PCR. The original DNA was derived from two different chromosomes of *A. thaliana* and the difference between the two controls was that the methylated DNA contained a single m^{5}C site and the cytosines in the unmethylated control were all unmethylated.

Two-fold serial dilutions of each DNA control were made from 100 ng-12.5 ng in 2-5 µL 0.1 M NaOH, denatured at 99°C for 5 min, placed on ice and neutralised with 6.6 M ammonium acetate (Figure 3.17). Two-fold serial dilutions of the RNA samples were made from 5 µg-625 ng. A higher amount of RNA was used due to the substantially lower amount of m^{5}C compared to DNA. The RNA was denatured at 75°C for 5 min

then placed on ice. Each sample was spotted onto a nylon membrane and dried until the spots disappeared to prevent quenching during the crosslinking procedure. The nucleic acids were UV crosslinked twice using the "auto-crosslink" function (120 mJ/cm^2). Following this, the membrane was incubated with the m^5C antibody in blocking buffer overnight at 4°C, washed in 1x PBS-T, incubated with an anti-mouse secondary antibody in blocking solution and washed again. The nucleic acids were detected using ECL.

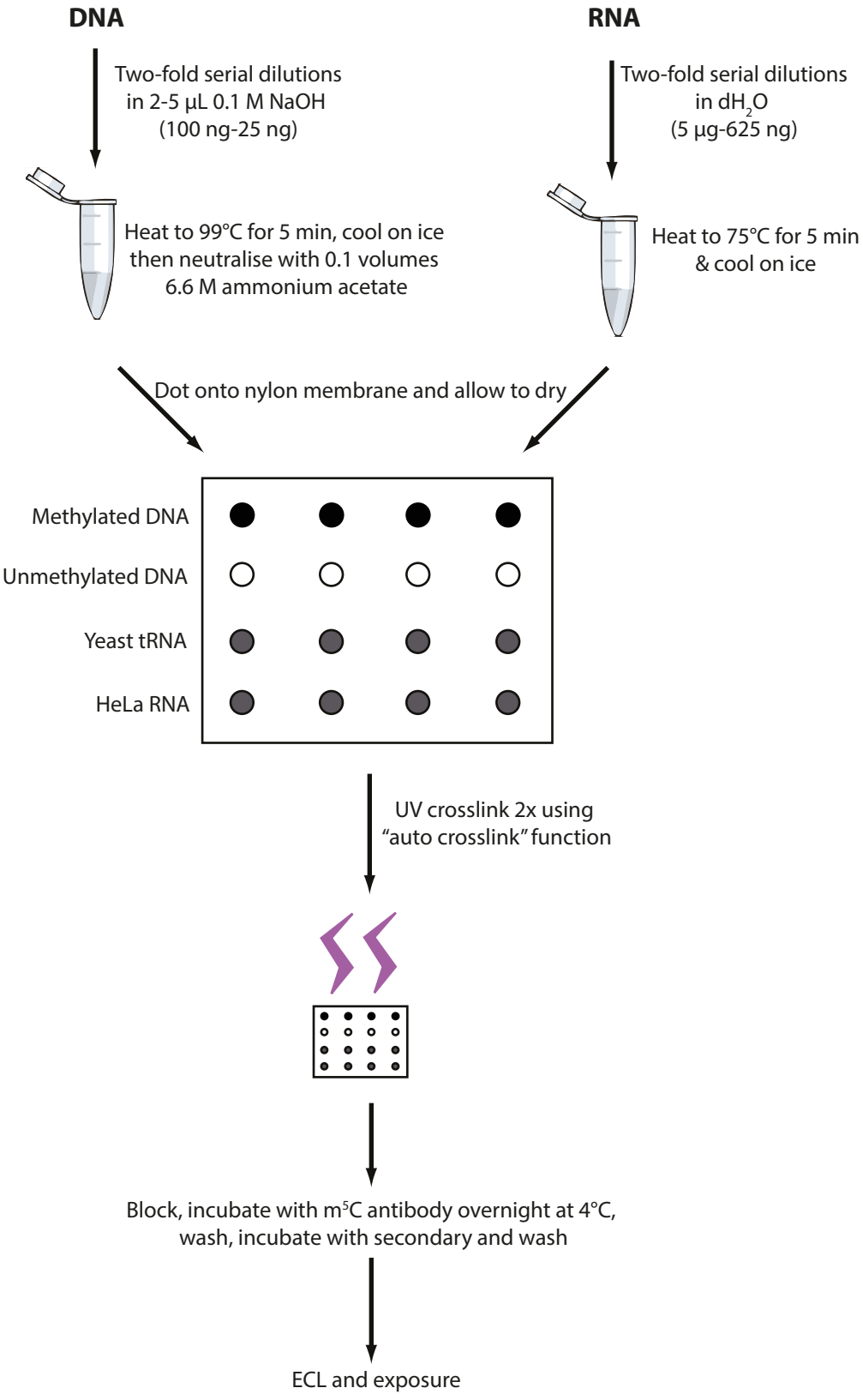

DNA
Two-fold serial dilutions
in 2-5 µL 0.1 M NaOH
(100 ng-25 ng)
Heat to 99°C for 5 min, cool on ice
then neutralise with 0.1 volumes
6.6 M ammonium acetate
RNA
Two-fold serial dilutions
in dH2O
(5 µg-625 ng)
Heat to 75°C for 5 min
& cool on ice
Dot onto nylon membrane and allow to dry
Methylated DNA
Unmethylated DNA
Yeast tRNA
HeLa RNA
UV crosslink 2x using
"auto crosslink" function
Block, incubate with m5C antibody overnight at 4°C,
wash, incubate with secondary and wash
ECL and exposure

Figure 3.17. Preparation of DNA and RNA dot blots for the detection of m^5C.
Two-fold serial dilutions of DNA (100ng-12.5 ng) and RNA (5 μg-625 ng) were made. DNA was denatured in 0.1 M NaOH at 99°C for 5 min, placed on ice then neutralised with 0.1 volumes of 6.6 M ammonium acetate. RNA was denatured at 75°C for 5 min then placing on ice. ~10 μL of the nucleic acids were spotted onto a positively charged nylon membrane and air-dried until spots were no longer visible. The membrane was UV crosslinked twice and blocked in blocking buffer for 1 h at room temperature. The membrane was incubated with the anti-m^5C antibody overnight at 4°C, washed with washing buffer, incubated with an anti-mouse secondary antibody and washed again. The membrane was exposed using ECL.

m^5C was detected at all concentrations in the methylated DNA control, and as expected no m^5C was detected in the unmethylated DNA control (Figure 3.18A). However, m^5C could not be detected in yeast tRNA or HeLa cell RNA at all concentrations. Staining of the membrane with methylene blue indicated successful hybridisation of all samples, suggesting that the antibody may have failed to detect m^5C in RNA due to inefficient denaturation (Figure 3.18B). It is possible that heat denaturation may not be sufficient to denature strong secondary structures, such as those in rRNAs and tRNAs, for the antibody to detect m^5C. To try and overcome this issue, the DNA and RNA samples were heat denatured in 2x RNA loading buffer containing formamide at 75°C for 5 min followed by snap cooling on ice. Formamide acts as both a stabiliser of RNA and denaturing agent by disrupting hydrogen bonds between nitrogen bases, preventing conformational changes. This, in combination with heat and cooling on ice, should aid in reducing strong secondary structures of RNA. As positive controls, the DNA controls were denatured in 0.1 M NaOH as described previously. As with the previous attempt, the antibody successfully detected m^5C in the methylated DNA control regardless of the denaturation process and despite reduced binding efficiency when denatured with formamide (Figure 3.19A). Despite the changes to the denaturation process, the antibody failed to detect m^5C in both RNA samples (Figure 3.19B).

Unfortunately the antibody could not successfully detect m^5C in RNA in this thesis, and therefore could not be used for MeRIP. The likely cause of the issue appeared to be inefficient denaturation of the RNA for the antibody to access m^5C sites.

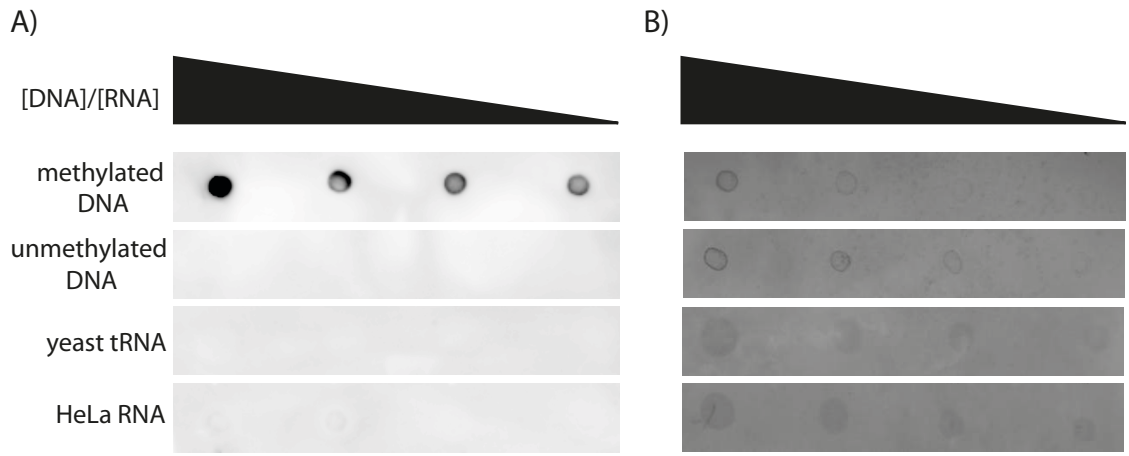


Figure 3.18. Dot blot assay for the detection of m^5C in DNA and RNA.

A) Dot blot to detect m^5C in DNA and RNA. Methylated and unmethylated DNA controls were denatured in 0.1 M NaOH at 99°C for 5 min, placed on ice and neutralised in 0.1 volumes of 6.6 M ammonium acetate. RNA was denatured at 75°C for 5 min then placed on ice. B) Staining of the membrane with 0.02% methylene blue indicated binding of the nucleic acids.

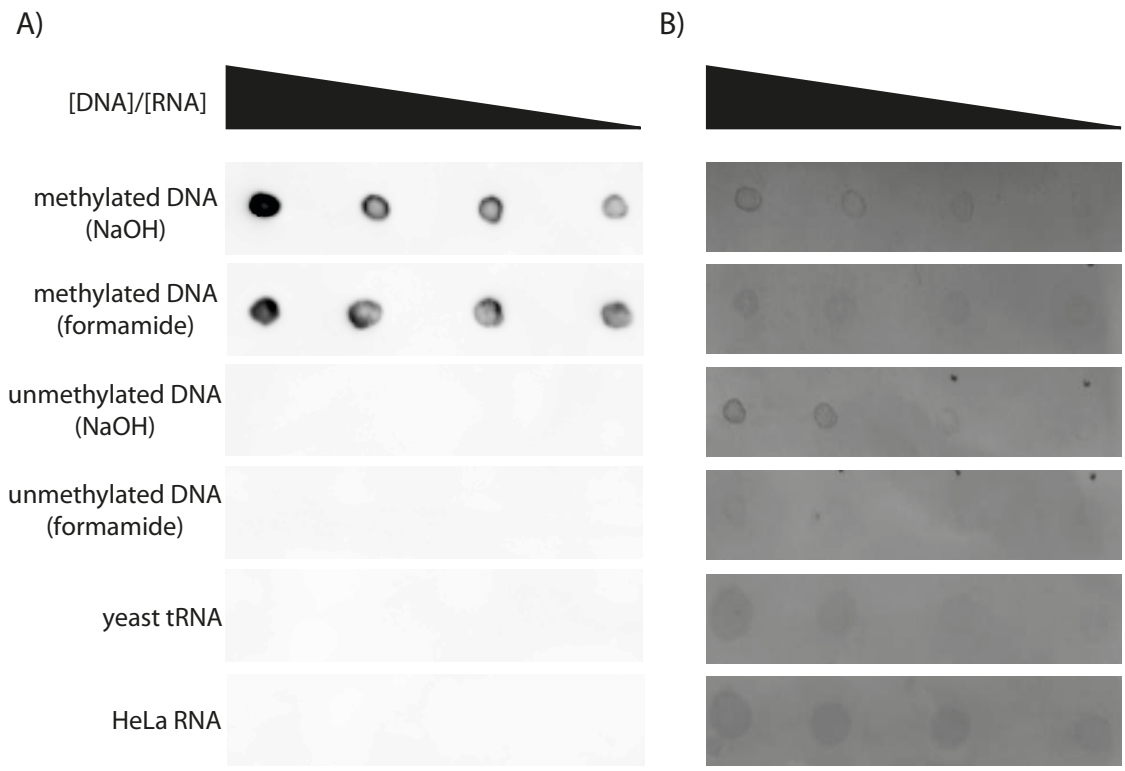


Figure 3.19. Dot blot assay for the detection of m^{5}C in DNA and RNA after denaturation in 2x RNA loading buffer.

A) Dot blot to detect m^{5}C in DNA and RNA. Methylated and unmethylated DNA controls were denatured either by heating with 0.1 M NaOH, as previously described, or heating to 75°C for 5 min in formamide then placing on ice. RNA was denatured in formamide at 75°C for 5 min. then placed on ice. B) Staining of the membrane with 0.02% methylene blue revealed binding of the nucleic acids.

3.2.5 Validation of m^{5}C sites using Illumina® MiSeq

The use of Sanger bisulfite sequencing for the validation of candidate m^{5}C sites is a laborious, expensive and low-throughput technique that requires sequencing of multiple clones. As a result, MiSeq amplicon sequencing was implemented for the validation of candidate m^{5}C sites in this study. This method enabled the high-throughput sequencing of multiple amplicons generated from bisulfite PCRs with high coverage, providing higher confidence in the validation of candidate m^{5}C sites. Furthermore, it allowed for multiplexing of samples so that they could be sequenced on a single run, making it a very cost-effective technique. Although this does rely on bisulfite conversion of RNA, it is significantly less laborious than Sanger bisulfite sequencing and implemented a different sequencing platform to that used for bsRNA-seq.

Unless stated otherwise, I performed all of the experimental work for the sequencing of the MiSeq libraries. Dr. Wenjia Qu from the Garvan Institute of Medical Research provided the protocol originally optimised for bisulfite-converted DNA. I optimised this protocol for use with bisulfite-converted RNA.

3.2.5.A Selecting candidate m^{5}C sites for validation

After determining candidate m^{5}C sites based on the mapping by B-SOLANA, a selection was made for validation. Candidates that were selected for validation had reasonable coverage across the transcript and ≥10% mean methylation across the control samples. To ensure selection of high-confidence candidates, sequencing reads spanning the candidate m^{5}C site were also inspected in Integrated Genomics Viewer (IGV) to ensure they were of high quality. Figure 3.20 displays an IGV snapshot of sequencing reads spanning the known m^{5}C site at C34 in tRNA$^{Leu(CAA)}$ in the HeLa NTC sample from the local mapping of the bsRNA-seq data. High confidence candidate m^{5}C sites exhibit high quality base calling and are present in multiple sequencing reads with different alignment start and end points (i.e. no clonal reads). PCR amplification during library preparation may cause clonal read biases, particularly for low abundant transcripts. This may result in the incorrect detection of candidate m^{5}C sites and therefore it was necessary to remove candidates that exhibited this.

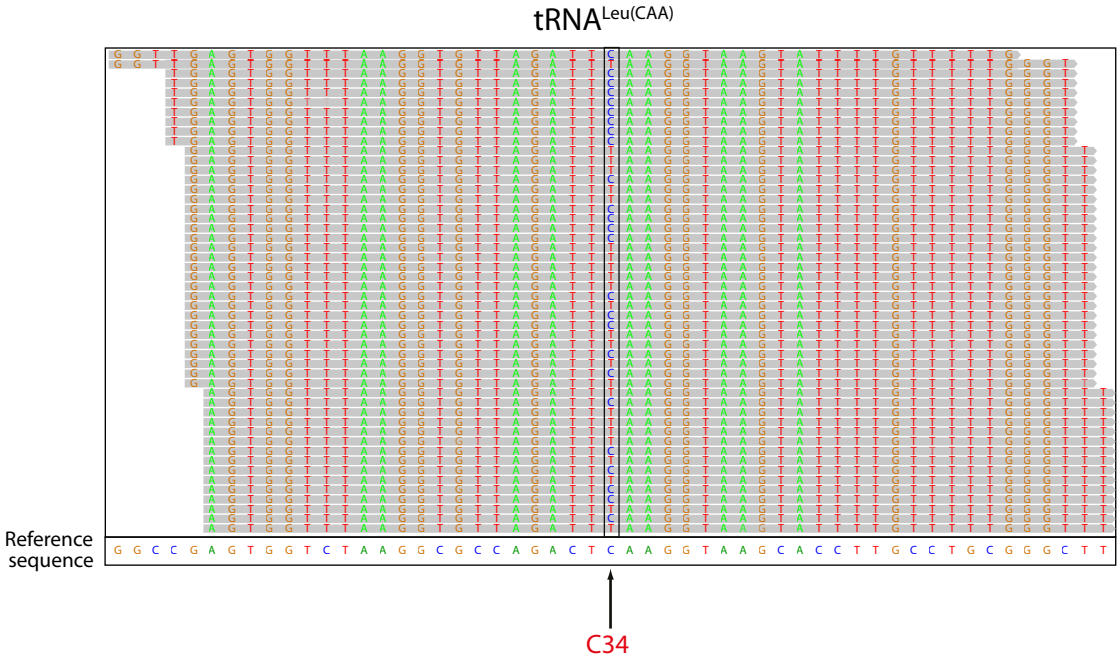
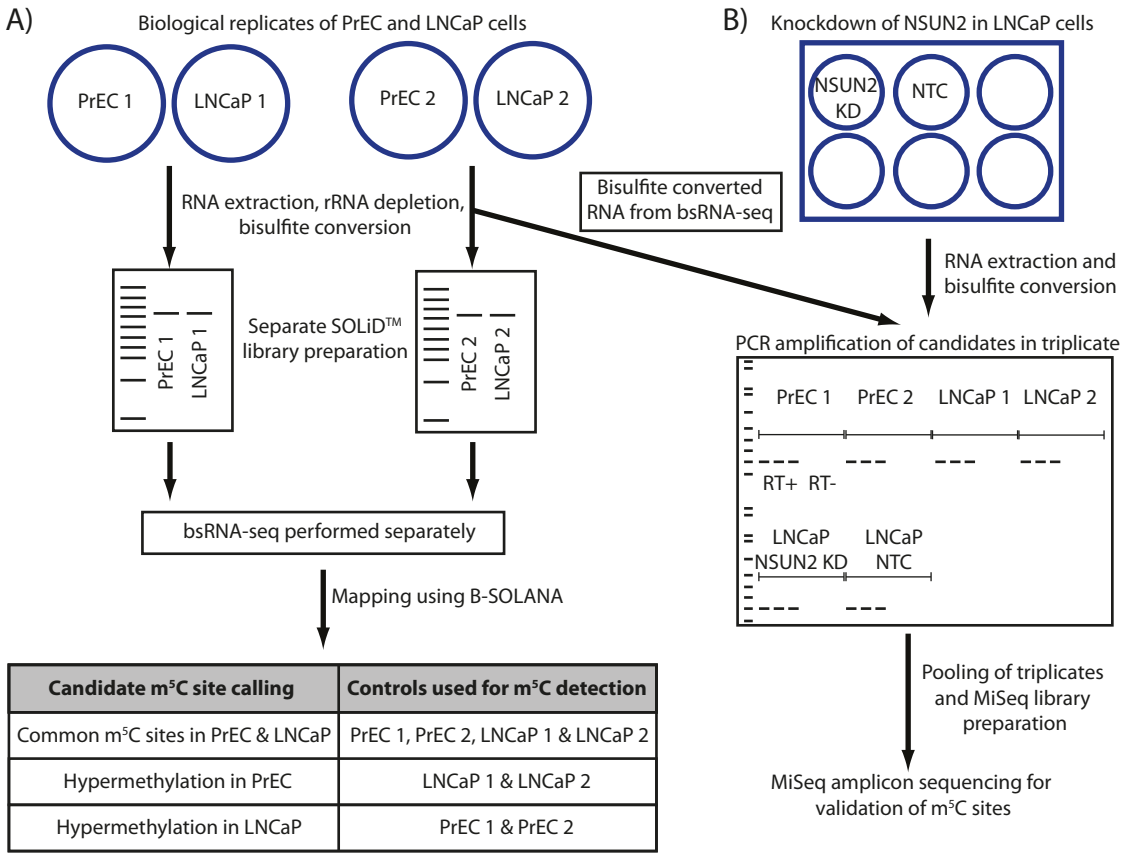


Figure 3.20. Representative snapshot from IGV depicting a subset of sequencing reads mapping to tRNA$^{Leu(CAA)}$.

The grey bars represent 50 nt sequencing reads mapped to part of the tRNA$^{Leu(CAA)}$ locus in the HeLa NTC sample based on the local mapping (chr1:249,168,062-249,168,114, + strand). Unconverted cytosines represent m^{5}C sites at C34. High quality detection of the m^{5}C site by bsRNA-seq is indicated by an opaque base. Low quality bases appear as transparent. Many sequencing reads containing the m^{5}C site exhibit different alignment start points, indicating high confidence m^{5}C sites.

3.2.5.B Preparation of libraries for MiSeq amplicon sequencing

As two replicates per cell line were sequenced for the prostate dataset, the bisulfite-converted RNA used for bsRNA-seq was also used for MiSeq amplicon sequencing. In addition, *NSUN2* was knocked down in LNCaP cells, which will be addressed in more detail in Chapter 6. Figure 3.21 shows a schematic of the PrEC and LNCaP sample preparation process for bsRNA-seq and MiSeq amplicon sequencing.



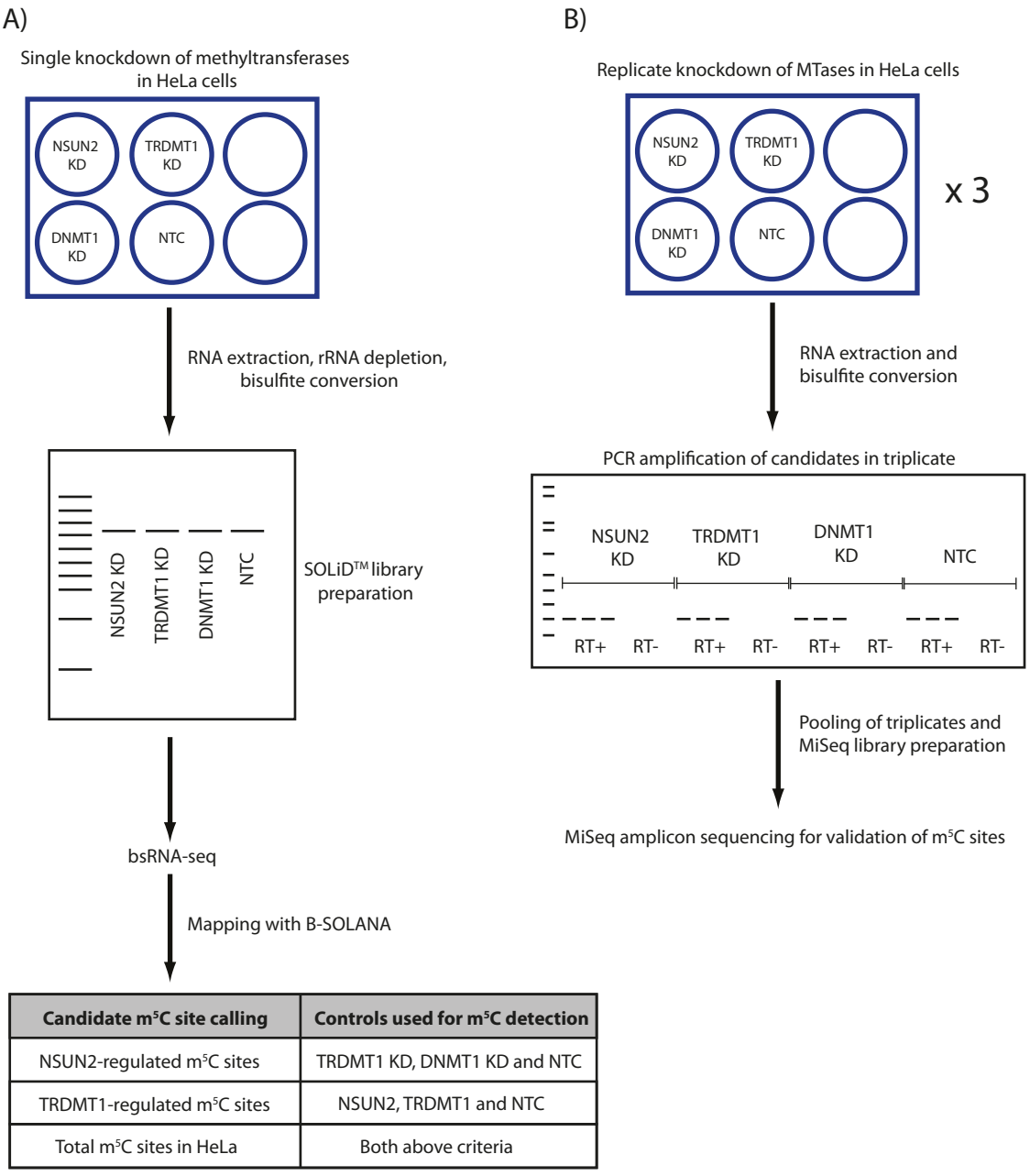
Candidate m^5C site calling	Controls used for m^5C detection
Common m^5C sites in PrEC & LNCaP	PrEC 1, PrEC 2, LNCaP 1 & LNCaP 2
Hypermethylation in PrEC	LNCaP 1 & LNCaP 2
Hypermethylation in LNCaP	PrEC 1 & PrEC 2

Figure 3.21. Preparation of PrEC and LNCaP cells for bsRNA-seq and MiSeq amplicon sequencing.

A) Each of the two replicates for PrEC and LNCaP cells were prepared for bsRNA-seq separately, including growth, RNA extraction, bisulfite conversion, library preparation and sequencing. The bioinformatics pipeline implemented was performed on both replicates simultaneously. B) The same replicates were used for MiSeq amplicon sequencing, however a new batch of LNCaP cells was transfected with *NSUN2* siRNA and NTC siRNA. The RNA from this was bisulfite converted and subjected to PCR amplification of candidates in triplicate alongside the PrEC and LNCaP replicates. MiSeq libraries were prepared and subjected to MiSeq amplicon sequencing.

As only a single sample of each HeLa methyltransferase knockdown sample was used for bsRNA-seq, replicate knockdowns of the methyltransferases were performed in triplicate (Figure 3.22). A single sample for each replicate knockdown was used for the validation of candidate m^5C sites by MiSeq amplicon sequencing. The knockdown of each methyltransferase was confirmed by RT-qPCR, western blotting and Sanger bisulfite sequencing of $tRNA^{Asp(GUC)}$ and $tRNA^{Leu(CAA)}$ (Figure 3.23). Specific antibodies were successfully obtained prior to the replicate methyltransferase knockdown and confirmed protein depletion (Figure 3.23B). The western blot that confirmed protein knockdown was performed by Mr. Andrew Shafik. Interestingly, although only ~40% of *DNMT1* mRNA was knocked down, almost complete depletion of DNMT1 protein was evident. This most likely is a result of the timing of the experiment, where depletion of mRNA occurred initially, followed by depletion of protein. It is possible that the time point selected for the harvesting of cells was the point at which mRNA levels were rising again, but protein levels were still relatively low.

As with the first samples used for bsRNA-seq, Sanger bisulfite sequencing was performed to confirm depletion of m^5C at known NSUN2 and TRDMT1 target sites in the biological replicates prior to MiSeq amplicon sequencing. 75% depletion of m^5C was evident at C38 in $tRNA^{Asp(GUC)}$ in the TRDMT1 knockdown sample compared to the NTC sample (Figure 3.23C). In the NSUN2 knockdown sample, 30% depletion of m^5C at C48 and 20% depletion of m^5C at C49 in $tRNA^{Asp(GUC)}$ was evident compared to the NTC sample. >50% depletion of m^5C was evident at C34 in $tRNA^{Leu(CAA)}$ in the NSUN2 knockdown sample compared to the NTC sample (Figure 3.23D).



Candidate m^{5}C site calling	Controls used for m^{5}C detection
NSUN2-regulated m^{5}C sites	TRDMT1 KD, DNMT1 KD and NTC
TRDMT1-regulated m^{5}C sites	NSUN2, TRDMT1 and NTC
Total m^{5}C sites in HeLa	Both above criteria

Figure 3.22. Preparation of RNAi-mediated knockdown of *NSUN2, TRDMT1* and *DNMT1* in HeLa cells for bsRNA-seq and MiSeq amplicon sequencing.

A) RNAi-mediated knockdown of *NSUN2*, *TRDMT1*, *DNMT1* and transfection of a NTC siRNA was performed. RNA was extracted, depleted of rRNA and bisulfite converted. SOLiD™ library preparation and bsRNA-seq were performed simultaneously for all samples. B) Replicate RNAi-mediated knockdown of the methyltransferases was performed in triplicate and RNA was extracted and bisulfite converted for one sample of each knockdown and control. PCR amplification of candidates was performed in triplicate and MiSeq libraries were prepared and subjected to MiSeq amplicon sequencing.

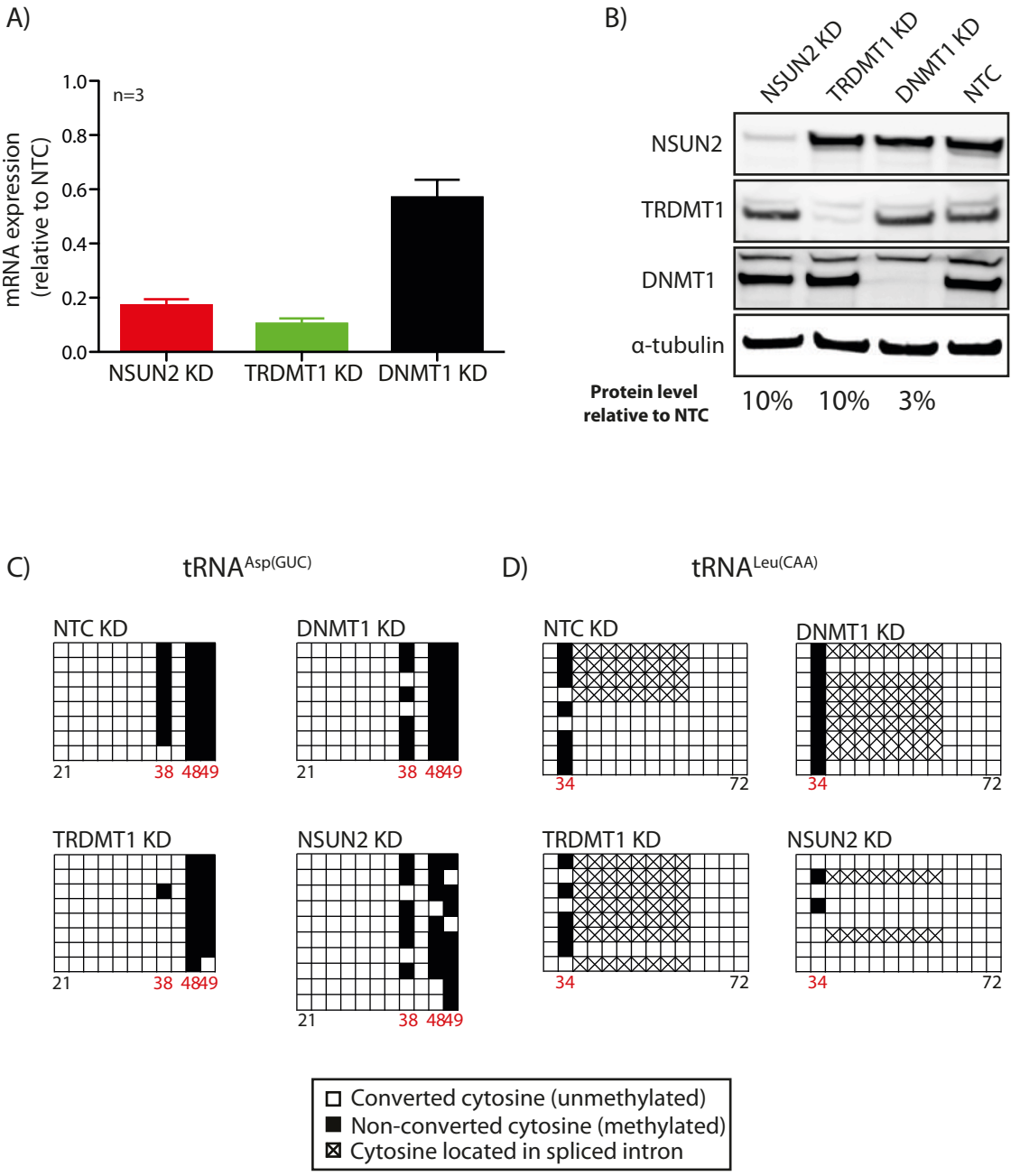


Figure 3.23. Confirmation of knockdown of *NSUN2*, *TRDMT1* and *DNMT1* in HeLa cells as biological replicates.

A) RT-qPCR confirmed knockdown of each methyltransferase on the mRNA level. The results were normalised to *HPRT* mRNA levels (n=3, error=SEM), B) Western blotting confirmed knockdown of each methyltransferase on the protein level. The amount of protein relative to the NTC sample after knockdown based on densitometry is indicated below the blot, C) Sanger bisulfite sequencing of tRNA$^{Asp(GUC)}$ showed depletion of m^{5}C at C38 in the TRDMT1 knockdown sample and C48 and C49 in the NSUN2 knockdown sample, D) Sanger bisulfite sequencing of tRNA$^{Leu(CAA)}$ showed depletion of m^{5}C at C34 in the NSUN2 knockdown sample.

Prior to MiSeq library preparation, 200 ng bisulfite-converted RNA was reverse transcribed using the SuperScript® III Reverse Transcriptase Kit. Bisulfite primers to amplify candidates were designed as described in section 3.2.1.C, and PCR amplification was performed in triplicate using Platinum® Taq DNA Polymerase in order to minimise PCR amplification biases. To confirm successful amplification of PCR products, 2.5 μL of each amplicon (all triplicates) was electrophoresed on a 2% agarose gel at 100 V for 40 min. Once confirmed, triplicate amplicons were pooled and purified using the Wizard SV Gel and PCR Clean-Up System according to the manufacturer's protocol and eluted in 15-30 μL dH_2O. The concentration of each amplicon was determined using the Qubit dsDNA BR Assay Kit according to the manufacturer's protocol and a minimum of 20 ng of each amplicon was required for successful sequencing on the MiSeq platform. If not enough was present, PCRs were repeated. For each sample, 20-30 ng of each amplicon was pooled into a single 1.5 mL tube and concentrated down to 55 μL.

To prepare amplicon sequencing libraries, the TruSeq DNA LT Sample Prep Kit v2 was used with some modifications. 50 μL of pooled amplicons were end-repaired and purified by AMPure XP beads for each sample (amplicons >100 nt). Initially, AMPure XP beads where used for the purification of all amplicons, which resulted in the loss of amplicons <100 nt. As a result, amplicons <100 nt were purified by the MinElute PCR Purification Kit. Following purification, the amplicons were subjected to A-tailing and adaptor ligation, of which indexed adaptors were used to allow multiplexing. All DNA fragments were then purified twice with AMPure XP beads followed by PCR to enrich the DNA fragments according to the manufacturer's protocol. PCR-enriched DNA fragments were purified once more with AMPure XP beads. For validation of the libraries and determining the average library size, 1 μL each library and 2 μL pooled amplicons prior to library preparation were electrophoresed on a 2% agarose gel at 100 V for 40 min (Figure 3.24). An increase in the length of the DNA fragments after library preparation was evident as a result of the ligated adaptors.

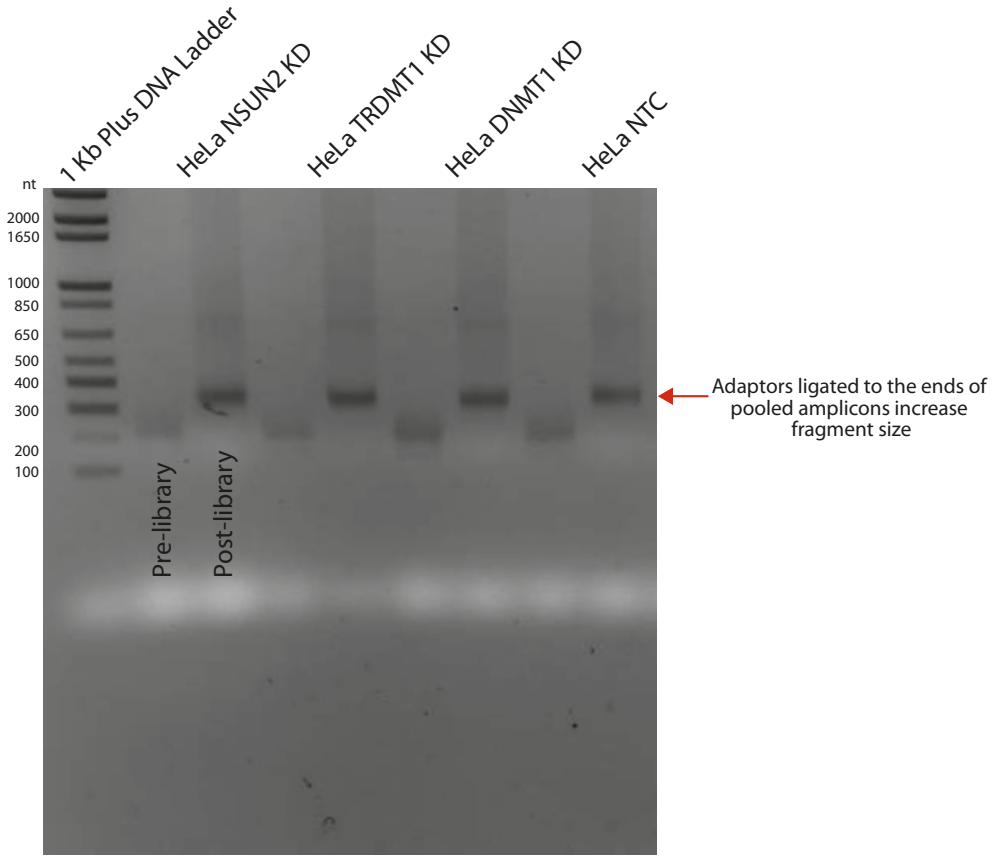


Figure 3.24. Validation of MiSeq amplicon libraries by agarose gel electrophoresis.
Pooled amplicons and final libraries for each sample were electrophoresed on a 2% agarose gel at 100 V for 40 min. In this case, libraries contained amplicons from transcripts that averaged ~200 nt in size prior to library preparation. After library preparation, the pooled amplicons were ~320 nt. As expected, the DNA fragments were more abundant after library preparation. The sizes of the 1 Kb Plus DNA Ladder are indicated.

3.2.5.C QUANTIFICATION AND LOADING OF MISEQ LIBRARIES

Importance of library concentration for cluster generation

Several methods were used to quantify the MiSeq libraries to determine the optimal concentration for the sequencing run. Prior to sequencing, cluster generation occurs on the flow cell. Clusters are small regions of amplified DNA with each containing a different amplified fragment of DNA. This enables enough DNA to be detected during sequencing. Cluster generation involves several major steps. First, the DNA fragments are hybridised to the flow cell and are amplified in a process known as “bridge amplification”. Second, the fragment is then linearised to prepare it for sequencing and blocked to prevent addition of nucleotides during the sequencing process. Third, the

sequencing primers are hybridised to the DNA fragments. The optimal cluster density for a sequencing run is 700 K-1000 K/mm^2, and this is highly dependent on the amount of library that is loaded into the cartridge. Loading too little DNA for the sequencing run can result in too few clusters, which can potentially make sequencing uneconomical. Overloading DNA results in higher cluster densities, leading to cluster overlap and a reduction in the amount of data obtained.

Library quantification methods

After library preparation and validation, the libraries were quantified using the Qubit dsDNA BR Assay Kit according to the manufacturer's protocol and diluted to ~50 nM based on the estimated library size. Either qPCR, a High Sensitivity DNA Chip or the Qubit dsDNA BR Assay Kit was used to quantify each library necessary for the sequencing run. For quantification by qPCR, the Illumina® Library Quantification Kit was used according to the manufacturer's protocol with a few modifications. This kit specifically amplifies DNA containing the Illumina® adaptor sequence. Two-fold serial dilutions of the libraries were made from 1:1,000 to 1:32,000 in library dilution buffer. The manufacturer's protocol recommended diluting to 1:8,000, but it was found that this dilution was not low enough to be in the range of the standard curve, therefore not allowing accurate quantification of the libraries. The concentration of the libraries based on qPCR was determined by averaging the results for the dilutions that were within the range of the standard curve. Quantification of the libraries using qPCR resulted in higher concentrations compared to the Qubit dsDNA BR Assay Kit.

In one instance, qPCR, a High Sensitivity DNA Chip and Qubit dsDNA BR Assay Kit were used for the quantification of the same libraries as a comparison. Interestingly, the Agilent Bioanalyzer yielded the lowest concentration, followed by the Qubit dsDNA BR Assay Kit and then qPCR (data not shown). In fact, all three quantification methods yielded vastly different results. Multiple MiSeq runs were performed for this study, with different quantification methods used to determine the amount to load into the cartridge. The libraries were equally pooled to 10 nM and diluted further to 2 nM at which point they were denatured using 0.2 M NaOH. The libraries were then diluted further to 7-9 pM prior to loading into the cartridge. In all instances, 150 bp paired-end sequencing was performed.

Spike-in control for sequencing

For each MiSeq sequencing run, a PhiX library was spiked into the final pooled libraries to a final proportion of 10%. The PhiX library is an adaptor-ligated library derived from the PhiX genome and is used as a control for sequencing runs. In this case it behaved as a troubleshooting control for cluster generation, enabling us to determine if any issues that arose were related to the sample preparation.

3.2.6 General features and mapping of the MiSeq libraries

3.2.6.A Mapping of MiSeq amplicon sequencing data

Prior to mapping, the Illumina® adaptors and poor quality reads were removed *in silico* using Trimmomatic in palindromic mode, allowing removal of 5' and 3' adaptors simultaneously. Once the adaptors were trimmed, the sequencing reads were aligned by Bismark, using Bowtie2 internally [249] (Figure 3.25). Bismark is an alignment algorithm designed for bisulfite-converted DNA sequencing data, performing both the alignment of single- and paired-end reads of directional and non-directional libraries, as well as the detection of m^{5}C sites. Initially, the sequencing reads were *in silico* converted to C-to-T and G-to-A versions, which were then aligned to equivalently converted reference sequences of the candidates interrogated. This alignment was performed in order to determine the single best location the sequencing reads mapped to; if this was ambiguous, the read was discarded. Only a single mismatch was allowed when aligning the reads. Following this, the proportion of methylation was determined by comparing the sequencing reads with the corresponding reference sequence, and the number of C and T calls at all C positions were extracted from the aligned reads. This allowed detection of m^{5}C sites in all sequence contexts.

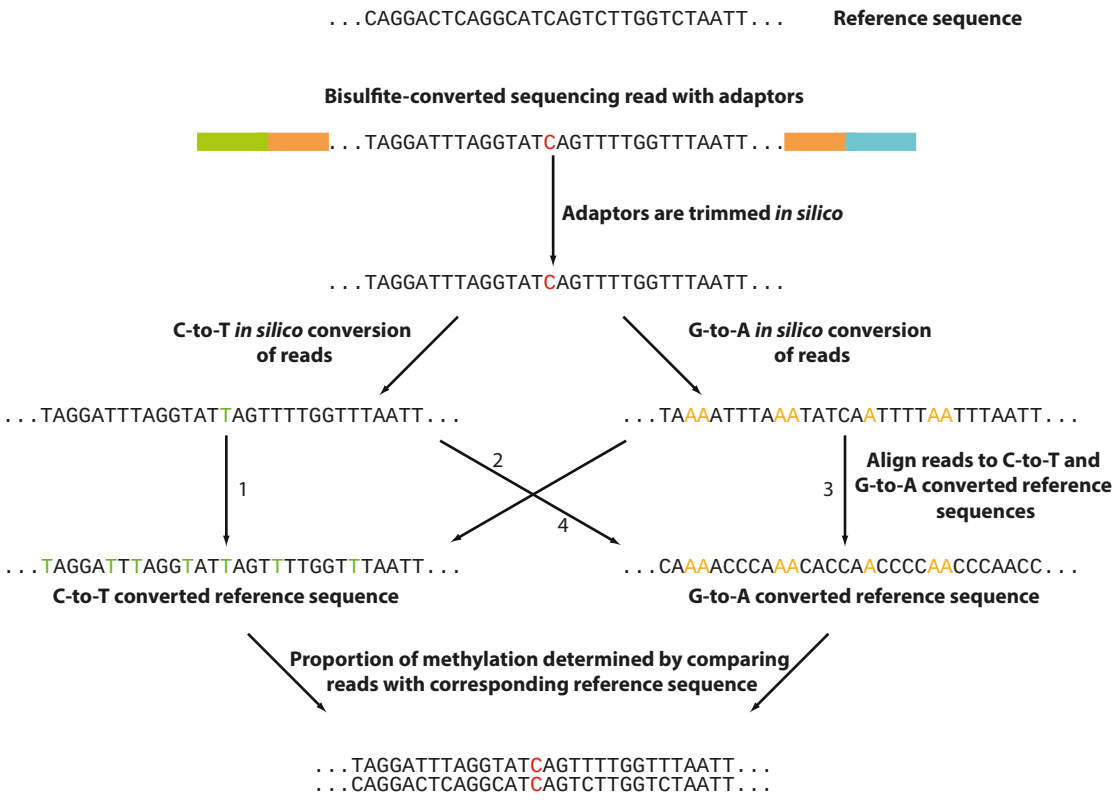


Figure 3.25. Schematic of the alignment of 150 bp paired-end reads from Illumina® MiSeq amplicon sequencing using Bismark.

Illumina® adaptors were trimmed *in silico* in palindromic mode using Trimmomatic. Sequencing reads were converted to C-to-T and G-to-A versions and were mapped to equivalent conversions of the reference sequence by Bismark to determine the single best location the sequencing reads mapped to. In this example, the best alignment is from alignment 1. Finally, to determine the proportion of m^{5}C, sequencing reads prior to the *in silico* conversions were compared to the original reference sequence. The red base represents the m^{5}C site, green bases represent C-to-T converted bases and orange bases represent G-to-A converted bases.

3.2.6.B General features of the MiSeq amplicon libraries

In this thesis, multiple MiSeq runs were performed for the validation of candidate m^{5}C sites. The initial MiSeq run was performed in Professor Susan Clark's lab at the Garvan Institute of Medical Research. I performed all experimental work including the preparation of the sequencing run. Library preparations for subsequent MiSeq sequencing runs were performed by myself, but the final dilution, spike in of the PhiX control and preparation of the sequencing runs were performed by Ms. Angela Higgins at the BRF at ANU.

Table 3.3 provides the general features of the libraries sequenced on the Illumina® MiSeq platform for this thesis, including the quantification method used to load the libraries, concentration of the library loaded for sequencing, cluster density, the proportion of reads with a quality score (Q)≥30 and the average proportion of mappable sequencing reads. The cluster density and proportion of Q≥30 are determined directly by the MiSeq platform and provide an indication of the quality of the sequencing run. The quality score is the probability of a base being incorrectly called and a Q≥30 indicates a 99.9% accuracy in base calling.

The quantification method used for the first MiSeq run performed at the Garvan Institute of Medical Research was based on qPCR and yielded a cluster density in the optimal range and over 74% mappable reads. For the following MiSeq sequencing run performed at the BRF, the library was quantified by a High Sensitivity DNA Chip. This provided a significantly lower concentration compared to the qPCR and yielded a higher cluster density and lower percentage of mappable reads. The subsequent sequencing runs were prepared based on the qPCR, which led to relatively optimal cluster densities and a high proportion of mappable reads.

Table 3.3. General features of the MiSeq amplicon libraries.

MiSeq run	Final quantification method	Concentration loaded for sequencing (pM)	Cluster density (K/mm^2)	% ≥Q30	% average mappable reads
1	qPCR	9	700	88.0	74.5
2	Bioanalyzer	8	1290	79.3	44.5
3	qPCR	7	596	97.2	88.7
4	qPCR	8	699	95.2	81.8
5	qPCR	9	734	97.0	88.8

Validation of candidate m^5C sites in RNA

Using this technique, a number of candidate m^5C sites detected from the bsRNA-seq were validated. For the HeLa knockdown samples, 8/9 m^5C sites in tRNAs, 5/5 m^5C sites in other ncRNAs and 9/10 m^5C sites in mRNAs were validated. This will be discussed in more detail in Chapter 4. For the PrEC and LNCaP samples, 7/7 m^5C sites in tRNAs, 2/2 m^5C sites in other ncRNAs and 7/10 m^5C sites in mRNAs were validated. This will be discussed in more detail in Chapter 6.

3.2.7 Bisulfite conversion and knockdown efficiency based on bsRNA-seq and MiSeq

In this section, I will describe the bisulfite conversion efficiency of all samples and the knockdown efficiency of the HeLa methyltransferase knockdown samples based on bsRNA-seq and MiSeq amplicon sequencing. The heatmaps I generated in this section and the remainder of this thesis using the R programming language demonstrate the percentage of non-conversion of cytosines within the segment of the transcript interrogated by bsRNA-seq or MiSeq amplicon sequencing. That is, the percentage of cytosines that were not converted to uracil, which may be attributed to m^5C sites or inefficient bisulfite conversion. However, the percentage of non-conversion at a known m^5C site or validated m^5C site will be described in terms of methylation percentage.

3.2.7.A R-Luc *in vitro* transcripts

Although the bisulfite conversion efficiency was assessed by Sanger bisulfite sequencing prior to library preparation, this was only examined for a small portion of the R-Luc *in vitro* transcripts. As such, the entire transcripts were examined from the bsRNA-seq for both the HeLa methyltransferase knockdown samples and PrEC and LNCaP samples. As expected, both *in vitro* transcripts exhibited high coverage and very little non-conversion (Figures 3.26 and 3.27). The mean cytosine conversion was calculated for each *in vitro* transcript, indicating >99.8% conversion in all samples. This reiterates that the bisulfite conversion efficiency was high for all samples (Table 3.4).

A small portion of the non-humanised R-Luc *in vitro* transcript was amplified and sequenced using MiSeq amplicon sequencing for validation, also revealing efficient

bisulfite conversion of the biological replicate HeLa methyltransferase knockdown samples and PrEC and LNCaP samples (Figures 3.26A and 3.27A; bottom panel). Examples of the typical read coverage throughout the amplicon and between samples from MiSeq amplicon sequencing are indicated for the R-Luc *in vitro* transcripts in Figures 3.26A and 3.27A. The coverage was relatively even throughout the amplicon and similar coverage was evident between samples, indicating even pooling. The range of coverage for subsequent MiSeq amplicon sequencing data will be indicated in the figure legends. In many cases, the range of coverage indicated in the following chapters may be quite large. This is chiefly due to variations in the concentrations for each MiSeq sequencing run that was performed.

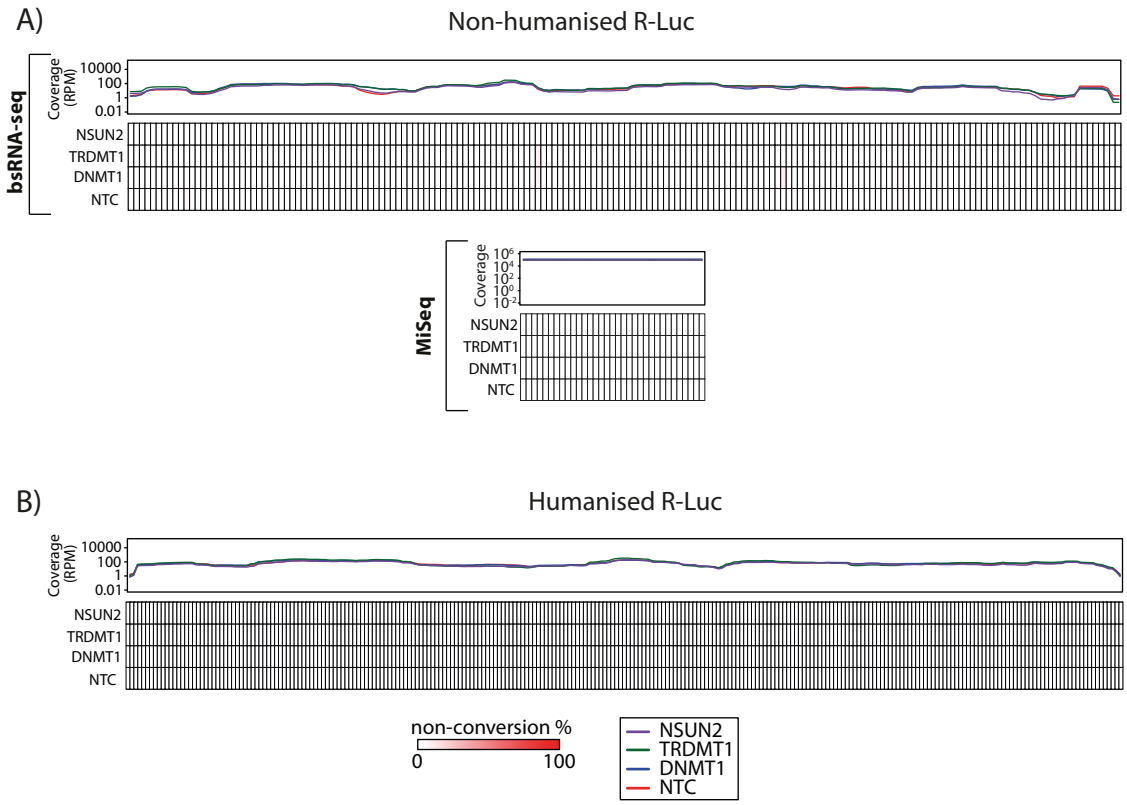


Figure 3.26. Bisulfite conversion efficiency of the R-Luc *in vitro* transcripts assessed by bsRNA-seq and MiSeq amplicon sequencing of the HeLa methyltransferase knockdown samples.

A) Non-humanised R-Luc *in vitro* transcript, B) Humanised R-Luc *in vitro* transcript. The top two panels show the coverage in RPM and the proportion of non-conversion based on the bsRNA-seq. The bottom two panels in A) show the coverage of a segment of the R-Luc *in vitro* transcript and proportion of non-conversion based on MiSeq amplicon sequencing.

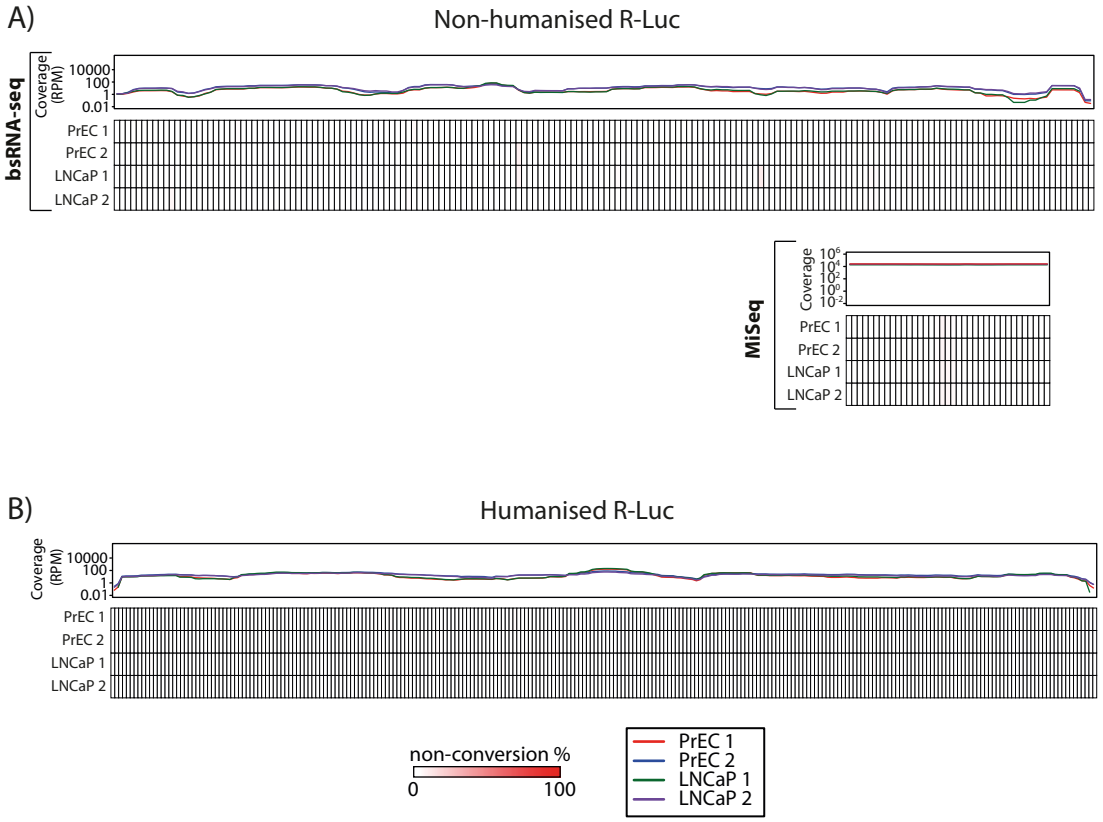


Figure 3.27. Bisulfite conversion efficiency of the R-Luc *in vitro* transcripts assessed by bsRNA-seq and MiSeq amplicon sequencing of PrEC and LNCaP cells.

A) Non-humanised R-Luc *in vitro* transcript, B) Humanised R-Luc *in vitro* transcript. The top two panels show the coverage in RPM and the proportion of non-conversion based on the bsRNA-seq. The bottom two panels in A) show the coverage of a segment of the R-Luc *in vitro* transcript and proportion of non-conversion based on MiSeq amplicon sequencing.

Table 3.4. Mean cytosine conversion of the R-Luc *in vitro* transcripts based on bsRNA-seq.

Sample	Mean conversion percentage (%)	
	Humanised R-Luc *in vitro* transcript	Non-humanised R-Luc *in vitro* transcript
NSUN2	99.87	99.87
TRDMT1	99.89	99.87
DNMT1	99.89	99.86
NTC	99.87	99.86
PrEC 1	99.87	99.90
PrEC 2	99.90	99.86
LNCaP 1	99.90	99.84
LNCaP 2	99.92	99.88

3.2.7.B Known m^5C sites in $tRNA^{Asp(GUC)}$ and $tRNA^{Leu(CAA)}$

The known m^5C sites of $tRNA^{Asp(GUC)}$ and $tRNA^{Leu(CAA)}$ were also examined in all samples using bsRNA-seq, and validated using MiSeq amplicon sequencing. Both tRNAs exhibited high coverage from the bsRNA-seq, which declined towards the 3' end of the tRNAs (Figures 3.28 and 3.29; top panels). Strong depletion of m^5C was observed at C48 and C49 of $tRNA^{Asp(GUC)}$ and C34 of $tRNA^{Leu(CAA)}$ in the NSUN2 knockdown sample according to the bsRNA-seq (Figure 3.28; middle panels). The depletion of m^5C at C48 and C49 in $tRNA^{Asp(GUC)}$ in the replicate NSUN2 knockdown was more modest, although strong depletion of methylation was observed at C34 of $tRNA^{Leu(CAA)}$ (Figure 3.28; bottom panels). Depletion of m^5C at C38 of $tRNA^{Asp(GUC)}$ in both TRDMT1 knockdown samples was also evident. High levels of methylation were observed in the DNMT1 knockdown and NTC samples, indicating that m^5C was not converted to uracil. While Sanger bisulfite sequencing provided an indication of enzyme knockdown efficiency, the high-throughput nature of bsRNA-seq and MiSeq amplicon sequencing provided higher confidence that NSUN2 and TRDMT1 were knocked down.

Both replicates of the PrEC and LNCaP samples showed high coverage of $tRNA^{Asp(GUC)}$ and $tRNA^{Leu(CAA)}$ through the majority of the transcripts and high methylation levels at all known m^5C sites (Figure 3.29). As with the HeLa methyltransferase knockdown samples, this indicated that conversion of the m^5C sites to uracil did not occur.

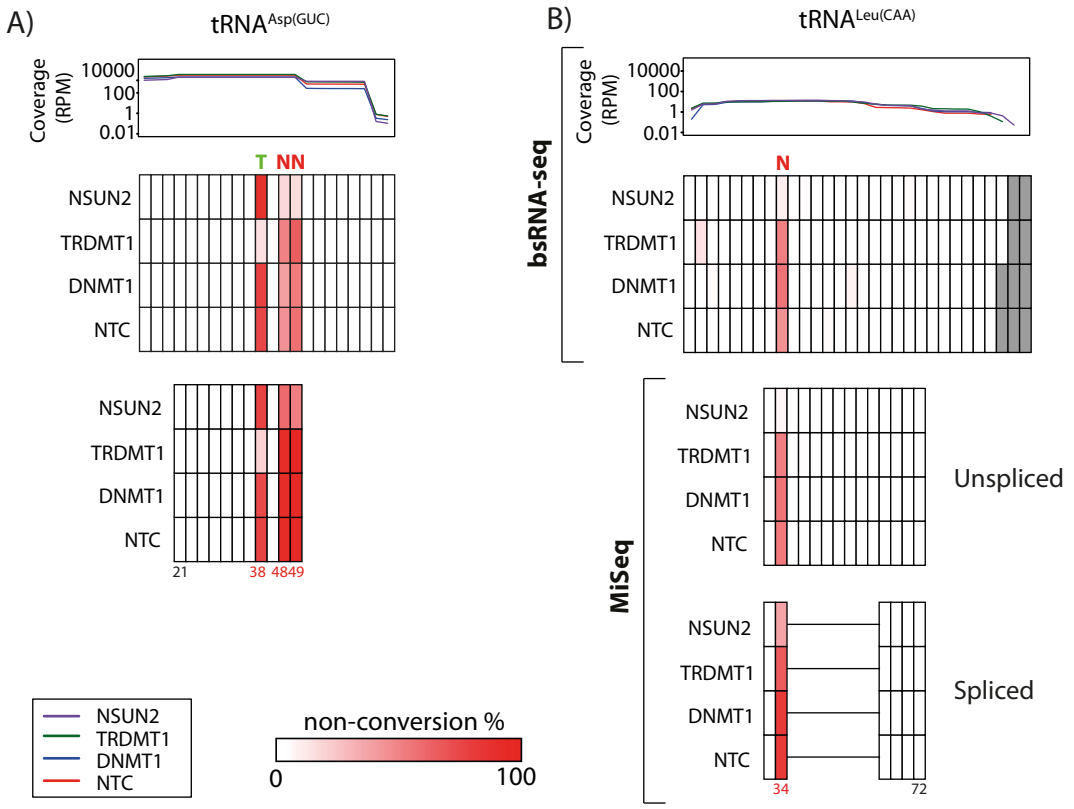


Figure 3.28. bsRNA-seq and MiSeq amplicon sequencing of tRNA$^{Asp(GUC)}$ and tRNA$^{Leu(CAA)}$ in the HeLa methyltransferase knockdown samples.

The top panel displays the coverage throughout the transcript in RPM based on bsRNA-seq, the middle panel displays the proportion of non-conversion based on bsRNA-seq, and the bottom panel displays the proportion of non-conversion in a segment of the transcript based on MiSeq amplicon sequencing. The bsRNA-seq data is based on the local mapping of tRNAs. A) tRNA$^{Asp(GUC)}$, B) tRNA$^{Leu(CAA)}$. Sequencing reads from MiSeq amplicon sequencing were mapped to spliced and unspliced tRNA$^{Leu(CAA)}$. The shades of grey in the middle panel indicate regions of the transcript where coverage was <5 reads based on bsRNA-seq. The line in the heatmap of the spliced tRNA$^{Leu(CAA)}$ indicates the region spliced. The tRNA structural positions with m^{5}C sites are indicated in red below the heatmap. N and T indicate m^{5}C sites mediated by NSUN2 and TRDMT1, respectively. The coverage from MiSeq amplicon sequencing ranged from 600-14,000 mappable sequencing reads.

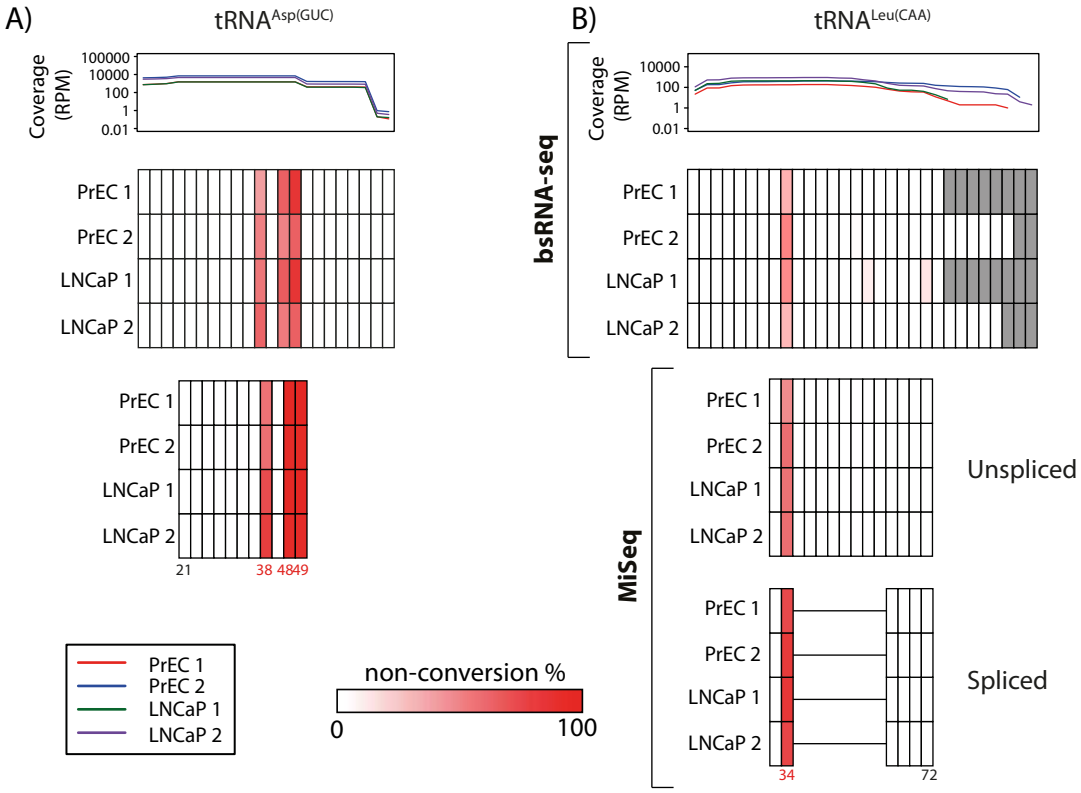


Figure 3.29. bsRNA-seq and MiSeq amplicon sequencing of tRNA$^{Asp(GUC)}$ and tRNA$^{Leu(CAA)}$ in PrEC and LNCaP cells.

The top panel displays the coverage throughout the transcript in RPM based on bsRNA-seq, the middle panel displays the proportion of non-conversion based on bsRNA-seq, and the bottom panel displays the proportion of non-conversion in a segment of the transcript based on MiSeq amplicon sequencing. The bsRNA-seq data is based on the local mapping of tRNAs. A) tRNA$^{Asp(GUC)}$. B) tRNA$^{Leu(CAA)}$. Sequencing reads from MiSeq amplicon sequencing were mapped to spliced and unspliced tRNA$^{Leu(CAA)}$. The shades of grey in the middle panel indicate regions of the transcript where coverage was <5 reads based on bsRNA-seq. The line in the heatmap of the spliced tRNA$^{Leu(CAA)}$ indicates the region spliced. The tRNA structural positions with the m^{5}C sites are indicated in red below the heatmap. The coverage from MiSeq amplicon sequencing ranged from 1,000-75,000 mappable sequencing reads.

3.3 Discussion

In order to examine the RNA m^5C profiles in several cell lines and the substrate range of NSUN2 and TRDMT1, it was necessary to improve and establish methods to detect m^5C in RNA. In addition, it was essential to prepare cells for bsRNA-seq and establish a knockdown model of NSUN2 and TRDMT1. Bisulfite conversion of RNA coupled with next-generation sequencing, termed bsRNA-seq, and a subsequent bioinformatics pipeline were implemented for the detection of candidate m^5C sites. For the validation of candidate m^5C sites, attempts were made to use an anti-m^5C antibody for MeRIP however the antibody failed to detect m^5C in RNA. I have improved the method for the validation of candidate m^5C sites using a high-throughput approach on the Illumina® MiSeq platform. In this section, I will discuss the outcomes of this work.

3.3.1 Knockdown of *NSUN2* and *TRDMT1* in HeLa cells

The process of knocking down *NSUN2* and *TRDMT1* was not trivial as in addition to reducing the amount of enzyme it was necessary to deplete m^5C at known target sites of these enzymes. Some transcript biotypes have half-lives of up to several days, with tRNAs and rRNAs being exceptionally long [251-253]. Thus, it was necessary to perform the RNAi-mediated knockdown of *NSUN2* and *TRDMT1* over a several day period. Dr. Jeffrey E. Squires previously performed a time course with a transfection period of 3, 4, 5 and 6 days. Sanger bisulfite sequencing of $tRNA^{Asp(GUC)}$ and $tRNA^{Leu(CAA)}$ at each time point revealed that the 6-day transfection period resulted in the best depletion of m^5C in these tRNAs, indicating that the majority of the tRNAs had turned over during a time at which the methyltransferase activities were substantially reduced. Interestingly, depletion of m^5C at C48 and C49 in $tRNA^{Asp(GUC)}$ in the NSUN2 knockdown samples was more modest compared to the depletion of m^5C observed at C34 in $tRNA^{Leu(CAA)}$. A recent study reported a similar result for $tRNA^{Gly(GCC)}$ when *NSUN2* was knocked down, where only partial reduction in methylation was observed at the NSUN2 target sites of C48, C49 and C50 [77]. It is possible that some tRNAs have a longer half-life than others, resulting in incomplete turnover of the transcript. Also, as the methyltransferases were only depleted in a transient manner, tRNAs that were transcribed later during the transfection period may still be present and methylated. This

is supported by a recent study that performed bisulfite sequencing on $tRNA^{Asp(GUC)}$ from *Nsun2* knockout mice and showed complete ablation of methylation at C48 and C49 [149]. As almost complete depletion of methylation was observed at C38 in the TRDMT1 knockdown sample, this suggests that knockdown of this enzyme was more efficient.

Although it was expected that there would be an underrepresentation of candidate m^5C sites called as NSUN2 and TRDMT1 targets from the bsRNA-seq due to their transient depletion, the fact that depletion of m^5C at the known target sites in tRNAs was evident indicates that m^5C sites mediated by NSUN2 and TRDMT1 in transcript biotypes with shorter half-lives should be called as NSUN2 and TRDMT1 targets in this thesis.

3.3.2 FEATURES OF THE BSRNA-SEQ LIBRARIES

3.3.2.A DISTRIBUTION OF SEQUENCING READS

The library construction was based on the size selection of cDNA products from 50-120 nt in order to capture both fragmented tRNA and mRNA sequences. While this method was successful in capturing tRNAs, other ncRNAs and mRNAs, ~40-75% of the sequencing reads in all libraries mapped to rRNAs. This was not unexpected as current rRNA depletion methods available are yet to completely remove rRNA due to its substantial abundance. We have previously detected a large proportion of sequencing reads from our previous bsRNA-seq study mapping to rRNA despite rRNA depletion and poly(A)+ selection of HeLa cell RNA [58]. Interestingly, the second replicate library preparation for the PrEC and LNCaP samples exhibited less sequencing reads mapping to rRNA compared to the first library preparation, which consisted up to 75% rRNA. This is most likely attributed to batch variation of the rRNA depletion kit. To remove unwanted transcripts such as rRNA in the future, a technique that involves probe-directed degradation (PDD) of unwanted RNA species could be used [254]. This technique relies on the hybridisation of DNA oligonucleotides to unwanted RNA species during the single-stranded cDNA library stage, followed by digestion with Duplex Specific Nuclease, which specifically digests dsDNA. This technique is advantageous in that it removes unwanted sequences during the cDNA library preparation stage, minimising RNA degradation. Furthermore, the flexibility in the

probe design means that other unwanted abundant transcripts can be removed. Removal of the unwanted abundant transcripts using PDD could potentially allow higher coverage of other transcript biotypes, including mRNAs and tRNAs.

Despite the substantial amount of rRNA in the libraries, the size selection used for bsRNA-seq library preparation successfully selected for mRNAs, tRNAs and other ncRNAs; ~13-23% of the sequencing reads mapped to mRNA, 2-4% to tRNA, and the remainder to various ncRNA. Thus, this library preparation enabled the sequencing of many transcript biotypes for candidate m^{5}C site detection.

3.3.2.B Alignment of sequencing reads and detection of candidate m^{5}C sites

The bioinformatics pipeline implemented in this study had been modified from the first transcriptome-wide detection of m^{5}C in RNA we performed [58]. Previously, SOCS-B was used, which is an alignment algorithm specific for bisulfite converted SOLiD™ sequencing reads. A limitation of SOCS-B is that it is computationally intensive and can be time consuming in aligning the reads; it is for this reason that B-SOLANA was chosen. B-SOLANA is an alignment algorithm specifically developed for the mapping of bisulfite-converted DNA in colour-space, like SOCS-B. For the alignment of DNA, two *in silico* bisulfite conversions of the reference genome are performed; one in which all cytosines are converted, and one in which all cytosines except those in a CpG context are converted, allowing detection of m^{5}C sites in CpG clusters. It was necessary to adapt B-SOLANA for the alignment and detection of candidate m^{5}C sites in RNA.

As m^{5}C sites in RNA have not yet been detected in clusters of CpGs and they do not necessarily occur in a CpG context, candidate m^{5}C sites were detected as mismatches. The Bowtie parameters used for the mapping of the bsRNA-seq data were set to a maximum of three colour-space mismatches (~1.5 bases) within the first 35 nt of the sequencing read (i.e. the higher-quality 5’ end of the read). Any sequencing reads that contained more mismatches within the first 35 nt were discarded. This could potentially be an issue for transcripts that exhibit multiple candidate m^{5}C sites in close proximity to each other, such as tRNAs, as reads derived from such a region may fail to be mapped (Figure 3.30). As a compromise, a Phred score sum of 200 was set as a threshold, which

allowed for more mismatches in the last 15 nt of the sequencing read and increased the possibility of detecting multiple candidate m^5C sites in the same sequencing read. A caveat of this approach, however, is the increased probability of detecting sequencing errors and low quality bases towards the 3' end of the sequencing read. However, in the case of $tRNA^{Asp(GUC)}$, these parameters allowed for the detection of all three known m^5C sites in all libraries (Figures 3.28A and 3.29A). Examination of the sequencing reads in IGV revealed that when the sequencing reads spanned $tRNA^{Asp(GUC)}$ such that the C38, C48 and C49 were located in the first 35 nt, no more than two m^5C sites were detected in a single sequencing read as a result of the mismatch threshold applied (data not shown). Therefore, the three m^5C sites were distributed over multiple sequencing reads. When C48 and C49 were located towards the 3' end of the sequencing read, all three m^5C sites could be detected as a result of the increased number of mismatches allowed towards the 3' end.

With the exception of tRNAs, multiple m^5C sites within close proximity have not yet been detected in other transcript biotypes. Previously, we found that 8.7% of the cytosines detected by bsRNA-seq in tRNAs contained m^5C, while only 1.2% of cytosines in other ncRNAs and 0.4% of cytosines in mRNAs contained m^5C [58]. Therefore, it was expected that the underrepresentation of candidate m^5C sites as a result of the alignment algorithm would be minimal in this thesis.

The issues associated with the detection of candidate m^5C sites in the bsRNA-seq data by B-SOLANA were not evident using Bismark for the alignment of the MiSeq amplicon sequencing data. Mapping of the MiSeq amplicon sequencing data by Bismark involved converting sequencing reads into C-to-T and G-to-A versions that were aligned to equivalently converted reference sequences. One of these conversions will result in no mismatches and align perfectly to the reference sequence (Figure 3.25). For example, if methylated Cs in a sequencing read are represented by a C, then this will be converted to Ts after the C-to-T conversion, and will map perfectly to the equivalent conversion of the reference. Following this, the sequencing reads prior to C-to-T and G-to-A conversion were aligned with the original reference sequence to detect m^5C sites regardless of the context of the site.

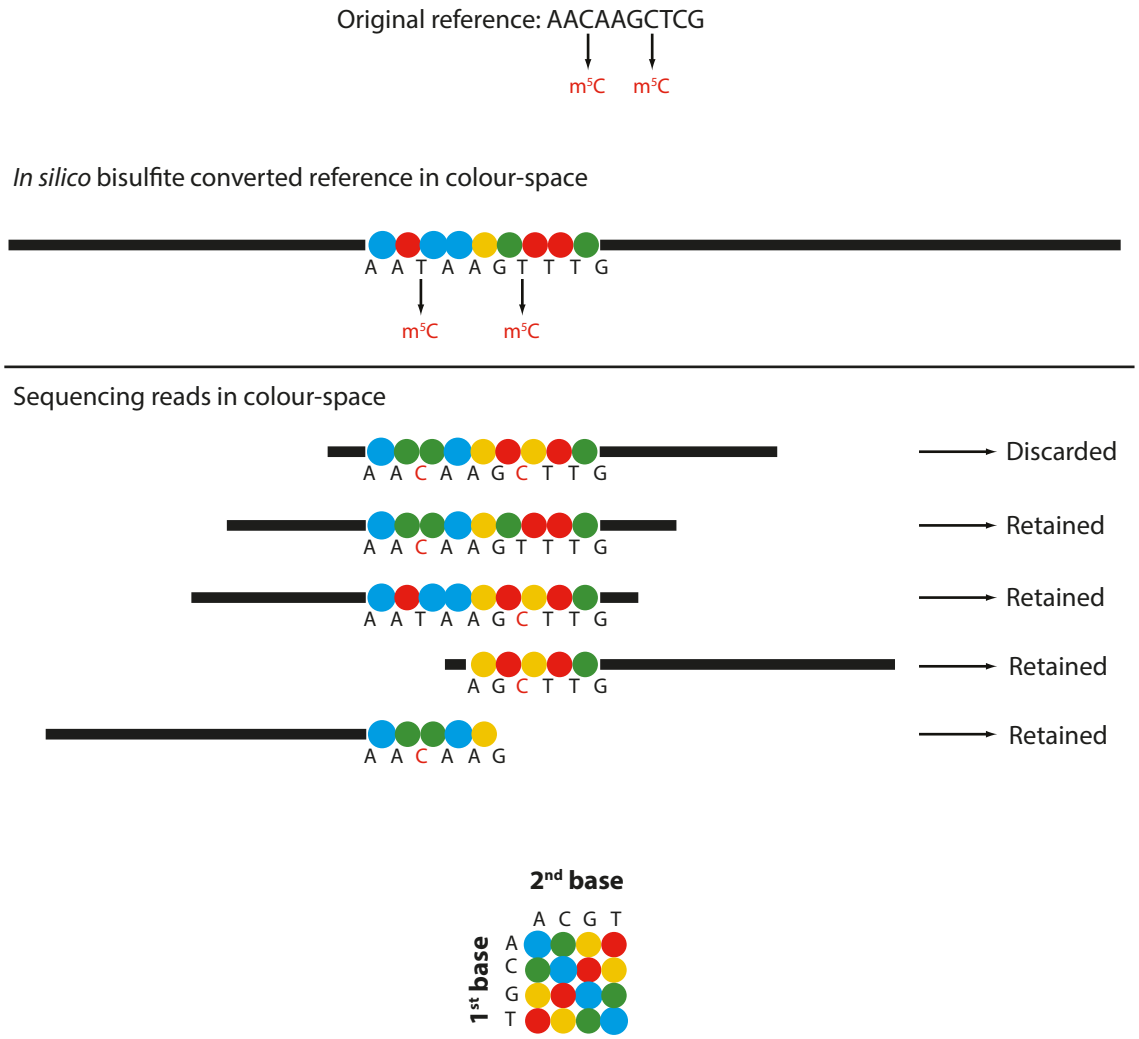


Figure 3.30. Schematic displaying the issues associated with multiple m^{5}C sites within a single 50 nt SOLiD™ colour-space sequencing read.

Prior to mapping, the reference sequence was *in silico* bisulfite converted. Sequencing reads that exhibited more than three colour-space mismatches within the first 35 nt of the read compared to the reference sequence were discarded. The m^{5}C sites are indicated in red.

The *in silico* non-CpG conversion of the references was not applied for the alignment of the bsRNA-seq data locally to rRNAs in order to limit the detection of non-converted cytosines attributed to incomplete denaturation of the secondary structure. The experimental procedures employed in this study were also performed in such a way to efficiently denature RNA while limiting its fragmentation, as too much fragmentation could affect the mapping of the sequencing reads. While these procedures were applied, it was expected that a low amount of clusters of non-conversion may still be detected as a result of the higher amount of mismatches allowed towards the 3' end of the

sequencing read. As expected, clusters of non-conversion were detected in 28S rRNA in this thesis, however the amount of clusters observed was substantially less than the amount detected in our previous study (see Chapter 4) [58].

Although these are drawbacks with the alignment of the bsRNA-seq data, it was necessary to adapt B-SOLANA for the alignment of bisulfite-converted RNA sequencing reads as no alignment algorithm has been developed specifically for the detection of m^{5}C sites in RNA. Thus, a compromise was made between limiting the alignment of sequencing reads with multiple errors and allowing the detection of candidate m^{5}C sites in RNA.

3.3.2.C EXPERIMENTAL DESIGN FOR THE DETECTION OF HIGH-CONFIDENCE CANDIDATE M^{5}C SITES

The experimental design for bsRNA-seq of the HeLa methyltransferase knockdown samples was set up as an initial screen for RNA targets of NSUN2 and TRDMT1, and as such a single sample was sequenced for each condition. Similarly, the experimental design for the bsRNA-seq of PrEC and LNCaP cells was set up as an initial comparison of candidate m^{5}C sites detected in different cellular contexts, which was enabled by the sequencing of two biological replicates. These biological replicates were performed at a later date to increase the accuracy of the initial screen.

Prior to candidate m^{5}C site detection, non-specific filtering was applied to remove low levels of non-conversion to exclude noise caused by spurious non-conversion. Additionally, candidate sites required ≥5 reads in all samples used in the model to prevent incorrect calling of a candidate differential m^{5}C site as a result of low expression. For example, if a candidate m^{5}C site is detected in a transcript in the TRDMT1 knockdown, DNMT1 knockdown and NTC samples, but the NSUN2 knockdown sample does not have sufficient coverage of that transcript, then this candidate m^{5}C site may be incorrectly assigned as a NSUN2-mediated site. Similarly, if a candidate m^{5}C site is detected in a transcript in LNCaP cells and there is insufficient coverage of this transcript in PrEC cells, this may be incorrectly assigned as a candidate m^{5}C site hypermethylated in LNCaP cells. To detect high-confidence candidate m^{5}C sites, a proportion statistic in combination with a moderated t-test was used to either

compare the proportion of m^{5}C between samples for the detection of candidate differential m^{5}C sites, or against 0% methylation to call candidate m^{5}C sites.

The use of three controls for the detection of candidate m^{5}C sites in the HeLa methyltransferase knockdown dataset allowed for the detection of high-confidence candidate m^{5}C sites. Although the control samples involved three different treatments, they behaved as "pseudo-replicates" for this study; DNMT1 does not target RNA for methylation and NSUN2 and TRDMT1 were found to target different sites. Therefore, for the detection of all candidate m^{5}C sites in HeLa cells, the NSUN2 knockdown, TRDMT1 knockdown and NTC samples, followed by the TRDMT1 knockdown, DNMT1 knockdown and NTC samples were used as controls. ≥10% mean methylation was required across these controls. For the detection of candidate m^{5}C sites mediated by NSUN2 or TRDMT1, three controls were used as described for the detection of all candidate m^{5}C sites, which required ≥10% mean methylation. The proportion of m^{5}C in the NSUN2 knockdown and TRDMT1 knockdown samples was then compared against the three controls for the detection of NSUN2 and TRDMT1 target sites. As one "experimental" sample was used for the detection of candidate m^{5}C sites mediated by NSUN2 or TRDMT1, it was expected there would be an underestimation of the number of candidate m^{5}C sites found to be targeted by these two enzymes in this thesis.

The sequencing of biological replicates of PrEC and LNCaP cells enabled the detection of candidate m^{5}C sites in common between each cell line by requiring ≥10% mean methylation across all four samples. For the detection of candidate differential m^{5}C sites in PrEC and LNCaP cells, ≥10% mean methylation was required in either both PrEC samples or both LNCaP samples.

To limit the amount of false positive sites detected using this bioinformatics pipeline, candidate m^{5}C sites that were not homogenous (i.e. <95% of the reads were not a C or T) and overlapped SNPs or editing sites based on the available databases were removed. Two issues could occur if a candidate m^{5}C site overlapped a SNP. The first is that if a C-to-T SNP existed at the site for half of the sequencing reads, then this would result in the underestimation of the proportion of m^{5}C at this site. The second issue is that for the detection of candidate differential m^{5}C sites in PrEC and LNCaP cells, a site may be incorrectly called if only one cell line exhibits a SNP.

3.3.2.D BISULFITE CONVERSION EFFICIENCY

It was important in this study to control for the bisulfite conversion efficiency in order to limit the amount of false positive m^5C sites detected due to lack of cytosine conversion. As such, non-human *in vitro* transcripts containing no m^5C sites were spiked into the RNA samples prior to bisulfite conversion. Sanger bisulfite sequencing of two segments of a R-Luc *in vitro* transcript prior to library preparation revealed high bisulfite conversion efficiency for those segments. The bisulfite conversion efficiency was also examined for the entire R-Luc *in vitro* transcripts based on bsRNA-seq, revealing >99.8% conversion of cytosines. Sequencing of $tRNA^{Asp(GUC)}$ and $tRNA^{Leu(CAA)}$ showed high levels of methylation at the known m^5C sites in all samples, with the exception of the NSUN2 and TRDMT1 knockdown samples where depletion of methylation at target sites was observed. This indicated that the bisulfite conversion technique employed in this study resulted in efficient conversion of unmethylated cytosines and did not result in conversion of known m^5C sites.

3.3.2.E RNA MODIFICATIONS RESISTANT TO BISULFITE CONVERSION

Although bisulfite sequencing of RNA is currently the gold standard for detecting m^5C, it does come with a potential drawback. A potential issue with bisulfite conversion is that there may be other modifications besides m^5C that are resistant to conversion, although the types of modifications that may be resistant still remain to be explored. Indications of this were made when bisulfite conversion was originally optimised for RNA, and N4, 2'-*O*-dimethylcytidine (m^4Cm) at position 1402 of bacterial 16S rRNA was detected by locus-specific bisulfite sequencing [59]. The equivalent region in mitochondrial 12S rRNA from the hamster is methylated in the motif Gm^4CCm^5CG [255]. However, we have previously investigated the orthologous region in humans by bsRNA-seq and Sanger bisulfite sequencing of HeLa cells and could only detect the m^5C site [58]. Furthermore, the bsRNA-seq data from this study also showed cytosine conversion at this site for all samples (see Chapter 4), suggesting that this modification is not resistant to the bisulfite conversion protocol employed in this thesis. Schaefer et al. also suggested that other modifications, such as N4-methylcytidine (m^4C) and 3-methylcytosine (m^3C), could affect the deamination process during bisulfite conversion. m^3C has been detected in human 18S and 28S rRNA by hydrolysis and

chromatography, although the location of these sites were never mapped [256]. Furthermore, m^3C has been detected in *S. cerevisiae* tRNAs at C32 of $tRNA^{Thr1}$ and $tRNA^{Ser1}$ [257, 258]. Recently, a study that involved predicting modifications in human tRNAs based on differences in RT incorporation found 13 m^3C sites [259]. Three of these m^3C sites were validated by m^3C immunoprecipitation, and were located at C3 of $tRNA^{Met(CAU)}$, C32 of mitochondrial $tRNA^{Met(CAU)}$ and $tRNA^{Thr(UGU)}$. Non-converted cytosines were not detected at these positions in this thesis (data not shown), suggesting that either this modification is not present in PrEC, LNCaP or HeLa cells or that this modification is not resistant to bisulfite conversion.

An advantage of this study is that sequencing of both the NSUN2 and TRDMT1 knockdown samples, which have the potential to target many sites for methylation, may provide enough information to narrow down the modification at a position. For example, if depletion of methylation is observed at a m^5C site in a transcript in the NSUN2 knockdown sample compared to the controls, then it is clear that the modification must be m^5C or potentially an oxidised form of this modification, such as 5-hydroxymethylcytosine (hm^5C), which is resistant to bisulfite conversion in DNA [260, 261]. However, the presence of hm^5C naturally in RNA remains elusive. One study thus far has shown that ten-eleven translocation (TET) hydroxylases, which are responsible for oxidising m^5C to hm^5C in DNA [262], also possess this activity for RNA *in vitro*, suggesting that this modification may also be present naturally in RNA [263]. Interestingly, hm^5C in DNA can be further oxidised to 5-formylcytosine (f^5C) followed by 5-carboxylcytosine (ca^5C) by the TET hydroxylases, both of which are not resistant to bisulfite conversion [264, 265]. Recently, a technique to distinguish between m^5C and hm^5C by taking advantage of the different chemical reactivity between these modifications has been developed by oxidising hm^5C to f^5C then performing bisulfite conversion, termed oxBS-seq [266]. In addition to this, standard bisulfite sequencing is performed, allowing distinction between m^5C and hm^5C by subtraction of the readouts. It would be worthwhile in the future to optimise this technique for RNA to not only distinguish between m^5C and hm^5C, but also provide the first transcriptome-wide detection of hm^5C.

3.3.3 Detection of m^{5}C using an antibody

To overcome potential issues associated with bisulfite conversion resistance of other modifications, attempts were made to optimise an anti-m^{5}C antibody for the purpose of MeRIP. This technique could be coupled with next-generation sequencing to identify m^{5}C regions transcriptome-wide, or RT-qPCR to detect enrichment of candidate m^{5}C sites relative to a negative control. This technique is well established in DNA, and as such anti-m^{5}C antibodies were readily available for use in RNA.

Several studies have used MeRIP-seq for the detection of m^{6}A in RNA. On the contrary, only a single study has used this technique for the validation of candidate m^{5}C sites detected by bsRNA-seq [76]. Although this technique does not provide single nucleotide resolution, the use of MeRIP would be favourable, both as a validation tool and for coupling with bsRNA-seq, which may help eliminate biases associated with both techniques. Therefore, the use of MeRIP for the detection of m^{5}C warranted further exploration.

The antibody implemented for this study was identical to that used in the m^{5}C MeRIP-seq study [76]. Optimisation of the anti-m^{5}C antibody for the detection of m^{5}C in RNA was initially performed using dot blot assays. Yeast tRNA was selected for the optimisation procedure due to its high abundance in m^{5}C. HeLa cell RNA was also used as a true biological sample and it was expected that the amount of m^{5}C detected in this sample would be substantially less than the highly modified tRNA. As controls, methylated DNA and unmethylated DNA were used, which were purchased from the manufacturer of the antibody. To ensure that the secondary structure did not impede the hybridisation of the antibody to RNA, the RNA samples were either heat denatured with or without formamide. The DNA controls were heat denatured with formamide and snap cooled on ice, or with 0.1 M NaOH, snap cooled on ice and neutralised with 6.6 M ammonium acetate. In all cases, m^{5}C was detected in the methylated DNA control and not in the unmethylated DNA control, indicating that the dot blot assay was successful. However, regardless of the denaturation procedure, m^{5}C was not detected in yeast tRNA or HeLa cell RNA. Staining of the membranes with methylene blue indicated that all samples hybridised successfully, suggesting that the secondary structure of the RNA may have impeded binding of the antibody to m^{5}C despite denaturation. That this

antibody failed to detect m^{5}C in RNA is unexpected considering it has been previously used for the detection of m^{5}C by MeRIP-seq in archaea [76]. Furthermore, the antibody was raised against m^{5}C conjugated to ovalbumin lacking the deoxyribose or ribose sugar, indicating that it could not distinguish between m^{5}C in DNA or RNA. To determine the source of the issue, I contacted the authors regarding the protocol that was used in this publication but specific details were not provided.

The dot blot assay used in this thesis was comparable to that used for detecting m^{6}A in RNA in a previous study, although the denaturation procedures differed slightly. For the m^{6}A study, the RNA was denatured by mixing glyoxal loading dye with the RNA and heating at 50°C for 20 min [56]. Glyoxal denatures the RNA by glyoxalating the bases, preventing base pairing and reducing RNA secondary structure formation in a similar manner to formamide. Further investigation into denaturation procedures would be important in order to use the m^{5}C antibody for MeRIP in the future. In addition to denaturing RNA, it may be beneficial to fragment the RNA prior to hybridisation to the membrane; this would minimise secondary structure formation, making the RNA more accessible for the antibody. As this procedure is necessary for MeRIP, it would be ideal to use this approach for optimisation.

3.3.4 Improvements to bisulfite sequencing validation techniques

Several bisulfite sequencing methods were implemented for the detection of known m^{5}C sites and validation of candidate m^{5}C sites in RNA. Sanger bisulfite sequencing was performed to ensure bisulfite conversion was efficient and knockdown of the methyltransferases was sufficient. We have also previously implemented Sanger bisulfite sequencing for validation of novel candidate m^{5}C sites in a number of transcript biotypes [58]. While this method proved to be useful, it was labour-intensive and low-throughput. Due to the larger number of samples in this study, it was necessary to implement a high-throughput approach that still involved locus-specific sequencing using the Illumina® MiSeq platform. This initially involved bisulfite PCR amplification of candidates in triplicate in order to reduce PCR amplification biases, followed by pooling of the amplicons, library preparation and high-throughput sequencing. There are a number of advantages to using this approach; it is less laborious and very high-

throughput, allowing multiplexing of samples. Furthermore, since this approach was used for locus-specific sequencing, the MiSeq platform was capable of sequencing up to hundreds of thousands of reads per amplicon, providing higher confidence in the validation of candidate m^5C sites compared to Sanger bisulfite sequencing.

A caveat of the MiSeq approach was the inconsistency in library quantification throughout the library preparation and library validation process, which can affect cluster generation and the success of the sequencing run. Although quantification using the Qubit fluorometer, DNA High Sensitivity Chip and qPCR yielded vastly different results, it was interesting that the former two methods provided lower concentrations than the qPCR. As mentioned previously, the qPCR specifically amplifies DNA containing the Illumina® adaptors. Thus, it would be expected that this method would yield a lower concentration in comparison to the Qubit fluorometer and Agilent Bioanalyzer, as these methods would quantify all species of DNA including those fragments that don't contain adaptors and adaptor primers.

Loading of the libraries based on a High Sensitivity DNA Chip resulted in a high cluster density beyond the optimal range. This was most likely a result of the Agilent Bioanalyzer quantification method underestimating the true concentration, leading to overloading of the libraries onto the flow cell. This in turn caused overlapping of clusters, reducing the quality of mappable reads. Quantification of Illumina® libraries is a renowned issue and a study performed at the The Wellcome Trust Sanger Institute has focused on improving quantification of Illumina® Genome Analyzer libraries and sequencing runs [267]. They reported that quantifying libraries by spectrophotometry resulted in inconsistencies in the cluster density that was typically lower than the optimal range. This is not unexpected as all DNA species, including those that do not contain adaptors would be quantified. Interestingly, it was reported that quantification of libraries using the Agilent Bioanalyzer resulted in much more consistent cluster densities that were in the desired range. However, there were a few situations in which this quantification method led to high cluster densities, which they explained was a result of ssDNA generated during the PCR enrichment step of the library preparation not being easily quantified when mixed with dsDNA. Perhaps this is the issue that arose for the MiSeq run in this thesis. Nonetheless, the Agilent Bioanalyzer may be useful in determining the average fragment size after library preparation, which is necessary for

the quantification of libraries. For this study, agarose gel electrophoresis was used to determine the average fragment size, and the library concentrations determined according to this fragment size by qPCR resulted in successful sequencing runs.

Despite the discrepancies associated with the quantification of the libraries, it is clear from this thesis that loading of the libraries based on concentrations determined by qPCR resulted in optimal cluster densities, read quality and high and consistent coverage of amplicons. The known m^5C sites in $tRNA^{Asp(GUC)}$ and $tRNA^{Leu(CAA)}$ were detected and validated as NSUN2 or TRDMT1-mediated m^5C sites. Thus, validation of candidate m^5C sites detected using bsRNA-seq by MiSeq amplicon sequencing proved to be a robust technique.

3.3.5 Conclusions

In this chapter, I have described the preparation of cell lines for the detection of m^5C in RNA, as well as the improvement of techniques to detect m^5C in RNA, both in terms of experimental procedures and bioinformatics pipelines. The RNAi-mediated knockdown of NSUN2 and TRDMT1 led to sufficient depletion of m^5C at the known target sites in transcripts exhibiting long half-lives. The bisulfite conversion procedure used in this study was efficient, with unmethylated cytosines converted to uracil and minimal conversion of m^5C to uracil based on the controls. The bsRNA-seq bioinformatics pipeline used for the alignment of sequencing reads and detection of candidate m^5C sites was adapted for use with RNA and was an important resource for the detection and comparison of m^5C sites in RNA under different conditions in this thesis. Finally, I have established a high-throughput and reliable procedure for the validation of candidate m^5C sites in RNA using the Illumina® MiSeq platform.

3.4 CHAPTER HIGHLIGHTS

- Two biological replicates of PrEC and LNCaP cells were prepared for bsRNA-seq.
- RNAi-mediated knockdown of *NSUN2*, *TRDMT1* and *DNMT1* was successfully performed in HeLa cells, resulting in depletion of the respective mRNAs and enzymes. A single sample of each knockdown was prepared for bsRNA-seq as an initial screen to identify targets of NSUN2 and TRDMT1.
- RNAi-mediated knockdown of *NSUN2* and *TRDMT1* resulted in depletion of m^{5}C at known target sites of NSUN2 and TRDMT1.
- Efficient bisulfite conversion was evident in all samples.
- bsRNA-seq libraries were prepared using the SOLiD™ Total RNA-Seq Kit and the library construction was based on the size selection of cDNA products from 50-120 nt in order to capture both fragmented tRNA and mRNA sequences.
- B-SOLANA, an alignment algorithm specific for bisulfite-converted DNA, was adapted for use with bisulfite-converted RNA.
- A proportion statistic and moderated t-test were applied to detect high confidence candidate m^{5}C sites. For the HeLa methyltransferase knockdown samples, this enabled the detection of candidate m^{5}C sites mediated by NSUN2 and TRDMT1, as well as all candidate m^{5}C sites in HeLa cells. For PrEC and LNCaP cells, this allowed for the detection of candidate differential m^{5}C sites and candidate m^{5}C sites in common between each cell line.
- Potential artefacts were limited by removing candidate m^{5}C sites that overlapped SNPs or editing sites or <95% of the bases were not a C or T.
- To be considered a candidate m^{5}C site by bsRNA-seq, ≥5 reads were required in all samples used in the model and ≥10% mean methylation was required across all control samples.
- Attempts were made to optimise an anti-m^{5}C antibody for MeRIP, however this antibody failed to detect m^{5}C in RNA based on dot blot assays.
- A high-throughput method for the validation of candidate m^{5}C sites was established using the Illumina® MiSeq platform, which allowed sequencing of bisulfite PCR amplicons with substantial read depth, increasing confidence in the validation of candidate m^{5}C sites.

CHAPTER FOUR

RNA m^5C PROFILING OF HELA CELLS DEPLETED OF NSUN2 OR TRDMT1

4.1 INTRODUCTION

The recent advancement in next-generation sequencing technologies for the detection of RNA modifications has prompted further investigations into their enzymology. Currently, two RNA methyltransferases are known to catalyse m^5C formation in humans; NSUN2 and TRDMT1. Although studies over the last several years have investigated the substrate range of these enzymes in certain transcript biotypes, this field still remains to be explored in further detail. Previously, we have used a candidate approach coupling RNAi-mediated knockdown of *NSUN2* and *TRDMT1* with Sanger bisulfite sequencing and discovered that NSUN2 has a wider substrate range than previously anticipated, targeting at least three structural positions in tRNAs as well as one ncRNA and three mRNAs [58]. More recent studies have employed next-generation sequencing to determine the RNA substrates of NSUN2 and to a lesser extent, TRDMT1, revealing various ncRNA targets of NSUN2 [77, 78, 149]. However, the understanding of m^5C in mRNA is still limited. One study has reported a single mRNA target [77], while another study detected 312 mRNA targets but did not investigate this any further [78]. Thus, it was of interest to explore NSUN2 RNA targets in more detail and to determine the specificity of NSUN2 for these transcripts.

The aim of this chapter was to perform an initial screen to identify RNA targets of NSUN2 and TRDMT1 based on the coupling of RNAi-mediated knockdown of *NSUN2* and *TRDMT1* with bsRNA-seq. Using this data, the distribution of all candidate m^5C sites in RNA in HeLa cells was also explored. Candidate m^5C sites were thoroughly investigated in tRNAs and the findings independently confirmed recently published work by others. Candidate m^5C sites in other ncRNAs and mRNAs were also extensively investigated, including the location bias, structural context and sequence context of this modification, providing a wealth of novel findings.

4.2 RESULTS

The following sections document the distribution, abundance, sequence and structural context of candidate m5C across transcript biotypes in HeLa cells, with a particular focus on candidate m5C sites called as NSUN2 targets. Mapping of the bsRNA-seq and MiSeq data were performed by Dr. Brian J. Parker, who also performed the subsequent global analyses in consultation with myself. I performed the experimental validations, secondary analyses and graphical representations generated by the R programming language.

4.2.1 M5C SITES DETECTED IN HELA CELLS

As detailed in Chapter 3, bsRNA-seq and the subsequent bioinformatics analyses were performed on a single sample for each control and knockdown condition. Sanger bisulfite sequencing and bsRNA-seq revealed efficient bisulfite conversion of all samples and sufficient knockdown of NSUN2 and TRDMT1 based on the depletion of m5C at known positions in $tRNA^{Asp(GUC)}$ and $tRNA^{Leu(CAA)}$. The single biological replicates used for the validation of candidate m5C sites by MiSeq amplicon sequencing also showed efficient bisulfite conversion and sufficient knockdown of the methyltransferases.

4.2.1.A TOTAL M5C SITES IN HELA CELLS

Using the proportion statistic and appropriate controls as described in Chapter 3, the total number of candidate m5C sites in HeLa cells was determined. The criteria for the detection of candidate m5C sites were defined as ≥5 reads in all samples used in the model and ≥10% mean methylation across the control samples, as well as a FDR≤1%. This yielded a total of 4,241 candidate m5C sites in a wide variety of transcript biotypes, including tRNAs, other ncRNAs and mRNAs (Figure 4.1).

As expected, a large number of candidate m5C sites were detected in tRNAs; 490 based on the global mapping and 185 based on the local mapping to tRNA sequences from tRNAdb. As detailed in section 3.2.3.A, local mapping of the data is a more accurate source for mature tRNA methylation and was thus used primarily in the following. As

the global mapping dataset also contains reads deriving from precursor tRNA (pre-tRNA) molecules, it was of particular use for analysing the temporal relationship between processing and methylation.

667 candidate m^{5}C sites were detected in other ncRNAs and were distributed across various transcript biotypes. Interestingly, the majority of the candidate m^{5}C sites detected were in mRNAs. The 1,564 candidate m^{5}C sites detected in mRNAs ranged over 1,229 transcripts, indicating an average of ~1.3 candidate m^{5}C sites per transcript. 1,486 candidate m^{5}C sites were also detected in unannotated intronic or intergenic regions that did not meet the criteria to be considered one of the other transcript biotypes. Transcripts in this category include previously unidentified lincRNAs and antisense transcripts of coding genes and intergenic regions. The complete list of candidate m^{5}C sites in HeLa cells is located in Appendix 8.3.

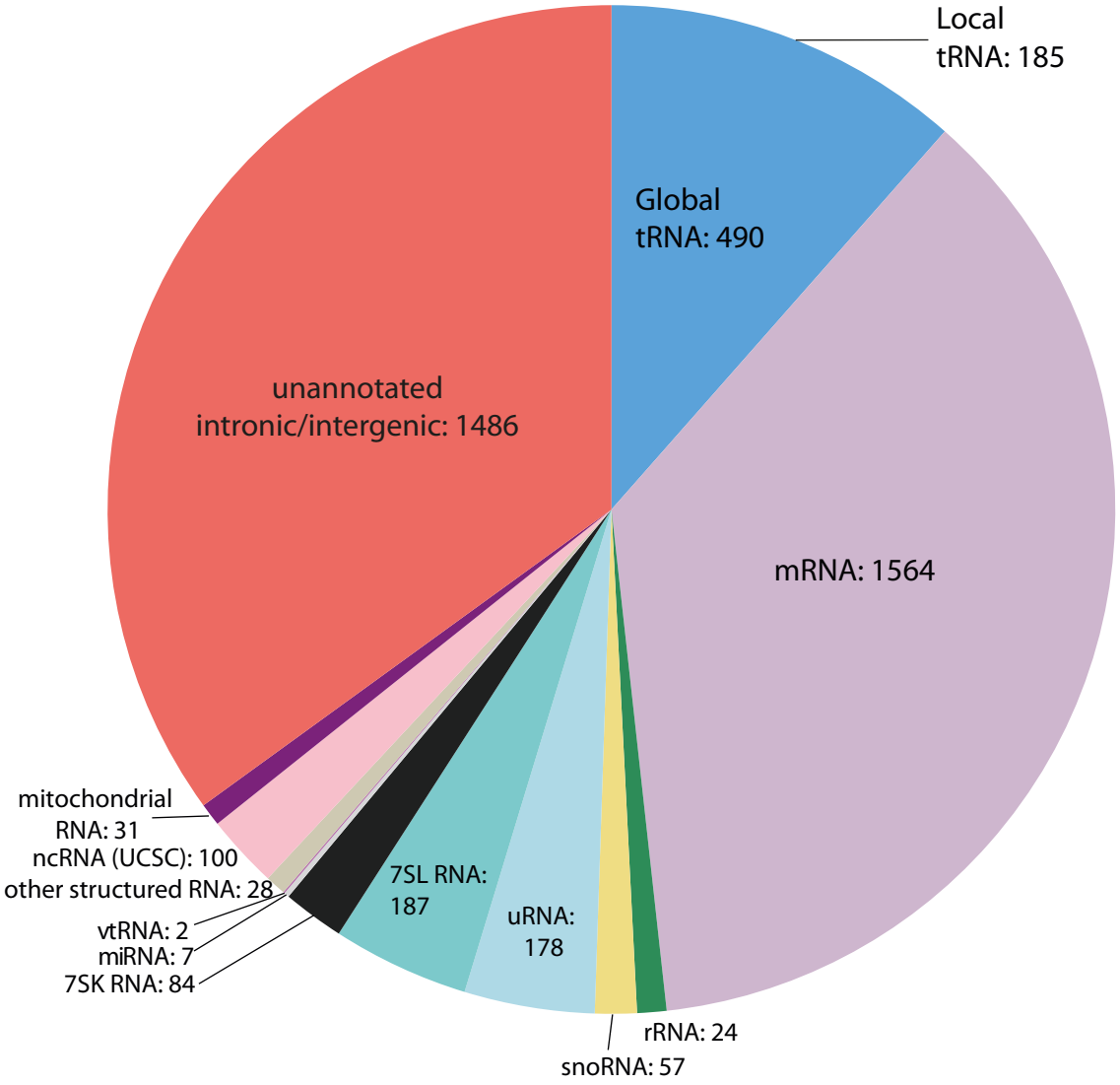


Figure 4.1. Distribution of candidate m^{5}C sites detected in HeLa cells across different transcript biotypes.

≥5 reads in all samples used in the model and ≥10% mean methylation across the control samples was required for the detection of candidate m^{5}C sites. FDR≤1% estimated by the Benjamini-Hochberg method. "Other structured RNA" refer to Rfam structured RNA entries and "ncRNA (UCSC)" refer to UCSC-annotated ncRNAs, which include long intergenic ncRNA (lincRNA) and unstructured RNA. The number of candidate m^{5}C sites detected in each transcript biotype is indicated.

In order to gain insight into the abundance of m^{5}C across transcript biotypes, the proportion of cytosines with ≥5 reads that exhibited candidate m^{5}C sites in each transcript biotype was determined (Table 4.1). As expected, m^{5}C in tRNAs was quite abundant, with 3.98% of cytosine positions being called as candidate m^{5}C sites from the local mapping and 5.73% in tRNAs mapped globally. The abundance of m^{5}C in mRNAs was much lower, with 0.08% of cytosines exhibiting candidate m^{5}C sites. Interestingly, several categories of ncRNA had >1% cytosines as m^{5}C sites, including rRNA, vault (vt) RNA, ubiquitous (u) RNA, 7SK RNA and 7SL RNA.

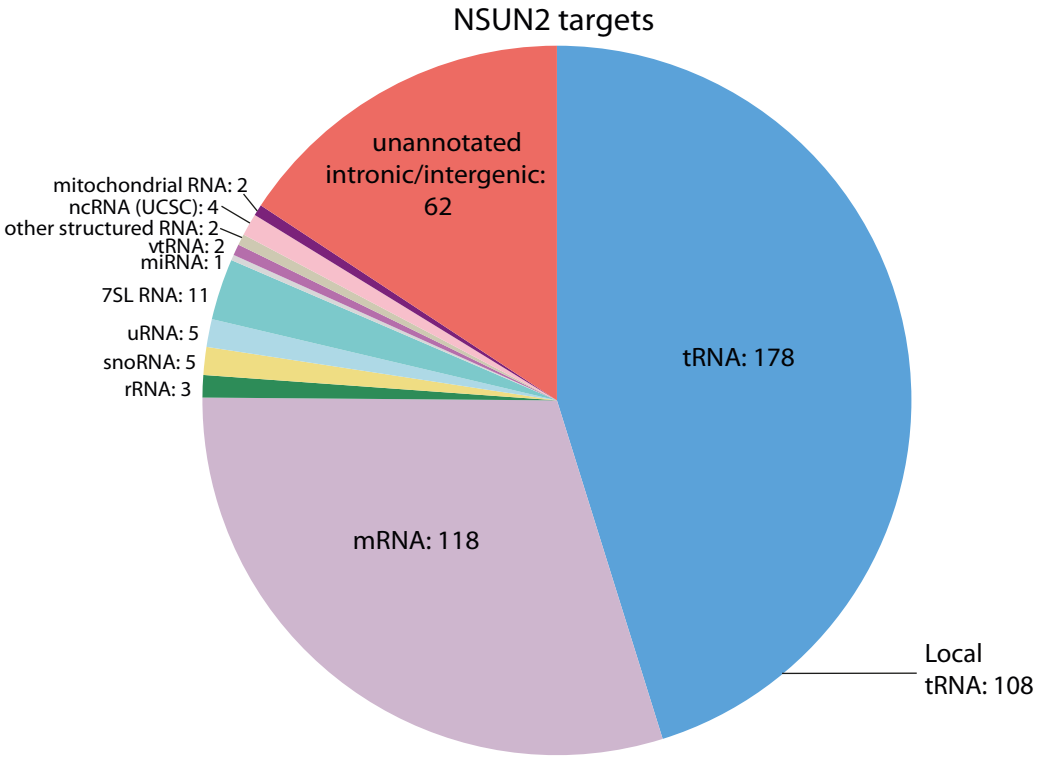
Table 4.1. The proportion of cytosines with ≥5 reads containing candidate m^{5}C sites across different transcript biotypes.

Sample	Number of m^{5}C sites	Total number of C with reads ≥ 5	% m^{5}C
Global tRNA	490	8553	5.73
Local tRNA	185	4645	3.98
mRNA	1564	1,966,054	0.08
rRNA	24	2229	1.08
snoRNA	57	5571	0.74
7SL RNA	187	9671	1.95
7SK RNA	84	5928	1.42
uRNA	178	9318	1.91
vtRNA	2	109	1.83
miRNA	7	1330	0.53
ncRNA (UCSC)	100	85,562	0.12
Mitochondrial RNA	31	2140	1.45

4.2.1.B m^{5}C SITES MEDIATED BY NSUN2 AND TRDMT1

Using the proportion statistic, the distribution of candidate m^{5}C sites across transcript biotypes called as NSUN2 or TRDMT1 targets based on bsRNA-seq were determined. The criteria for enzyme target site detection were defined as ≥5 reads in all samples and ≥10% mean methylation across all control samples, as well as a FDR≤30%. This yielded a total of 393 candidate m^{5}C sites called as NSUN2 targets and 120 called as TRDMT1 targets, with both RNA methyltransferases targeting various transcript biotypes (Figure 4.2).

As expected, both enzymes were responsible for many candidate m^{5}C sites in tRNAs, with NSUN2 targeting 108 candidate m^{5}C sites and TRDMT1 targeting 15 candidate m^{5}C sites based on the local mapping of tRNAs. Interestingly, 118 candidate m^{5}C sites in mRNAs were called as NSUN2 targets and 53 candidate m^{5}C sites in mRNAs were called as TRDMT1 targets, substantially extending what was previously known about the substrate range of these enzymes. Additionally, 33 candidate m^{5}C sites in other ncRNAs were called as NSUN2 targets, and 9 candidate m^{5}C sites in other ncRNA were called as TRDMT1 targets. The complete lists of candidate m^{5}C sites called as NSUN2 or TRDMT1 targets are located in Appendix 8.3.


NSUN2 targets
unannotated intronic/intergenic: 62
mitochondrial RNA: 2
ncRNA (UCSC): 4
other structured RNA: 2
vtRNA: 2
miRNA: 1
7SL RNA: 11
uRNA: 5
snoRNA: 5
rRNA: 3
tRNA: 178
mRNA: 118
Local tRNA: 108

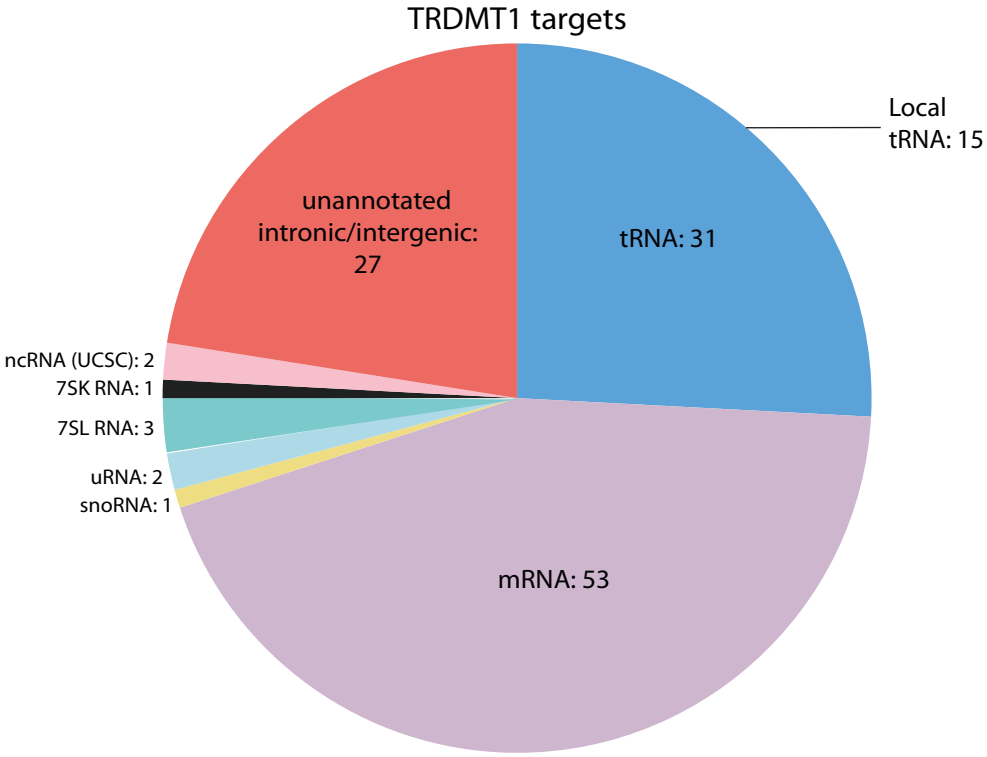

TRDMT1 targets
Local tRNA: 15
unannotated intronic/intergenic: 27
tRNA: 31
ncRNA (UCSC): 2
7SK RNA: 1
7SL RNA: 3
uRNA: 2
snoRNA: 1
mRNA: 53

Figure 4.2. The distribution of candidate m^{5}C sites across different transcript biotypes called as NSUN2 or TRDMT1 targets.

≥5 reads in all samples used in the model and ≥10% mean methylation across the control samples was required for the detection of candidate m^{5}C sites called as NSUN2 or TRDMT1 targets. FDR≤30% estimated by the Benjamini-Hochberg method. The number of candidate m^{5}C sites detected in each transcript biotype is indicated.

Overall, of the 4,241 candidate m^{5}C sites that were detected in HeLa cells, only 513 were called as NSUN2 or TRDMT1 targets. This is due to a combination of factors, although as this study was performed as an initial screen it is likely that there was under-calling of methyltransferase target sites due to incomplete methyltransferase knockdown.

In order to understand the contribution of NSUN2 and TRDMT1 to methylation, the proportion of candidate m^{5}C sites in each transcript biotype called as NSUN2 or TRDMT1 targets was determined and the bulk methylation levels in each transcript biotype were also determined for each sample (Figure 4.3). The proportion of candidate m^{5}C sites called as NSUN2 or TRDMT1 targets varied for each transcript biotype (Figure 4.3; pie charts). The two candidate m^{5}C sites detected in vtRNAs were called as NSUN2 targets, while 34-96% of candidate m^{5}C sites detected in other transcript biotypes were not called as NSUN2 or TRDMT1 targets. To calculate the bulk methylation level in each transcript biotype for each sample, the mean proportion of methylation was calculated in each sample for all candidate m^{5}C sites detected in HeLa cells (Figure 4.3; bar charts). Knockdown of NSUN2 resulted in at least some depletion of methylation across all transcript biotypes, indicating that this enzyme has a very broad substrate range. The knockdown of TRDMT1 had a more modest and selective effect in that reduced mean methylation was only evident in tRNA, mRNA and 7SL RNA.

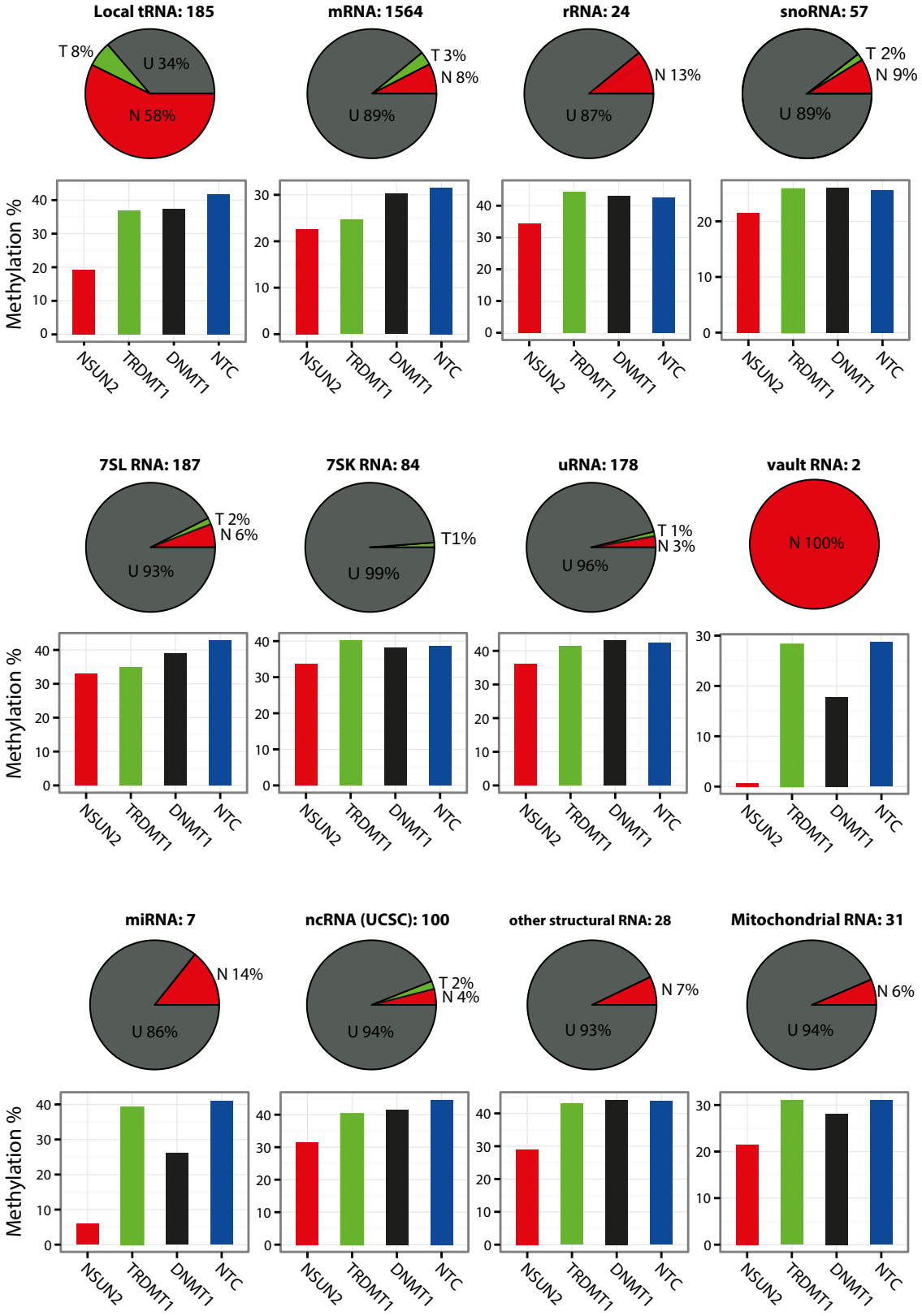

Local tRNA: 185
T 8%
U 34%
N 58%
mRNA: 1564
T 3%
N 8%
U 89%
rRNA: 24
N 13%
U 87%
snoRNA: 57
T 2%
N 9%
U 89%
Methylation %
NSUN2
TRDMT1
DNMT1
NTC
7SL RNA: 187
T 2%
N 6%
U 93%
7SK RNA: 84
T1%
U 99%
uRNA: 178
T 1%
N 3%
U 96%
vault RNA: 2
N 100%
miRNA: 7
N 14%
U 86%
ncRNA (UCSC): 100
T 2%
N 4%
U 94%
other structural RNA: 28
N 7%
U 93%
Mitochondrial RNA: 31
N 6%
U 94%

Figure 4.3. Contribution of NSUN2 and TRDMT1 to m^{5}C across transcript biotypes based on bsRNA-seq.

The proportion of candidate m^{5}C sites called as NSUN2 or TRDMT1 targets (pie charts) and the mean methylation of candidate m^{5}C sites in each sample within each transcript biotype (bar charts) were determined based on bsRNA-seq. NSUN2 targets=N, TRDMT1 targets=T and undetermined=U.

4.2.2 m^{5}C in tRNA

It was of interest to document the candidate m^{5}C sites detected in tRNAs in HeLa cells, as well as determine the substrate specificity of NSUN2 and TRDMT1. A focus was placed on the tRNA structural positions, isotypes and isoacceptors that exhibited candidate m^{5}C sites. The tRNA structural positions refer to the positions within the secondary structure of the tRNA. Currently, there are seven structural positions known to contain m^{5}C, including C34, C38, C40, C48-C50 and C72 (see Figure 4.4). Isoacceptors refer to the sequence within the anti-codon of a tRNA that translates into an amino acid for translation. Isotypes refer to those tRNAs that translate into the same amino acid. For example, tRNAs that contain the sequences AAA and GAA within the anti-codon translate into the same amino acid of phenylalanine, but are different isoacceptors. Thus, the aim of this section was to investigate the substrate specificity of NSUN2 and TRDMT1 in tRNA structural positions, isotypes and isoacceptors. This analysis primarily focused on the candidate m^{5}C sites detected from the local mapping of tRNAs and conclusions were drawn from this. However, for completeness, the analyses were also performed on candidate m^{5}C sites detected from the global mapping of tRNAs. Our results were juxtaposed with those of two recent studies that have used complementary transcriptome-wide techniques to detect tRNA targets of NSUN2 [77, 150] and TRDMT1 [77], which will be detailed in the discussion section of this chapter.

An investigation was also made to confirm that m^{5}C in tRNAs can occur prior to splicing and whether m^{5}C can occur prior to 5’ and 3’ end processing. I implemented a candidate approach to assess whether m^{5}C can be detected in tRNAs prior to splicing. As the global mapping of tRNAs enabled mapping of a small proportion of sequencing reads to pre-tRNAs, this, in addition to a candidate approach, was used to investigate whether tRNAs are modified prior to 5’ and 3’ end processing.

4.2.2.A tRNA structural positions and isotypes targeted by NSUN2 and TRDMT1

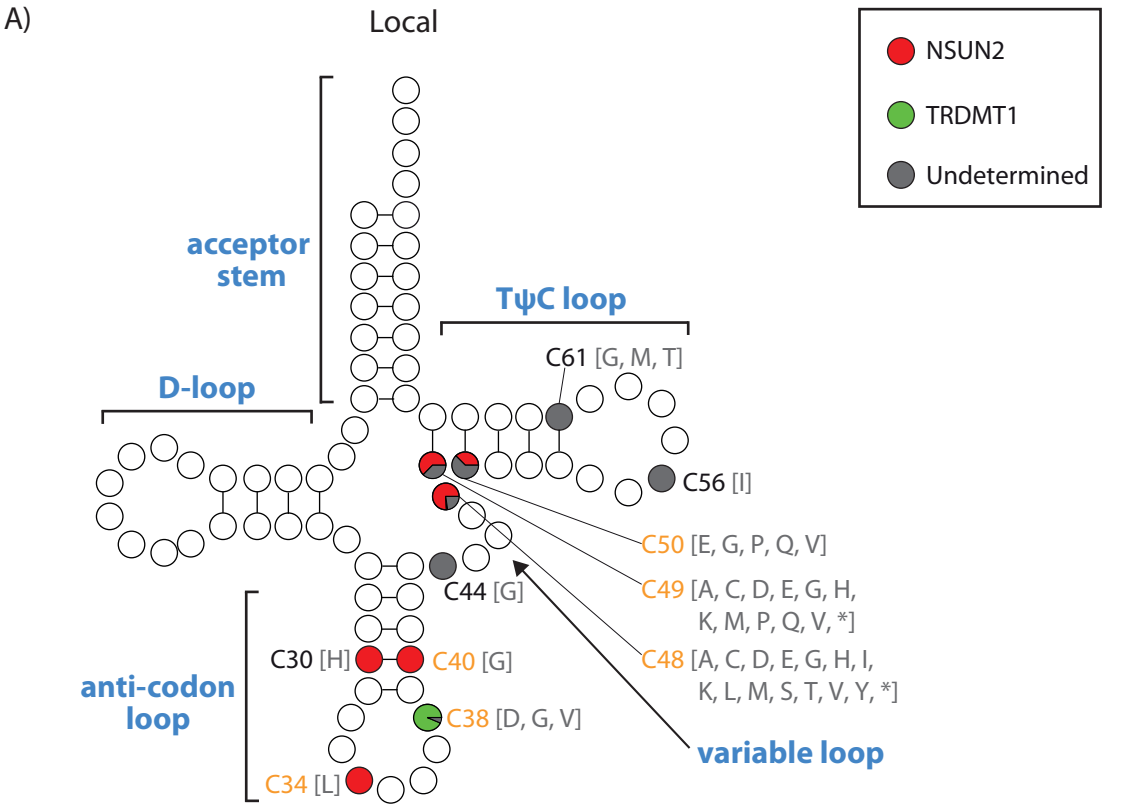
In order to determine the tRNA structural positions and isotypes that exhibited candidate m^{5}C sites and the specificity of NSUN2 and TRDMT1, the bsRNA-seq data was mined and cloverleaf structures indicative of the tRNA secondary structure were generated. The cloverleaf structures detail the location of candidate m^{5}C sites within the structural positions and the tRNA isotypes exhibiting candidate sites in these positions (Figure 4.4). In addition, the proportion of candidate m^{5}C sites that were called as NSUN2 or TRDMT1 targets for each structural position was determined, which are indicated by the pie charts.

Candidate m^{5}C sites were detected at all known structural positions except C72 (Figure 4.4A). The lack of sites at C72 was possibly due to the declining coverage towards the 3' end of the tRNAs. NSUN2 was found to exclusively mediate m^{5}C at C34 in tRNA^{Leu} and C40 at tRNA^{Gly}. >60% of candidate m^{5}C sites detected at C48 and C49 were called as NSUN2 targets with the remainder undetermined, and candidate m^{5}C sites at these positions were in a wide range of tRNA isotypes. In addition, there was high concordance in the tRNA isotypes exhibiting candidate m^{5}C at these two positions; 10 tRNA isotypes containing candidate m^{5}C sites at C48 and C49 overlapped. Interestingly, candidate m^{5}C sites were detected at C48 and C49 of a suppressor tRNA, which is an entirely novel finding. By contrast, only 5 tRNA isotypes exhibited candidate m^{5}C sites at C50, and 35% of the tRNA isotypes exhibiting a candidate m^{5}C site at this position were called as NSUN2 targets. Unlike NSUN2, TRDMT1 was found to almost exclusively modify C38 in tRNA^{Asp}, tRNA^{Gly} and tRNA^{Val}. A candidate m^{5}C site at C48 was detected in tRNA^{Cys} as a TRDMT1 target, but this would require validation to determine if this is a genuine site.

Four non-canonical structural positions were found to exhibit candidate m^{5}C sites including C30, C44, C56 and C61. The site at C30 was specific to tRNA^{His} and was called as a NSUN2 target. The candidate m^{5}C site at C44 was found exclusively in tRNA^{Gly} and at C56 in tRNA^{Ile}. Finally, the candidate m^{5}C site at C61 was detected in tRNA^{Gly}, tRNA^{Met} and tRNA^{Thr}.

The structural distribution pattern of candidate m^5C sites in tRNAs based on the global mapping was broadly similar to that based on the local mapping (Figure 4.4B). Here, all known structural positions exhibited candidate m^5C sites, including C72. The ability to map the immature tRNAs by global mapping most likely resulted in more coverage at the 3' end of the tRNA, allowing detection of this site. Three of the non-canonical structural positions that exhibited candidate m^5C sites based on the local mapping were also detected in the global mapping, which included C30, C44 and C56. ~25% $tRNA^{His}$ transcripts containing a candidate m^5C site at C30 were mediated by NSUN2. Additional non-canonical structural positions that exhibited candidate m^5C sites were also detected, including C17, C25, C51, C62, C63, C65, C67 and C68.

In summary, 87% candidate m^5C sites detected in tRNAs from the local mapping were in the anticipated structural positions; 99% of candidate m^5C sites called as NSUN2 targets were in the anticipated structural positions and 93% of candidate m^5C sites called as TRDMT1 targets were in the anticipated structural positions. In addition, it is clear from this analysis that NSUN2 is responsible for the majority of m^5C sites in tRNAs, targeting several different structural positions in a broad range of tRNA isotypes, while TRDMT1 almost exclusively mediated m^5C at C38 in three tRNA isotypes.


A)
Local
NSUN2
TRDMT1
Undetermined
acceptor stem
TψC loop
D-loop
C61 [G, M, T]
C56 [I]
C50 [E, G, P, Q, V]
C49 [A, C, D, E, G, H, K, M, P, Q, V, *]
C48 [A, C, D, E, G, H, I, K, L, M, S, T, V, Y, *]
C44 [G]
C30 [H]
C40 [G]
C38 [D, G, V]
C34 [L]
anti-codon loop
variable loop

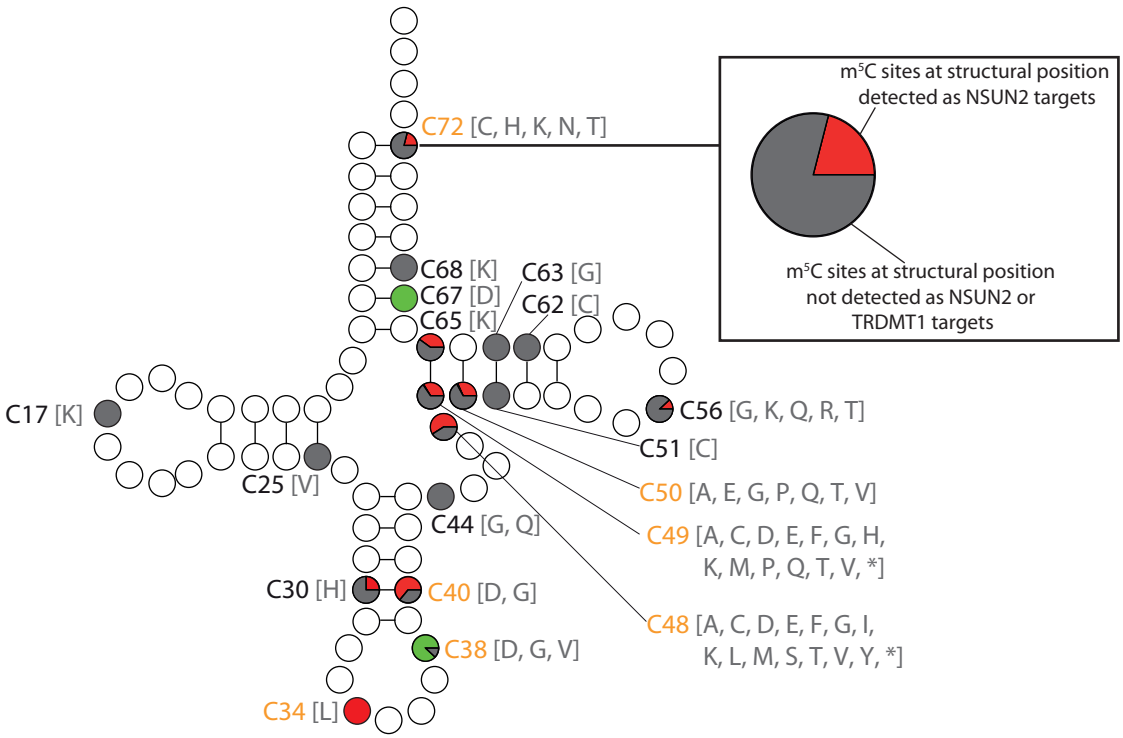

B)
Global
m5C sites at structural position detected as NSUN2 targets
m5C sites at structural position not detected as NSUN2 or TRDMT1 targets
C72 [C, H, K, N, T]
C68 [K]
C67 [D]
C65 [K]
C63 [G]
C62 [C]
C17 [K]
C25 [V]
C56 [G, K, Q, R, T]
C51 [C]
C50 [A, E, G, P, Q, T, V]
C49 [A, C, D, E, F, G, H, K, M, P, Q, T, V, *]
C48 [A, C, D, E, F, G, I, K, L, M, S, T, V, Y, *]
C44 [G, Q]
C30 [H]
C40 [D, G]
C38 [D, G, V]
C34 [L]

Figure 4.4. tRNA structural positions and isotypes exhibiting candidate m^{5}C sites called as NSUN2 and TRDMT1 targets.

A) Generated using local tRNA mapping. B) Generated using global tRNA mapping. Red=NSUN2 targets, green=TRDMT1 targets, grey=undetermined. A particular structural position may contain multiple colours depending on whether the candidate m^{5}C site was called as a NSUN2 target, TRDMT1 target or was undetermined. tRNA isotypes exhibiting candidate m^{5}C sites in these positions are indicated by the amino acid abbreviations. * indicates a suppressor tRNA. The different regions of the tRNA secondary structure are indicated in blue. The seven structural positions known to exhibit m^{5}C sites are indicated in orange.

4.2.2.B tRNA ISOACCEPTORS TARGETED BY NSUN2 AND TRDMT1

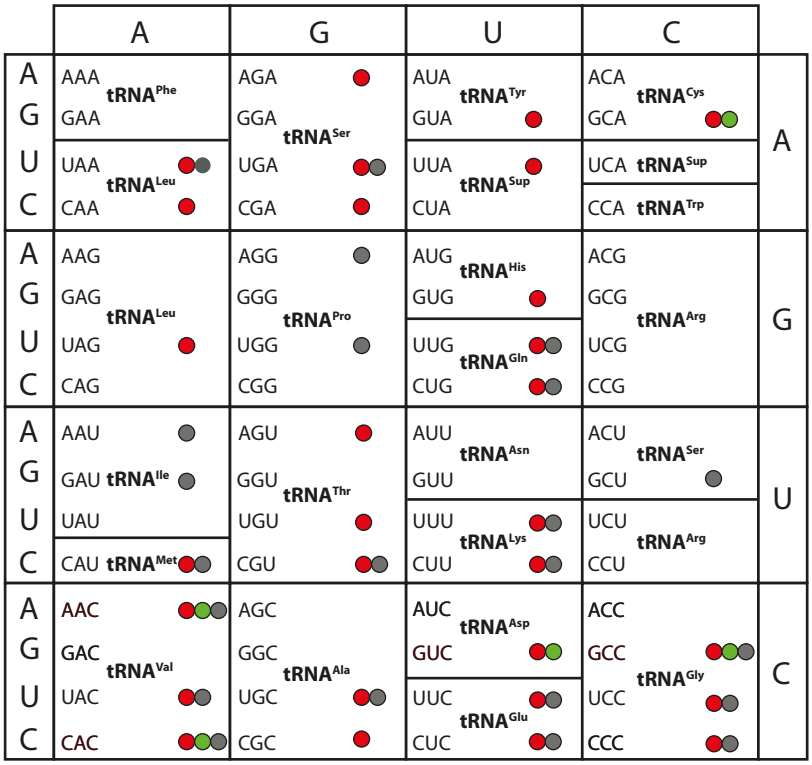
The bsRNA-seq data was also mined for tRNA isoacceptors exhibiting candidate m^{5}C sites, and a table was generated displaying the tRNA isoacceptors exhibiting candidate m^{5}C sites that were called as NSUN2 targets, TRDMT1 targets or were not undetermined. This table was generated to understand the specificity of NSUN2 and TRDMT1 for tRNA isoacceptors.

In total, 29 tRNA isoacceptors were called as NSUN2 targets based on the local mapping, and 32 based on the global mapping (Figure 4.5). There was also high concordance between both methods of mapping, with 26 tRNA isoacceptors called as NSUN2 targets in common. Interestingly, there was no defined trend for tRNA isoacceptors that were targeted by NSUN2.

In contrast to NSUN2, TRDMT1 was much more specific, targeting tRNA$^{Asp(GUC)}$, tRNA$^{Gly(GCC)}$, tRNA$^{Val(AAC)}$, tRNA$^{Val(CAC)}$ and tRNA$^{Cys(GCA)}$. The candidate m^{5}C site targeted by TRDMT1 in tRNA$^{Cys(GCA)}$ has not been previously identified, and this was also not at C38 known to be targeted by TRDMT1 (see section 4.2.2.A). The 14 candidate m^{5}C sites detected as TRDMT1 targets at C38 in these four tRNA isoacceptors were derived from different sequences of each tRNA isoacceptor. The tRNA isoacceptors exhibiting candidate m^{5}C sites called as TRDMT1 targets in the global mapping were largely in agreement with the local mapping, with tRNA$^{Asp(GUC)}$, tRNA$^{Gly(GCC)}$, tRNA$^{Val(AAC)}$ and tRNA$^{Val(CAC)}$ all detected.

Thus, in addition to targeting a wide range of tRNA structural positions and isotypes, NSUN2 exhibits broad substrate specificity for different tRNA isoacceptors, while TRDMT1 was specific to four tRNA isoacceptors.

A) Local



B) Global

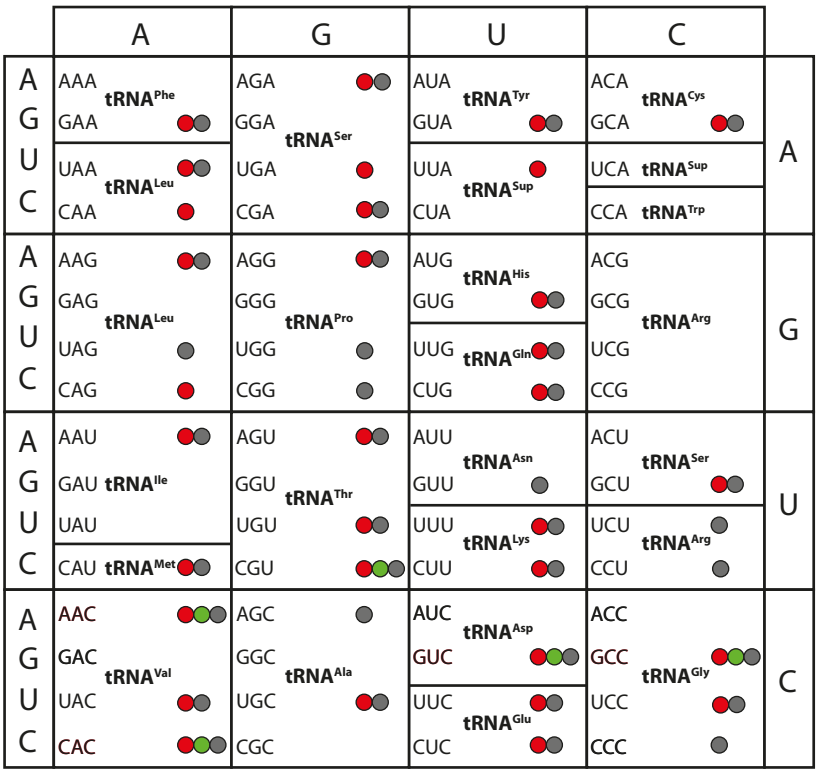


Figure 4.5. tRNA isoacceptors exhibiting candidate m^{5}C sites called as NSUN2 and TRDMT1 targets.

A) Generated using local tRNA mapping. B) Generated using global tRNA mapping. Red=NSUN2 targets, green=TRDMT1 targets, grey=undetermined.

4.2.2.C Validation of candidate m^{5}C sites in tRNA

Given that the majority of candidate m^{5}C sites measured by bsRNA-seq were detected in the anticipated structural positions, our additional MiSeq amplicon sequencing validation efforts were directed towards revealing the technical limitations of our mapping methods. Specifically, it was expected that several m^{5}C sites in close proximity may escape detection due to the mismatch limits imposed in B-SOLANA (see section 3.3.2.B). To examine this further, five tRNAs exhibiting candidate m^{5}C sites called as NSUN2 or TRDMT1 targets based on the local mapping were selected for validation. High coverage of the selected tRNAs was evident in the bsRNA-seq data, however a reduction in coverage in the 3' end of the tRNA was observed (Figure 4.6; top panel). Indeed, the MiSeq validation revealed several additional candidate m^{5}C sites that were not detected by bsRNA-seq (Figure 4.6; middle and bottom panels). Candidate m^{5}C sites at C38, C40 and C49 in tRNA$^{Gly(GCC)}$ were detected by bsRNA-seq, but MiSeq amplicon sequencing detected additional m^{5}C sites at C48 and C50. Similarly, a single candidate m^{5}C site was detected at C38 in tRNA$^{Val(AAC)}$ based on bsRNA-seq, with MiSeq amplicon sequencing detecting additional m^{5}C sites at C48 and C49.

The candidate m^{5}C site at C40 in tRNA$^{Gly(GCC)}$ that was called as a NSUN2 target by bsRNA-seq was detected at low levels, and this site was detected at even lower levels by MiSeq amplicon sequencing. The proportion of methylation at this position based on MiSeq amplicon sequencing was 2-5% in the control samples, and <1% in the NSUN2 knockdown sample. All candidate m^{5}C sites detected at C38 and C48-50 in the selected tRNAs by bsRNA-seq were also validated by MiSeq amplicon sequencing. The candidate m^{5}C sites detected at C38 in tRNA$^{Gly(GCC)}$ and tRNA$^{Val(AAC)}$ exhibited depletion in the TRDMT1 knockdown sample by both bsRNA-seq and MiSeq amplicon sequencing, indicating that these sites are mediated by TRDMT1. The candidate m^{5}C

sites detected at C48-50 in tRNA$^{Gly(GCC)}$, tRNA$^{Glu(CUC)}$, tRNA$^{Thr(UGU)}$, tRNA$^{Val(AAC)}$ and tRNA$^{Val(CAC)}$ exhibited depletion in the NSUN2 knockdown sample by bsRNA-seq and/or MiSeq amplicon sequencing, indicating that these sites are mediated by NSUN2.

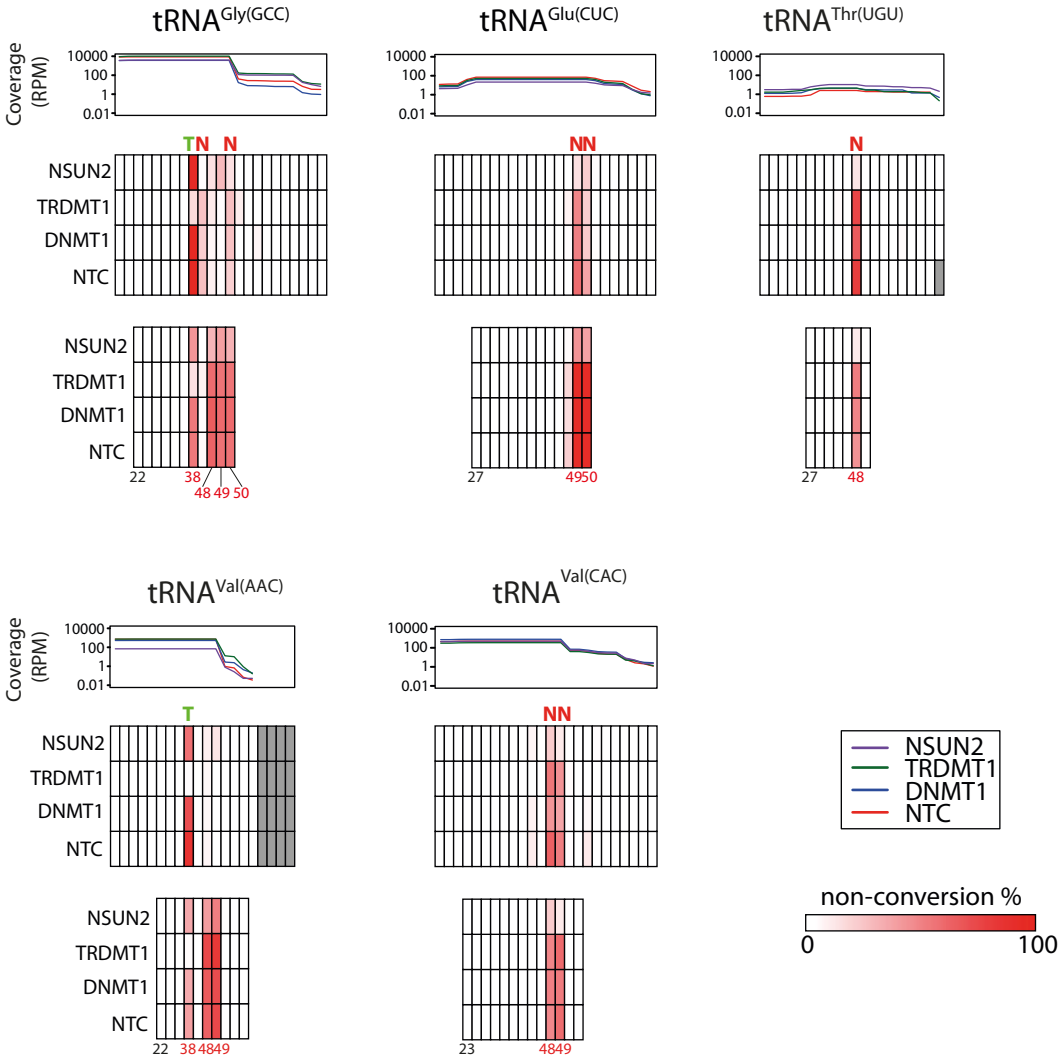


Figure 4.6. Validation of candidate m^{5}C sites in tRNAs by MiSeq amplicon sequencing.
The top panel displays the coverage throughout the transcript in RPM based on bsRNA-seq, the middle panel displays the proportion of non-conversion based on bsRNA-seq, and the bottom panel displays the proportion of non-conversion in a segment of the transcript based on MiSeq amplicon sequencing. The bsRNA-seq data is based on the local mapping of tRNAs. The shades of grey in the middle panel indicate regions of the transcript where coverage was <5 reads based on bsRNA-seq. The tRNA structural positions with candidate m^{5}C sites are indicated in red below the heatmap. The N and T indicate sites that were called as NSUN2 or TRDMT1 targets based on bsRNA-seq, respectively. The bsRNA-seq data is based on the local mapping of tRNAs. The MiSeq coverage ranged from 1,000-121,000 mappable sequencing reads.

4.2.2.D METHYLATION OF TRNAS PRIOR TO PROCESSING

tRNAs are transcribed in the nucleolus and undergo end processing in the nucleolus and nucleoplasm, followed by addition of the 3' CCA, splicing and final export into the cytoplasm (for a review see [268]), although the order of end processing and splicing can be reversed in several species [269, 270]. NSUN2 is predominantly localised in the nucleolus [99, 208] while TRDMT1 has been reported to localise predominantly in the nucleus [166]. As such, it was of interest to determine whether m^{5}C in tRNAs can occur early in the tRNA biogenesis pathway.

Two studies have reported that m^{5}C at C34 of tRNA$^{Leu(CAA)}$ specifically occurs prior to splicing, but this employed *in vitro* methylation assays [100, 148]. We have previously demonstrated that methylation at this position can occur prior to splicing based on low-throughput Sanger bisulfite sequencing [58]. Furthermore, investigation into methylation prior to 5' and 3' end processing is lacking. Thus, it was of interest to independently validate that methylation of tRNAs can occur prior to splicing and also determine whether m^{5}C can occur prior to 5' and 3' end processing of tRNAs.

Methylation of tRNAs prior to splicing

Out of the 631 tRNAs in the human genome, only 32 contain introns [271]. Several studies have shown that the nucleoside modification of tRNA$^{Tyr(GUA)}$ in yeast and tRNA$^{Leu(CAA)}$ in yeast and humans occurs only prior to the removal of the intron. ψ at position 35 of tRNA$^{Tyr(GUA)}$ was reported to occur specifically prior to splicing [272]. Similarly, the formation of m^{5}C at C34 in tRNA$^{Leu(CAA)}$ was found to only occur in the presence of the intron in yeast and HeLa extract using *in vitro* methylation assays [100, 148]. In this study, candidate m^{5}C sites were only detected in two intron-containing tRNAs; at C34 in tRNA$^{Leu(CAA)}$ and C48 in tRNA$^{Tyr(GUA)}$. As the bsRNA-seq coverage was low for tRNA$^{Tyr(GUA)}$, and the m^{5}C site at C34 in tRNA$^{Leu(CAA)}$ is a known site and was validated in this study (see Chapter 3), a focus was placed on tRNA$^{Leu(CAA)}$.

In order to independently validate that m^{5}C at C34 of tRNA$^{Leu(CAA)}$ can occur prior to splicing, sequencing reads from MiSeq amplicon sequencing were mapped to spliced and unspliced tRNA$^{Leu(CAA)}$ (see Figure 3.28B). Indeed, both spliced and unspliced

tRNA$^{Leu(CAA)}$ were methylated and the proportion of methylation was higher in the spliced tRNA$^{Leu(CAA)}$ compared to unspliced tRNA$^{Leu(CAA)}$, indicating that methylation of tRNAs can occur prior to splicing (Table 4.2).

Table 4.2. Proportion of m^{5}C at C34 in unspliced and spliced tRNA$^{Leu(CAA)}$ based on MiSeq amplicon sequencing.

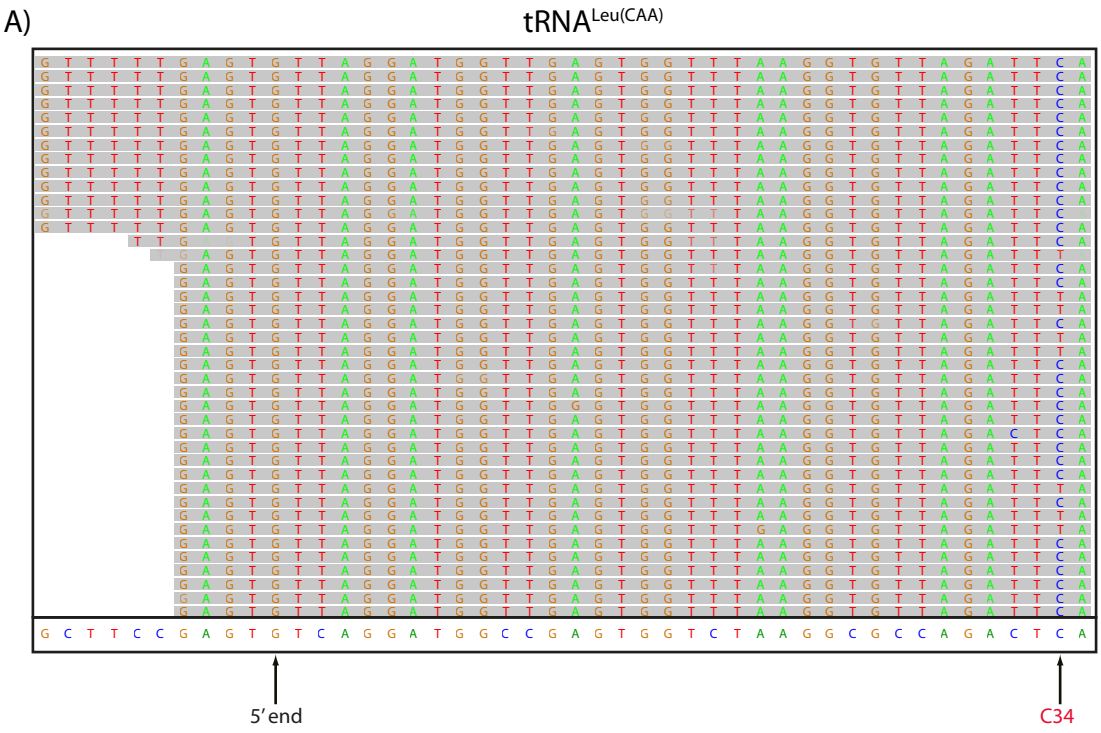
Sample	% m^{5}C for unspliced tRNA$^{Leu(CAA)}$	% m^{5}C for spliced tRNA$^{Leu(CAA)}$
NSUN2	4.78%	31.32%
TRDMT1	47.27%	58.12%
DNMT1	52.67%	79.33%
NTC	50.08%	78.05%

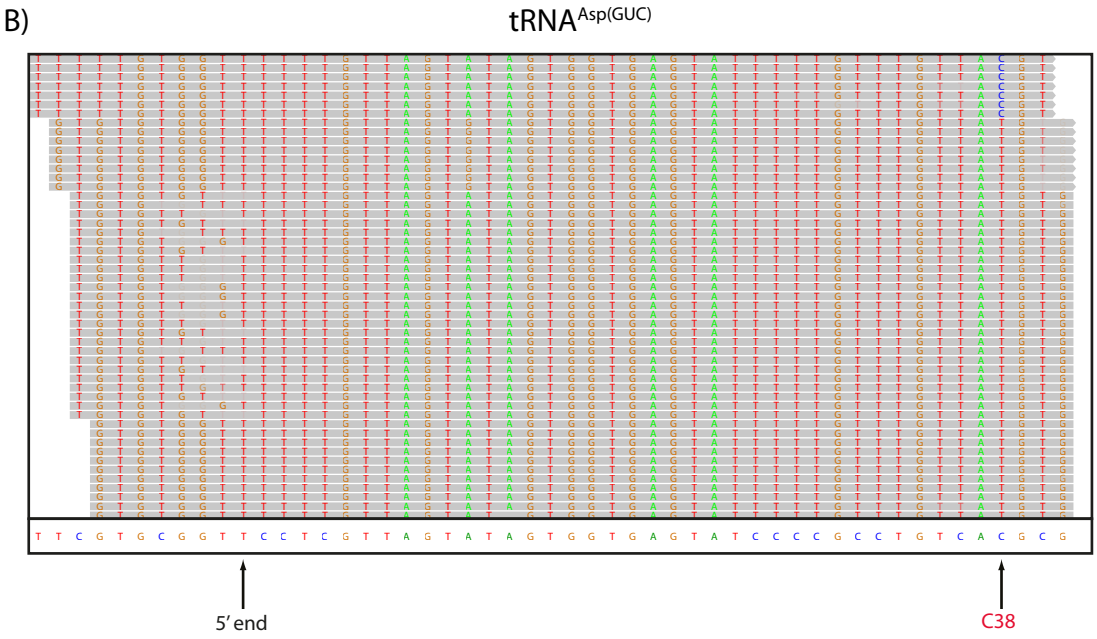
Methylation of tRNAs prior to 5’ and 3’ end processing

The major benefit of mapping tRNAs genome-wide to the full Rfam database was that it enabled the mapping of sequencing reads to pre-tRNAs. Interestingly, a small proportion of the 50 nt sequencing reads mapping to tRNAs were found to extend beyond the mature tRNA sequence by approximately 10-12 nt, indicating that a small amount of pre-tRNAs were sequenced. Thus, it was of interest to determine whether m^{5}C can occur prior to 5’ and 3’ end processing. If candidate m^{5}C sites were present in sequencing reads that extended beyond the mature ends of the tRNA, then this was suggestive of methylation occurring prior to end processing.

Initially a candidate approach was used, focusing on the known m^{5}C sites in tRNA$^{Leu(CAA)}$ and tRNA$^{Asp(GUC)}$. As C34 in tRNA$^{Leu(CAA)}$ is located close to the 5’ end of the tRNA and mediated by NSUN2, an examination was made into whether the sequencing reads spanning this m^{5}C site also extended beyond the 5’ end. Indeed, m^{5}C was evident in reads extending beyond the 5’ end of the tRNA, suggesting that NSUN2 can mediate m^{5}C prior to 5’ end processing of the tRNA (Figure 4.7A). C38 in tRNA$^{Asp(GUC)}$ is also located towards the 5’ end of the tRNA and is mediated by TRDMT1. Several sequencing reads exhibiting the m^{5}C site extended beyond the 5’ end of the tRNA, suggesting that TRDMT1 can mediate m^{5}C prior to 5’ end processing (Figure 4.7B). The m^{5}C sites at C48 and C49 of tRNA$^{Asp(GUC)}$ are located towards the 3’ end of the tRNA and mediated by NSUN2, and it was evident that many of the

sequencing reads that extended beyond the 3' end of the tRNA exhibited m^{5}C sites at these two positions (Figure 4.7C). This suggests that NSUN2 can mediate m^{5}C prior to 3' end processing of tRNAs.





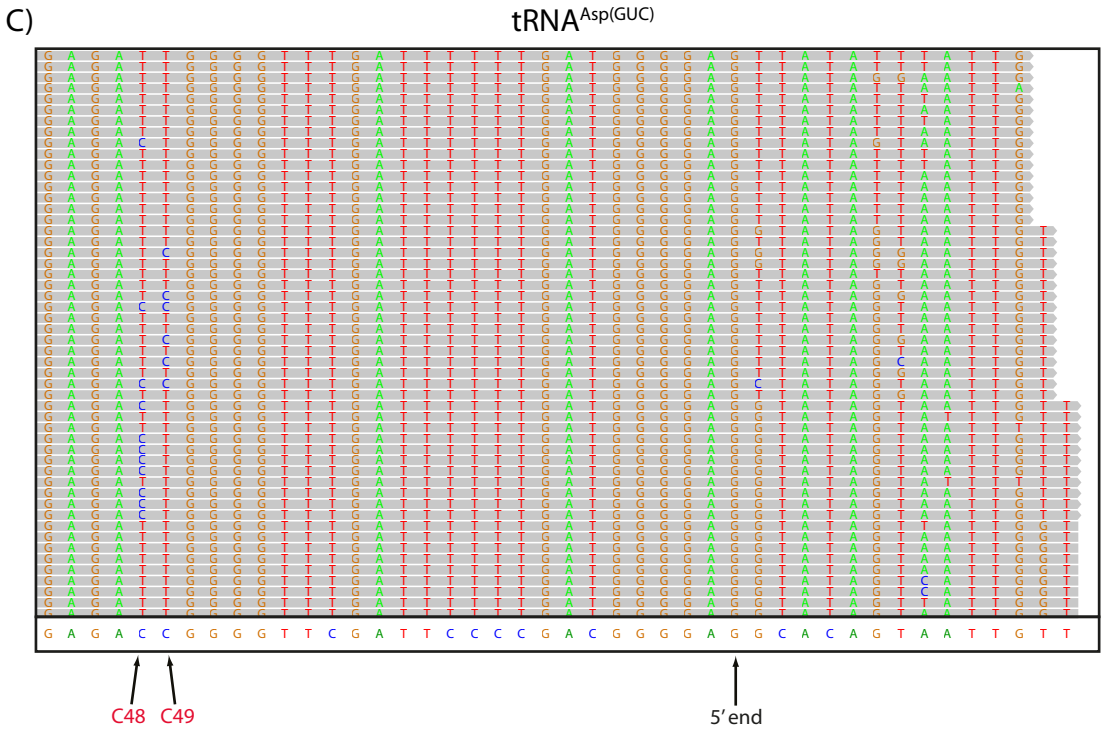


Figure 4.7. Representative snapshots from IGV depicting a subset of sequencing reads from the global tRNA mapping of tRNA$^{Leu(CAA)}$ and tRNA$^{Asp(GUC)}$ in the NTC sample.

The grey bars represent 50 nt sequencing reads mapped to part of the locus. Unconverted cytosines represent m^{5}C sites. A) m^{5}C at C34 in tRNA$^{Leu(CAA)}$ (chr1:249,168,044-249,168,089, + strand) in sequencing reads extending beyond the 5' end. B) m^{5}C at C38 in tRNA$^{Asp(GUC)}$ (chr12:96,429,789-96,429,839, + strand) in sequencing reads extending beyond the 5' end. C) m^{5}C at C48 and C49 in tRNA$^{Asp(GUC)}$ (chr12:96,429,841-96,429,884) in sequencing reads extending beyond the 3' end.

As methylation prior to end processing was observed using a candidate approach, a global approach was used in which all tRNAs that exhibited candidate m^{5}C sites based on the global mapping were examined for methylation prior to processing. The sequencing reads from the global mapping of the NTC sample were divided into two subsets; one containing no mismatches, and those containing ≥1 T-to-C mismatch corresponding to a candidate m^{5}C site. Mean coverage plots of the two subsets of tRNAs were generated, which were anchored at either the 5' or 3' end of the tRNA. Modest coverage of sequencing reads that did not contain mismatches or candidate m^{5}C sites was observed beyond the 5' end of tRNAs, and coverage of sequencing reads containing ≥1 T-to-C mismatch beyond the 5' end of the tRNA was also evident (Figure

4.8). This was also the case for the 3' end (Figure 4.9). Thus, this suggests that m^{5}C in tRNAs can occur prior to 5' and 3' end processing.

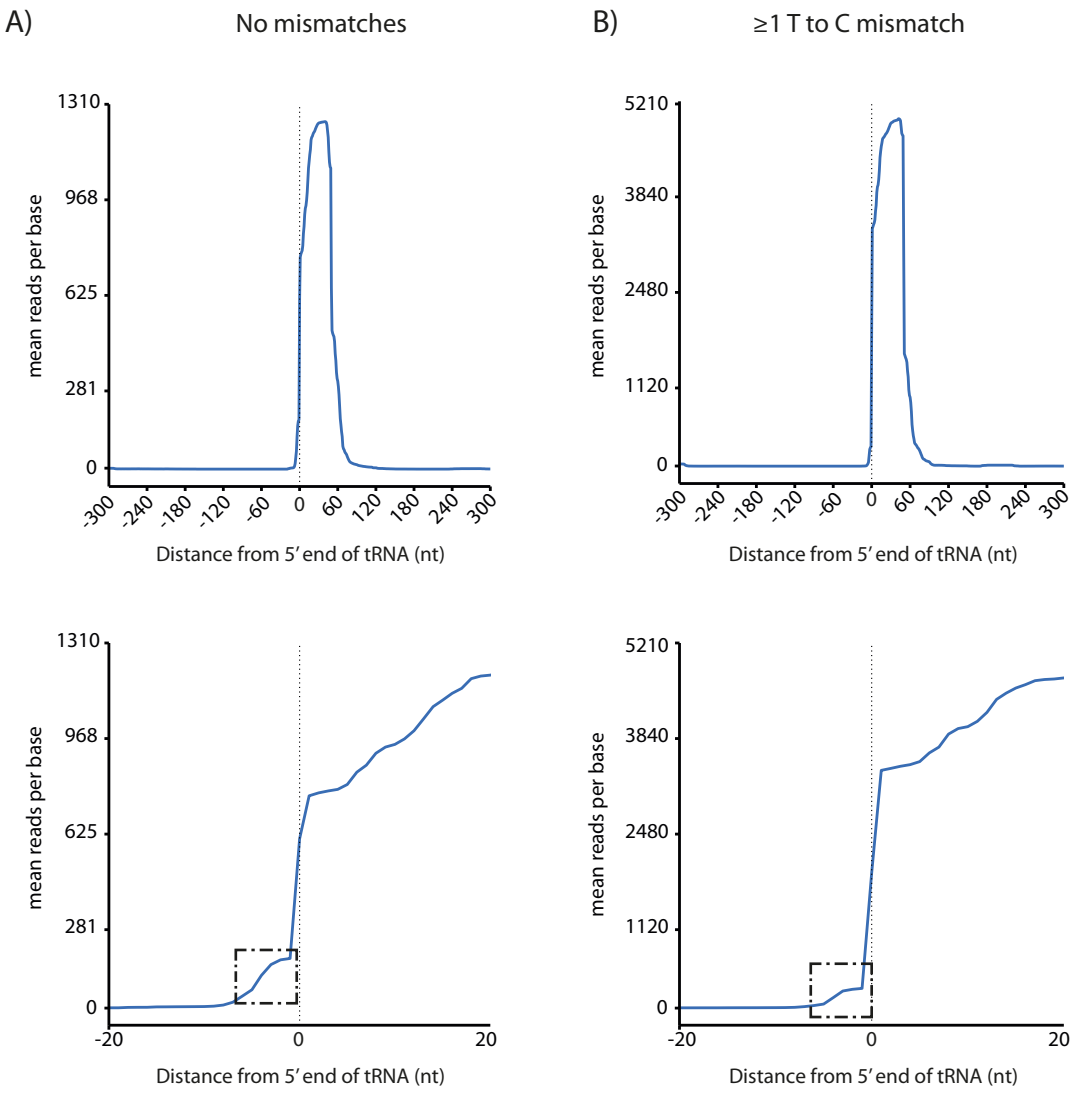


Figure 4.8. Mean read coverage of all tRNAs containing candidate m^{5}C sites anchored at the 5' end of the mature tRNA in the NTC sample.

All tRNAs found to contain candidate m^{5}C sites based on the global mapping were divided into two subsets containing either A) no mismatches or B) ≥1 T-to-C mismatch. Mean coverage plots were anchored at the 5' end of the mature tRNA sequences. The bottom panels show a zoomed view to 20 nt flanking either side of the 5' end of the tRNA. Sequencing reads extending beyond the 5' end of the tRNA are shown in the dotted rectangles.

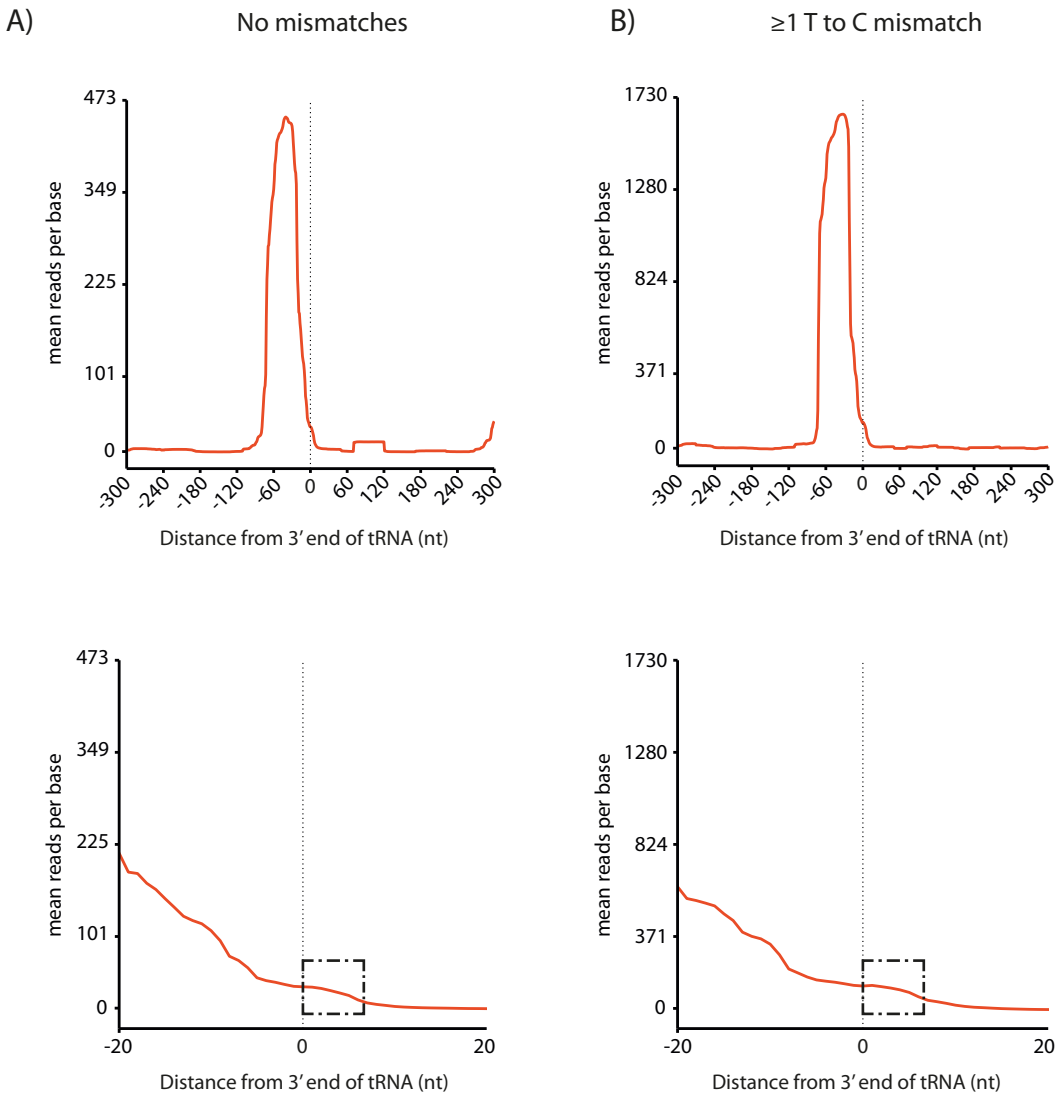


Figure 4.9. Mean read coverage of all tRNAs containing candidate m^{5}C sites anchored at the 3' end of the mature tRNA in the NTC sample.

All tRNAs found to contain candidate m^{5}C sites based on the global mapping were divided into two subsets containing either A) no mismatches or B) ≥1 T-to-C mismatch. Mean coverage plots were anchored at the 3' end of the mature tRNA sequences. The bottom panels show a zoomed view to 20 nt flanking either side of the 3' end of the tRNA. Sequencing reads extending beyond the 3' end of the tRNA are shown in the dotted rectangles.

4.2.3 m^{5}C IN OTHER ncRNA

4.2.3.A m^{5}C IN rRNA

40-75% of the sequencing reads from the bsRNA-seq mapped to rRNA (see section 3.2.3.A). It is well known that no depletion method will completely remove rRNA and thus the bsRNA-seq data was mined for candidate m^{5}C sites in rRNA. Applying our usual threshold criteria, a total of 24 candidate m^{5}C sites were detected by bsRNA-seq in nuclear-encoded rRNA and 5 were detected in mitochondrial-encoded rRNA. Of the 24 candidate m^{5}C sites detected in nuclear-encoded rRNA, 20 were in 28S rRNA, 3 were in 18S rRNA and 1 was in 5S rRNA. The previously identified m^{5}C sites at C3782 and C4447 in 28S rRNA were among these [2, 58, 59, 92, 273]. The coverage surrounding these known sites was substantial, with >700 reads in each sample. Additionally, >85% methylation was detected at both sites in all samples, suggesting that NSUN2 or TRDMT1 does not mediate these sites, in agreement with our previous study (Figure 4.10A) [58]. Interestingly, the m^{5}C sites at these positions are equivalent to the m^{5}C sites at C2278 and C2870 in yeast 25S rRNA, which are mediated by Rcm1 [86, 92] and Nop2 [92], respectively (see section 1.3). The human orthologues of Rcm1 and Nop2 are NSUN5A and NSUN1, respectively. Although catalytic activity of these human RNA:m^{5}C methyltransferases have not been confirmed, it is possible that they mediate the m^{5}C sites in 28S rRNA.

Analysis of the rRNA-mapped reads is more likely to be affected by non-conversion artefacts due to inefficient denaturation, as rRNA folds into particularly strong and extensive structures. Thus, we applied additional thresholds to further enrich for genuine sites. As both known m^{5}C sites exhibited substantial coverage and methylation in all samples, ≥500 reads and ≥70% methylation was set as additional thresholds for candidate m^{5}C sites in rRNA. After applying these thresholds, one novel candidate m^{5}C site remained, which exhibited >800 reads and >85% methylation in all samples. Examination of the sequencing reads surrounding the one novel candidate m^{5}C site revealed 6 sites of non-conversion within a 10 nt window, suggesting that even this candidate is a result of inefficient denaturation of the strong secondary structure.

Of the 5 candidate m^{5}C sites detected in mitochondrial-encoded rRNA, 4 were in 12S mitochondrial rRNA and 1 was in 16S mitochondrial rRNA. The known m^{5}C site at C841 in 12S mitochondrial rRNA [58, 255] was detected by bsRNA-seq, exhibiting ≥100 reads and ~40% methylation in all samples. This suggests that NSUN2 or TRDMT1 does not mediate this site, in agreement with our previous bsRNA-seq study (Figure 4.10B) [58]. It was recently reported that NSUN4 mediates m^{5}C at the equivalent position in mice, thus it is highly likely that NSUN4 mediates this site in humans [97]. As with nuclear-encoded rRNAs, additional thresholds were applied to enrich for genuine sites. The thresholds used were ≥100 reads and ≥20% methylation in all samples, and no candidate sites remained after applying these thresholds. In agreement, the two candidate m^{5}C sites detected upstream of the known site in 12S mitochondrial rRNA were not validated by MiSeq amplicon sequencing (performed by Mr. Andrew Shafik- data not shown), suggesting that these two sites were a result of non-conversion due to incomplete denaturation (Figure 4.10B). Similarly, the single candidate m^{5}C site detected at 16S mitochondrial rRNA by bsRNA-seq was also not validated (performed by Mr. Andrew Shafik- data not shown).

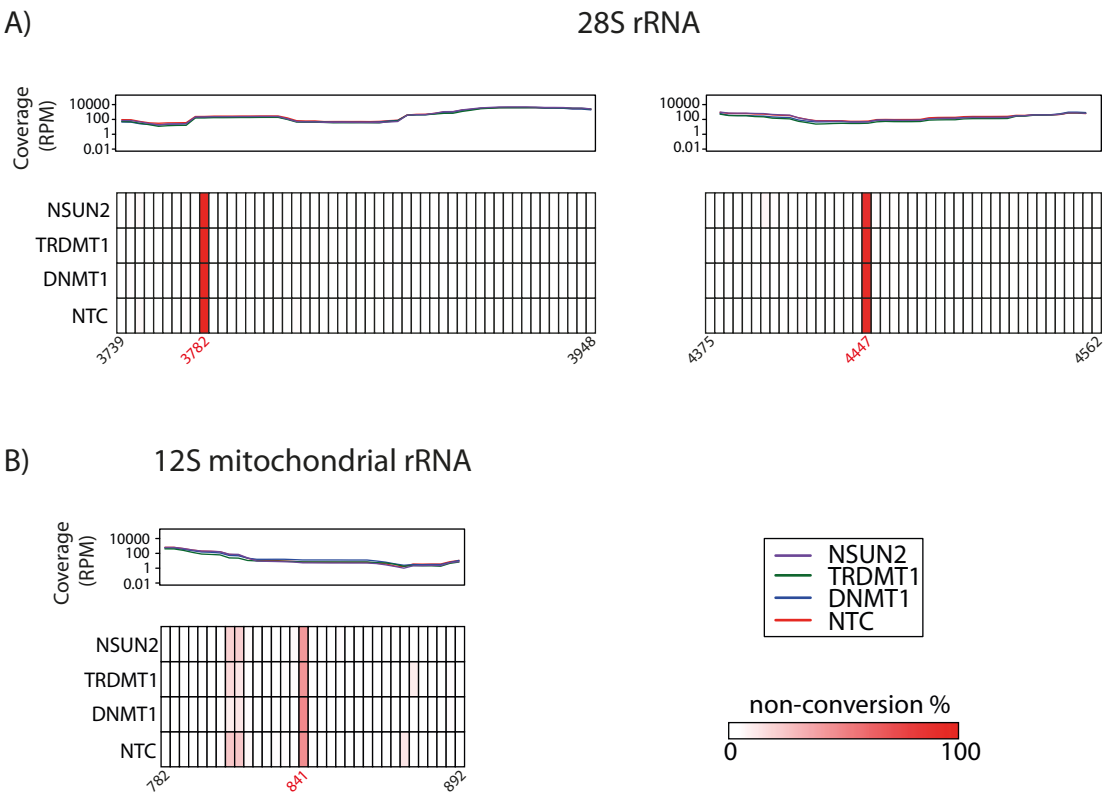


Figure 4.10. The coverage and proportion of m5C at known sites in 28S rRNA and 12S mitochondrial rRNA.

28S rRNA contains two known m5C sites at C3782 and C4447 and 12S mitochondrial rRNA contains one known m5C sites at C841. The top panel displays the coverage in RPM based on bsRNA-seq and the bottom panel displays the proportion of non-conversion based on bsRNA-seq. The locations of the m5C sites relative to the start of the transcript are indicated in red below the heatmap.

In summary, bsRNA-seq clearly detects the three known m5C sites in rRNA species, in addition to suggesting several additional candidate sites. Applying more conservative thresholds for such sites, as well as independent tests using MiSeq amplicon sequencing suggests that none of the additional sites are genuine. This highlights a residual sensitivity of bsRNA-seq to incomplete RNA denaturation.

4.2.3.B m5C in other ncRNA

643 candidate m5C sites were detected in a wide variety of ncRNAs in HeLa cells by bsRNA-seq, which encompassed miRNA, snoRNA, 7SK RNA, 7SL RNA, uRNA, vtRNA and other types of ncRNA. Of this, 30 were detected as NSUN2 targets and 9 were detected as TRDMT1 targets. As NSUN2 was found to target more sites, a focus was placed on ncRNAs targeted by this enzyme.

Five ncRNAs exhibiting candidate m5C sites detected as NSUN2 targets were selected for validation by MiSeq amplicon sequencing. These included the ncRNA component of RNase P (*RPPH1*), Small Cajal Body-Specific RNA (*SCARNA2*), snoRNA U62 (*SNORD62*), *NSUN5* pseudogene 2 (*NSUN5P2*) and a pseudogene located within the intron of *CAGE1*, all of which contained a single candidate m5C site. *RPPH1* is the RNA component of RNase P, which is involved in the cleavage of the 5’ end of pre-tRNAs along with a number of protein subunits [274], and has been reported to be required for efficient RNA Polymerase III transcription [275]. *SCARNA2* and *SNORD62* are C/D box guide RNAs that are part of small nucleolar ribonucleoproteins (snoRNPs) predicted to guide 2’-*O*-methylation of G25 and C61 in U2 snRNA [276] and A590 in 18S rRNA [277, 278], respectively. The function of *NSUN5P2* and the pseudogene located in the intron of *CAGE1* remain unknown.

We have previously discovered and validated the m^{5}C site in *RPPH1* [58], and two other studies have also validated this site [77, 78]. The m^{5}C site in *SCARNA2* has been previously discovered and validated in HeLa cells by Aza-IP [77]. The candidate m^{5}C sites detected in *SNORD62* and the pseudogenes were entirely novel. The proportion of methylation detected in each transcript varied from ~50%-90% in the TRDMT1 knockdown, DNMT1 knockdown and NTC samples, and almost complete depletion of methylation was evident in the NSUN2 knockdown samples by bsRNA-seq and MiSeq amplicon sequencing, indicating that these sites were mediated by NSUN2 (Figure 4.11). Thus, it is clear that m^{5}C occurs in various ncRNA, including snoRNAs, lincRNAs and pseudogenes, and at least a subset of these are mediated by NSUN2.

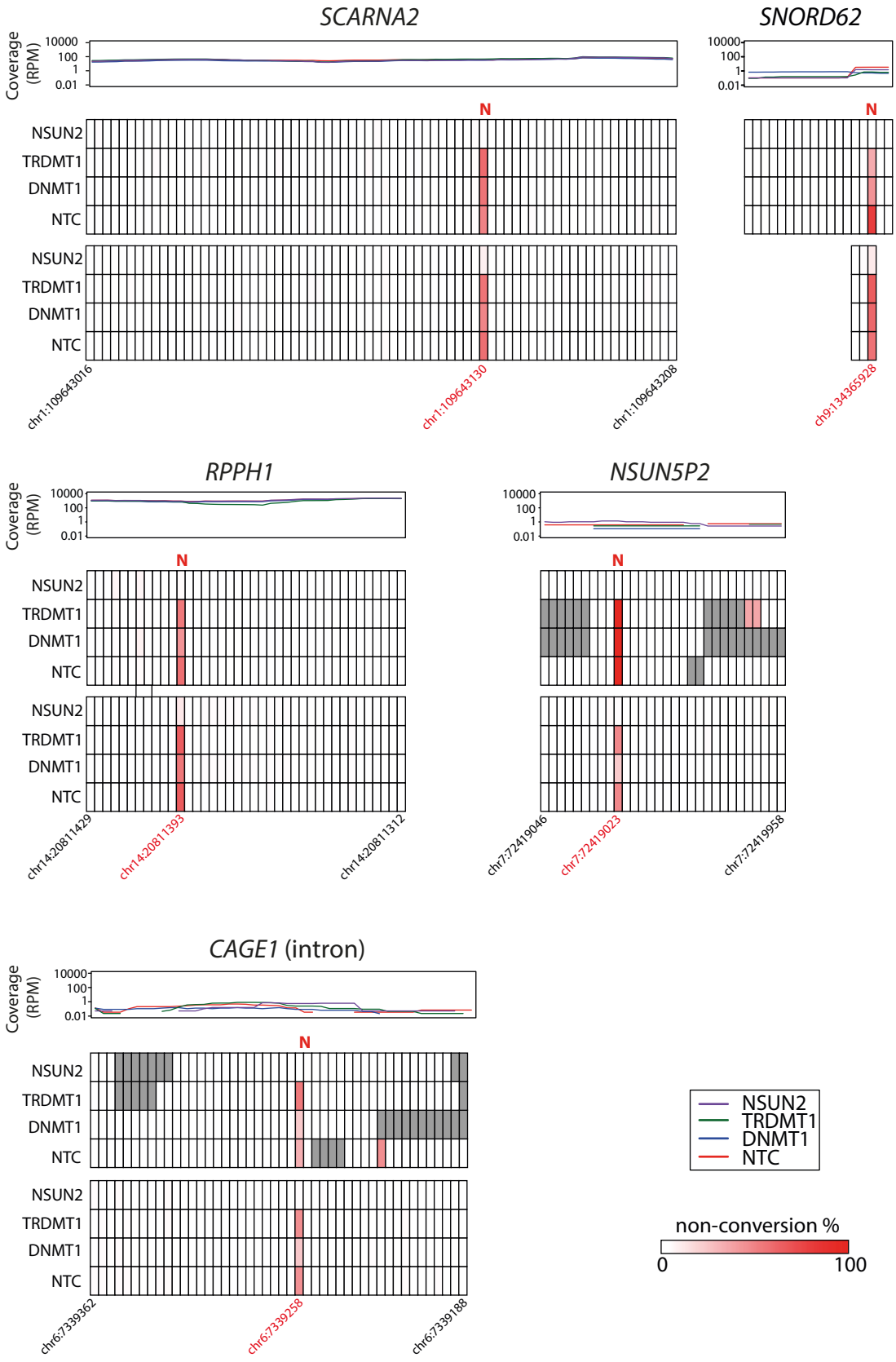

SCARNA2
SNORD62
RPPH1
NSUN5P2
CAGE1 (intron)
Coverage (RPM)
10000
100
1
0.01
N
NSUN2
TRDMT1
DNMT1
NTC
chr1:109643016
chr1:109643130
chr1:109643208
chr9:134365928
chr14:20811429
chr14:20811393
chr14:20811312
chr7:72419046
chr7:72419023
chr7:72419958
chr6:7339362
chr6:7339258
chr6:7339188
non-conversion %
0
100

Figure 4.11. Validation of candidate m^{5}C sites in other ncRNA by MiSeq amplicon sequencing.

The top panel displays the coverage in RPM based on bsRNA-seq, the middle panel displays the proportion of non-conversion based on bsRNA-seq, and the bottom panel displays the proportion of non-conversion based on MiSeq amplicon sequencing. The shades of grey in the middle panel indicate regions of the transcript where coverage was <5 reads based on bsRNA-seq. Chromosomal locations of the candidate m^{5}C sites are indicated in red below the heatmap. The N and T indicate sites that were called as NSUN2 or TRDMT1 targets based on bsRNA-seq, respectively. The MiSeq coverage ranged from 40-110,000 mappable sequencing reads.

4.2.3.C Sequence and structural context of m^{5}C in other ncRNA

Several studies have investigated the role of NSUN2 in mediating m^{5}C sites in various ncRNA, although limited research has been performed to understand the sequences and structures that NSUN2 recognises to mediate m^{5}C. Due to the low number of candidate m^{5}C sites detected in each class of ncRNA, a global approach to investigate this was not applied here. Instead, a focus was placed on the context features around m^{5}C sites in ncRNAs that have been validated as NSUN2 targets in this study or previous studies. To gain insight into the context features of m^{5}C in the validated ncRNAs, secondary structures were obtained from the RNase P database [279] Rfam database [280] or RNA STRAND database [281]. Secondary structures were obtained for *RPPH1*, *SNORD62* and *SCARNA2*, but were not available for the pseudogenes (Figure 4.12A-C). The known features and domains for these ncRNAs are indicated in Figure 4.12.

The m^{5}C site in *RPPH1* is located in the P12 domain, which has been proposed to be important for efficient transcription [274], suggesting that this site may regulate the role of *RPPH1* in transcription. The site is located in a stem-loop junction in a GCA context, a structure that is similar to that exhibited in the variable loop junction around C48-50 in tRNAs that NSUN2 targets.

The canonical C/D boxes and novel C'/D' snoRNA-like boxes in *SCARNA2* and *SNORD62* and the predicted binding sites of the substrate RNAs are indicated in Figure 4.12B and C [276, 277]. The m^{5}C site in *SNORD62* is located within the loop of a stem-

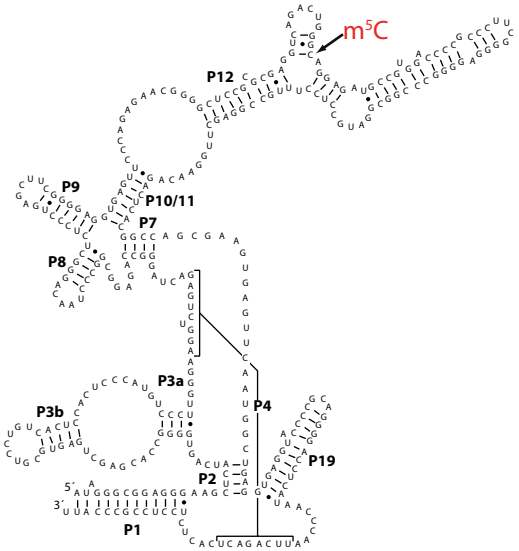
loop structure, and as with *RPPH1* is in a GCA context (Figure 4.12B). The m^{5}C site in *SCARNA2* is located a few nucleotides upstream of a stem-loop structure in a CpG context (Figure 4.12C). Interestingly, the m^{5}C sites in both guide RNAs are not located at the C/D boxes, C'/D' boxes or the predicted binding site of the substrate RNAs, suggesting that m^{5}C may not directly affect the modification of substrate RNAs of *SCARNA2* and *SNORD62*.

Secondary structures were also obtained for two vtRNAs, vtRNA1-1 and vtRNA1-3, which were detected as NSUN2 targets by bsRNA-seq (Figure 4.12D and E). Although independent validation of the candidate m^{5}C sites in the vtRNAs was not attempted in this thesis, both have been previously reported and validated as NSUN2 targets [77, 78]. vtRNAs are evolutionarily conserved non-coding transcripts present in vault particles of eukaryotic cells [282]. These are cytoplasmic ribonucleoproteins of which their function remains elusive. In both cases, the candidate m^{5}C site was located in the stem near the stem-loop junction. This structure is similar to that exhibited in tRNAs around positions C48-50 that NSUN2 targets and C30 in the anti-codon loop junction that NSUN2 was also found to target in this thesis. The candidate m^{5}C sites in both vtRNAs were in a TCG context, which has been previously noted [78]. Interestingly, Hussain et al. recently reported that loss of NSUN2-mediated m^{5}C sites in vtRNAs caused aberrant processing into small vtRNA fragments, termed svRNAs [78].

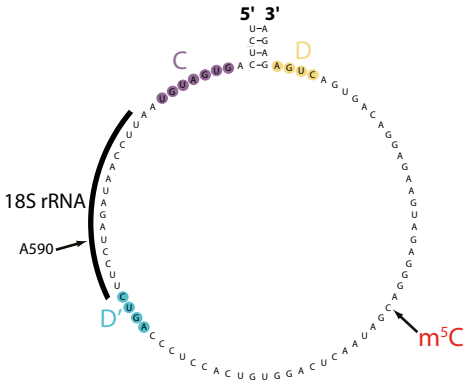
Although secondary structures for *NSUN5P2* and the pseudogene located within the *CAGE1* were not available, the m^{5}C sites were in a CpA and CpT context respectively.

In summary, NSUN2 can recognise several different sequence and structural motifs in various ncRNA. Several of these structures resemble tRNA structural positions, however it is clear that NSUN2 is capable of recognising structures different to those exhibited in tRNAs, highlighting the multisite specificity of this enzyme.

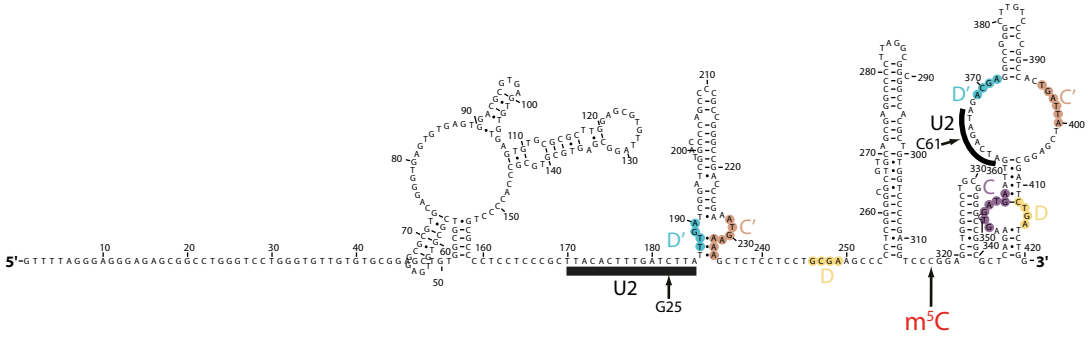
A) *RPPH1*



B) *SNORD62*



C) *SCARNA2*



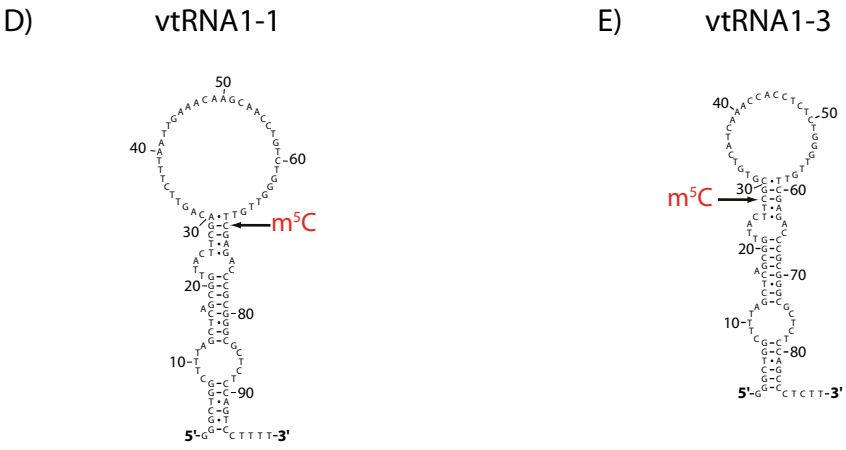


Figure 4.12. Secondary structures of ncRNAs with validated m^{5}C sites obtained either from the RNase P database, Rfam database or RNA STRAND.

A) *RPPH1*, B) *SNORD62* and C) *SCARNA2* were detected and validated in this study. The domains in *RPPH1* and C/D box domains and C'/D' box snoRNA-like domains are indicated. The binding sites of the substrate RNAs that *SCARNA2* and *SNORD62* are predicted to guide are also indicated in black. D) vtRNA1-1 and E) vtRNA1-3 were detected by bsRNA-seq but validated by two published studies. The m^{5}C sites are indicated in red.

4.2.4 m^{5}C in mRNA

The distribution and abundance of m^{5}C in mRNA has recently come to the fore with our previous bsRNA-seq study [58]. However, an understanding of the substrate specificity of NSUN2 in mRNAs has remained elusive. The miCLIP study detected 312 mRNA targets of NSUN2, however this was not followed up in any detail [78]. The Aza-IP study detected only a single mRNA target of TRDMT1 [77]. As a result, this motivated further investigation into the distribution and structural and sequence context of m^{5}C in mRNA.

4.2.4.A Distribution and location bias of m^{5}C in mRNA

Candidate m^{5}C sites in mRNA in HeLa cells were distributed throughout all regions of mRNA; 38% of the candidate m^{5}C sites were in the 3' UTR, 37% in the CDS, 23% in introns and 3% in the 5' UTR (Figure 4.13A). To assess the location bias of candidate m^{5}C sites in the 3' UTR, 5' UTR and CDS, the observed number of candidate m^{5}C sites in these regions was compared with the expected number of candidate m^{5}C sites detected in each region for cytosines with ≥5 reads. Candidate m^{5}C sites were enriched

in the 3' UTR ($p=1.1x10^{-34}$, binomial test) and depleted in the 5' UTR ($p=1.0x10^{-19}$, binomial test) and CDS ($p=1.3x10^{-10}$, binomial test) than expected at random (Figure 4.13B). This highlights the non-random distribution of m^{5}C in mRNA.

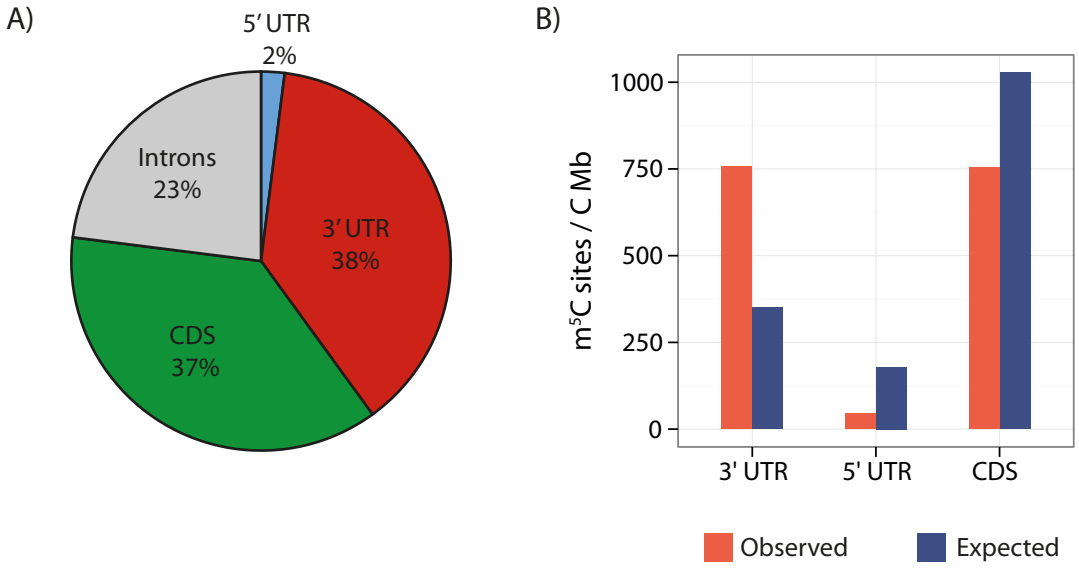


Figure 4.13. The distribution and location bias of candidate m^{5}C sites in mRNA.

The distribution and enrichment of candidate m^{5}C sites was determined for the total number of candidate m^{5}C sites detected in mRNAs in HeLa cells. A) Distribution of candidate m^{5}C sites in different regions of mRNA. B) Analysis of location bias of candidate m^{5}C sites within mRNAs. The observed number of candidate m^{5}C sites in the 3' UTR, 5' UTR and CDS in HeLa cells was compared with the expected number of candidate m^{5}C sites for cytosines ≥5 reads. Candidate m^{5}C sites were enriched in the 3' UTR ($p=1.1x10^{-34}$, binomial test) and depleted in the 5' UTR ($p=1.0x10^{-19}$, binomial test) and CDS ($p=1.3x10^{-10}$, binomial test).

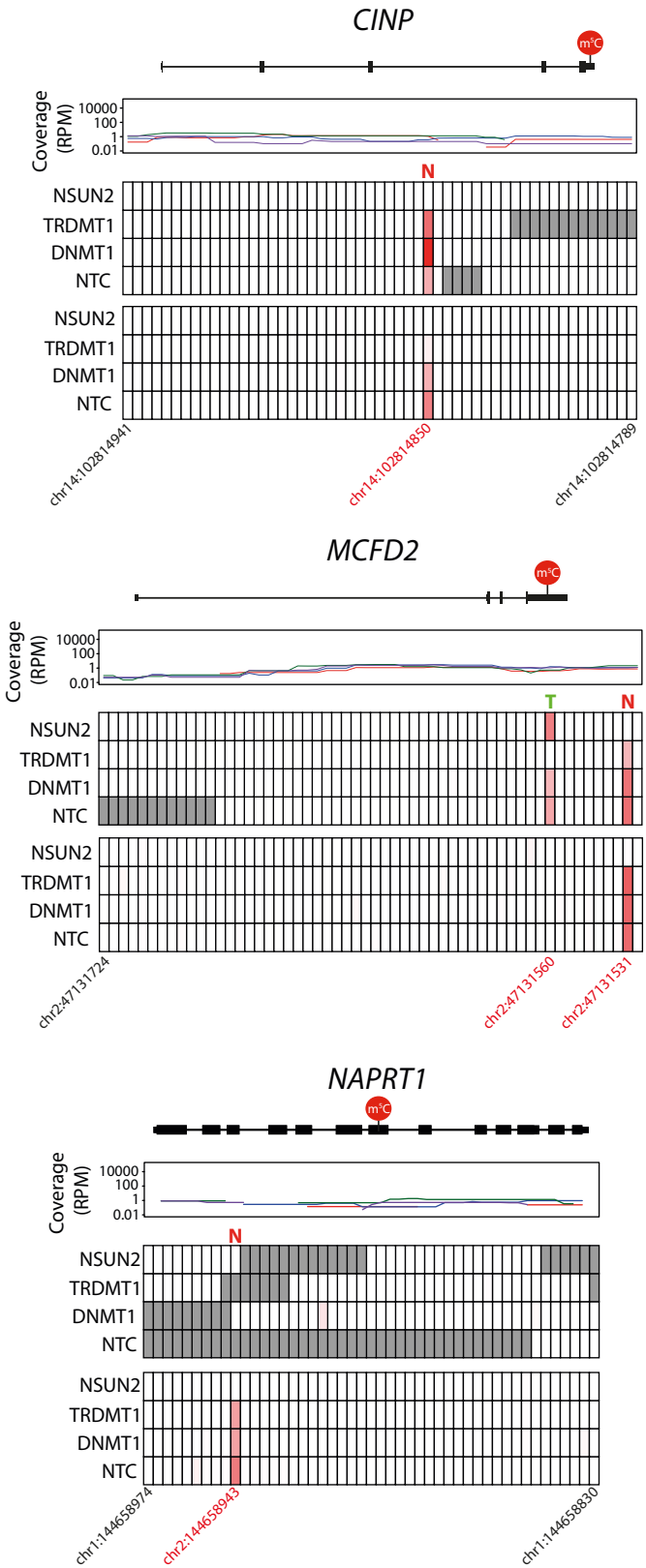
4.2.4.B Validation of m^{5}C sites in mRNA

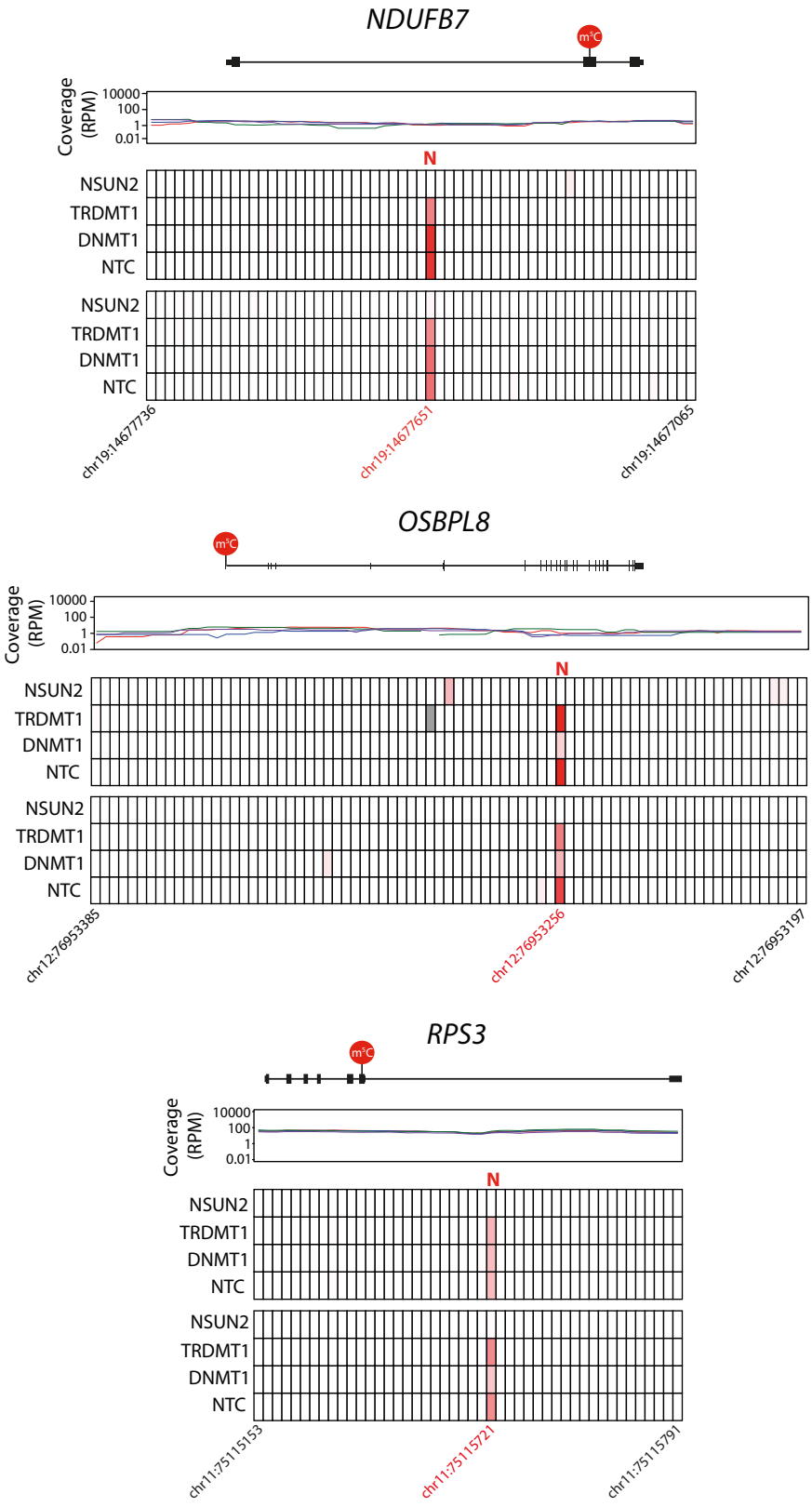
In order to validate that NSUN2 mediates m^{5}C in mRNA, seven novel candidate m^{5}C sites and three known m^{5}C sites across nine mRNAs were selected for independent validation by MiSeq amplicon sequencing. A focus was made on candidate m^{5}C sites targeted by NSUN2 in different regions of the mRNA, including the 5' UTR, 3' UTR and CDS.

Sites in three mRNAs, *CINP*, *NAPRT1* and *NDUFB7*, have been validated by Sanger bisulfite sequencing in our previous study [58]. bsRNA-seq again correctly called the known m^{5}C site in *NDUFB7* as a NSUN2 target. Interestingly, the known m^{5}C site in *CINP* was not called as a NSUN2 target in the present study. The site was detected by bsRNA-seq in HeLa cells, however the NSUN2 knockdown sample did not have sufficient coverage across the site to meet our NSUN2 target selection criteria. The known m^{5}C site in *NAPRT1* had insufficient coverage for most samples in this study, and as such was not detected. All three known sites were selected for independent validation using MiSeq amplicon sequencing.

Of the seven novel candidate m^{5}C sites selected for validation, four were called as NSUN2 targets (*MCFD2*, *RPS3*, *SCO1* and *SZRD1*), and two were not assigned to either methyltransferase (*OSBPL8* and *RTN3*) based on bsRNA-seq. Interestingly, *MCFD2* exhibited two candidate m^{5}C sites in close proximity to each other, one called as a NSUN2 target and one as a TRDMT1 target, and as such both candidate sites were examined. Of the ten sites examined, nine were validated, and depletion of methylation was evident in the NSUN2 knockdown sample indicating that these m^{5}C sites were mediated by NSUN2 (Figure 4.14). Interestingly, the candidate m^{5}C sites in *OSBPL8* and *RTN3* were validated as NSUN2-mediated sites despite not being called as a target of NSUN2 or TRDMT1 by bsRNA-seq. The candidate m^{5}C site in *MCFD2* called as a TRDMT1 target did not validate in this study. Examination of the sequencing reads spanning this candidate m^{5}C site in *MCFD2* revealed that this site was of low quality due to uncertainty in the colour calling by the SOLiD^{TM} platform. In contrast, the sequencing reads that did not exhibit the candidate m^{5}C sites were of higher quality (data not shown).

Overall, the MiSeq results emphasise the importance of NSUN2 as a mRNA modifying enzyme. Proper assessment of whether TRDMT1 does so too would require independent validation of its candidate sites.




NDUFB7
m5C
Coverage (RPM)
10000
100
1
0.01
N
NSUN2
TRDMT1
DNMT1
NTC
NSUN2
TRDMT1
DNMT1
NTC
chr19:14677736
chr19:14677651
chr19:14677065
OSBPL8
m5C
Coverage (RPM)
10000
100
1
0.01
N
NSUN2
TRDMT1
DNMT1
NTC
NSUN2
TRDMT1
DNMT1
NTC
chr12:76953385
chr12:76953256
chr12:76953197
RPS3
m5C
Coverage (RPM)
10000
100
1
0.01
N
NSUN2
TRDMT1
DNMT1
NTC
NSUN2
TRDMT1
DNMT1
NTC
chr11:75115153
chr11:75115721
chr11:75115791

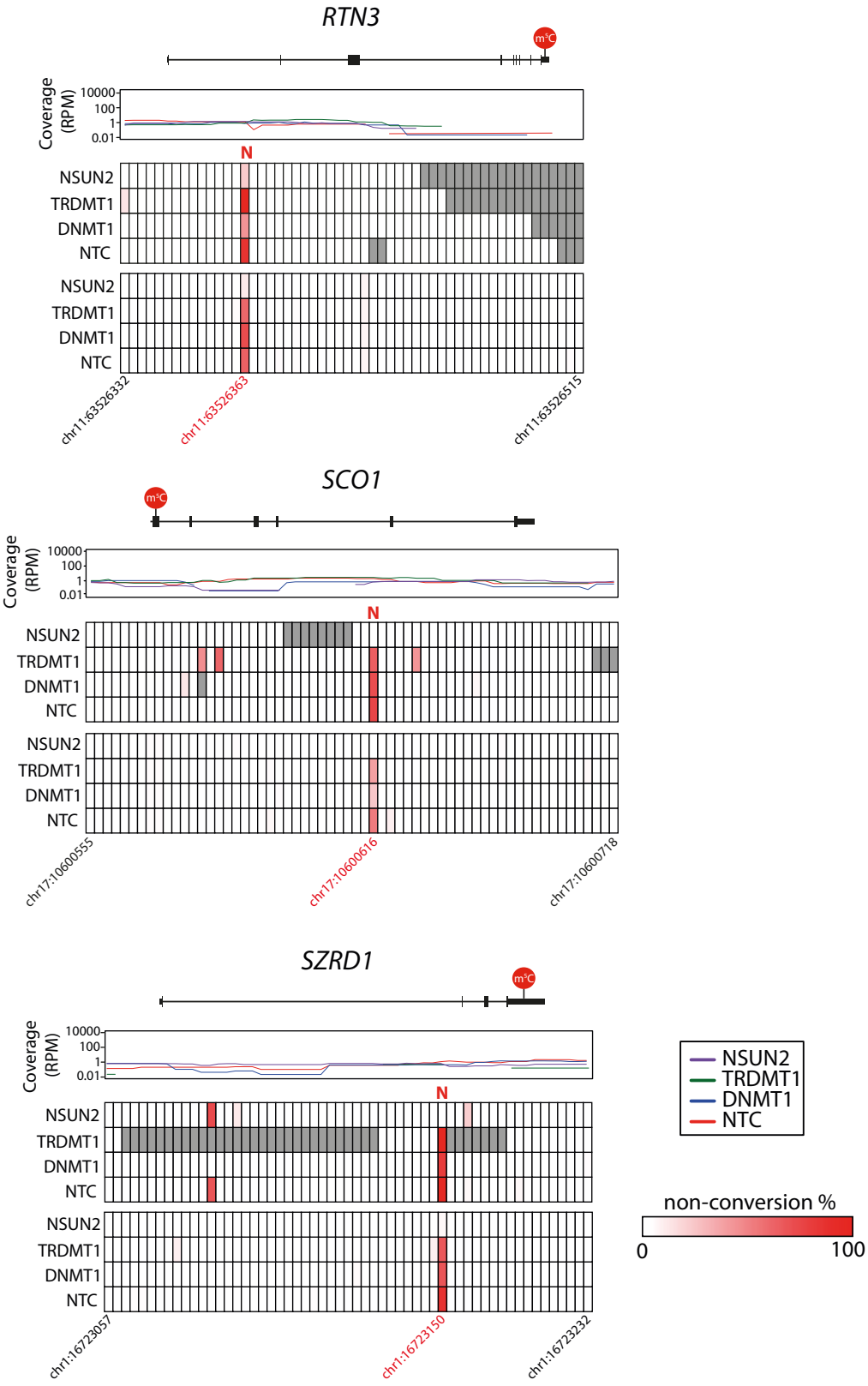

RTN3
m5C
Coverage (RPM)
10000
100
1
0.01
N
NSUN2
TRDMT1
DNMT1
NTC
NSUN2
TRDMT1
DNMT1
NTC
chr11:63526332
chr11:63526363
chr11:63526515
SCO1
m5C
Coverage (RPM)
10000
100
1
0.01
N
NSUN2
TRDMT1
DNMT1
NTC
NSUN2
TRDMT1
DNMT1
NTC
chr17:10600555
chr17:10600616
chr17:10600718
SZRD1
m5C
Coverage (RPM)
10000
100
1
0.01
N
NSUN2
TRDMT1
DNMT1
NTC
NSUN2
TRDMT1
DNMT1
NTC
chr1:16723057
chr1:16723150
chr1:16723232
NSUN2
TRDMT1
DNMT1
NTC
non-conversion %
0
100

Figure 4.14. Validation of candidate m^{5}C sites in mRNA by MiSeq amplicon sequencing. The location of the m^{5}C site relative to the regions in the mRNA is indicated at the top. The top panel displays the coverage in RPM based on bsRNA-seq, the middle panel displays the proportion of non-conversion based on bsRNA-seq, and the bottom panel displays the proportion of non-conversion based on MiSeq amplicon sequencing. The shades of grey in the middle panel indicate regions of the transcript where coverage was <5 reads based on bsRNA-seq. Chromosomal locations of the candidate m^{5}C sites are indicated in red below the heatmap. The N and T indicate sites that were called as NSUN2 or TRDMT1 targets based on bsRNA-seq, respectively. The MiSeq coverage ranged from 4,000-110,000 mappable sequencing reads.

4.2.4.C Structural and sequence context of m5C in mRNA

In order to gain more insight into the structural and sequence context of m5C in mRNA and the substrate specificity of NSUN2, a preliminary investigation was made using the bsRNA-seq data. As there was a substantial number of candidate m5C sites detected in mRNAs in HeLa cells and >100 candidate m5C sites were called as NSUN2 targets, a computational search for enriched structural and sequence motifs was performed using the bsRNA-seq data.

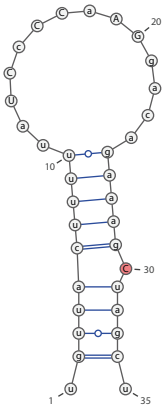
Structural context

The structural context of m5C in mRNAs was investigated using CMFinder [241], with regions flanking 100 nt either end of subsets of candidate m5C sites in mRNAs. The structural context of m5C was investigated for all candidate m5C sites in mRNAs in HeLa cells, as well as candidate m5C sites called as NSUN2 targets. The final structures that were selected required significant enrichment, >30% of candidate m5C sites showing the structure and the distribution of candidate m5C sites localised on or surrounding the structural motif.

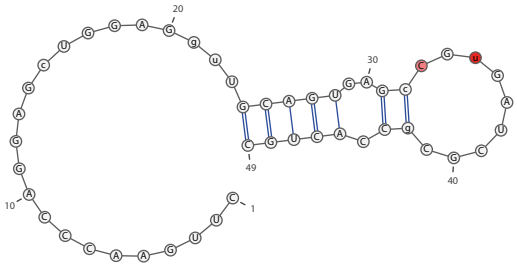
No significant structural motif was evident when investigating sites in all regions of the mRNA together. For candidate m5C sites specifically in the CDS, a stem-loop motif was evident, with candidate m5C sites predominantly in a bulge midway along the stem (Figure 4.15A). Interestingly, candidate m5C sites in the 3' UTR also yielded a stem-loop motif, with sites predominantly located in the loop (Figure 4.15B).

Several stem-loop motifs were enriched for candidate m5C sites targeted by NSUN2, with sites located either in the loop or the stem (Figure 4.15C-E). These enriched structures are similar to those detected for candidate m5C sites in ncRNAs found in this study. For example, the candidate m5C sites located in the loop of a stem-loop in mRNAs are reminiscent of structures exhibiting the m5C site in *SNORD62*.

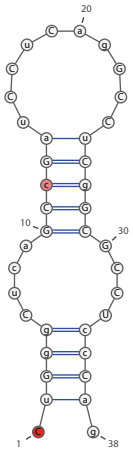
A) CDS



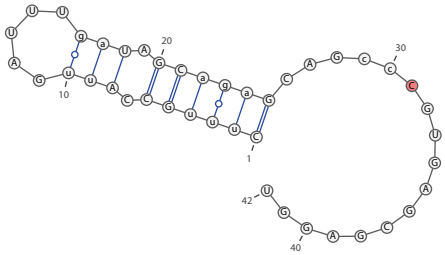
B) 3′ UTR



C) NSUN2



D) NSUN2



E) NSUN2

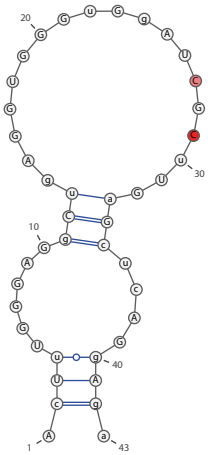


Figure 4.15. Significant structural motifs identified by CMFinder for candidate m^{5}C sites detected in mRNA.

The red bases represent locations where candidate m^{5}C sites were predominantly detected, with dark red indicating more enriched locations. Upper case letters indicate high frequency bases occurring at that location. Note that candidate m^{5}C sites are not necessarily indicated by a "C" as other bases may be found there. A) All candidate m^{5}C sites detected in the CDS, B) All candidate m^{5}C sites detected in the 3' UTR, C)-E) Candidate m^{5}C sites in mRNAs called as NSUN2 targets.

Sequence context

To determine the sequence context of m^{5}C in mRNA, sequence logos were generated using the seqLogo R package [242]. For the generation of sequencing logos, the nucleotides at each position were scaled according to their frequency and the entire stack of nucleotides at a given position was represented as the information content. When displayed, the bases were stacked on top of each other in increasing order of their frequencies. The sequence logos were generated for 25 nt flanking regions of candidate m^{5}C sites in mRNAs, and were determined for all candidate m^{5}C sites detected in mRNA in HeLa cells and candidate m^{5}C sites in mRNA called as NSUN2 targets. As a comparison, sequence logos were also generated for candidate m^{5}C sites in tRNAs called as NSUN2 targets. Furthermore, as a precaution and to remove any bias associated with aligning the bsRNA-seq data to non-CpG *in silico* bisulfite converted reference sequences, this analysis was performed on candidate m^{5}C sites detected by aligning the sequencing reads to complete *in silico* bisulfite converted reference sequences only.

In the case of all candidate m^{5}C sites detected in HeLa cells, the position directly downstream of the candidate sites displayed no base preference, although the prevalence of a string of three Gs in the following positions was evident (Figure 4.16A). For candidate m^{5}C sites called as NSUN2 targets, the preference for Gs at positions 2-4 nt downstream was even stronger, and a mild preference for m^{5}CpG or m^{5}CpA over m^{5}CpC or m^{5}CpT was evident (Figure 4.16B). Interestingly, the sequence logo generated for candidate m^{5}C sites in tRNAs called as NSUN2 targets also revealed a string of three Gs further downstream of the site, although in this case m^{5}C

predominantly occurred in a CpC or CpG context (Figure 4.16C). The preference for m^{5}CpC and m^{5}CpG in tRNAs is most likely influenced by m^{5}C sites at C48-50 in tRNAs, where Cs and Gs are commonly located. The string of Gs evident in tRNAs is most likely influenced by the nucleotides at positions 51-53, of which ~60% tRNAs exhibiting candidate m^{5}C sites in this thesis contained Gs at all three positions (see Appendix 8.5 for the tRNA Rfam structural alignment).

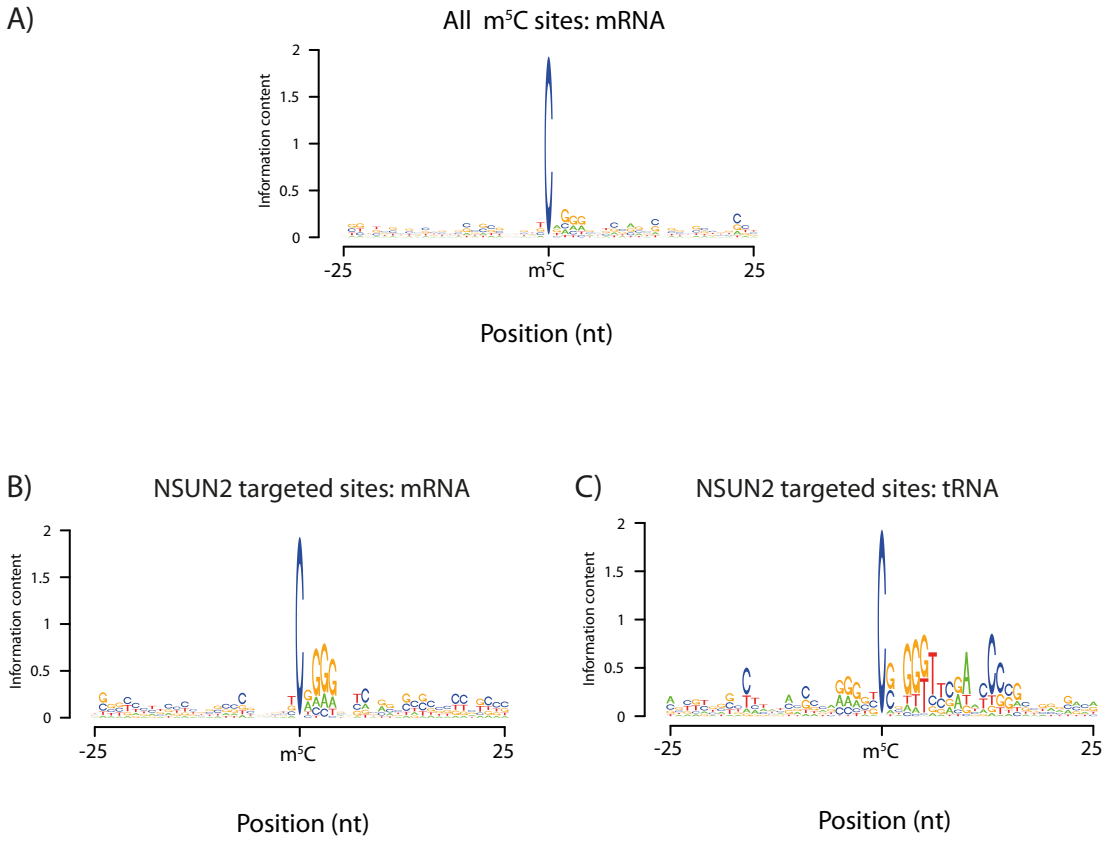


Figure 4.16. Sequence logos surrounding candidate m^{5}C sites in mRNA and tRNA.
Sequence logos were generated using the seqLogo R package for 25 nt regions flanking the candidate m^{5}C site. A) All candidate m^{5}C sites in mRNAs, B) Candidate m^{5}C sites in mRNAs called as NSUN2 targets, C) Candidate m^{5}C sites in tRNAs called as NSUN2 targets.

The trend of NSUN2-mediated m^{5}C sites in mRNAs occurring in a CpG and CpA context was also reflected in the mRNAs that were validated by MiSeq amplicon sequencing in this study. Out of the nine m^{5}C sites validated as NSUN2 targets, three were in a CpG context (*OSBPL8*, *SCO1*, *SZRD1*), five were in a CpA context (*CINP*, *MCFD2*, *NAPRT1*, *NDUFB7*, *RTN3*), and one was in a CpT context (*RPS3*) (Table 4.3). Additionally, seven of the validated candidates exhibited at least two Gs positions 2-4 downstream of the m^{5}C site, in concordance with the sequence logo.

Table 4.3. Sequence context of m^{5}C sites discovered and validated as NSUN2 targets in this thesis.

mRNA	Sequence downstream of m^{5}C site
CINP	m^{5}CGGGC
MCFD2	m^{5}CAGGT
NAPRT1	m^{5}CAGGT
NDUFB7	m^{5}CAAGC
OSBPL8	m^{5}CGGGG
RPS3	m^{5}CTGGG
RTN3	m^{5}CAGGG
SCO1	m^{5}CGAGG
SZRD1	m^{5}CGCCC

In summary, candidate m^{5}C sites in mRNAs are present around stem-loop structures, with m^{5}C sites located in both the stem and the loop. Candidate m^{5}C sites called as NSUN2 targets demonstrated a slight preference for m^{5}CpG or m^{5}CpA and there was a propensity for sites to be upstream of Gs. This, once again, highlights the multisite specificity of NSUN2.

4.3 Discussion

In this chapter, I have investigated the distribution and abundance of m^{5}C in RNA and the substrate range of NSUN2 and TRDMT1. I have extensively investigated the substrate specificity of NSUN2 and TRDMT1 in tRNAs and features of m^{5}C in tRNAs, which have both confirmed previous reports and provided novel findings. I have also explored the specificity of NSUN2 in other ncRNAs and mRNAs, which have highlighted the versatility of this enzyme. In this section, I will discuss the outcomes of these findings.

4.3.1 m^{5}C sites in HeLa cells

4.3.1.A Total m^{5}C sites in HeLa cells

In this study, a total of 4,241 candidate m^{5}C sites were detected in HeLa cells, which were distributed over tRNAs, other ncRNAs and mRNAs. Similarly, our previous bsRNA-seq study performed in HeLa cells revealed candidate m^{5}C sites distributed over tRNAs, other ncRNA and mRNAs, however the number of candidate m^{5}C sites detected was substantially more than this study, totalling to 10,530 sites [58]. Our previous bsRNA-seq study used SOCS-B to map the bsRNA-seq data, and ≥10 read coverage and ≥20% methylation were defined as thresholds to detect all candidate m^{5}C sites. In this present study, B-SOLANA and a proportion statistic were used to align the data and detect candidate m^{5}C sites. For detection of all candidate m^{5}C sites in HeLa cells, ≥5 reads in all samples used in the model and ≥10% mean methylation across the control samples were defined as the thresholds, as well as a FDR≤1%. Thus, it was expected that the variations in the mapping and candidate m^{5}C site detection methodology used for each dataset would result in differences in the number of candidate m^{5}C sites detected in RNA. While the use of a proportion statistic and more stringent criteria for the detection of candidate m^{5}C sites in this thesis may have resulted in the underestimation of candidate m^{5}C sites, it has most likely reduced the amount of false positives. Despite the different bioinformatics pipelines implemented for each study, both studies have detected candidate m^{5}C sites in many transcript biotypes, and have demonstrated that m^{5}C is abundant in mRNAs, significantly extending the known distribution of this modification in RNA.

In our previous HeLa bsRNA-seq study, tRNAs were uniquely mapped in a similar manner to the local mapping applied in this study, revealing 255 candidate m^{5}C sites, a comparable number to this thesis. The proportion of cytosines exhibiting candidate m^{5}C sites in tRNAs in our previous HeLa bsRNA-seq study was 8.7%, in comparison to 3.98% based on the local mapping in this study. Interestingly, the coverage of cytosines in this thesis was substantially higher than that detected in our previous bsRNA-seq study, resulting in the lower density of m^{5}C in tRNAs in this thesis. Furthermore, the global mapping of tRNAs yielded a total of 5.73% cytosines exhibiting candidate m^{5}C sites, somewhat higher than the local mapping. The higher density of m^{5}C in globally mapped tRNAs mapped is most likely a result of differences in the mapping. tRNAs that were mapped locally were mapped to a single representative sequence, resulting in the lower number of candidate m^{5}C sites detected and cytosines with ≥ 5 reads in the tRNAs (see Table 4.1). As the global mapping could differentiate between identical tRNA sequences due to sequencing reads extending beyond the mature ends of the tRNA in some instances, a substantially higher number of candidate m^{5}C sites were detected, although the number of cytosines with ≥ 5 reads was only double the amount detected for the local mapping.

A total of 667 candidate m^{5}C sites were detected in a wide variety of ncRNAs in this thesis. Our previous HeLa bsRNA-seq study yielded a total of 1,780 candidate m^{5}C sites in various ncRNAs. However, the density of candidate sites in both studies was comparable, with 1.4% of cytosines in ncRNA containing m^{5}C sites on average in this study, and 1.2% in the previous HeLa bsRNA-seq study.

A total of 1,564 candidate m^{5}C sites were detected in mRNAs in this study, in comparison to 8,495 candidate m^{5}C sites detected in our previous HeLa bsRNA-seq study. The proportion of cytosines in mRNA containing candidate m^{5}C sites in this study was 0.08%, lower than the 0.4% determined in our previous bsRNA-seq. The coverage of mRNAs in both studies were similar, however the stringent thresholds applied for the detection of candidate m^{5}C sites in this present study resulted in a lower number of candidate m^{5}C sites detected and therefore the lower density of m^{5}C in mRNAs. Nevertheless, the thesis revealed ~1.3 m^{5}C sites/mRNA, substantially more than previously thought based on chromatographic techniques. Investigation of hamster

cell mRNA nearly four decades ago revealed ~0.45 m5C sites/mRNA [26]. This indicates that bsRNA-seq is quite sensitive, detecting candidate m5C in mRNA at similar levels to m6A, where ~1.7 m6A peaks/mRNA were detected [57]. Both of our studies have substantially extended the previously known distribution of m5C in mRNA.

4.3.1.B m5C SITES MEDIATED BY NSUN2 AND TRDMT1

A total of 393 candidate m5C sites were detected as NSUN2 targets, and 120 as TRDMT1 targets in a wide variety of transcript biotypes, significantly extending what was previously known about the substrate specificity of these enzymes.

NSUN2 was found to be primarily responsible for m5C sites in tRNAs, in agreement with the literature. Additionally, 118 candidate m5C sites in mRNAs and 34 candidate m5C sites in various ncRNAs were called as NSUN2 targets. Recent complementary next-generation sequencing studies have reported NSUN2 to target tRNAs and some ncRNAs, but limited research has been performed to investigate mRNA substrates of NSUN2 [77, 78, 150]. The substrate range of TRDMT1 determined in this thesis was similar to that of NSUN2, but with lower numbers. The number of candidate m5C sites in tRNAs targeted by TRDMT1 was much smaller than that of NSUN2, which was expected as it has previously been shown to target only C38 in several different tRNAs [73, 77]. A total of 53 candidate m5C sites in mRNAs and 9 other ncRNAs were targeted by TRDMT1 based on the bsRNA-seq, potentially extending what was previously known about the substrate range of this enzyme. The candidate m5C sites detected in ncRNAs and mRNAs called as TRDMT1 targets would require validation, as this thesis was primarily focused on NSUN2.

Out of the 4,241 candidate m5C sites detected in HeLa cells, only 513 were called as NSUN2 or TRDMT1 targets. This is due to a combination of factors, chief of which is that we were very likely massively under-calling methyltransferase target sites as this study was performed by (incomplete) methyltransferase knockdown and it was only powered as an initial screen. Indeed, a recent study in cells from *Nsun2* and *Trdmt1* double knockout mice demonstrated that NSUN2 and TRDMT1 are responsible for all m5C sites in tRNAs in mice, and likely also in humans [149]. Thus, the ~34% of

candidate m^{5}C sites in tRNAs detected in HeLa cells that were not called as NSUN2 or TRDMT1 targets represent a false negative discovery rate, at least for this transcript biotype. Candidate m^{5}C sites in the other transcript biotypes that were detected in HeLa cells but not called as NSUN2 or TRDMT1 targets likely represent a mixture of false negative calls as well as genuine targets of other cellular RNA:m^{5}C methyltransferases that are yet to be confirmed in humans.

Implementation of cells exhibiting a permanent deletion of NSUN2 or TRDMT1 would be of benefit to overcome the incomplete depletion of the enzymes. To validate transcripts detected as NSUN2 targets by miCLIP, Blanco et al. and Hussain et al. performed bisulfite sequencing on RNA derived from human skin fibroblasts carrying a homozygous mutation of NSUN2 [78, 150]. Complete ablation of m^{5}C was evident at NSUN2-mediated sites in NSUN2$^{-/-}$ human fibroblasts. Use of such a deletion cell line would increase the sensitivity of detection of candidate m^{5}C sites by bsRNA-seq.

4.3.1.C Global comparison of m^{5}C sites mediated by NSUN2 and TRDMT1 with the literature

Several recent published studies have used complementary next-generation sequencing techniques to investigate m^{5}C sites in RNA (see section 1.2.3). One of these studies used a technique known as Aza-IP for the detection of m^{5}C sites that are targeted by NSUN2 and TRDMT1 in HeLa cells [77]. The second study used a combination of bisulfite sequencing and miCLIP focusing on NSUN2-mediated m^{5}C sites in human fibroblasts and a number of other human and mouse tissues [78, 150]. Both studies relied on the irreversible covalent bond formed between the RNA methyltransferases and the RNA target, which were then immunoprecipitated and subjected to next-generation sequencing, revealing NSUN2 and TRDMT1 target sites at single nucleotide resolution. In this section, I will make a comparison of the candidate m^{5}C sites called as NSUN2 or TRDMT1 targets from the global mapping in this thesis with the Aza-IP [77] and miCLIP studies [78, 150]. In most cases, comparisons were made between this thesis and the Aza-IP study, which provided single nucleotide resolution for m^{5}C sites called as NSUN2 or TRDMT1 targets. Although miCLIP and bisulfite sequencing are at single nucleotide resolution, the extended data for these publications did not provide

candidate m^{5}C sites at single nucleotide resolution. Furthermore, the miCLIP studies only focused on NSUN2 targets. However, there were several individual examples that were the prime focus of the miCLIP studies that were compared.

Initially, a comparison was made between the number of tRNAs called as NSUN2 targets in all three studies. Both the Aza-IP and miCLIP studies mapped the tRNAs genome-wide and therefore the candidate m^{5}C sites detected in tRNAs based on the global mapping in this thesis were compared to these studies. Almost 50% of the tRNAs called as NSUN2 targets in this thesis were in agreement with the Aza-IP and miCLIP studies, and nearly 60% were in agreement with at least one of these studies (Figure 4.17).

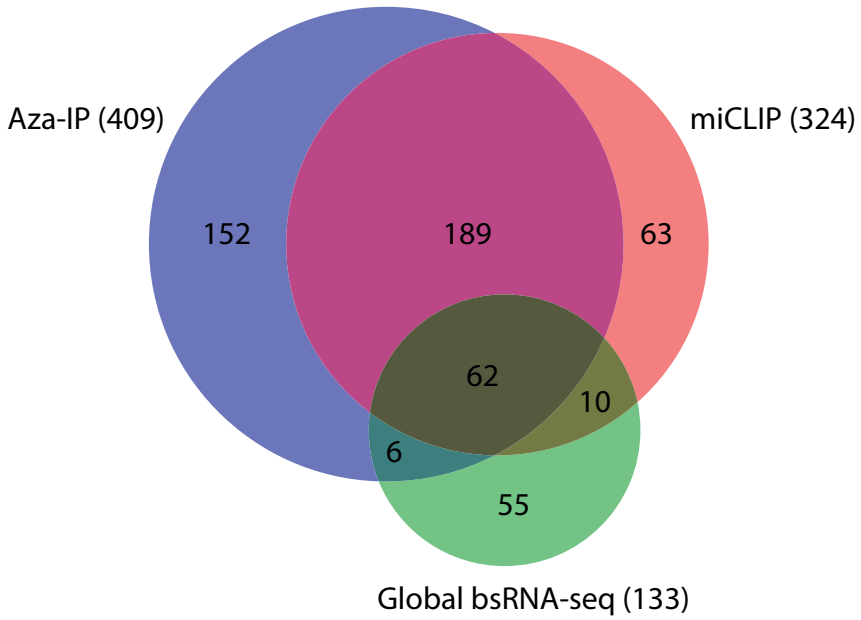


Figure 4.17. Comparison of the number of tRNAs exhibiting candidate m^{5}C sites called as NSUN2 targets by bsRNA-seq, miCLIP and Aza-IP.

The global mapping of tRNAs was used for this analysis. Comparisons were made with the miCLIP study [150] and Aza-IP study [77] The global mapping was used for this analysis. The total number of tRNAs containing candidate m^{5}C sites called as NSUN2 targets in each study is indicated in brackets.

Further comparisons in the number of candidate m^{5}C sites detected in tRNAs, other ncRNAs and mRNAs called as NSUN2 or TRDMT1 targets were made with the Aza-IP study. Of the 178 candidate m^{5}C sites in tRNAs that were called as NSUN2 targets in this thesis, 60% overlapped with the candidate m^{5}C sites detected by Aza-IP (Figure 4.18A). Of the 31 candidate m^{5}C sites detected in tRNAs as TRDMT1 targets, 77% were also detected by Aza-IP (Figure 4.18B). Thus, it is clear that there is high concordance between the two studies despite implementing very different techniques.

Interestingly, the overlap between other ncRNAs and mRNAs was very low or non-existent. Of the 33 candidate m^{5}C sites detected as NSUN2 targets in other ncRNAs in this thesis, only 4 overlapped with the Aza-IP study, which included *RPPH1*, *SCARNA2*, vtRNA1-1 and vtRNA1-3 (Figure 4.18C). The same candidate m^{5}C sites in vtRNA1-1 and vtRNA1-3 were also detected by miCLIP [78]. No candidate m^{5}C sites in ncRNAs called as TRDMT1 targets in this thesis overlapped with the Aza-IP study, which only detected a single candidate m^{5}C site in the pseudogene, *KRT18P10* (Figure 4.18D). NSUN2 was not reported to target any mRNAs by Aza-IP, however a single candidate m^{5}C site in the mRNA, *KRT18*, was detected as a target of TRDMT1; this candidate m^{5}C site was not detected in this thesis.

It is interesting that the Aza-IP technique, which relies on irreversible binding of the RNA methyltransferases to the RNA, did not detect many candidate m^{5}C sites in mRNAs. A possible reason for this is that mRNA may be present at much lower levels than tRNA and some other ncRNA, so the enzymes may bind preferentially to tRNAs and other ncRNAs, thus enriching them in comparison to mRNA. Additionally, HeLa cell lines vary across different laboratories and thus may also contribute to the differences observed in this thesis. Nevertheless, the limited detection of candidate m^{5}C sites in mRNA by Aza-IP did not deter further investigation into this transcript biotype in this thesis. Despite the low overlap of candidate m^{5}C sites detected in other ncRNAs in this thesis with the Aza-IP study, a subset of these sites were successfully validated by Aza-IP.

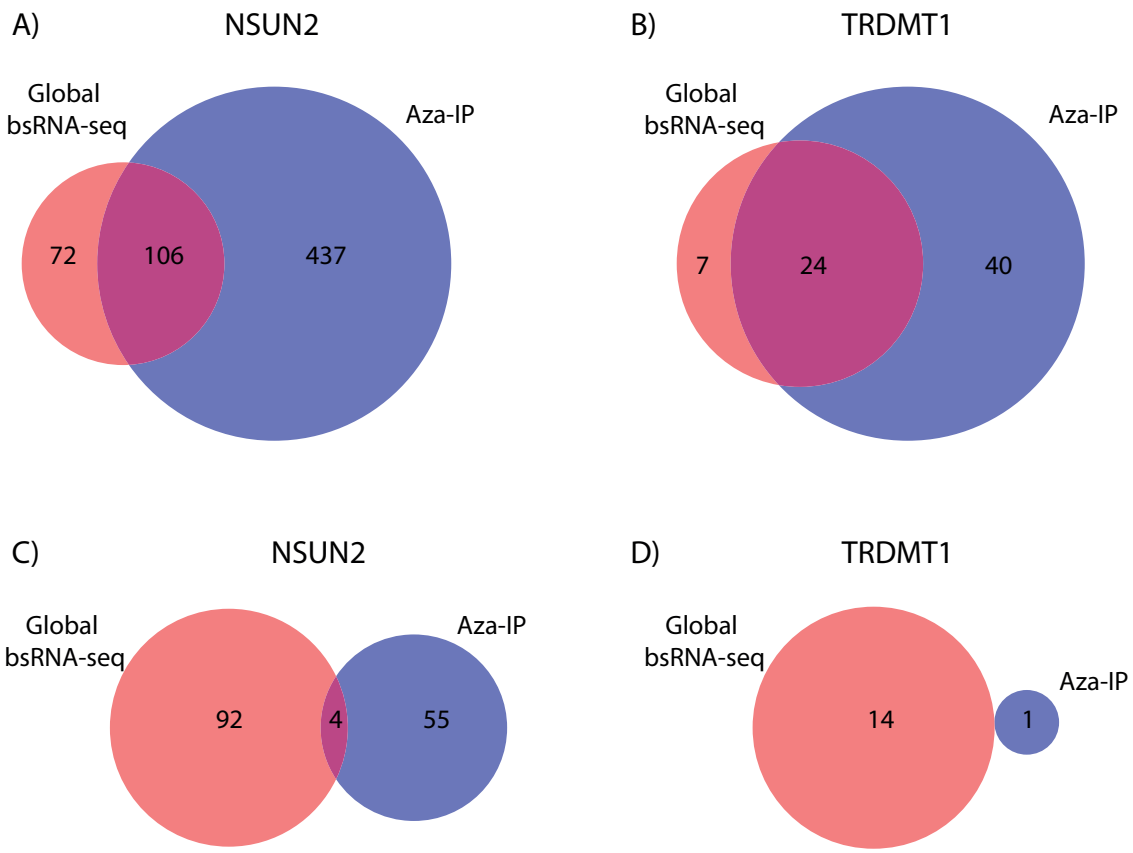


Figure 4.18. Comparison of candidate m^{5}C sites in tRNAs and ncRNAs called as NSUN2 or TRDMT1 targets by bsRNA-seq and Aza-IP.

Comparisons were made with the Aza-IP study [77]. A) Candidate m^{5}C sites in tRNAs detected as NSUN2 targets, B) m^{5}C sites in tRNAs detected as TRDMT1 targets, C) Candidate m5C sites in other ncRNAs detected as NSUN2 targets, D) Candidate m^{5}C sites in other ncRNAs detected as TRDMT1 targets.

4.3.2 m^{5}C SITES IN tRNA

The recent Aza-IP and miCLIP studies have focused heavily on m^{5}C sites in tRNAs. The Aza-IP study investigated the tRNA structural positions targeted by NSUN2 and TRDMT1 and the miCLIP study focused primarily on tRNA isoacceptors targeted by NSUN2. In this thesis, I have thoroughly examined the tRNA structural positions, isotypes and isoacceptors that were called as NSUN2 and TRDMT1 targets by bsRNA-seq. In the following sections, I will discuss the findings from this analysis and collate the findings from the Aza-IP and miCLIP studies for comparisons between all three studies. The tRNA structural positions targeted for methylation were compared to the Aza-IP study, and the tRNA isotypes and isoacceptors targeted for methylation were compared to both the Aza-IP and miCLIP studies.

4.3.2.A tRNA STRUCTURAL POSITIONS AND ISOTYPES TARGETED BY NSUN2 AND TRDMT1

In this thesis, candidate m^{5}C sites were detected at all known structural positions except C72 by the local mapping, while candidate m^{5}C sites were detected at all seven known structural positions from the global mapping. The lack of detection of candidate m^{5}C sites at C72 from the local mapping was possibly a result of the declining coverage towards the 3' end of the tRNA, which was exacerbated by specifically mapping to mature tRNA sequences. At least a proportion of the tRNAs exhibiting candidate m^{5}C sites at the known structural positions were detected as NSUN2 or TRDMT1 targets, suggesting that NSUN2 and TRDMT1 are solely responsible for mediating m^{5}C sites at these positions. This is highlighted by the study that reported that the knockout of *Nsun2* and *Trdmt1* in mice resulted in complete ablation of m^{5}C in tRNAs (see section 4.3.1.B) [149].

Focusing on candidate m^{5}C sites that were detected from the local mapping, NSUN2 was found to mediate m^{5}C at C34 exclusively in tRNA$^{Leu(CAA)}$, C40 exclusively in tRNA$^{Gly(GCC)}$ and C48-C50 in a wide range of tRNA isotypes. This is largely in agreement with previous literature, where NSUN2 has been shown to mediate m^{5}C at C34 exclusively in tRNA$^{Leu(CAA)}$ [100], C40 exclusively in tRNA$^{Gly(GCC)}$ in yeast [98], mice [149] and more recently in humans [77], and C48-50 in several tRNA isotypes

[58, 98, 148, 149]. In contrast, TRDMT1 was found to mediate m^{5}C sites at C38 in tRNA$^{Asp(GUC)}$, tRNA$^{Gly(GCC)}$, tRNA$^{Val(AAC)}$ and tRNA$^{Val(CAC)}$ only. TRDMT1 has been shown previously to mediate m^{5}C at C38 in tRNA$^{Asp(GUC)}$ [58, 59, 159] in addition to tRNA$^{Gly(GCC)}$ and tRNA$^{Val(AAC)}$ [73, 77], and more recently in tRNA$^{Val(CAC)}$ [77] in eukaryotes. A single candidate m^{5}C site was detected at C48 in a single tRNA$^{Cys(GCA)}$ transcript as a TRDMT1 target, however this site is most likely mediated by NSUN2 and was picked up as an artefact in this thesis. Overall, 87% of the candidate m^{5}C sites detected in tRNAs in this thesis were in the anticipated structural positions, with 99% and 93% of NSUN2 and TRDMT1 targets located in these positions, respectively.

The m^{5}C sites detected in the known structural positions were also validated by MiSeq amplicon sequencing for five tRNAs. TRDMT1 was confirmed to mediate m^{5}C at C38 in tRNA$^{Asp(GUC)}$ (see Chapter 3), tRNA$^{Gly(GCC)}$ and tRNA$^{Val(AAC)}$ and NSUN2 was confirmed to mediate C34 in tRNA$^{Leu(CAA)}$ (see Chapter 3) and C48-50 in several different tRNA isotypes. Interestingly, methylation at C40 in tRNA^{Gly} was detected at low levels by MiSeq amplicon sequencing. The bsRNA-seq data revealed approximately 15% methylation in the control samples that was depleted in the NSUN2 knockdown sample, however MiSeq amplicon sequencing revealed only 2-5% methylation in the control samples that was depleted to <1% in the NSUN2 knockdown sample. The Aza-IP study is the only other study to report m^{5}C at C40 in tRNA$^{Gly(GCC)}$ in humans as well as being mediated by NSUN2 [77]. However, subsequent validation by Sanger bisulfite sequencing revealed only 15% methylation, which was depleted in the NSUN2 knockdown sample. Thus, it is possible that tRNA$^{Gly(GCC)}$ is methylated consistently at low levels in HeLa cells.

Candidate m^{5}C sites in several non-canonical structural positions were also detected by bsRNA-seq in this thesis. The local mapping detected candidate m^{5}C sites in 4 non-canonical structural positions, while the global mapping detected candidate m^{5}C sites in 11 non-canonical structural positions. The candidate m^{5}C sites detected at C44 in tRNA^{Gly} and C61 in tRNA^{Gly}, tRNA^{Met} and tRNA^{Thr} were not called as NSUN2 or TRDMT1 targets. In contrast, the candidate m^{5}C site detected at C30 in tRNA^{His} was called as a NSUN2 target. The candidate m^{5}C site at C56 was detected in both the local and global mapping, although the enzyme responsible for this site could not be

determined from the local mapping. The global mapping revealed that a small proportion of the tRNAs exhibiting the candidate m^{5}C site at this position were called as NSUN2 targets, suggesting that NSUN2 may mediate this site.

Global comparison with the Aza-IP and miCLIP studies

The Aza-IP study also investigated the tRNA structural positions targeted by NSUN2 and TRDMT1 in HeLa cells [77]. Although several of the structural positions with candidate m^{5}C sites could not be determined as sole NSUN2 or TRDMT1 targets by bsRNA-seq in this thesis, the fact that a proportion of tRNA isotypes exhibiting candidate m^{5}C sites in these positions were targeted by NSUN2 or TRDMT1 suggested that they were exclusively responsible for mediating methylation at these positions. As such, Figure 4.19 displays the structural positions that were targeted by NSUN2 or TRDMT1 in at least one tRNA in this thesis.

The Aza-IP study revealed that six of the seven known structural positions were targeted by NSUN2 for methylation, and C38 was targeted by TRDMT1, which is in agreement with this thesis. The Aza-IP study reported 23 non-canonical structural positions exhibiting candidate m^{5}C sites targeted by NSUN2, however the candidate sites in these positions are yet to be validated. Despite the larger number of non-canonical structural positions targeted by NSUN2 in the Aza-IP study compared to this study, there was some overlap in these positions. The candidate m^{5}C site at C30 in tRNA^{His} was detected as a NSUN2 target in both studies, and the candidate m^{5}C site at C56 was also detected in both studies and this was found to be a NSUN2 target from the global mapping and Aza-IP.

That candidate m^{5}C sites were detected by bsRNA-seq and Aza-IP at C30 and C56 suggest that these may be additional structural positions that NSUN2 targets, extending the currently known targets of NSUN2 in tRNAs and highlighting that this enzyme is multisite-specific.

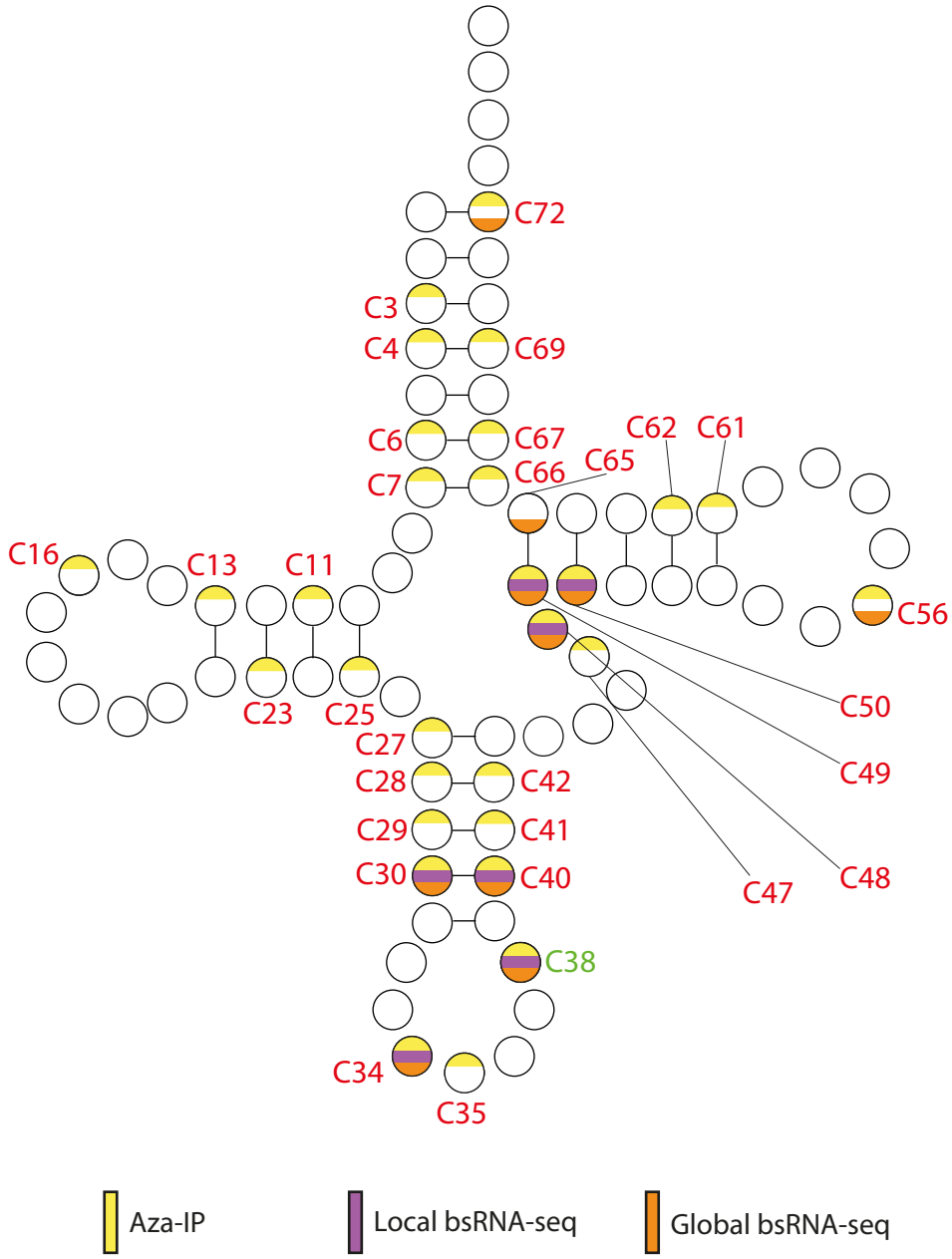


Figure 4.19. Comparison of candidate m^{5}C sites in tRNA structural positions called as NSUN2 or TRDMT1 targets by bsRNA-seq and Aza-IP.

Comparisons were made between the global bsRNA-seq mapping, local bsRNA-seq mapping and Aza-IP study [77]. For a structural position to be considered a NSUN2 or TRDMT1 target, at least one tRNA must exhibit a candidate m^{5}C site called as a NSUN2 or TRDMT1 target at the given position. The structural positions detected as NSUN2 and TRDMT1 targets by either study are indicated in red and green, respectively.

In addition to comparing the structural positions targeted for methylation by NSUN2 and TRDMT1, a comparison was made between the tRNA isotypes that were called as NSUN2 and TRDMT1 targets in this thesis with the Aza-IP and miCLIP studies.

The tRNA isotypes that were called as NSUN2 targets in all three studies were quite broad; of the 22 tRNA isotypes, 16 were called as NSUN2 targets by miCLIP, 18 by Aza-IP, 15 from the local mapping and 19 from the global mapping (Figure 4.20). In addition, there was high concordance in the tRNA isotypes called as NSUN2 targets in all studies, with 13 isotypes overlapping. This indicates that NSUN2 has a broad substrate range for tRNA isotypes in addition to being multisite-specific. In all studies, $tRNA^{Asn}$ and $tRNA^{SeC}$ were not called as NSUN2 targets. These tRNAs do not exhibit cytosines at positions 48-50, which may partly explain the lack of m^5C sites mediated by NSUN2 in these tRNAs. The Aza-IP study was the only study to report candidate m^5C sites in $tRNA^{Trp}$ that were mediated by NSUN2.

Interestingly, this thesis was the first to report candidate m^5C sites in $tRNA^{Sup}$, a suppressor tRNA, which has not been previously identified as exhibiting m^5C sites. Suppressor tRNAs are naturally occurring tRNAs that form from a mutation in the anti-codon, resulting in the read-through of the stop codon and an amino acid being inserted instead, preventing truncation of the protein [283]. The subsequent result is the production of multiple polypeptides from a single gene. The existence of natural suppressor tRNAs provides an additional regulatory mechanism for gene expression and the presence of m^5C in these tRNAs may regulate their function. Interestingly, Blanco et al. have reported that depletion of NSUN2-mediated m^5C sites in tRNAs causes angiogenin-cleavage and accumulation of 5' tRNA fragments [150]. Thus, the presence of m^5C in $tRNA^{Sup}$ may stabilise this tRNA, allowing it to function in regulating gene expression. There is the possibility of m^5C affecting tRNA charging by aminoacyl tRNA synthetases, however other modifications have been reported to affect this. Methylation of G37 (m^1G37) of tRNAs in *S. typhimurium* has been shown to be important in promoting binding to aminoacyl tRNA synthetase [284]. Furthermore, this modification in *S. cerevisiae* prevents the mischarging of $tRNA^{Asp}$ by arginyl-tRNA synthetase [285].

The tRNA isotypes that were called as TRDMT1 targets were quite narrow, indicating that this particular enzyme has narrow specificity for tRNA structural positions and isotypes. In both this these and the Aza-IP study, TRDMT1 targeted tRNA^{Asp}, tRNA^{Gly} and tRNA^{Val} for methylation. Interestingly, the local mapping revealed candidate m^{5}C sites mediated by TRDMT1 at tRNA^{Cys}, while the global mapping revealed candidate m^{5}C sites mediated by TRDMT1 in tRNA^{Thr}. As the different mapping methods resulted in inconsistencies with these two tRNA isotypes, this suggests that these may be artefacts.

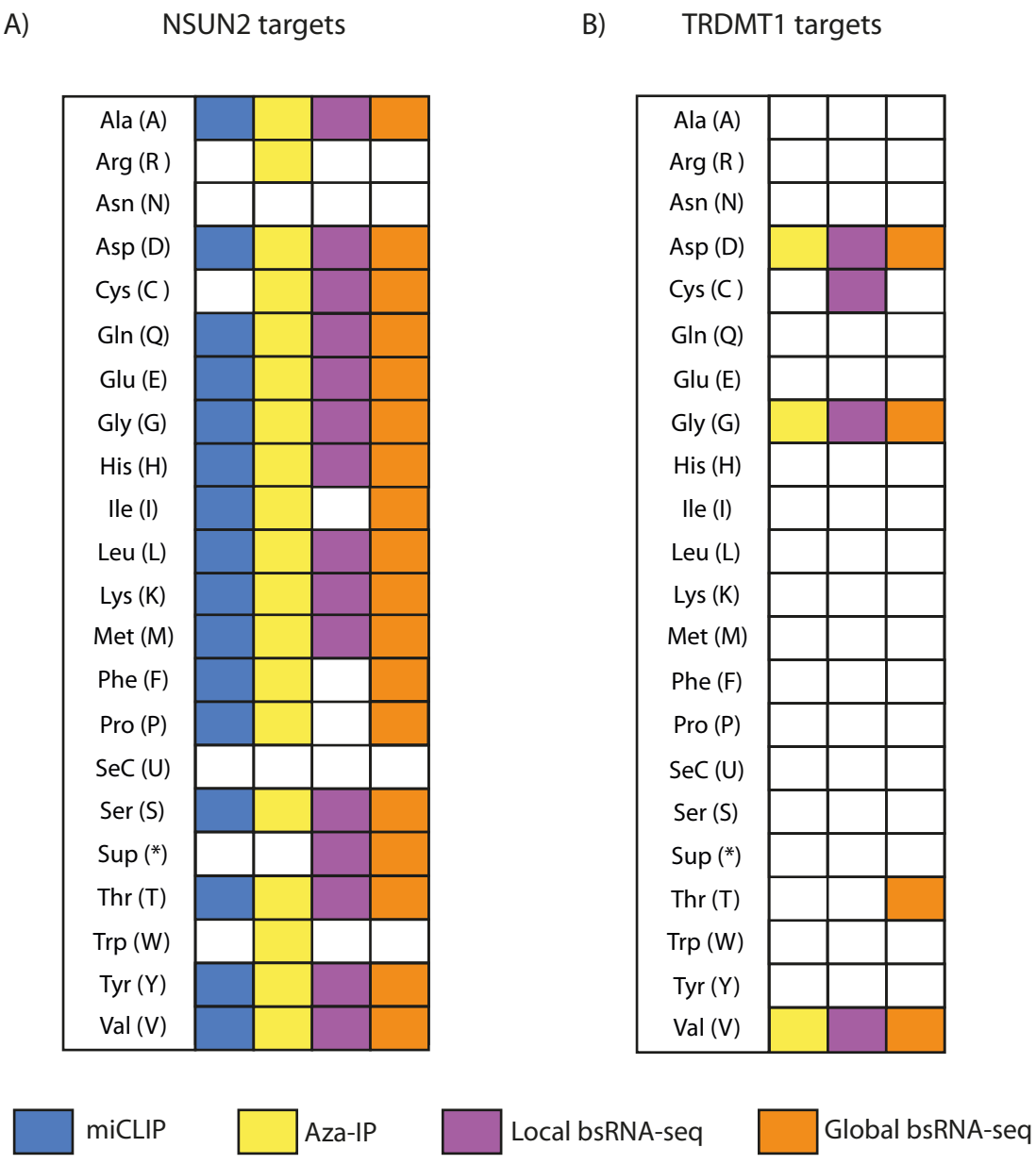


Figure 4.20. Comparison of tRNA isotypes called as NSUN2 and TRDMT1 targets by bsRNA-seq, Aza-IP and miCLIP.

Comparisons were made with the miCLIP study [150] and Aza-IP study [77]. A) Isotypes targeted by NSUN2, B) Isotypes targeted by TRDMT1. Comparisons were made between the local bsRNA-seq mapping, global bsRNA-seq mapping, miCLIP study and Aza-IP study.

4.3.2.B tRNA isoacceptors targeted by NSUN2 and TRDMT1

Investigation into the tRNA isoacceptors that were targeted by each enzyme revealed that NSUN2 targeted not only a broad range of tRNA isotypes, but also a broad range of different isoacceptors within those isotypes. TRDMT1 was much more specific for isoacceptors in addition to structural positions and isotypes, targeting specifically tRNA$^{Asp(GUC)}$, tRNA$^{Gly(GCC)}$, tRNA$^{Val(AAC)}$ and tRNA$^{Val(CAC)}$ in both the local and global mappings (see section 4.3.2.A).

Global comparison with the Aza-IP and miCLIP studies

Figure 4.21 illustrates the comparison of tRNA isoacceptors called as NSUN2 and TRDMT1 targets. The comparisons of tRNA isoacceptors called as NSUN2 targets were made between all studies and the comparison of tRNA isoacceptors targeted by TRDMT1 was made with the Aza-IP study.

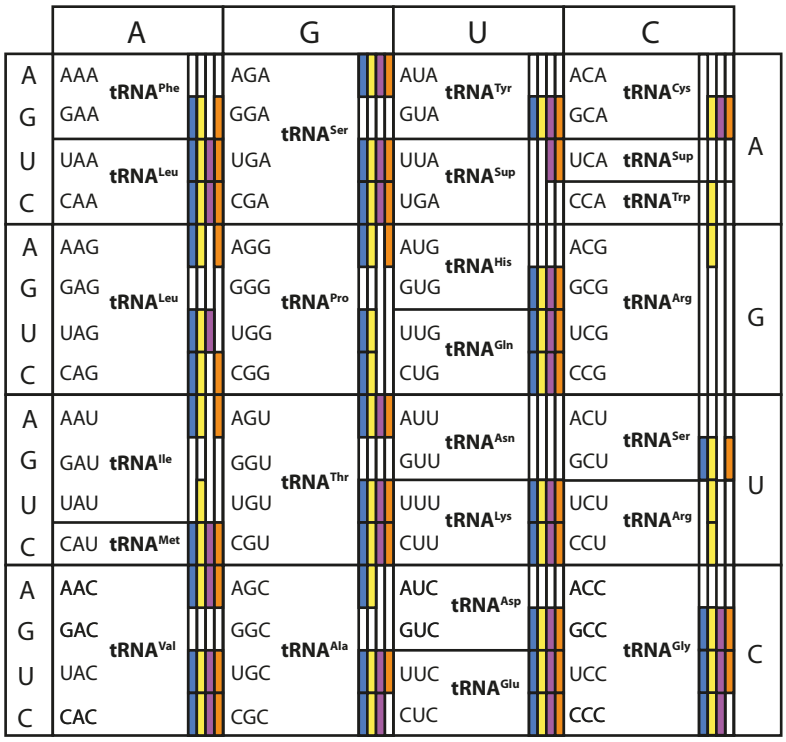
The miCLIP study detected 36 isoacceptors [150] and the Aza-IP detected 42 isoacceptors [77] that were targeted by NSUN2. There was high concordance between these two studies, with 36 isoacceptors overlapping. Of the 29 isoacceptors called as NSUN2 targets from the local mapping, 28 overlapped with the Aza-IP study and 27 overlapped with the miCLIP study. Of the 32 isoacceptors called as NSUN2 targets from the global mapping, 31 overlapped with the Aza-IP study and 30 overlapped with the miCLIP study. A total of 24 isoacceptors overlapped between all studies, indicating very high concordance between all datasets (Figure 4.21A). The only isoacceptor detected in this thesis that did not overlap with either of these published studies was tRNA$^{Sup(UUA)}$. The overlap in tRNA isoacceptors called as TRDMT1 targets in this thesis and Aza-IP was high, with 4 out of 5 isoacceptors overlapping with the Aza-IP study (Figure 4.21B).

Early studies into the specificity of NSUN2 in tRNAs suggested that this was narrow, with *in vitro* methylation assays suggesting that NSUN2 could not introduce a methyl group to a number of human and yeast tRNAs at C48 and C49, while Trm4 in yeast could target these positions [100]. Our previous bsRNA-seq study in HeLa cells revealed that NSUN2 is responsible for m^{5}C sites at C48 and C49 in tRNA$^{Asp(GUC)}$,

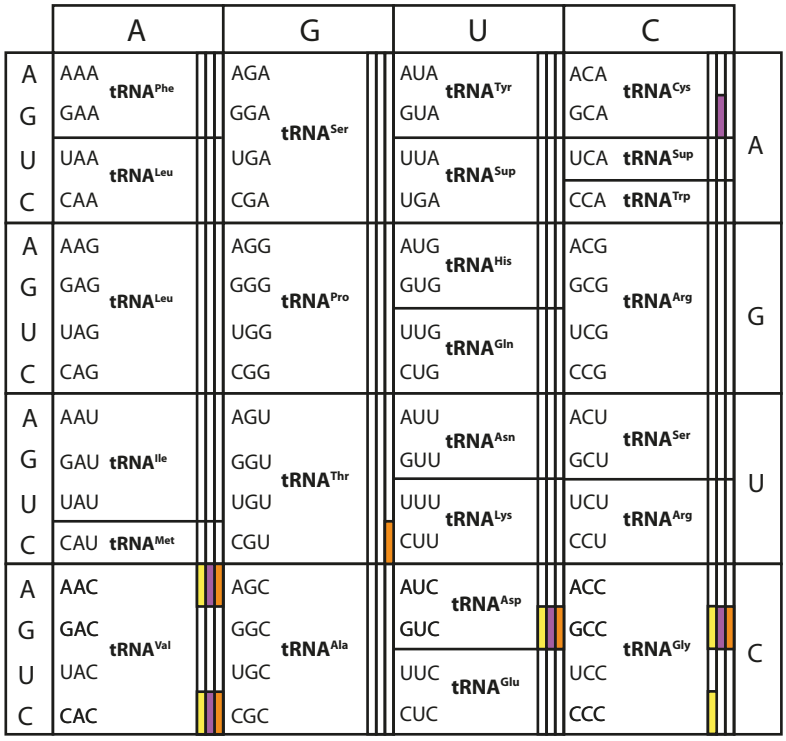
indicating a broader structural specificity in tRNAs [58]. The work in this thesis has demonstrated the specificity of NSUN2 and TRDMT1 on three different levels; tRNA structural positions, isotypes and isoacceptors. NSUN2 is a multisite-specific enzyme, targeting at least six tRNA structural positions with the possibility of targeting more, including at C30 and C56. NSUN2 was also found to target the majority of tRNA isotypes and a wide range of tRNA isoacceptors, highlighting its broad substrate range and confirming recently published studies. In contrast, the substrate specificity of TRDMT1 was found to be narrow in this thesis, almost exclusively targeting C38 in only four tRNA isoacceptors.

The high concordance in the tRNA structural positions, isotypes and isoacceptors targeted for methylation by NSUN2 and TRDMT1 in this study with the literature indicates that the bsRNA-seq technique is sensitive in detecting m^5C sites in tRNAs. The fact that these two published studies have employed complementary techniques to bsRNA-seq reiterates this point. The similarities highlight that the bsRNA-seq method is efficient in detecting m^5C sites in various transcript biotypes.

A)


NSUN2
A
G
U
C
AAA tRNA Phe GAA
AGA GGA tRNA Ser UGA CGA
AUA tRNA Tyr GUA
ACA tRNA Cys GCA
UAA tRNA Leu CAA
UUA tRNA Sup UGA
UCA tRNA Sup CCA tRNA Trp
A
AAG GAG tRNA Leu UAG CAG
AGG GGG tRNA Pro UGG CGG
AUG GUG tRNA His
UUG tRNA Gln CUG
ACG GCG tRNA Arg UCG CCG
G
AAU GAU tRNA Ile UAU
CAU tRNA Met
AGU GGU tRNA Thr UGU CGU
AUU GUU tRNA Asn
UUU tRNA Lys CUU
ACU GCU tRNA Ser
UCU tRNA Arg CCU
U
AAC GAC tRNA Val UAC CAC
AGC GGC tRNA Ala UGC CGC
AUC GUC tRNA Asp
UUC tRNA Glu CUC
ACC GCC tRNA Gly UCC CCC
C

B)


TRDMT1
A
G
U
C
AAA tRNA Phe GAA
AGA GGA tRNA Ser UGA CGA
AUA tRNA Tyr GUA
ACA tRNA Cys GCA
UAA tRNA Leu CAA
UUA tRNA Sup UGA
UCA tRNA Sup CCA tRNA Trp
A
AAG GAG tRNA Leu UAG CAG
AGG GGG tRNA Pro UGG CGG
AUG GUG tRNA His
UUG tRNA Gln CUG
ACG GCG tRNA Arg UCG CCG
G
AAU GAU tRNA Ile UAU
CAU tRNA Met
AGU GGU tRNA Thr UGU CGU
AUU GUU tRNA Asn
UUU tRNA Lys CUU
ACU GCU tRNA Ser
UCU tRNA Arg CCU
U
AAC GAC tRNA Val UAC CAC
AGC GGC tRNA Ala UGC CGC
AUC GUC tRNA Asp
UUC tRNA Glu CUC
ACC GCC tRNA Gly UCC CCC
C

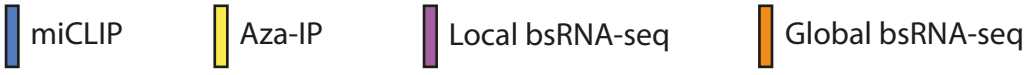

miCLIP
Aza-IP
Local bsRNA-seq
Global bsRNA-seq

Figure 4.21. Comparison of tRNA isoacceptors called as NSUN2 or TRDMT1 targets by bsRNA-seq, miCLIP and Aza-IP.

Comparisons were made with the miCLIP study [150] and Aza-IP study [77]. A) tRNA isoacceptors targeted by NSUN2, B) tRNA isoacceptors targeted by TRDMT1.

4.3.2.C Methylation of tRNA prior to processing

One aspect of m^{5}C in tRNAs that has had very little attention is the point at which methylation of tRNAs occurs during its biogenesis. The biogenesis of tRNAs is a very complex process that is still not well-understood, particularly in humans. Many of the studies investigating this process have focused on yeast, although many of the processes apply to higher eukaryotes (for a review, see [268]). Nuclear tRNAs in eukaryotes are transcribed by RNA Polymerase III as a pre-tRNA within the nucleus. Once transcribed, the pre-tRNA is bound by the La autoantigen at the 3' end, protecting the 3' end from endonucleolytic cleavage. This La-bound pre-tRNA is the substrate for RNase P, which processes the 5' end of the pre-tRNA by a single endonucleolytic cleavage (for a review, see [286]). The location of 5' end processing within the cell in yeast and humans is conflicting, with studies suggesting that RNase P RNA is localised to both the nucleolus and cytoplasm, in addition to the perinucleolar compartments that surround the edge of the nucleolus in several human cell lines [287, 288]. Following this, 3' end processing is performed by tRNase Z within the nucleoplasm, followed by the addition of the 3' CCA, splicing and export to the cytoplasm. Interestingly, the order of end processing and splicing can be reversed in several species [269, 270], although it is unknown whether splicing precedes end processing in humans. During this entire process, a number of nucleoside modifications take place at various stages, although the understanding of this is limited. That NSUN2 is localised in the nucleolus and TRDMT1 is localised in the nucleus suggest that methylation of tRNAs may occur prior to splicing and end processing, motivating further investigation.

We detected two tRNAs containing introns that exhibited m^{5}C sites, which included C34 in tRNA$^{Leu(CAA)}$ and C48 in tRNA$^{Tyr(GUA)}$. Due to the low coverage of tRNA$^{Tyr(GUA)}$, this tRNA was not explored further. The intron of tRNA$^{Leu(CAA)}$ is 22 nt long and located 3' to the anti-codon, resulting in an elongated anti-codon stem [100, 289]. Two studies thus far have reported that m^{5}C at C34 in tRNA$^{Leu(CAA)}$ is mediated

by NSUN2 only prior to splicing in yeast and humans [100, 148]. This was found to be specific to the sequence and structural context that NSUN2 recognises, as intron removal and mutations surrounding the cytosine resulted in lack of methylation in most cases. However, this relied on *in vitro* methylation assays and no studies have investigated this *in vivo*. In previous work, we have detected methylation of tRNA$^{Leu(CAA)}$ prior to splicing using Sanger bisulfite sequencing, but this was not investigated any further [58]. In this thesis, the MiSeq amplicon sequencing data was mapped to spliced and unspliced versions of tRNA$^{Leu(CAA)}$, and m^{5}C was detected at C34 in both versions of the tRNA, indicating that methylation of this tRNA can occur prior to splicing. In addition, a higher proportion of methylation was detected in the spliced tRNA compared to the unspliced tRNA, suggesting that not all of the precursor transcripts have undergone methylation, while the mature transcript is mostly methylated. It is highly likely that methylation of this tRNA occurs immediately prior to splicing, however the possibility of methylation occurring after splicing cannot be discounted despite *in vitro* methylation assays indicating otherwise. It is possible that the intron may somehow affect m^{5}C formation in the tRNA, but this has not been investigated further.

Methylation of other tRNAs or at other structural positions prior to splicing has not been explored in great detail. One study has investigated m^{5}C at C48 of tRNA$^{Leu(CAA)}$ and reported that this can occur either before or after splicing based on *in vitro* methylation assays [148], however we did not detect this site in our present study. It is not unexpected that NSUN2 can mediate m^{5}C at C48 in spliced and unspliced tRNAs, as the structure surrounding this region does not change upon splicing, unlike the anti-codon region that C34 is located in [100, 272]. This suggests that the m^{5}C site detected in this thesis at C48 in tRNA$^{Tyr(GUA)}$ may occur before or after splicing. However, as m^{5}C at C34 in tRNA$^{Leu(CAA)}$ can occur prior to splicing and the structure of the tRNA surrounding C48 does not change upon maturation, it is highly likely that m^{5}C at C48 predominantly occurs prior to splicing.

Although the understanding of the order end processing and splicing of tRNAs in humans is limited, we were interested in whether m^{5}C can also occur prior to 5’ and 3’ end processing. This study was the first to explore this in detail. The global mapping of

tRNAs without the 3' CCA addition allowed the mapping of tRNAs that were not completely processed, although it was expected that the levels of pre-tRNAs would be low. Examination of sequencing reads mapping to tRNA$^{Leu(CAA)}$ and tRNA$^{Asp(GUC)}$ in IGV revealed that a small proportion of the reads extended beyond the 5' and 3' ends of the mature tRNA and that a small proportion of these sequencing reads contained m^{5}C sites mediated by NSUN2 and TRDMT1. This indicated that NSUN2 and TRDMT1 can mediate m^{5}C prior to end processing of tRNAs.

A global approach was also used where all tRNAs exhibiting candidate m^{5}C sites based on the global mapping were divided into a subset containing no mismatches and those containing ≥1 T-to-C mismatch. It should be noted that while a T-to-C mismatch is indicative of candidate m^{5}C sites, it could also potentially be non-conversion due to the RNA secondary structure, or possibly an error associated with colour-space sequencing. However, we have not detected these issues previously in tRNAs. T-to-C mismatches were detected in a small proportion of sequencing reads extending beyond the 5' and 3' ends of the tRNAs, suggesting that methylation can occur prior to end processing.

It seems reasonable to contend that if methylation can occur prior to splicing and end processing of tRNAs, then it most likely will exclusively occur prior to splicing and end processing. It should be noted, however, that these results do not formally exclude that some methylation might also occur after maturation.

4.3.3 m^{5}C IN ncRNA

In this thesis, candidate m^{5}C sites were detected in a variety of ncRNA, encompassing miRNA, rRNA, snoRNA, uRNA, vtRNA, 7SK RNA, 7SL RNA and other ncRNA. A subset of candidate m^{5}C sites detected in these transcripts were called as NSUN2 targets, and an even smaller subset were called as TRDMT1 targets. Recent studies have discovered candidate m^{5}C sites in several different ncRNA including vtRNA, snoRNA and 7SK RNA that were mediated by NSUN2 [58, 77, 78], however investigation into the sequence and substrate specificity of NSUN2 is lacking. In this section, I will discuss in detail the m^{5}C sites detected in ncRNA, the sequence and structural specificity of NSUN2 in these transcripts and possible molecular implications of m^{5}C in these transcripts.

4.3.3.A m5C in rRNA

In this thesis, I examined the candidate m5C sites detected in nuclear-encoded and mitochondrial-encoded rRNA. 24 candidate m5C sites were detected in nuclear-encoded rRNA, with the majority detected in 28S rRNA. The two known m5C sites at C3782 and C4447 in 28S rRNA were detected with high coverage and >85% methylation in all samples, suggesting that NSUN2 or TRDMT1 do not mediate these sites. Interestingly, the m5C sites at these positions are equivalent to the m5C sites at C2278 and C2870 in yeast 25S rRNA, which are mediated by Rcm1 [86, 92] and Nop2 [92], respectively. The human homologues of Rcm1 and Nop2 are NSUN5A and NSUN1, respectively. Although catalytic activity of these human RNA:m5C methyltransferases have not been confirmed, it is possible that they mediate the two known m5C sites in 28S rRNA.

As previous studies have not yet detected other m5C sites in nuclear-encoded rRNA and clusters of non-conversion were evident in the rRNAs in this study, an investigation was made into the remaining candidate m5C sites detected in this study. As m5C at C3782 and C4447 were detected with high coverage and high levels of methylation in all samples, additional stringent thresholds were applied to enrich for genuine candidate m5C sites. This revealed no additional candidate m5C sites in nuclear-encoded rRNA in this study.

5 candidate m5C sites were called in mitochondrial-encoded rRNA, including 4 in 12S mitochondrial rRNA and 1 in 16S mitochondrial rRNA. The known m5C site at C841 in 12S mitochondrial rRNA was detected with high coverage and ~40% methylation in all samples, suggesting that NSUN2 or TRDMT1 does not mediate this site. It was recently reported that the mitochondrial-localised enzyme, NSUN4, mediates m5C formation at the equivalent position in mice [97]. As with NSUN1 and NSUN5A, the catalytic activity of the human orthologue remains to be explored, however it is highly likely that NSUN4 is responsible for m5C at C841 in 12S mitochondrial rRNA.

Two of the candidate m5C sites detected in 12S mitochondrial rRNA in this thesis could not be validated, suggesting that these may be false positives as a result of incomplete denaturation. The one remaining candidate m5C site in 12S mitochondrial rRNA

exhibited low and inconsistent levels of methylation in all samples, suggesting that this may be also an artefact. Additional m^{5}C sites in mitochondrial-encoded rRNA are yet to be confirmed. Several studies have proposed that NSUN4 mediates m^{5}C at an unknown residue in 16S mitochondrial rRNA [95, 290, 291], however we were not able to validate the candidate m^{5}C site detected in 16S mitochondrial rRNA in this thesis. In agreement with our finding, m^{5}C was only detected at the one known site in 12S rRNA in mice by bisulfite sequencing [97].

The inability to detect depletion of m^{5}C in the NSUN2 and TRDMT1 knockdown samples at these positions in human rRNA and that the yeast and mice orthologues of NSUN1, NSUN4 and NSUN5A mediate m^{5}C at equivalent positions highlights the possibility of other active RNA:m^{5}C methyltransferases in humans. It would be of benefit to explore these RNA:m^{5}C methyltransferases in more detail in the future to understand their substrate range and specificity.

4.3.3.B SEQUENCE AND STRUCTURAL CONTEXT OF M^{5}C IN OTHER NCRNA

In this thesis, I used a candidate approach to investigate the sequence and structures that NSUN2 may recognise to mediate m^{5}C in ncRNA. I validated two previously discovered m^{5}C sites in *RPPH1* and *SCARNA2*, and novel m^{5}C sites in *SNORD62*, *NSUN5P2* and a pseudogene in the intron of *CAGE1*. These m^{5}C sites were also validated as NSUN2 targets, indicating that NSUN2 is capable of mediating m^{5}C in a broad range of transcript biotypes.

To understand the context features around the m^{5}C sites and the specificity of NSUN2 in these transcripts, predicted and published secondary structures were obtained for *RPPH1*, *SCARNA2* and *SNORD62*. Additionally, secondary structures were obtained for vtRNA1-1 and vtRNA1-3, which were detected by bsRNA-seq as NSUN2 targets and were previously detected and validated in HeLa cells, HEK293 cells and human fibroblasts [77, 78].

RPPH1, *SNORD62* and *SCARNA2* are all involved in the processing and modification of different transcript biotypes. *RPPH1*, *SNORD62* and *SCARNA2* and the vtRNAs are either synthesised or function in the nucleolus where NSUN2 is predominantly localised. *RPPH1*, *SNORD62* and vtRNAs are transcribed by RNA Polymerase III within the nucleolus [292], and it is speculated that RNase P and therefore *RPPH1* localises to the nucleolus (see section 4.3.2.C). *SCARNA2* is independently transcribed by RNA Polymerase II but is predicted to function in the nucleolus [293].

The m^{5}C site in *RPPH1* was in a GCA context and located in the stem-loop junction, a structure reminiscent of that surrounding C48-50 in tRNAs that NSUN2 targets. *RPPH1* is the ncRNA component of RNase P that functions in the 5' end processing [274] and has been shown to be important for RNA Polymerase III transcription [274]. Interestingly, the RNA component of RNase P in bacteria and some archaea has been shown to mediate tRNA cleavage in the absence of RNase P protein subunits [294]. Recently, it has been reported that *RPPH1* is capable of mediating cleavage of four pre-tRNAs in the absence of protein *in vitro*, although the rate of cleavage was substantially lower than bacterial RNase P RNA [295]. This was suggested to be a result of *RPPH1* not exhibiting the P15/P16/P17 region, which plays an important role in substrate binding and catalysis in bacteria [296-298]. Additionally, it is suspected that the additional protein subunits of the RNase P complex function in enhancing this catalytic activity [295]. The m^{5}C site detected and validated in this thesis was located in the P12 domain. Interestingly, targeted cleavage of the P12 domain of *RPPH1* by RNase H was reported to cause a substantial reduction in transcription of 7SL RNA, 5S rRNA and several tRNAs, suggesting that this domain is important for transcription [274]. Thus, it is possible that the m^{5}C site in this domain may regulate RNA Polymerase III transcription, which could potentially provide an additional layer of gene regulation.

The m^{5}C site in *SCARNA2* was in a CpG context and located between two stem-loops and the m^{5}C site in *SNORD62* was in GCA context and located within the loop of a stem-loop structure. The structural context of both of these sites thus appears to be different to those targeted by NSUN2 in tRNAs. The m^{5}C sites in *SCARNA2* and *SNORD62* were not located within the C/D box and C'/D' box domains that are predicted to guide modification of U2 and 18S rRNA, respectively. In addition, the m^{5}C

sites were not located at the predicted binding sites of the substrate RNAs. This suggests that the m^{5}C sites in these guide RNAs may not directly affect recognition and binding of the substrate RNA, although this remains to be explored further.

The candidate m^{5}C sites in the two vtRNAs were in a TCG context and located in the stem near the stem-loop junction, a structure consistent with that exhibited around C48-50 in the variable loop junction and C30 in the anti-codon loop junction in tRNAs. It was recently reported that loss of NSUN2-mediated m^{5}C sites in vtRNAs caused aberrant processing into svRNAs, with one fragment of vtRNA1-1 exhibiting a significant decrease in the absence of NSUN2 [78]. Previous studies have reported that svRNAs can induce silencing of transcripts in a similar manner to miRNAs [299]. This fragment was demonstrated to associate with both Ago2 and Ago3, and two potential target mRNAs of this fragment demonstrated increased expression in the absence of this fragment. This strongly suggests that m^{5}C in vtRNAs regulates its processing into svRNAs that in turn may mediate gene expression.

The m^{5}C sites in *NSUN5P2* and the pseudogene located within the *CAGE1* intron were in a CpA and CpT context, respectively. Little information is available for the processed pseudogene located in the intron of *CAGE1*. However, *NSUN5P2* is one of the two *NSUN5* pseudogenes located in the same region as *NSUN5*. The function of this pseudogene is unknown, however a deletion of the region of chromosome 7 that constitutes *NSUN5* and both pseudogenes was observed in patients suffering from William-Beuren syndrome, a disease characterised by a complex array of symptoms, including neurodevelopment disorders [300]. To this end it is not known which gene contributes to the phenotype, although this example suggests that pseudogenes may play an important role in biology and disease, and m^{5}C may function in regulating these transcripts.

Based on this analysis, it is clear NSUN2 can mediate m^{5}C in different sequence and structural contexts. There is a clear propensity for NSUN2 to target sites near stem-loop structures that may resemble the stem-loop structures in tRNAs. However, it is evident that NSUN2 can target structures that are different to those exhibited in tRNAs. This suggests that other features besides the surrounding sequence and structure of the target

site may possibly aid in the recognition by NSUN2 to mediate m^{5}C. This is in contrast to the narrow specificity of NSUN2 previously reported for unspliced tRNA$^{Leu(CAA)}$ at C34, which was attributed to the surrounding sequence and structural context [98]. The presence of m^{5}C in ncRNAs may regulate their role in processing, but this remains to be explored in further detail.

4.3.4 m^{5}C in mRNA

Early studies of RNA modifications using chromatographic techniques revealed that m^{5}C existed at low levels in cellular and viral mRNAs [26, 30, 47, 48]. Due to their low levels, it was initially thought that this was simply contamination and was never followed up. With the improvement to techniques, the presence of m^{5}C in mRNA is coming to the fore, revealing its abundance in this transcript biotype [58]. With the exception of our previous HeLa bsRNA-seq study, no studies have focused on the abundance of m^{5}C in mRNA and the substrate specificity of NSUN2 in these transcripts. In this thesis, I have shown that m^{5}C is abundant in mRNAs and that a proportion of these candidates were called as NSUN2 or TRDMT1 targets. I have validated previously discovered and novel candidate m^{5}C sites as targets of NSUN2, and have investigated the sequence and structural specificity of NSUN2 in mRNAs, demonstrating that NSUN2 is an important mRNA modifying enzyme.

4.3.4.A Distribution of m^{5}C in mRNA

Using the candidate m^{5}C sites detected in mRNAs in HeLa cells, the distribution of m^{5}C was determined throughout different regions of mRNA. Candidate m^{5}C sites were distributed throughout all regions of mRNAs, including introns. The fact that candidate m^{5}C sites were detected in introns suggests that methylation can occur prior to splicing, although further examination of sequencing reads would be required to determine if this is the case. Overall, a large number of candidate m^{5}C sites were located in the 3’ UTR and CDS and were significantly enriched in the 3’ UTR, while candidate m^{5}C sites were depleted in the CDS and 5’ UTR compared to the expected levels. The enrichment of candidate m^{5}C sites in the 3’ UTR is consistent with our previous study, however sites were also enriched in the 5’ UTR in our previous study [58]. The fact that both studies

revealed a clear 3' UTR enrichment of candidate m5C sites highlights the non-random distribution of this modification.

Recent next-generation sequencing studies investigating the distribution and role of m6A in mRNA have also reported non-random distribution, being predominantly enriched near stop codons and the 5' UTR, and were not present in poly(A) tails [22, 56, 57]. Several studies have also investigated the molecular function of m6A in mRNA, reporting a potential role in mRNA stability and alternative splicing [57, 69, 105, 107, 120]. Thus, the non-random distribution of m5C in mRNA is highly suggestive of an important regulatory role for this modification. Furthermore, the enrichment of m5C in the 3' UTR implies a possible regulatory role through association with Ago and miRNAs. We have previously detected an enrichment of Ago footprints upstream of candidate m5C sites, using a publicly available HEK293 Ago PAR-CLIP dataset [58]. Although the PAR-CLIP was performed in a different human cell line, it is highly indicative of a role for m5C in post-transcriptional gene regulation through association with Ago.

4.3.4.B Validation of m5C sites in mRNA

The candidate m5C sites selected for validation were located in different regions of the mRNA, the majority of which were called as NSUN2 targets based on bsRNA-seq. Three of the m5C sites selected for validation were discovered and validated as NSUN2 targets in our previous study and included *CINP*, *NAPRT1* and *NDUFB7* [58]. In this present study, *NDUFB7* was called as a NSUN2 target by bsRNA-seq consistent with our previous validation by Sanger bisulfite sequencing [58]. In contrast, the m5C site in *CINP* was detected but was not called as a NSUN2 target due to insufficient coverage in the NSUN2 knockdown sample. The m5C site in *NAPRT1* was not detected at all in this present study due to insufficient coverage in most samples. All three previously discovered and validated m5C sites were validated as NSUN2 targets in this study. In addition, the novel m5C sites detected in *OSBPL8* and *RTN3* were not called as NSUN2 or TRDMT1 targets by bsRNA-seq in this study, but were validated as NSUN2-mediated sites. This indicates that the number of candidate m5C sites detected by bsRNA-seq in this thesis as NSUN2 and TRDMT1 targets was an underestimation.

The novel m^{5}C sites detected in *RPS3*, *SCO1* and *SZRD1* were discovered and validated as NSUN2-directed sites in this thesis. Interestingly, *MCFD2* exhibited two candidate m^{5}C sites, with each enzyme targeting this transcript. Although a focus was placed on candidate m^{5}C sites targeted by NSUN2, the candidate m^{5}C site called as a TRDMT1 target was in close proximity to the NSUN2 target site and as a result was included in the validation process. The m^{5}C site called as a NSUN2 target was validated, however the candidate m^{5}C called as a TRDMT1 target was not detected by MiSeq amplicon sequencing. Examination of the sequencing reads spanning this unvalidated candidate m^{5}C site revealed that the site was of low quality due to uncertainty in the colour calling by the SOLiD^{TM} platform.

The validated m^{5}C sites were in mRNAs that encoded proteins with varying functions in the cell, including components of the DNA replication complex (CINP [301]), blood coagulation (MCFD2 [302]), intracellular lipid receptors (OSBPL8), ribosomal proteins (RPS3), regulation of the production of beta-amyloid (RTN3 [303]) as well as proteins that function in electron transport and energy production (NAPRT1, NDUFB7 and SCO1). Thus, m^{5}C in these transcripts may have a role in post-transcriptional gene regulation that may affect a number of different cellular processes.

4.3.4.C Structural and sequence context of m^{5}C in mRNA

Preliminary global analyses were performed to identify enriched structural and sequence motifs surrounding candidate m^{5}C sites in mRNA that may be recognised by NSUN2. These analyses were performed for all candidate m^{5}C sites detected in HeLa cells and candidate m^{5}C sites called as NSUN2 targets. Understanding of the structural and sequence specificity of NSUN2 in mRNAs may elucidate the features of mRNA that may help achieve selective methylation.

Candidate m^{5}C sites detected in the CDS and 3’ UTR in HeLa cells were significantly enriched around stem-loop motifs, consistent with the structural motifs identified in the subset of ncRNAs investigated in this thesis. The candidate m^{5}C sites in the CDS were enriched in the stem, where the location was slightly bulged, and those in the 3’ UTR were enriched in the loop. Similarly, candidate m^{5}C sites called as NSUN2 targets were

enriched around stem-loop motifs, with sites located in the stem and the loop. Those candidate m^{5}C sites enriched in the stem a few nucleotides upstream of the stem-loop junction were in a similar structural context to the candidate m^{5}C sites called as NSUN2 targets at C30 in the anti-codon junction of tRNAs and in the stem near the stem-loop junction in vtRNAs. The candidate m^{5}C sites enriched in the loop are in a similar structural context to m^{5}C site in *SNORD62*. This indicates that the structural motifs that are recognised by NSUN2 are broad although can be similar between transcript biotypes, and once again highlights the multisite specificity of this enzyme.

No defined sequence motif was identified for candidate m^{5}C sites in mRNAs, although prevalence of a string of Gs was evident 2-4 nt downstream for candidate m^{5}C sites detected in HeLa cells and candidate m^{5}C sites called as NSUN2 targets. In addition, there was a mild preference for m^{5}CpG and m^{5}CpA over m^{5}CpT and m^{5}CpC for candidate m^{5}C sites called as NSUN2 targets. These features were highlighted for the m^{5}C sites validated in this study, which demonstrated a higher occurrence of m^{5}CpG and m^{5}CpA and a prevalence of Gs further downstream. Similarly, candidate m^{5}C sites in tRNAs called as NSUN2 targets by bsRNA-seq showed a prevalence of Gs, although this was 3-5 nt downstream of the site. It is possible that NSUN2 may recognise the multiple Gs to mediate m^{5}C, however as two of the validated mRNAs did not exhibit these Gs, this suggests that NSUN2 is capable of mediating m^{5}C in different sequence contexts. Previously, we discovered increased frequency of C and G as flanking bases [58]. Furthermore, the NSUN2 miCLIP study performed in human fibroblasts also showed an enrichment of a CG dinucleotides at the NSUN2 target site, however this was detected as an overall enrichment in all transcript biotypes [78].

In agreement with the findings for ncRNAs and tRNAs, these analyses highlight the multisite specificity of NSUN2. NSUN2 is capable of mediating m^{5}C sites in different sequence and structural contexts in mRNA, suggesting that other features of the mRNA substrate may be required to achieve such selective methylation.

4.3.5 Conclusions

In this chapter, I have described the abundance and distribution of candidate m5C sites in HeLa cells as well as the substrate range of NSUN2 and TRDMT1 and specificity of NSUN2 in particular. I have shown that candidate m5C sites are present in tRNAs, other ncRNAs and mRNAs and that NSUN2 and TRDMT1 are responsible for targeting at least a subset of these transcripts. I have thoroughly investigated m5C in tRNAs and have independently confirmed findings from recently published studies. I have also provided evidence of m5C occurring prior to splicing and 5' and 3' end processing of tRNAs, the latter of which is a novel finding. I have validated novel m5C sites in ncRNAs and mRNAs as NSUN2 targets and have also shown that candidate m5C sites in mRNAs are enriched in the 3' UTR, highlighting its non-random distribution that is suggestive of an important regulatory function. Furthermore, preliminary investigation into the substrate specificity of NSUN2 in ncRNAs and mRNAs have highlighted that NSUN2 is a multisite-specific RNA methyltransferase. These finding highlight the prevalence of m5C in mRNA and demonstrate for the first time that NSUN2 has a broad role in modifying mRNAs, in addition to some ncRNAs.

4.4 Chapter highlights

- This study was performed as an initial screen to identify novel RNA targets of NSUN2 and TRDMT1.
- 4,241 candidate m5C sites were detected in HeLa cells and distributed across tRNAs, other ncRNA and mRNAs, with the majority of candidate m5C sites detected in mRNAs.
- 393 candidate m5C sites were called as NSUN2 targets and 120 as TRDMT1 targets in tRNAs, other ncRNAs and mRNAs.

4.4.1 tRNA

- tRNA-focused bsRNA-seq data analysis confidently detected 185 candidate m5C sites in tRNAs, 87% of which were in one of seven anticipated structural positions: C34, C38, C40, C48-50 and C72.

- 108 of these sites were assigned to NSUN2. 107 of these sites were situated in positions, C34, C40, C48-50 across many different tRNA isotypes and isoacceptors. Thus, all positions known to be targeted by NSUN2 except C72 were revealed here. This demonstrated that NSUN2 is responsible for the large majority of m^{5}C sites in tRNAs, while displaying selectivity for several structural positions.
- TRDMT1 was assigned to 15 sites, with 14 of these sites at a single structural position, C38, in four tRNA types: tRNA$^{Asp(GUC)}$, tRNA$^{Gly(GCC)}$, tRNA$^{Val(AAC)}$ and tRNA$^{Val(CAC)}$. Thus, TRDMT1 is much more restricted in its tRNA substrate range.
- The above findings on NSUN2 and TRDMT1 tRNA substrate range and specificity revealed several newly detected individual sites of m^{5}C in tRNA, while broadly concurring with a large body of published work, including the recent transcriptome-wide studies by us and others.
- MiSeq amplicon sequencing of biological replicates confirmed that NSUN2 mediates m^{5}C sites at C48-50 of several tRNAs, and TRDMT1 mediates m^{5}C at C38 of tRNA$^{Gly(GCC)}$ and tRNA$^{Val(CAC)}$.
- MiSeq amplicon sequencing demonstrated that m^{5}C in tRNAs can occur prior to splicing, independently confirming previously published work.
- Analysis of bsRNA-seq reads mapping to pre-tRNA sequences across all tRNA loci exhibiting candidate m^{5}C sites showed, for the first time, that cytosines in tRNAs are modified prior to 5’ and 3’ end processing.

4.4.2 Other ncRNA

- bsRNA-seq detected 667 candidate m^{5}C sites in other ncRNAs, 33 and 9 of which were assigned to NSUN2 and TRDMT1, respectively.
- The bsRNA-seq data was selectively mined for m^{5}C sites in ncRNAs.
- It confidently detected the previously known m^{5}C sites at C3782 and C4447 in 28S rRNA and C841 in 12S mitochondrial RNA, which were refractory to NSUN2 or TRDMT1 knockdown. This matched with expectations that the m^{5}C:RNA methyltransferases NSUN5A, NSUN1 and NSUN4, respectively, are responsible for these sites, as shown in yeast and/or mice.

- NSUN2-mediated m^{5}C sites in *SCARNA2* and *RPPH1* were detected by bsRNA-seq and validated by MiSeq amplicon sequencing, independently confirming recently published work.
- Similarly, entirely novel m^{5}C sites in *SNORD62*, *NSUN5P2* and the pseudogene in the intron of *CAGE1* were discovered and validated as NSUN2 targets.
- Visual inspection of the local context around the NSUN2-mediated m^{5}C sites in *RPPH1*, *SCARNA2*, *SNORD62*, vtRNA1-1 and vtRNA1-3 indicated variability in secondary structure and sequence, highlighting the multisite specificity of this enzyme.

4.4.3 mRNA

- bsRNA detected 1,564 candidate m^{5}C sites in mRNAs, 118 and 53 of which were assigned to NSUN2 and TRDMT1, respectively.
- Candidate m^{5}C sites were enriched in the 3' UTR of mRNAs, which is a hotspot for regulatory motifs, suggesting a role for m^{5}C in post-transcriptional gene control.
- Focusing on NSUN2 targets, MiSeq amplicon sequencing of biological replicates confirmed previously known NSUN2 target sites in *CINP*, *NAPRT1* and *NDUFB7*, as well as validating the novel NSUN2 target sites in *MCFD2*, *OSBPL8*, *RPS3*, *RTN3*, *SCO1* and *SZRD1*.
- Motif searches in the vicinity of all mRNA candidate m^{5}C sites, or the NSUN2-targeted subset, revealed that stem-loop containing structures were enriched as was a string of Gs 2-4 nt downstream of the m^{5}C site. There was also a slight preference for m^{5}CpG and m^{5}CpA. Some of these features are reminiscent of NSUN2 sites in tRNA, hinting at a unifying principle underlying NSUN2 substrate recognition in different RNA biotypes that warrants further investigation.
- Overall, these findings highlight the unexpected prevalence of m^{5}C sites in mRNAs and demonstrate for the first time that NSUN2 has a broad role in modifying mRNAs.

Chapter Five

Exploring the molecular functions of m^5C in mRNA

5.1 INTRODUCTION

The availability of next-generation sequencing technologies that has enabled transcriptome-wide mapping of several RNA modifications has begun to profoundly change the understanding of their prevalence in the transcriptome. Nucleoside modifications including m^{5}C, m^{6}A and ψ are not confined to specific ncRNAs of the translational machinery but are further non-randomly distributed throughout the transcriptome, suggesting that they serve multiple regulatory functions. There has been a flurry of research in this field of "epitranscriptomics", and an initial understanding of the molecular functions of these nucleoside modifications in mRNA is beginning to emerge. For instance, a number of studies have implemented transcriptome-wide sequencing technologies for the detection of m^{6}A, revealing its abundance in mRNA [56, 57, 68, 69]. This has enabled the investigation of the molecular function of m^{6}A in mRNA, with an emerging consensus that it is involved in regulating mRNA stability and alternative splicing [57, 69, 105, 107, 120].

The investigation of the molecular functions of m^{5}C in RNA has primarily focused on rRNAs, tRNAs and recently also vtRNAs. In tRNAs, m^{5}C mediated by TRDMT1 at C38 in tRNA$^{Asp(GUC)}$, tRNA$^{Gly(GCC)}$ and tRNA$^{Val(AAC)}$ in *D. melanogaster* protects the tRNAs against ribonuclease cleavage, and loss of this modification results in decreased viability under stress conditions [73]. In humans, angiogenin mediates stress-induced cleavage of tRNAs and *in vitro* cleavage of tRNAs in *D. melanogaster* by human angiogenin was reduced when TRDMT1 was overexpressed [73, 304]. Recently, it was reported that disruption of NSUN2 and TRDMT1 in mice led to a decrease in tRNA steady-state levels and an overall decrease in the rate of protein synthesis [149]. A recent study has also revealed that the absence of NSUN2, and therefore ablation of m^{5}C in vtRNAs and the majority of m^{5}C sites in tRNAs, resulted in aberrant processing into svRNAs [78, 150]. svRNAs can associate with Ago and regulate gene expression in a similar manner to miRNAs, suggesting that m^{5}C in vtRNAs may regulate gene expression of mRNAs [299].

The role of m^{5}C in mRNAs remains elusive, however we have previously reported enrichment of Ago footprints upstream of m^{5}C sites in mRNA, while another post-

transcriptional regulator, Pumilio 2 (PUM2), showed no enrichment in these regions [58]. This points to a role for m^{5}C in mRNA in post-transcriptional gene regulation, motivating further research into this area.

The aim of this chapter was to investigate the molecular functions of m^{5}C in mRNA in further detail. The roles of m^{5}C in mRNA may be diverse and context-dependent, thus an investigation was made into several possible molecular functions for this modification. We achieved this using bioinformatics analysis of the newly generated HeLa methyltransferase knockdown bsRNA-seq data as well as experimental investigations using a candidate approach and focusing on NSUN2 mRNA targets.

5.2 Results

The following sections detail investigations into the molecular functions of m^{5}C in mRNA. Mapping of the MiSeq amplicon sequencing data was performed by Dr. Brian J. Parker, who also performed the global analyses of the bsRNA-seq data in consultation with myself. I performed the experimental investigations including RT-qPCR and sucrose density gradient centrifugation. I also performed the statistical analyses of the experimental data with the help of Dr. Teresa Neeman using GenStat at the Statistical Consulting Unit.

5.2.1 Gene set enrichment and GO analyses of methylated mRNAs

To gain insight into the molecular functions of m^{5}C in mRNA, gene set enrichment analyses (GSEA) and GO analyses were performed on the bsRNA-seq data. GO terms are grouped into three ontologies; biological process, molecular function and cellular component [305]. GO analysis was performed using the topGO R package [247] to determine which biological processes were significant for mRNAs exhibiting candidate m^{5}C sites. The elim method was used for this analysis, which investigated the GO terms from bottom to top to ensure the most specific terms were scored. This method also iteratively removes genes that mapped to significant GO terms before scoring the parent terms. An example of a parent term is “membrane organisation and biogenesis”, which

can have the progeny “membrane fusion”, which is more specific. GSEA provides a more general analysis focusing on related gene sets involved in certain pathways or complexes using the Kyoto Encyclopedia of Genes and Genomes (KEGG). This provides a link between a set of genes in the genome with a network of interacting molecules such as a pathway or complex [306, 307]. Pathways include metabolism, genetic information processing, environmental information processing, cellular processes, organismal systems and human diseases. Thus, it represents a higher order of biological function. For this analysis, the gage R package was used [246].

Both sets of analyses were performed on all mRNAs exhibiting candidate m^5C sites in HeLa cells, as well as mRNAs with candidate m^5C sites called as NSUN2 and TRDMT1 targets, based on the bsRNA-seq data. The background set used to detect enrichment were all expressed genes from the bsRNA-seq, which was defined as at least one cytosine within a gene having ≥5 reads.

GSEA using the gage R package did not reveal enrichment of particular pathways or complexes for all candidate m^5C site-containing mRNAs in HeLa cells, as well as mRNAs with candidate m^5C sites called as NSUN2 and TRDMT1 targets. For the set of all candidate m^5C site-containing mRNAs in HeLa cells, GO terms relating to a diverse range of cellular processes were enriched (Table 5.1). GO terms enriched for the list of NSUN2- or TRDMT1-targeted transcripts were similarly diverse (Tables 5.2 & 5.3). Interestingly, several GO terms relating to mRNA metabolism were enriched in all three lists of genes (coloured in green in the following tables). The process of mRNA translation was particularly prominent, with all three phases including initiation, elongation and termination being represented. The full list of enriched GO terms are in Appendix 8.7.

Table 5.1. GO term enrichment of mRNAs containing candidate m^{5}C sites in HeLa cells based on topGO.

The top 30 biological processes in order of significance are listed.

Term ID	Term Description	Number of significant transcripts	P-value
GO:0006415	translational termination	71	5.7E-30
GO:0044281	small molecule metabolic process	759	1.4E-28
GO:0044419	interspecies interaction between organisms	71	1.1E-25
GO:0015031	protein transport	224	1.1E-19
GO:0007165	signal transduction	333	1E-18
GO:0051301	cell division	66	7.6E-18
GO:0000084	S phase of mitotic cell cycle	42	8.5E-18
GO:0007596	blood coagulation	74	1.2E-17
GO:0007411	axon guidance	42	8.5E-17
GO:0030154	cell differentiation	197	2.2E-16
GO:0008284	positive regulation of cell proliferation	50	2.5E-16
GO:0006457	protein folding	54	2.5E-16
GO:0042493	response to drug	39	1.8E-15
GO:0016032	viral reproduction	132	2.3E-15
GO:0008283	cell proliferation	132	1.9E-14
GO:0000216	M/G1 transition of mitotic cell cycle	32	2.3E-14
GO:0002479	antigen processing and presentation of exogenous peptide antigen via MHC class I, TAP-dependent	32	2.3E-14
GO:0048011	nerve growth factor receptor signaling pathway	34	3.4E-14
GO:0016071	mRNA metabolic process	179	6.8E-14
GO:0045944	positive regulation of transcription from RNA polymerase II promoter	37	9.5E-14
GO:0030168	platelet activation	33	9.6E-14
GO:0055085	transmembrane transport	75	1.3E-13
GO:0007049	cell cycle	178	2.1E-13
GO:0006886	intracellular protein transport	164	2.2E-13
GO:0006916	anti-apoptosis	45	2.2E-13
GO:0042981	regulation of apoptosis	156	8.7E-13
GO:0006810	transport	430	9.3E-13
GO:0000398	nuclear mRNA splicing, via spliceosome	48	3E-12
GO:0019221	cytokine-mediated signaling pathway	43	3.7E-12
GO:0045892	negative regulation of transcription, DNA-dependent	62	3.7E-12

Table 5.2. GO term enrichment of NSUN2-targeted mRNAs in HeLa cells based on topGO.

The top 30 biological processes in order of significance are listed.

Term ID	Term Description	Number of significant transcripts	P-value
GO:0006413	translational initiation	16	0.00000014
GO:0006414	translational elongation	14	0.00000036
GO:0044281	small molecule metabolic process	82	0.0000049
GO:0006200	ATP catabolic process	6	0.000016
GO:0048513	organ development	26	0.000022
GO:0006987	activation of signaling protein activity involved in unfolded protein response	6	0.000023
GO:0051084	'de novo' posttranslational protein folding	6	0.000032
GO:0006415	translational termination	11	0.000044
GO:0008283	cell proliferation	19	0.000054
GO:0051384	response to glucocorticoid stimulus	5	0.000079
GO:0000184	nuclear-transcribed mRNA catabolic process, nonsense-mediated decay	11	0.000082
GO:0019538	protein metabolic process	62	0.000086
GO:0019083	viral transcription	11	0.000093
GO:0055085	transmembrane transport	11	0.00012
GO:0019221	cytokine-mediated signaling pathway	8	0.00014
GO:0006614	SRP-dependent cotranslational protein targeting to membrane	11	0.00015
GO:0030168	platelet activation	7	0.00017
GO:0008150	biological_process	135	0.00027
GO:0009888	tissue development	12	0.00027
GO:0006355	regulation of transcription, DNA-dependent	20	0.00042
GO:0010605	negative regulation of macromolecule metabolic process	14	0.00043
GO:0071310	cellular response to organic substance	24	0.00057
GO:0001701	in utero embryonic development	5	0.00066
GO:0010646	regulation of cell communication	12	0.00074
GO:0019752	carboxylic acid metabolic process	16	0.00079
GO:0009057	macromolecule catabolic process	24	0.00081
GO:0071495	cellular response to endogenous stimulus	10	0.00091
GO:0009607	response to biotic stimulus	8	0.00095
GO:0006810	transport	47	0.0011
GO:0006120	mitochondrial electron transport, NADH to ubiquinone	4	0.00113

Table 5.3. GO term enrichment of TRDMT1-targeted mRNAs in HeLa cells based on topGO.

The top 30 biological processes in order of significance are listed.

Term ID	Term Description	Number of significant transcripts	P-value
GO:0044238	primary metabolic process	36	0.000015
GO:0006457	protein folding	6	0.000023
GO:0000184	nuclear-transcribed mRNA catabolic process, nonsense-mediated decay	6	0.00016
GO:0006468	protein phosphorylation	8	0.00034
GO:0006413	translational initiation	6	0.00047
GO:0051495	positive regulation of cytoskeleton organization	2	0.00067
GO:0018279	protein N-linked glycosylation via asparagine	3	0.0007
GO:0006259	DNA metabolic process	8	0.00072
GO:0044237	cellular metabolic process	36	0.00095
GO:0006936	muscle contraction	3	0.00099
GO:0006415	translational termination	5	0.00103
GO:0010564	regulation of cell cycle process	5	0.0011
GO:0032413	negative regulation of ion transmembrane transporter activity	2	0.0011
GO:0048468	cell development	8	0.00114
GO:0006414	translational elongation	5	0.00131
GO:0019083	viral transcription	5	0.00147
GO:0045787	positive regulation of cell cycle	2	0.00164
GO:0051208	sequestering of calcium ion	2	0.00164
GO:0055007	cardiac muscle cell differentiation	2	0.00164
GO:0030029	actin filament-based process	4	0.00171
GO:0006614	SRP-dependent cotranslational protein targeting to membrane	5	0.00183
GO:0009968	negative regulation of signal transduction	4	0.00184
GO:0071310	cellular response to organic substance	6	0.00191
GO:0043687	post-translational protein modification	3	0.00199
GO:0043161	proteasomal ubiquitin-dependent protein catabolic process	4	0.00213
GO:0007059	chromosome segregation	3	0.00251
GO:0006974	response to DNA damage stimulus	5	0.00261
GO:0007276	gamete generation	5	0.00264
GO:0010604	positive regulation of macromolecule metabolic process	8	0.00264
GO:0033554	cellular response to stress	9	0.00274

5.2.2 Investigating the association of m^{5}C with Ago and miRNAs

The role of a given m^{5}C site in mRNA is likely context-dependent, suggesting that this modification may serve different molecular functions (see Figure 1.5). Previously, we found enrichment of Ago footprints upstream of candidate m^{5}C sites in mRNAs in addition to enrichment of m^{5}C sites in the 3' UTR, suggesting a relationship between m^{5}C and Ago [58]. These analyses were done with Ago PAR-CLIP data from HEK293 cells, as no direct HeLa cell data was available at the time. As discussed in Chapter 4, candidate m^{5}C sites were again found to be enriched in the 3' UTR based on the newly generated HeLa methyltransferase knockdown dataset. To investigate the potential relationship between m^{5}C sites and Ago further, the newly generated HeLa methyltransferase knockdown data was integrated with HeLa Ago2 HITS-CLIP data that is now publicly available [308]. The published study performed HITS-CLIP on HeLa cells transfected with a miRNA mimic and a non-targeting control. For this analysis, the control HITS-CLIP dataset was integrated with the bsRNA-seq data. The bsRNA-seq data was also integrated with a publicly available HeLa HuR PAR-CLIP dataset [244]. As a negative control, candidate m^{5}C sites detected were assigned random positions within the same genomic region.

This analysis was performed on all candidate m^{5}C sites detected in HeLa cells, as the datasets for the NSUN2 and TRDMT1 knockdown samples were too small and resulted in substantial noise. Genomic coordinates from the bsRNA-seq data were aligned with those in the Ago2 and HuR datasets and the footprint density of Ago2 and HuR were examined across regions in the 3' UTR and CDS flanking candidate m^{5}C sites. A 2.8-fold enrichment of Ago2 footprints relative to the random shuffled control was evident in the 3' UTR within 25 nt upstream on average of candidate m^{5}C sites ($p \approx 0$, permutation test, 10 repetitions) (Figure 5.1A). Similarly, a 2.3-fold enrichment of Ago2 footprints relative to the random shuffled control was evident in the CDS within 25 nt upstream on average of candidate m^{5}C sites ($p \approx 0$, permutation test, 10 repetitions) (Figure 5.1B). No enrichment of HuR was detected near candidate m^{5}C sites in the 3' UTR ($p=0.0910$, permutation test, 10 repetitions) or CDS ($p=0.2727$, permutation test, 10 repetitions) relative to the random shuffled control.

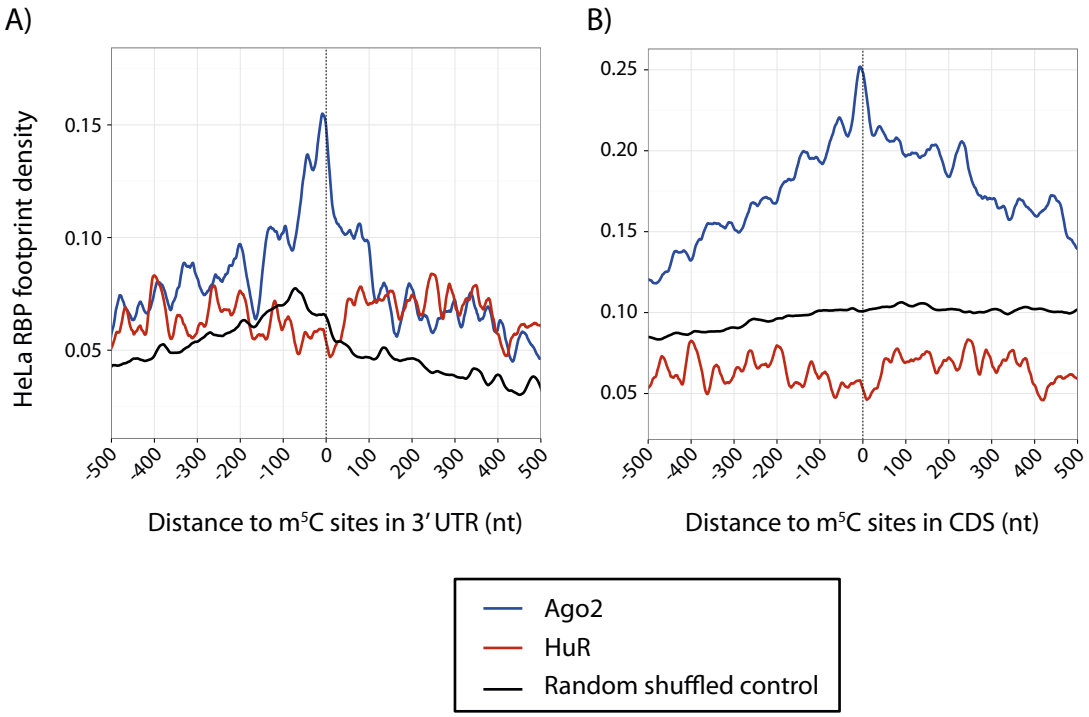


Figure 5.1. HeLa Ago2 and HuR footprint density surrounding candidate m^{5}C sites in mRNA.

The HeLa methyltransferase knockdown dataset was integrated with the HeLa Ago2 HITS-CLIP dataset [308] and HuR PAR-CLIP dataset [244]. Spatial distribution summary plots were generated by anchoring at candidate m^{5}C sites in mRNAs in the 3' UTR and CDS. The random shuffled control was generated by randomly reassigning the genomic location of candidate m^{5}C sites to another position within each region. 500 nt regions flanking the candidate m^{5}C sites are shown. A) Significant enrichment of Ago2 footprints was evident upstream of candidate m^{5}C sites in the 3' UTR relative to the random shuffled control ($p \approx 0$, permutation test, 10 repetitions). No enrichment of HuR footprints was evident near candidate m^{5}C sites in the 3' UTR ($p=0.0910$, permutation test, 10 repetitions). B) A significant enrichment of Ago2 footprints was evident upstream of candidate m^{5}C sites in the CDS relative to the random shuffled control ($p \approx 0$, Permutation test, 10 repetitions). No enrichment of HuR footprints was evident near candidate m^{5}C sites in the CDS ($p=0.2727$, Permutation test, 10 repetitions).

To investigate the relationship of Ago2 and m^5C in more detail, the density of conserved miRNA target sites surrounding all candidate m^5C sites in HeLa cells was determined in the 3' UTR and CDS. 7-mer motifs were selected genome-wide that showed a PhyloP conservation score of >1.3 and were divided into subsets based on whether they were the reverse complement of miRNA seed regions in miRBase v19. The PhyloP score is a measure of how conserved a base is, and a score of >1.3 retains only those motifs that are moderately to highly conserved. The conserved 7-mer motifs that were the reverse complement of miRNA seed regions in miRBase v19 were used as the final set of conserved miRNA target sites. The conserved 7-mer motifs that were not the reverse complement of miRNA seed regions in miRBase v19 were used as a negative control. As an additional negative control, candidate m^5C sites were assigned random positions within the same genomic region.

The 3' UTR showed a 55% decrease of 7-mer miRNA targets at candidate m^5C sites relative to the random shuffled control ($p \approx 0$, permutation test, 10 repetitions) (Figure 5.2A). A 36% decrease of 7-mer miRNA targets at candidate m^5C sites relative to the random shuffled control was evident in the CDS ($p \approx 0$, permutation test, 10 repetitions) (Figure 5.2B). The spatial distribution summary plot for candidate m^5C sites in the 3' UTR was noisier than the CDS; this is attributed to the smaller length of the 3' UTR compared to the CDS and the smaller number of data points for this analysis.

Depletion of non-miRNA 7-mer motifs was not observed in either the 3' UTR or CDS, although a slight enrichment was evident in the CDS. To determine whether the depletion of miRNA motifs was a result of a lower cytosine frequency within these regions, miRNA target density was determined surrounding unmethylated cytosines, with no depletion evident.

The analysis performed in this section indicates a potential role for m^5C in post-transcriptional gene regulation through association with Ago2.

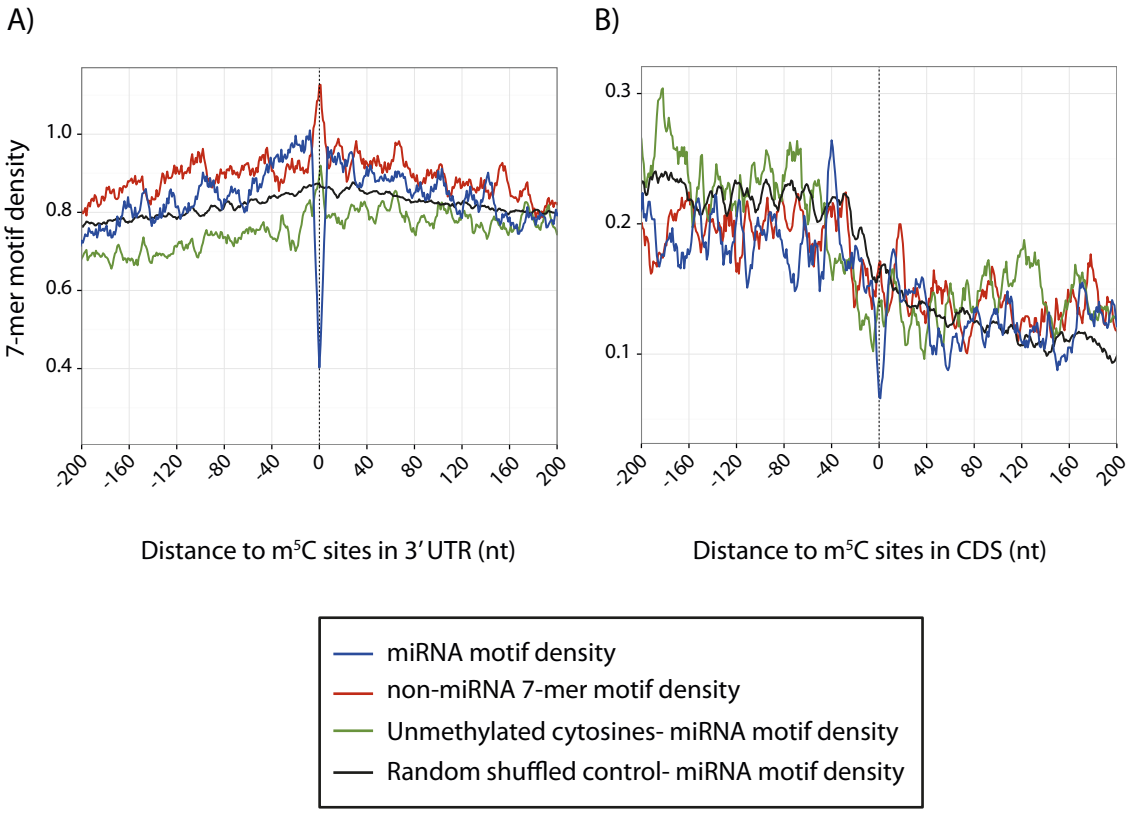


Figure 5.2. Density of conserved 7-mer motifs surrounding candidate m^{5}C sites in mRNA. Conserved miRNA target sites were determined by ensuring they exhibited a PhyloP conservation score of >1.3 and were the reverse complement of miRNA seed regions in miRBase v19. Spatial distribution summary plots were generated by anchoring at candidate m^{5}C sites in the 3' UTR and CDS with the conserved 7-mer motif density on the y-axis. The density of those conserved 7-mer motifs that were not the reverse complement of miRNA seed regions in miRBase v19 were also determined as a negative control. To determine if the depletion of conserved 7-mer miRNA motifs at candidate m^{5}C sites was a result of biases in the distribution of bases, the density of conserved 7-mer miRNA motifs was plotted for unmethylated cytosines. The random shuffled control was generated by randomly reassigning the genomic location of candidate m^{5}C sites to another position within each region. 200 nt regions flanking the candidate m^{5}C sites are shown. A) Depletion of miRNA target sites was evident at candidate m^{5}C sites in the 3' UTR relative to the random shuffled control ($p \approx 0$, permutation test, 10 repetitions). B) Depletion of miRNA target sites was evident at candidate m^{5}C sites in the CDS relative to the random shuffled control ($p \approx 0$, permutation test, 10 repetitions).

5.2.3 Analysis of global steady-state levels of methylated mRNAs

It was of interest to perform an initial screen to determine if depletion of m^5C in mRNAs resulted in a change in their steady-state level, by making additional use of the HeLa methyltransferase knockdown bsRNA-seq data. As NSUN2 was found to target a larger number of mRNAs compared to TRDMT1, this analysis was performed on NSUN2-targeted mRNAs and the fold-change determined between the NSUN2 knockdown sample and TRDMT1 knockdown and NTC samples as controls. The DNMT1 knockdown sample was excluded from this analysis on the basis that DNMT1 causes substantial pleiotropic changes in gene expression.

The analysis was performed using Voom weighting of raw read counts over UCSC canonical genes and the differential expression determined using the Limma pipeline [240]. The Limma pipeline uses a moderated t-test to detect significantly differentially expressed transcripts. As the data used for this analysis was based on next-generation sequencing, Voom weighting was applied where those transcripts with higher read counts (higher confidence) were given a higher weight.

The mean expression level in counts per million (CPM) of NSUN2-targeted mRNAs was determined for the NSUN2 knockdown sample relative to the controls. In addition, to determine if the knockdown of NSUN2 resulted in an indirect global effect on expression such as affecting transcription, the mean expression levels were determined for mRNAs without detectable methylation for the NSUN2 knockdown sample relative to the controls. Compared to the mean expression level of unmethylated mRNAs, the NSUN2-targeted mRNAs exhibited an 11.9% decrease in mean steady-state expression levels in the NSUN2 knockdown sample ($p=2 \times 10^{-4}$, Wilcoxon test) (Figure 5.3). Thus, as NSUN2 knockdown is unlikely to selectively affect transcription of its target, this analysis suggests a role for NSUN2-mediated m^5C sites in stabilising mRNA.

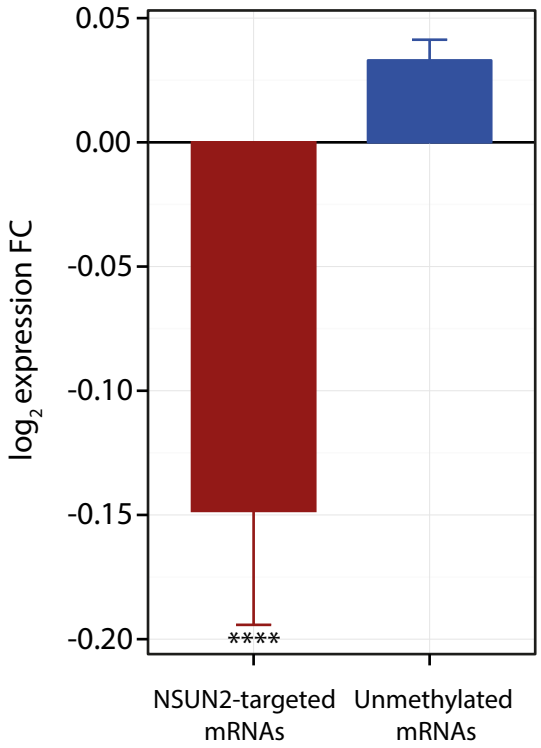


Figure 5.3. Global analysis of the average steady-state level of mRNAs containing candidate m^{5}C sites called as NSUN2 targets based on bsRNA-seq.

Expression over each category was determined using the Limma pipeline and Voom weighting of raw read counts. The mean expression level of mRNAs called as NSUN2 targets was determined in CPM and the fold-change (FC) between the NSUN2 knockdown sample and the TRDMT1 knockdown and NTC samples was calculated. The mean expression fold-change was also calculated for unmethylated mRNAs as a control. A significant decrease in the mean expression of NSUN2-targeted transcripts was evident compared to unmethylated mRNAs ($p=2 \times 10^{-4}$, Wilcoxon test). Data represented as mean fold-change of expression of all mRNAs in each category with SEM.

5.2.4 Analysis of steady-state levels of methylated mRNAs using RT-qPCR

A 11.9% decrease in the mean steady-state level of NSUN2-targeted mRNAs in the NSUN2 knockdown sample might be driven by a small reduction in levels consistently for many mRNAs among the group, or a more diverse response with a pronounced change for some examples. A consistent but small change would be hard to detect using RT-qPCR, nevertheless, a candidate approach using RT-qPCR was implemented to explore whether evidence for the latter scenario could be obtained.

Two approaches were used to examine the effect of m^5C on mRNA steady-state levels. First, an investigation was made into whether the knockdown of NSUN2 and therefore the depletion of m^5C in mRNA resulted in a decrease in expression of each individual mRNA. As a second approach, an investigation was made into whether a decrease in the mean steady-state level in all mRNAs that were interrogated was evident, similar to the global analysis used for the bsRNA-seq data.

5.2.4.A Determining housekeeping genes for normalisation of RT-qPCR data

Careful selection of housekeeping genes for the normalisation of the RT-qPCR data was made to ensure they did not contain candidate m^5C sites, which may affect their steady-state levels. A list of housekeeping genes was determined from [309] and were filtered to remove any transcripts that exhibited candidate m^5C sites or a large standard deviation in expression between samples based on the bsRNA-seq data (see Appendix 8.8). Ultimately, three housekeeping genes were selected based on these selection criteria and included the transcripts *MRPL9, H2AFV* and *TCF25*. The geometric mean of these three housekeeping genes was calculated for final normalisation of the expression data.

5.2.4.B Steady-state level analysis of candidate mRNAs targeted by NSUN2

The expression levels of six mRNAs that were validated to contain m^{5}C sites mediated by NSUN2 were investigated in the NSUN2 knockdown, TRDMT1 knockdown and NTC samples. These included *MCFD2*, *NAPRT1*, *NDUFB7*, *OSBPL8*, *RTN3* and *SZRD1*. As with the global analysis, the DNMT1 knockdown sample was removed from this investigation. Both the TRDMT1 knockdown and NTC samples were used as controls in which mRNA expression levels of NSUN2 targets were expected to remain unchanged.

Six biological replicates of the RNA methyltransferase knockdown samples were used for this analysis. This included the three biological replicates described in Chapter 3 of which one was used for the validation of candidate m^{5}C sites in Chapter 4, and an additional three biological replicate knockdowns that were performed by myself. Bisulfite conversion and knockdown efficiency were assessed for two of three new biological replicates by MiSeq amplicon sequencing. A segment of the non-humanised R-Luc *in vitro* transcript and the depletion of m^{5}C at the known sites in tRNA$^{Asp(GUC)}$ and tRNA$^{Leu(CAA)}$ were assessed for bisulfite conversion efficiency and knockdown efficiency, respectively.

In tRNA$^{Asp(GUC)}$, depletion from ~70% methylation in the NTC samples to ~7% in the TRDMT1 knockdown samples was evident at C38 (Figure 5.4A). Depletion from ~94% methylation in the NTC samples to ~45% in the NSUN2 knockdown samples was evident at C48 and C49. In unspliced tRNA$^{Leu(CAA)}$, depletion from ~50% methylation in the NTC samples to ~7% in the NSUN2 knockdown samples was evident at C34, while in the spliced transcript, depletion from ~85% methylation in the NTC samples to ~20% in the NSUN2 knockdown samples was evident (Figure 5.4B). Thus, this indicates sufficient knockdown of NSUN2 and TRDMT1. Additionally, efficient bisulfite conversion was evident in all samples (Figure 5.4C).

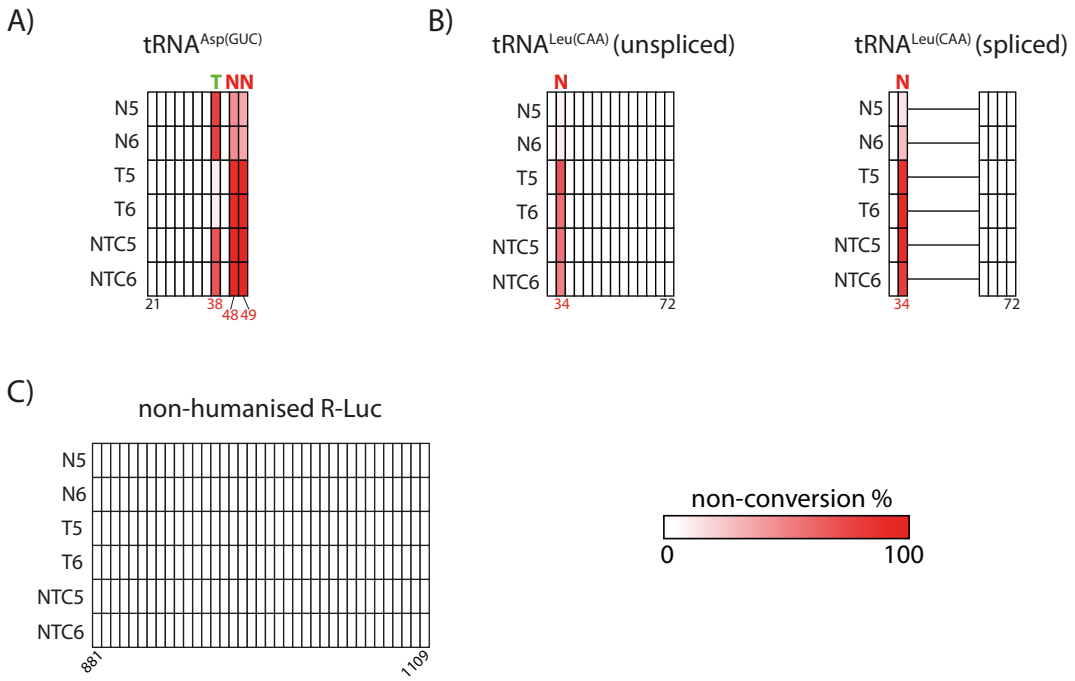


Figure 5.4. Confirmation of depletion of m^{5}C at known target sites of NSUN2 and TRDMT1 in the replicate NSUN2 and TRDMT1 knockdown samples.

A) MiSeq amplicon sequencing of tRNA$^{Asp(GUC)}$ showed depletion of m^{5}C at C38 in the TRDMT1 knockdown samples and C48 and C49 in the NSUN2 knockdown samples. B) MiSeq amplicon sequencing of tRNA$^{Leu(CAA)}$ showed depletion of m^{5}C at C34 in the NSUN2 knockdown samples. The tRNA structural positions with known m^{5}C sites are indicated in red below the heatmap. The N and T indicate sites that are directed by NSUN2 and TRDMT1. C) MiSeq amplicon sequencing of a segment of the non-humanised R-Luc *in vitro* transcript showed efficient bisulfite conversion. Numbers below the heatmap indicate the segment of the non-humanised R-Luc *in vitro* transcript interrogated relative to the start of the transcript. MiSeq amplicon sequencing was performed on two out of three of the biological replicate knockdowns. The MiSeq amplicon sequencing coverage ranged from 35,000-244,000 mappable sequencing reads.

To ensure that all six mRNAs that were interrogated for steady-state levels by RT-qPCR exhibited the expected amounts of methylation, MiSeq amplicon sequencing was performed for two out of three new biological replicates of the NSUN2 knockdown, TRDMT1 knockdown and NTC samples. Thus, out of the six biological replicates used for this analysis, m^{5}C sites were examined in three biological replicates, including the knockdown samples used for the validation of candidate m^{5}C sites in Chapter 4. Figure 5.5 shows a schematic of the samples used for the analysis of candidate steady-state levels.

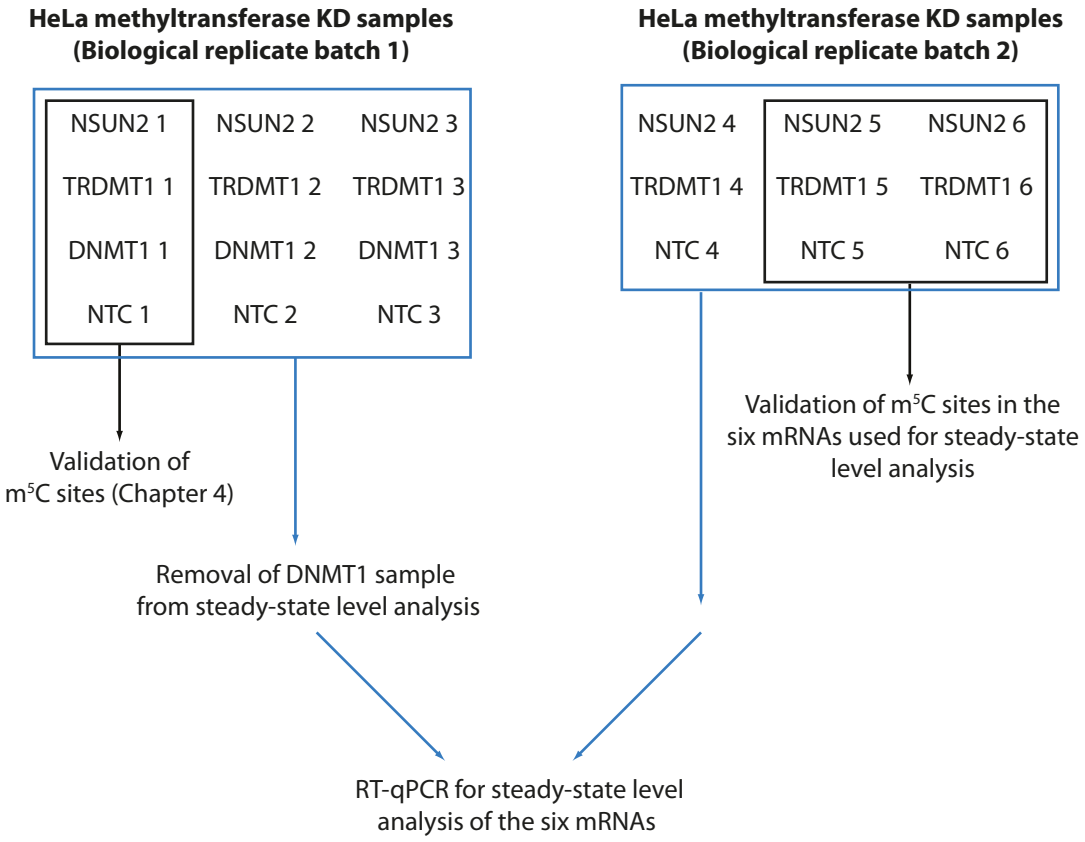


Figure 5.5. Schematic demonstrating the samples used for the validation of candidate m^{5}C sites and those used for the analysis steady-state levels.

RNAi-mediated knockdown (KD) of the methyltransferases were performed twice in triplicate. One replicate from biological replicate batch 1 was used for the validation of candidate m^{5}C sites in Chapter 4. Two replicates from biological replicate batch 2 were used for the validation of candidate m^{5}C sites in the six mRNAs interrogated for steady-state levels. All six replicates were used for the steady-state level analysis in this thesis.

All six mRNAs exhibited methylation in the NTC samples and depletion of methylation in the NSUN2 knockdown samples, confirming that these m^5C sites were mediated by NSUN2 in at least three of the six biological replicates (Figure 5.6) One replicate of the TRDMT1 knockdown samples exhibited methylation in all six mRNAs, while the other only exhibited methylation in five mRNAs, with *MCFD2* not containing the m^5C site in this replicate. As a result, the expression level of *MCFD2* was not examined in this replicate. In addition, the MiSeq coverage of *MCFD2* in one replicate of the NTC sample was quite low, with only 16 reads. This may be a result of a low and incorrect amount of amplicon being pooled prior to MiSeq library preparation.

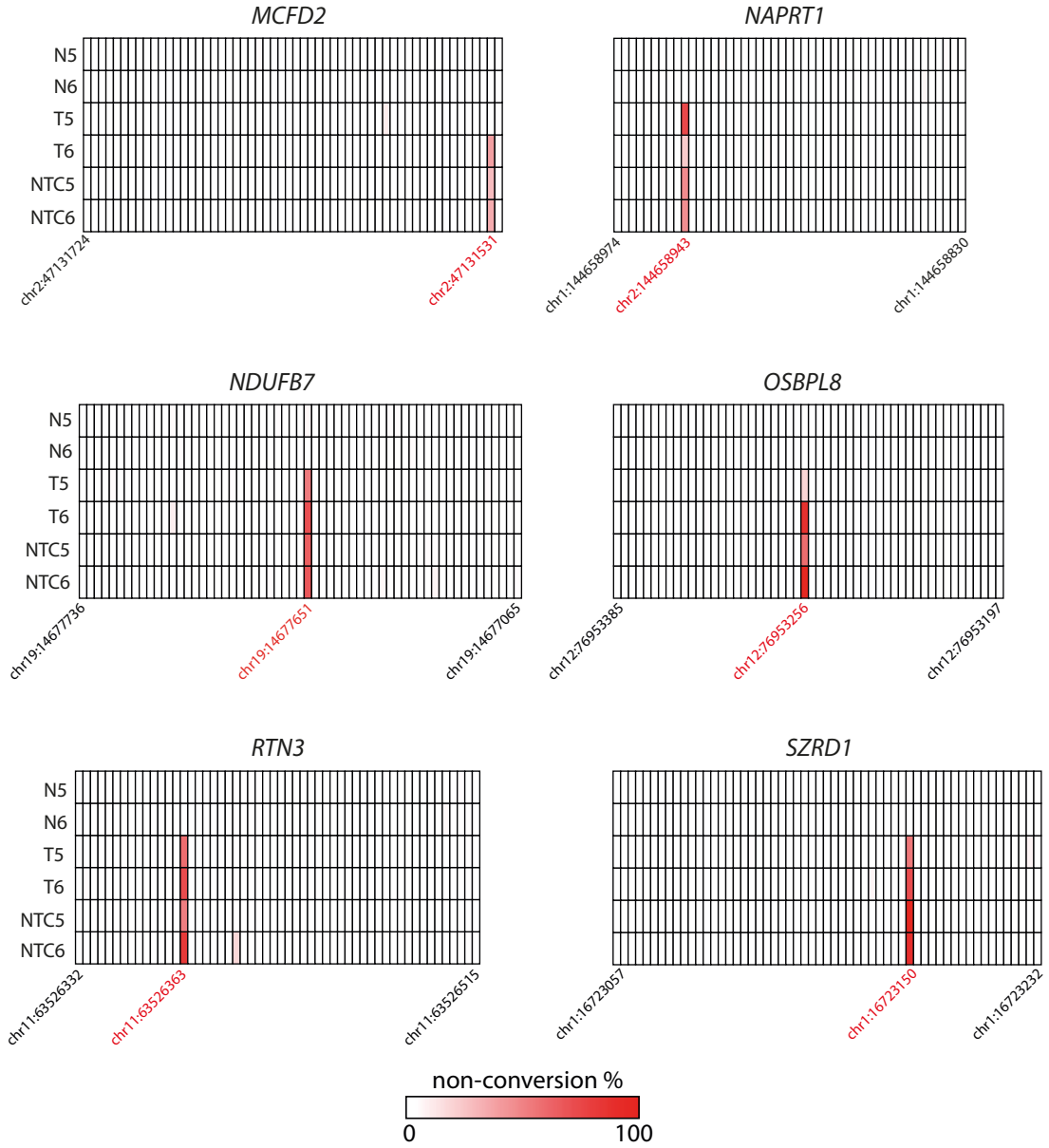


Figure 5.6. Validation of m^{5}C sites in six mRNAs in which steady-state levels were investigated by RT-qPCR in the replicate NSUN2 and TRDMT1 knockdown samples.

MiSeq amplicon sequencing was performed on two out of three of the new biological replicates of the NSUN2 knockdown (N), TRDMT1 knockdown (T) and NTC samples. Chromosomal locations of the m^{5}C sites are indicated in red below the heatmap. The MiSeq amplicon sequencing coverage ranged from 16-230,000 mappable sequencing reads.

As the m^5C sites in the six mRNAs were confirmed as NSUN2 targets in several replicates, RT-qPCR was performed to examine the expression levels of these transcripts. The RT-qPCR revealed that there was no significant difference in the steady-state levels of these transcripts in the NSUN2 knockdown sample compared to the controls (Figure 5.7). Interestingly, although it was hypothesised that knockdown of TRDMT1 would not result in a change in the expression of these transcripts, a modest yet significant increase in expression of *MCFD2* and *NDUFB7* relative to both the NSUN2 knockdown and NTC samples was observed. *RTN3* showed increased expression in the TRDMT1 knockdown sample relative to the NSUN2 knockdown sample and *SZRD1* showed decreased expression in the TRDMT1 knockdown sample relative to the NSUN2 knockdown and NTC samples.

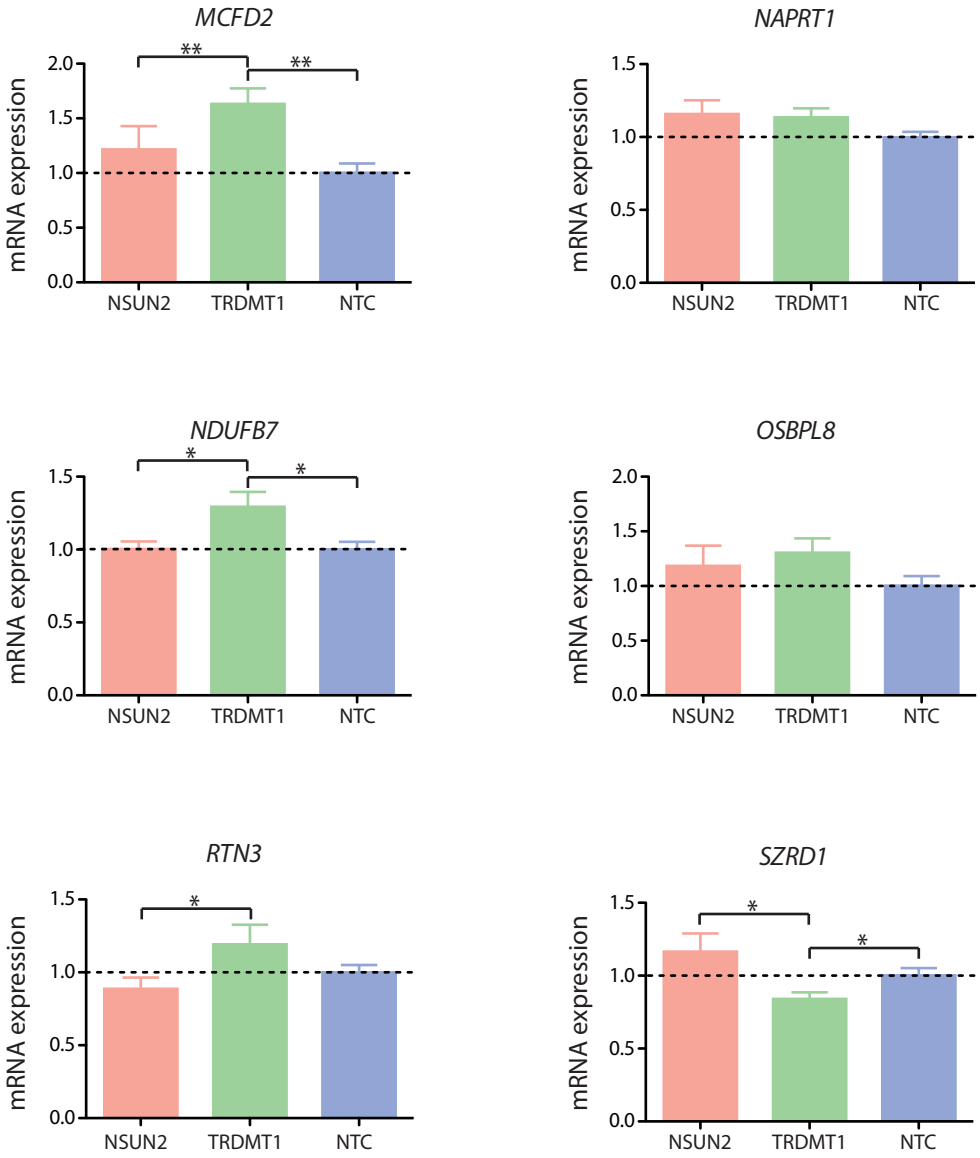


Figure 5.7. Steady-state level of mRNAs validated to contain NSUN2-mediated m^{5}C sites in the HeLa methyltransferase knockdown cells.

Expression levels of mRNAs validated to contain m^{5}C sites regulated by NSUN2 were determined using RT-qPCR in NSUN2 knockdown sample as well as TRDMT1 knockdown and NTC samples, the latter two which were controls. Expression levels were normalised to the geometric mean of *H2AFV*, *MRPL9* and *TCF25* mRNA levels. Data shown as mean fold-change relative to NTC with SEM (n=5-6). There was no significant difference in mRNA steady-state levels in the NSUN2 knockdown sample compared to the NTC sample. A significant increase in *MCFD2* and *RTN3* expression levels and significant decrease in *SZRD1* expression levels was evident in the TRDMT1 knockdown sample (** $p<0.01$, * $p<0.05$, two-way ANOVA with Fisher's protected least significant difference post-hoc analysis).

For the statistical analysis, a two-way ANOVA was used due to the batch variation between the biological replicates (Figure 5.8). Batch variation is common with studies where different batches of cells were used. In this case, the siRNA knockdowns were performed in triplicate in one batch of HeLa cells, while the second set of knockdowns were performed in triplicate in another batch of HeLa cells. A two-way ANOVA took into account two variables, including the batch variation, which was removed from this statistical analysis, and variation due to siRNA knockdown, which was the variation of interest. The initial analysis of variance determined whether there was a difference between all samples by determining the means. In the case where a variance was detected ($p<0.05$), a post-hoc analysis using Fisher's protected least significance test was performed to determine which of the samples varied (Table 5.4). The complete output from the statistical analysis is located in Appendix 8.9.

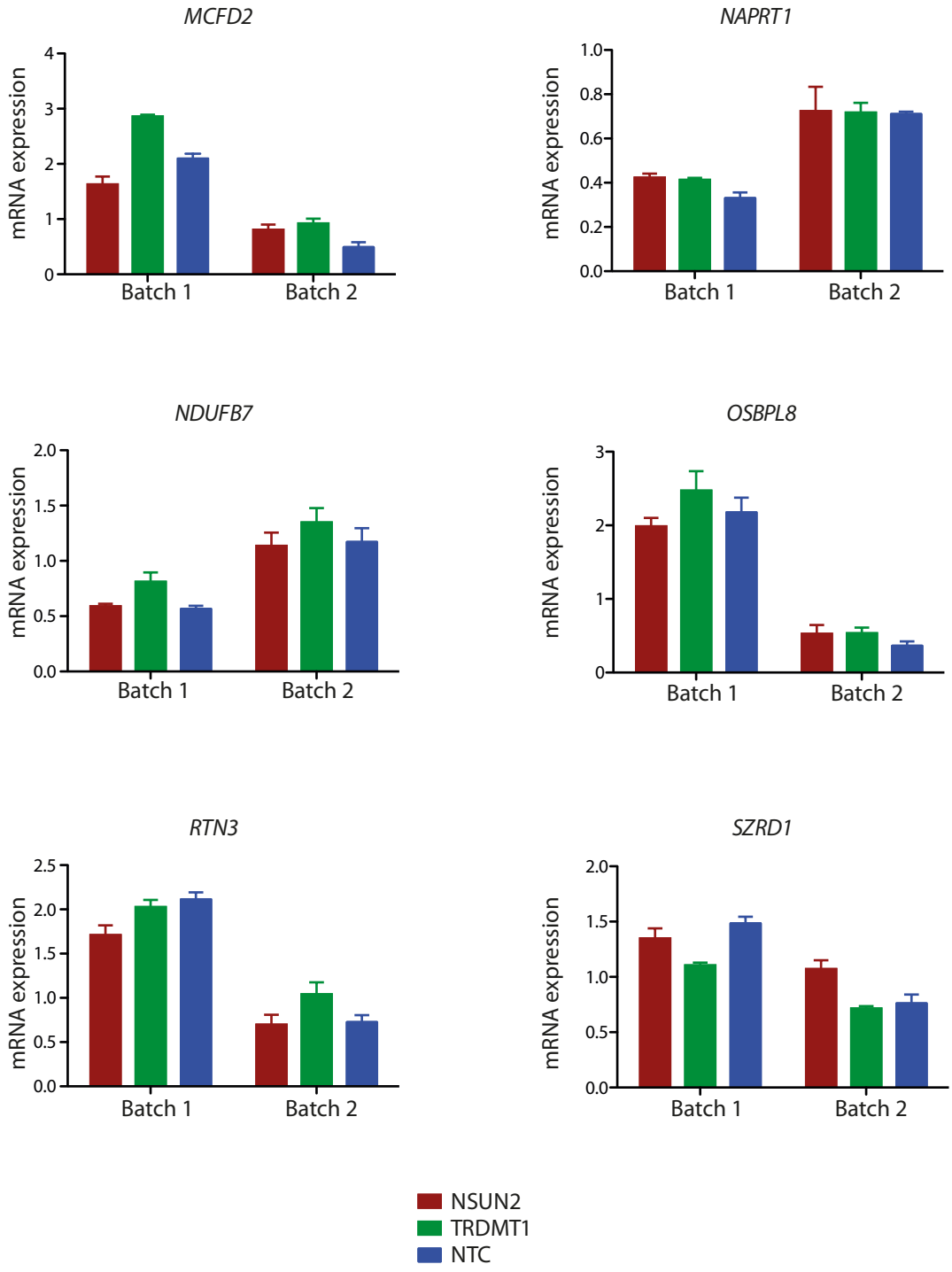


Figure 5.8. Batch variation evident after RT-qPCR of mRNAs validated to contain NSUN2-mediated m^{5}C sites.

Expression levels are normalised to the geometric mean of *H2AFV*, *MRPL9* and *TCF25*. Data is **not** shown as mean fold-change relative to NTC to illustrate the batch variation. Data shown as mean with SEM (n=2-3).

Table 5.4. Two-way ANOVA statistical analysis to determine differences in expression levels of NSUN2-targeted mRNAs.

Transcript	Analysis of variance p-value	Least significant difference of means	Mean of mRNA expression			Significant against TRDMT1 or NTC based on post-hoc analysis?	
			NSUN2	TRDMT	NTC	TRDMT1	NTC
MCFD2	0.003	0.376	1.219	1.893	1.293	Yes	No
NAPRT1	0.468	N/A	N/A			N/A	
NDUFB7	0.049	0.193	0.862	1.080	0.869	Yes	No
OSBPL8	0.255	N/A	N/A			N/A	
RTN3	0.032	0.239	1.204	1.531	1.421	Yes	No
SZRD1	0.011	0.189	1.211	0.906	1.120	Yes	No

5.2.4.C Analysis of overall steady-state level of candidate mRNAs targeted by NSUN2

Similar to the analysis used on the bsRNA-seq data, the mean steady-state level over all six NSUN2-targeted mRNAs was investigated in the NSUN2 knockdown sample. In order to do this, a secondary two-way ANOVA analysis was performed to determine whether the trend detected in the bsRNA-seq was also detected using the RT-qPCR candidate approach. There was no significant decrease in the mean steady-state level of the candidate mRNAs detected in the NSUN2 knockdown sample compared to the controls (Table 5.5). However, TRDMT1 showed a significant increasing trend compared to both the NSUN2 knockdown and NTC samples.

Table 5.5. Two-way ANOVA statistical analysis to determine differences in overall steady-state levels of NSUN2-targeted mRNAs.

Treatment	Analysis of variance p-value	Least significant difference of means	Mean of all transcripts	Mean of all transcripts in NTC	Significant against NTC based on post-hoc analysis
NSUN2	<0.001	0.134	1.05	1.08	No
TRDMT1			1.25		Yes

The inability to detect a change in the steady-state levels of the six NSUN2-targeted mRNAs upon NSUN2 knockdown is most likely a result of the limitations of the sensitivity of the RT-qPCR assay to detect the subtle decrease found in the global analysis. In addition, the global analysis revealed a decrease over all NSUN2-targeted mRNAs, suggesting the possibility of a small group of these transcripts exhibiting a pronounced decrease that may have not been exhibited in the six candidates examined by RT-qPCR. The RT-qPCR assay thus failed to provide evidence for an effect of m^{5}C on the steady-state levels of the six chosen candidate mRNAs.

5.2.5 Investigating the role of m^{5}C in translation

An investigation was made into whether m^{5}C may function in post-transcriptional gene regulation through regulation of mRNA translation. Depending on the context of the m^{5}C site, translation may be regulated in different ways. We hypothesised that if m^{5}C is involved in regulation of translation, the proportion of methylated mRNA molecules should be different when comparing translated with non-translated molecules. As detailed below, sucrose density gradient centrifugation was employed to separate mRNAs with validated m^{5}C sites into populations of different translation state.

5.2.5.A Assessing the reproducibility of bisulfite conversion and PCR amplification

The candidate approach employed for the investigation of m^{5}C in regulating translation involved PCR amplification of validated candidates at different translation states followed by MiSeq amplicon sequencing to determine the proportion of methylation at each translation state. One of the caveats of this approach is the possibility of the abundance of a transcript affecting the bisulfite conversion efficiency or causing PCR amplification biases that may result in variability in the proportion of methylation. To address this, an experiment was designed in which aliquots from a single HeLa RNA sample were bisulfite converted at different times with fresh sodium bisulfite solution. MiSeq amplicon sequencing was performed for various transcript biotypes with validated m^{5}C sites in order to compare the proportion of methylation within each transcript of each replicate. PCR of each amplicon was performed in triplicate for each replicate, as previously described (Figure 5.9).

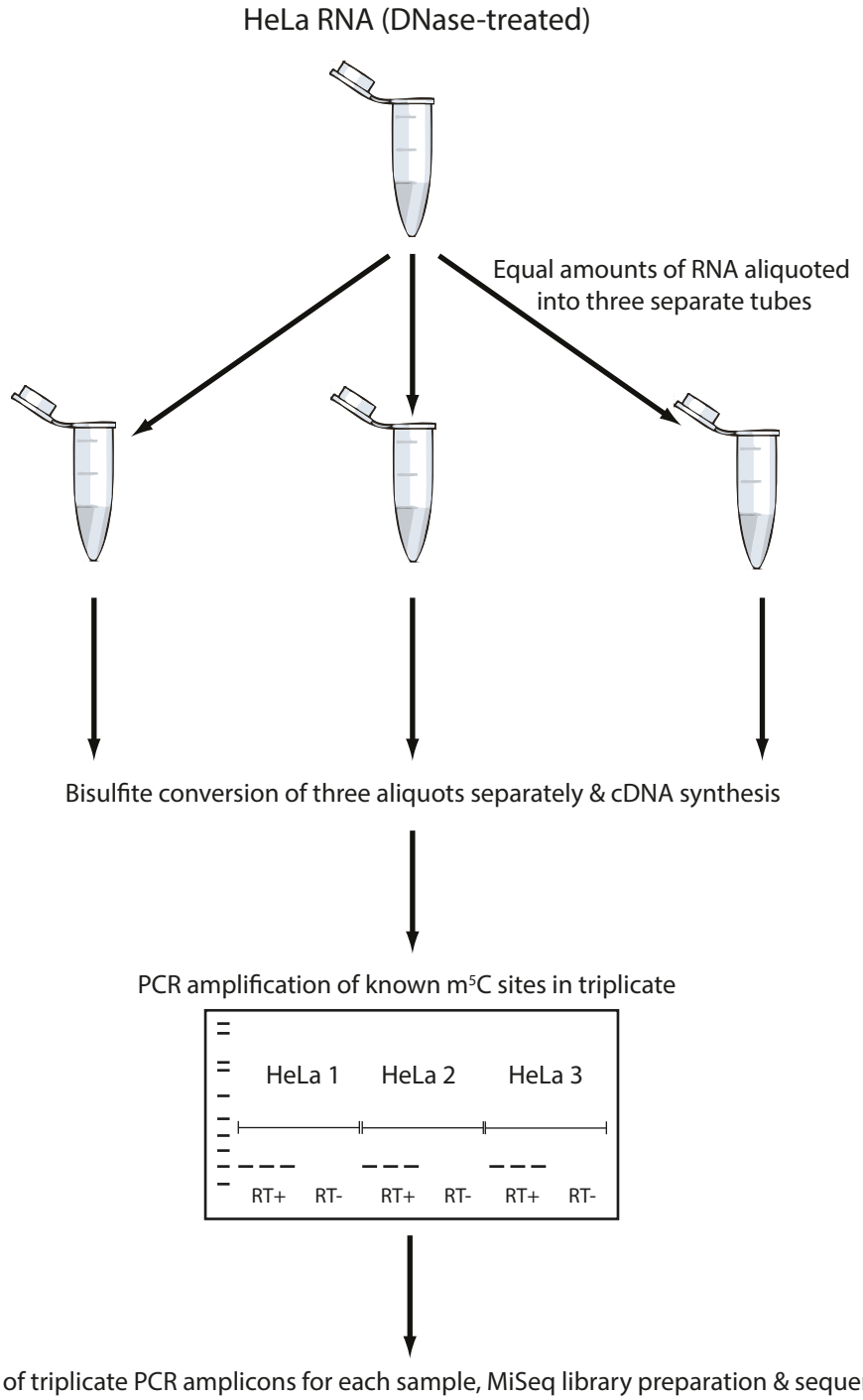


Figure 5.9. Experimental design to investigate the reproducibility of bisulfite conversion and PCR amplification.

A single HeLa RNA sample that had been previously DNase-treated was divided into three separate tubes and bisulfite converted at different times using freshly made sodium bisulfite solution. The bisulfite-converted RNA samples were reverse transcribed as previously described and PCR amplification of known m^{5}C sites were performed in triplicate for each HeLa sample. The triplicate PCR products were pooled and MiSeq amplicon libraries were prepared then sequenced.

Several transcript biotypes were investigated, including mRNAs, tRNAs and another ncRNA. These included *CINP*, *MCFD2*, *NAPRT1*, *NDUFB7*, *RTN3*, *SZRD1*, *RPPH1*, $tRNA^{Asp(GUC)}$ and $tRNA^{Leu(CAA)}$, which all contain validated m^5C sites. In addition to sequencing these transcripts, a segment of the non-humanised R-Luc *in vitro* transcript was sequenced to ensure efficient bisulfite conversion; >99% mean cytosine conversion was evident in all HeLa cell replicates (Figure 5.10 and Table 5.6).

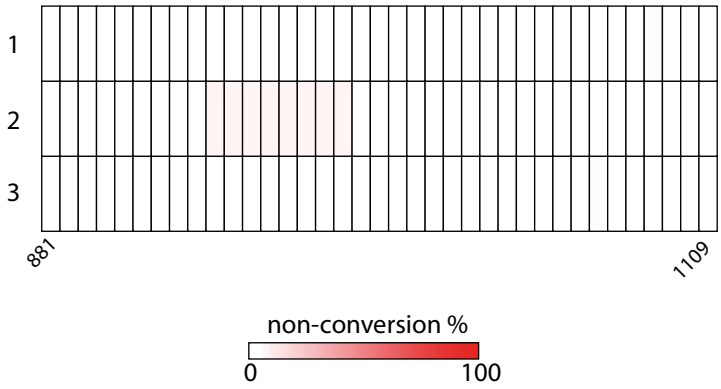


Figure 5.10. Confirmation of bisulfite conversion in each HeLa technical replicate by MiSeq amplicon sequencing.

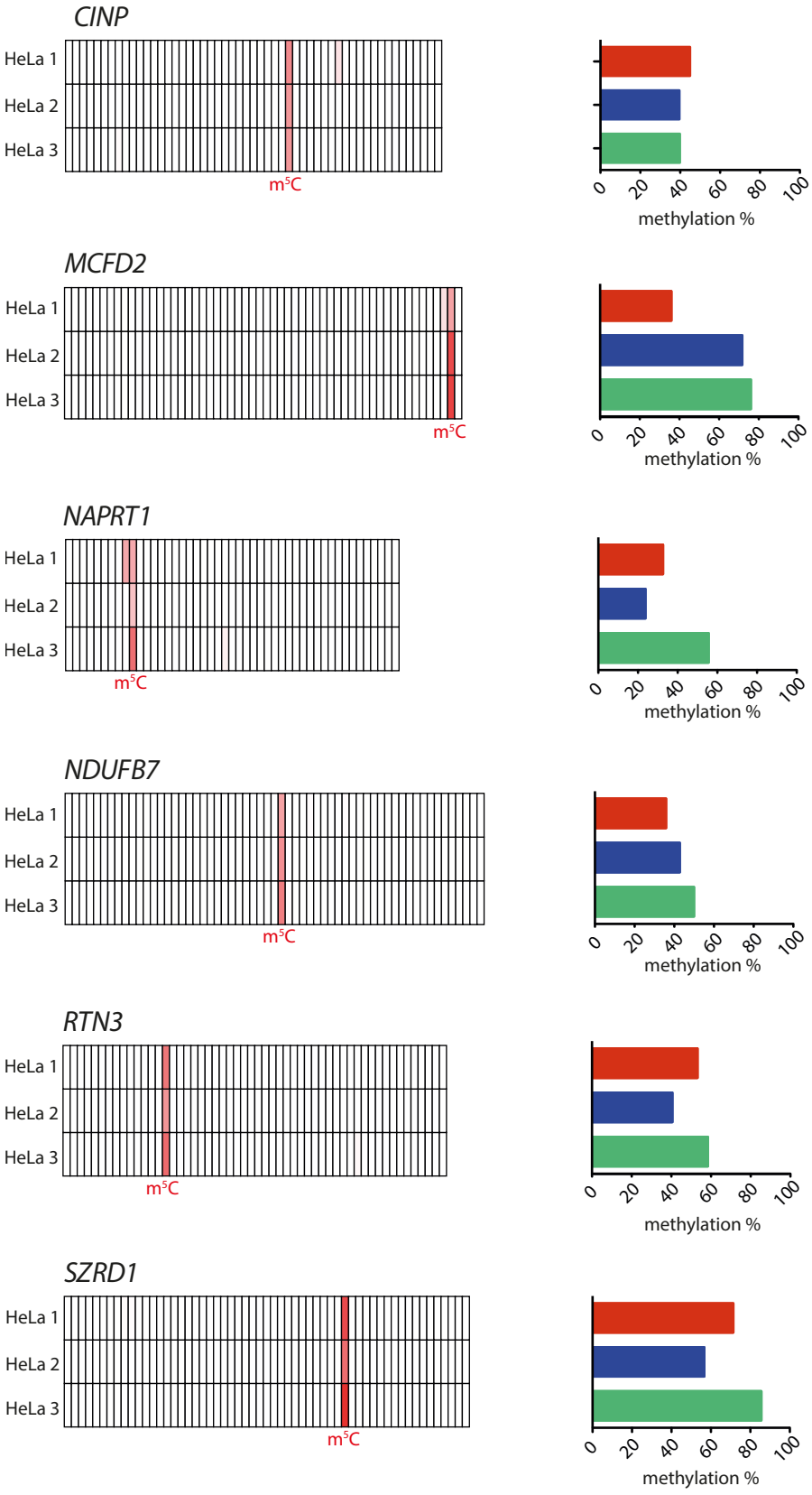
One segment of the non-humanised R-Luc *in vitro* transcript was amplified and subjected to MiSeq amplicon sequencing. Numbers below the heatmap indicate the segment of the transcript interrogated relative to the start of the transcript. The MiSeq amplicon sequencing coverage ranged from 34,000-42,000 mappable sequencing reads.

Table 5.6. Mean cytosine conversion of the non-humanised R-Luc *in vitro* transcript.

Replicate	Mean cytosine conversion (%)
1	99.83
2	99.20
3	99.86

Examination of the proportion of methylation for the mRNAs revealed some variability between the technical replicates (Figure 5.11). The proportion of methylation in *CINP*, *NDUFB7*, *RTN3* and *SZRD1* was relatively consistent between the replicates. The proportion of methylation in *MCFD2* was consistent between two replicates, however the third replicate exhibited approximately half the proportion of methylation compared

to first two replicates. There was also some variability between the replicates in *NAPRT1*. In contrast, the proportion of methylation in *RPPH1*, $tRNA^{Asp(GUC)}$ and $tRNA^{Leu(CAA)}$ were consistent between all replicates. As *MCFD2* and *NAPRT1* showed greater variability in the proportion of methylation in each replicate, these transcripts were removed from the translation analysis.


CINP
HeLa 1
HeLa 2
HeLa 3
m5C
0 20 40 60 80 100
methylation %
MCFD2
HeLa 1
HeLa 2
HeLa 3
m5C
0 20 40 60 80 100
methylation %
NAPRT1
HeLa 1
HeLa 2
HeLa 3
m5C
0 20 40 60 80 100
methylation %
NDUFB7
HeLa 1
HeLa 2
HeLa 3
m5C
0 20 40 60 80 100
methylation %
RTN3
HeLa 1
HeLa 2
HeLa 3
m5C
0 20 40 60 80 100
methylation %
SZRD1
HeLa 1
HeLa 2
HeLa 3
m5C
0 20 40 60 80 100
methylation %

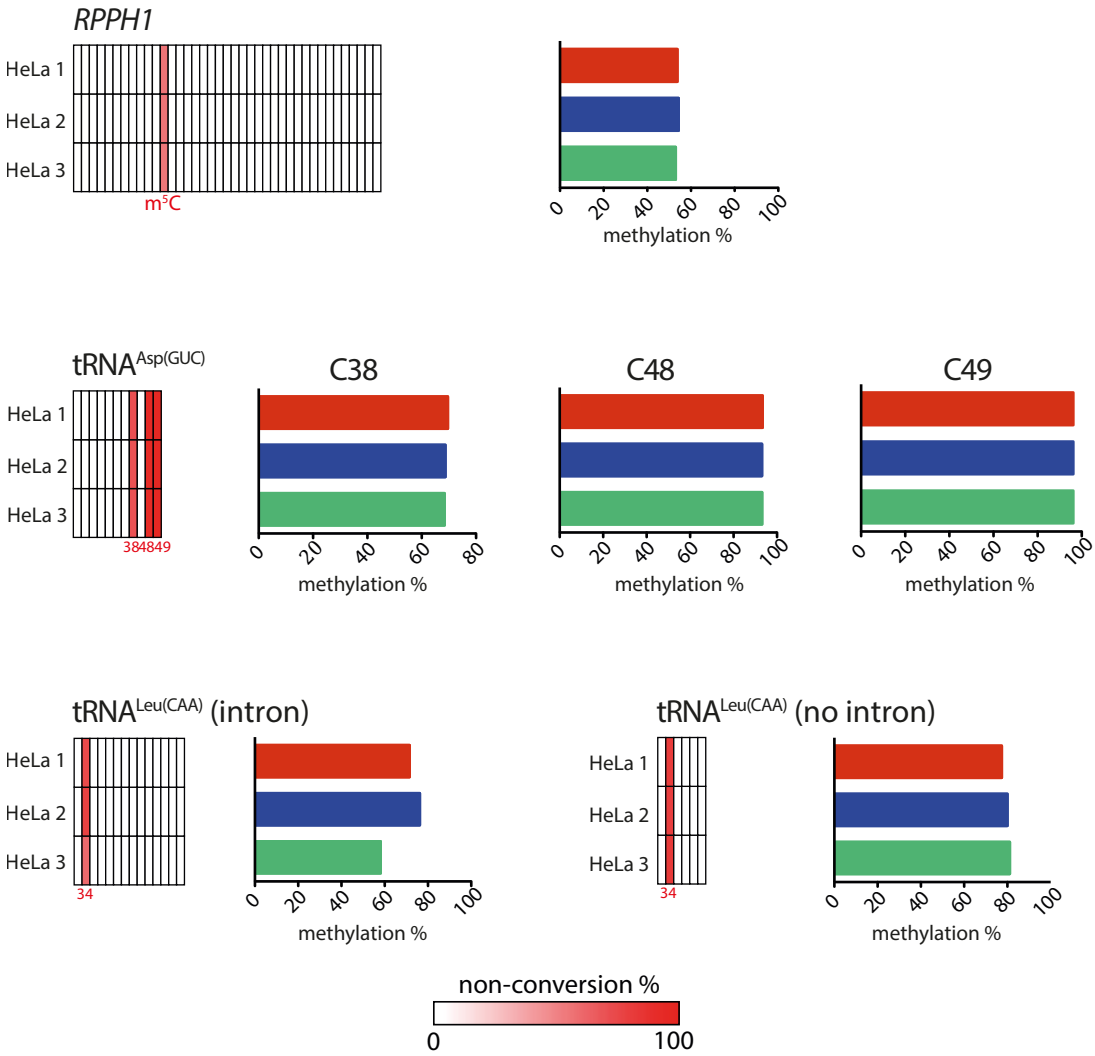


Figure 5.11. The proportion of methylation at m^{5}C sites in three HeLa cell technical replicates as determined by MiSeq amplicon sequencing.

The location of the m^{5}C sites is indicated below the heatmap. The bar charts to the right represent the proportion of methylation at m^{5}C sites in each transcript for the three HeLa cell technical triplicates. The MiSeq amplicon sequencing coverage ranged from 6,000-246,000 mappable sequencing reads.

5.2.5.B Sucrose density gradient centrifugation and polysome fractionation

In order to investigate the role of m^{5}C in regulating translation, sucrose density gradient centrifugation was performed using two biological replicates of HeLa cells. A 17.5%-50% sucrose gradient was generated, which, after ultracentrifugation separated mRNAs into those co-sedimenting with multiple ribosomes (polysomes) from those that co-sediment with a single ribosome. Those transcripts associated with polysomes, and therefore more actively translated, settle down to the higher density sucrose, whereas those transcripts that are associated with a single ribosome are not highly translated and are within the lower density sucrose.

HeLa cells were seeded at a density of 7 x 10^6 cells per 15 cm cell culture dish and incubated with cycloheximide to a final concentration of 0.1 mg/mL for 3 min at 37°C in order to inhibit translation. Cell culture media was removed and cells were washed twice with ice-cold PBS containing 0.1 mg/mL cycloheximide. Following this, cells were lysed and the cleared lysate was loaded onto the sucrose gradient and centrifuged at 35,000 rpm for 2 h 15 min at 4°C. The polysome gradient was then fractionated into 24 tubes while measuring the absorbance at 254 nm (Figure 5.12).

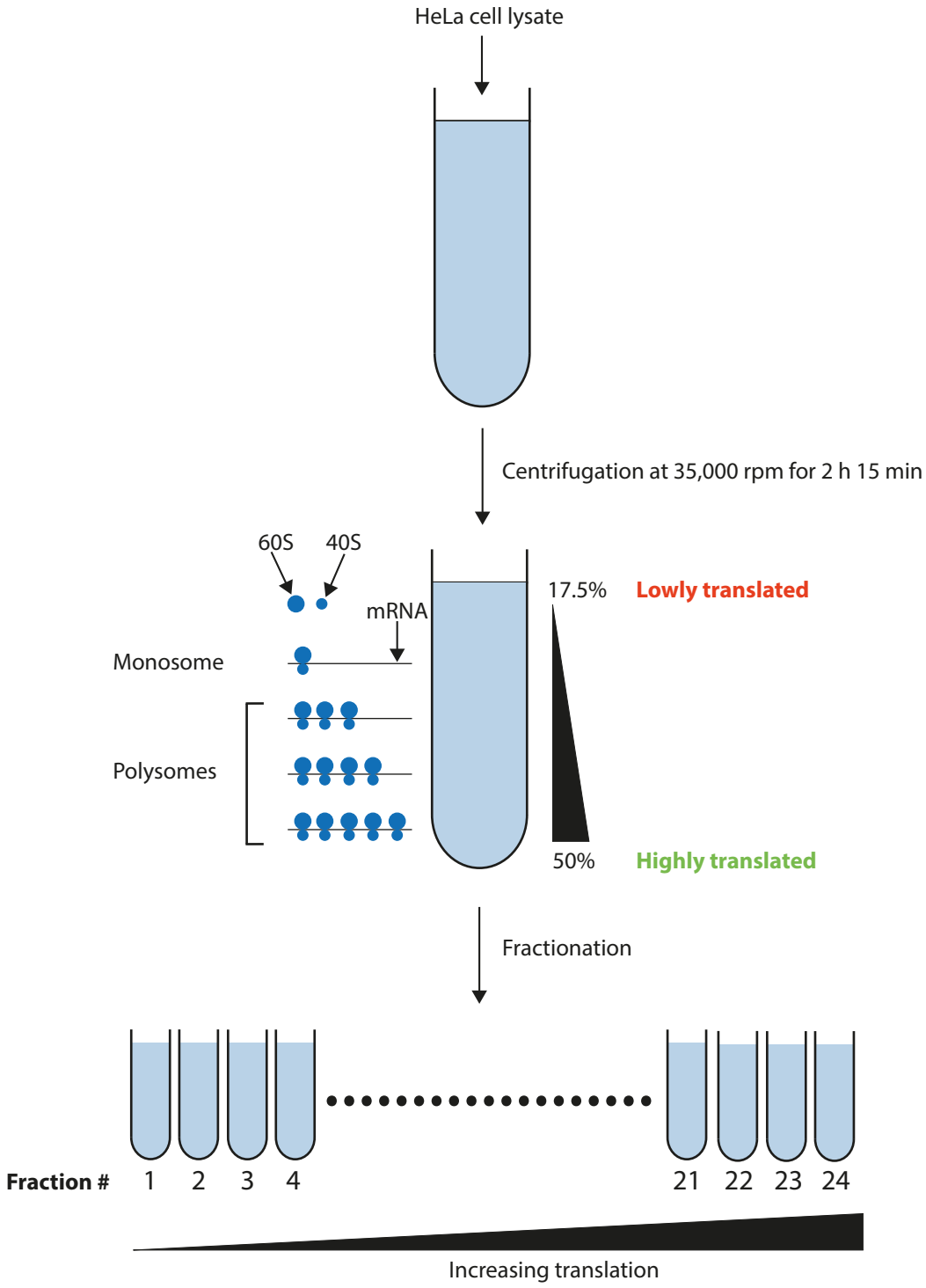


Figure 5.12. Schematic of sucrose density gradient centrifugation.

The HeLa cell lysate was carefully pipetted onto at 17.5-50% sucrose gradient in a polyallomer tube and centrifuged at 35,000 rpm for 2 h 15 min at 4°C. mRNA associated with polysomes settle to the higher density sucrose while mRNA associated with a single ribosome are within the lower density sucrose. The gradient was then fractionated into 24 tubes while measuring absorbance at 254 nm.

As expected, three peaks corresponding to the 40S subunit, 60S subunit and the 80S ribosome were observed (Figure 5.13A). Polysomes were clearly visible and more abundant than the ribosome and ribosome subunits. Collected fractions were precipitated overnight with three volumes of ethanol and 4 µL glycogen in preparation for RNA extraction. A technical challenge with RNA extraction from sucrose gradients is achieving uniform quality and efficiency for each fraction as a result of varying amounts of RNA in each fraction, and the high concentration of sucrose in the heavier fractions interfering with phase separation in typical phenol-based extraction steps. To address the issue of differences in amounts of RNA, 1 ng of R-Luc 3xB pA *in vitro* transcript was spiked into each fraction prior to overnight precipitation to correct for differences in RNA recovery and reverse transcription efficiency. The sucrose was removed by precipitation of the total nucleic acid and protein with ethanol, and washing the precipitate with 70% ethanol. The RNA was then extracted using Tri Reagent®, the fractions were combined to 12 fractions and the RNA examined on a RNA 6000 Nano Chip (Figure 5.13B). It was clear that tRNAs were predominantly present in the first three fractions, and 18S and 28S rRNA were visible from the fourth fraction. To reduce the number of samples further, the fractions were combined again to 6 fractions. For the remainder of this chapter, the fractions will be referred to by their final combined number.

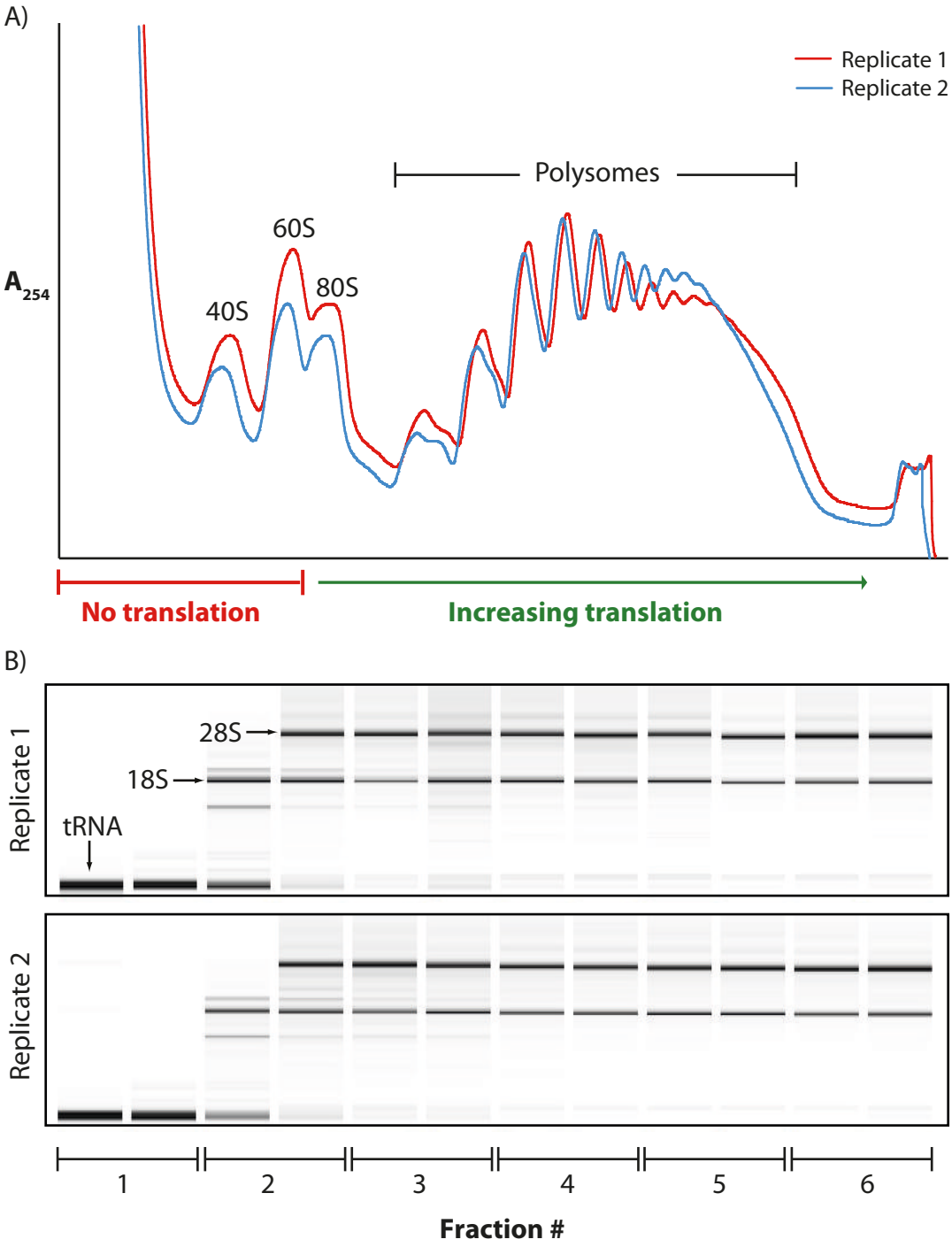

A)
Replicate 1
Replicate 2
Polysomes
60S
40S
80S
A254
No translation
Increasing translation
B)
Replicate 1
28S
18S
tRNA
Replicate 2
1
2
3
4
5
6
Fraction #

Figure 5.13. Absorbance trace at 254 nm of HeLa cells and analysis of the RNA integrity after polysome gradient fractionation.

HeLa cells were seeded at a density of 7 x 10^6 cells per 15 cm dish and harvested 24 h later in lysis buffer, and subjected to density gradient ultracentrifugation. A) The absorbance profile at 254 nm was recorded and fractions collected. The polysomes, ribosome (80S) and ribosome subunits (40S and 60S) are indicated. B) After spiking in R-Luc 3xB pA *in vitro* transcript, RNA was extracted from each fraction and its integrity analysed on a RNA 6000 Nano Chip. Equal amounts of RNA were loaded onto the RNA 6000 Nano Chip. Final fraction numbers are indicated at the bottom. Bands corresponding to tRNA as well as 18S and 28S rRNA are indicated.

5.2.5.C Translation efficiency of candidate mRNAs

The mRNAs that were focused on for the investigation of the relationship between m^{5}C and translation state were those that were validated in Chapter 4. As *CINP*, *NDUFB7*, *RTN3* and *SZRD1* showed minimal variability in the proportion of methylation in technical replicates as described in the previous section, these candidates were investigated. Additional candidates with validated m^{5}C sites were also selected out of interest, including *OSBPL8*, *RPS3* and *SCO1*. The gradient distribution of these mRNAs was examined using RT-qPCR to determine the translation efficiency of the transcripts (Figure 5.16; line graph). The RT-qPCR data was normalised to R-Luc 3xB pA and represented as a percentage of the total amount of each mRNA collected from the fractionation. Most mRNAs were predominantly associated with polysomes and thus were translated well. The gradient distribution of *RPS3* showed two peaks and was more variable between replicates, which most likely is a result of differences in the cell growth rate between cell populations. As expected, minimal amounts of mRNA were present in fraction 1, and as a result this fraction was removed from this investigation.

5.2.5.D PREPARATION OF FRACTIONS FOR MISEQ AMPLICON SEQUENCING

To assess the methylation levels of each mRNA at different translation states, the fractions were bisulfite converted, reverse transcribed and amplicons of the mRNAs exhibiting validated m^5C sites were PCR amplified in triplicate. As a negative control for methylation and to ensure high bisulfite conversion efficiency, two segments of the non-humanised R-Luc *in vitro* transcript were amplified. Determining the best positive control for methylation was not trivial; tRNAs have been previously used but as can be seen from Figure 5.13, the distribution of tRNAs was primarily limited to fractions 2 and 3. PCR amplification of $tRNA^{Asp(GUC)}$ and $tRNA^{Leu(CAA)}$ was attempted but as expected, amplification was only successful in fractions 2 and 3. As a result, PCR amplification of 28S rRNA was performed to examine methylation at the known m^5C sites at C3782 and C4447. Although limited 28S rRNA was visible in the first two fractions based on the RNA 6000 Nano Chip, it was expected that rRNA would still be highly abundant in these fractions. PCR amplification of 28S rRNA was successful in all fractions and was therefore used as a positive control for bisulfite conversion. Using 28S rRNA as a positive control for bisulfite conversion was beneficial in that both m^5C sites exhibited high levels of methylation based on bsRNA-seq (see section 4.2.3.A). However, it was expected that the level of methylation in 28S rRNA in the earlier fractions may be lower as a result of the requirement for m^5C at these sites for translation [92].

After pooling the amplicons for each fraction and performing the library preparation as previously described, the libraries were subjected to MiSeq amplicon sequencing. High and even coverage was evident throughout the two segments of the non-humanised R-Luc *in vitro* transcripts and high bisulfite conversion efficiency was evident, with >99.5% conversion evident in both segments (Figure 5.14 and Table 5.7).

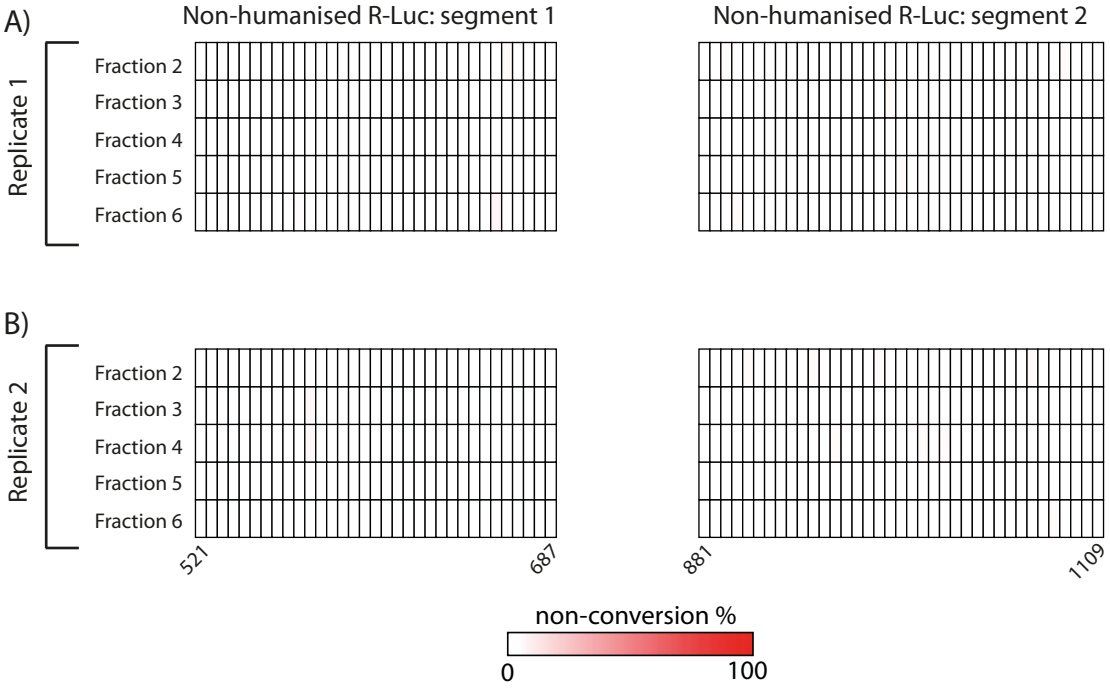


Figure 5.14. Confirmation of bisulfite conversion in polysome fractions for each replicate by MiSeq amplicon sequencing.

Two different segments of the non-humanised R-Luc *in vitro* transcript were amplified and subjected to MiSeq amplicon sequencing. Numbers below the heatmap indicate the segment of the transcript interrogated relative to the start of the transcript. The MiSeq amplicon sequencing coverage ranged from 80,000-274,000 mappable sequencing reads.

Table 5.7. Mean cytosine conversion of the non-humanised R-Luc *in vitro* transcript.

Sample		Mean conversion percentage (%)	
		Non-humanised R-Luc *in vitro* transcript segment 1	Non-humanised R-Luc *in vitro* transcript segment 2
Replicate 1	Fraction 2	99.82	99.77
	Fraction 3	99.92	99.88
	Fraction 4	99.92	99.93
	Fraction 5	99.89	99.86
	Fraction 6	99.76	99.88
Replicate 2	Fraction 2	99.67	99.59
	Fraction 3	99.82	99.69
	Fraction 4	99.82	99.65
	Fraction 5	99.92	99.59
	Fraction 6	99.93	99.83

Examination of the proportion of methylation at C3782 in 28S rRNA revealed 80%-100% methylation throughout all fractions, indicating that conversion of m^5C to uracil did not occur at this site (Figure 5.15A). Clusters of non-conversion were evident in the segment containing m^5C at C4447 in several fractions of both replicates, which is most likely a result of incomplete denaturation of the secondary structure of 28S rRNA (Figure 5.15B). The m^5C site at this location was distinguishable, revealing a high proportion of methylation in the majority of fractions. Interestingly, fraction 2, corresponding to free ribosomes and ribosomal subunits, showed a clear trend towards lower rRNA methylation in comparison to the remaining fractions containing actively translating ribosomes, consistent with a requirement for m^5C at these sites for translation [92]. Despite clear detection of the m^5C site at all fractions, it is difficult to interpret whether the methylation proportion is truly an indication of m^5C, or if it is a result of the non-conversion.

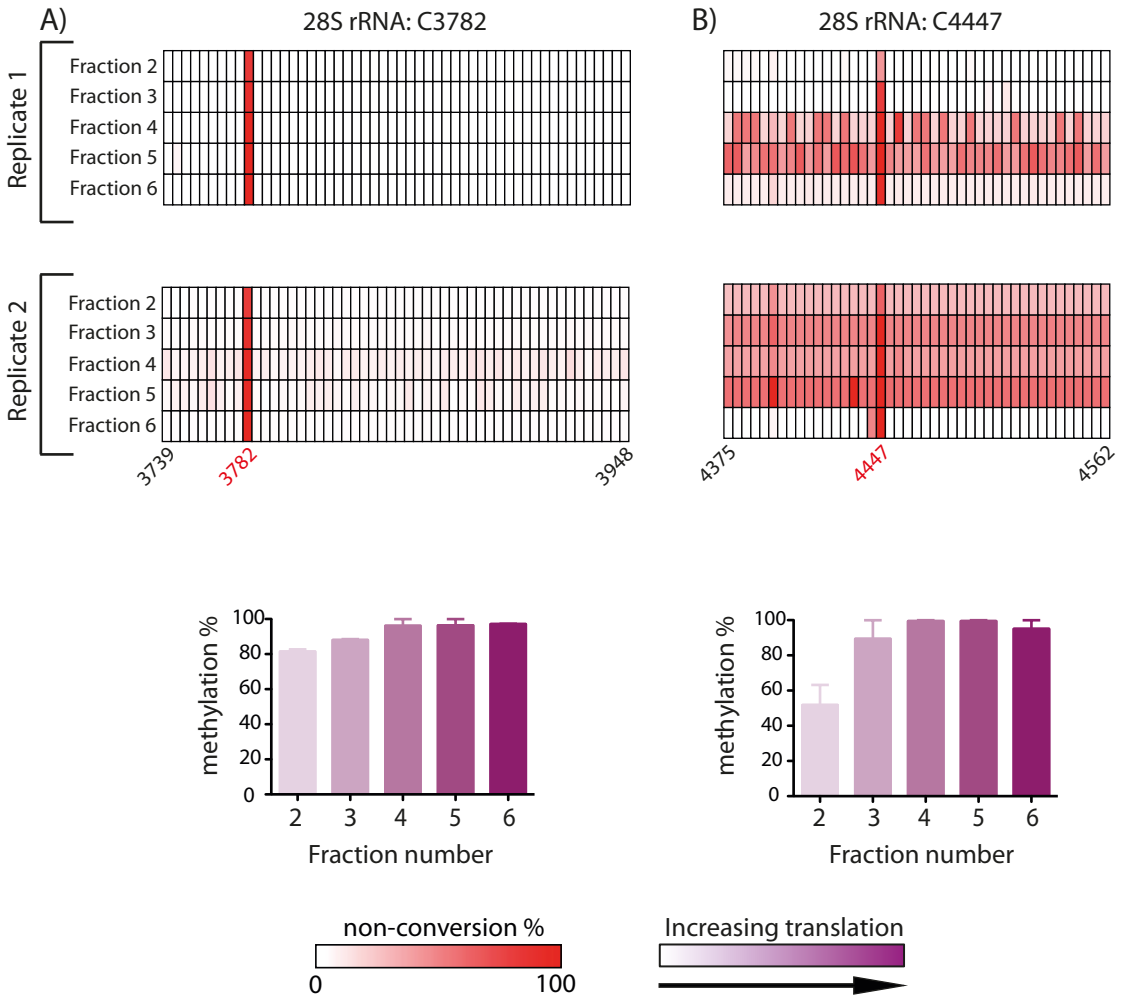


Figure 5.15. Confirmation of known m^{5}C sites in 28S rRNA by MiSeq amplicon sequencing.

A) m^{5}C site at C3782 in 28S rRNA and B) m^{5}C site at C4447 in 28S rRNA. The proportion of methylation in each fraction is indicated in the bottom bar graphs as the mean with SEM (n=2). Numbers below the heatmap indicate the segment of the transcript interrogated relative to the start of the transcript. The MiSeq amplicon sequencing coverage ranged from 92,000-320,000 mappable sequencing reads.

5.2.5.E Investigating the relationship between m^{5}C in mRNA and translation

Next, the proportion of methylation at the m^{5}C sites in the mRNAs was determined at different translation states (Figure 5.16; bar graphs). As mentioned previously, it is possible that the role of m^{5}C in regulating translation may be context-dependent and therefore the location of the m^{5}C site in these transcripts was also taken into account. *CINP*, *NDUFB7*, *SCO1* and *SZRD1* had relatively similar proportions of methylation throughout different translation states, with the m^{5}C sites being located in either the CDS or 3' UTR. *OSBPL8* exhibited a m^{5}C site in the 5' UTR and the proportion of methylation was slightly variable, which may be a result of PCR amplification biases.

Interestingly, although the m^{5}C site located in the CDS of *RPS3* was not highly methylated, a decreasing trend in methylation with increasing translation was evident; the methylation decreased from ~12% in fraction 2 to ~6% in fraction 6. This m^{5}C site is located 10 nt downstream of an Ago2 footprint based on the HeLa Ago2 HITS-CLIP dataset [308], consistent with the global finding that Ago2 footprints are enriched upstream of m^{5}C sites in the CDS (see section 5.2.2). The proportion of methylation at the m^{5}C site located in the 3' UTR of *RTN3* showed an increasing trend with increasing translation state, with the proportion increasing from ~50% in fraction 2 to ~80% in fraction 6. This m^{5}C site is ~500 nt downstream of putative binding sites of miR-206 and miR-218 based on TargetScan [310]. Furthermore, this site was located 219 nt upstream of an Ago2 footprint, and 673 nt downstream of another Ago2 footprint.

In summary, the majority of mRNAs examined in this investigation did not exhibit a trend in methylation with translation state. However, the m^{5}C sites in *RPS3* and *RTN3* exhibited clear trends in methylation with mRNA translation state, suggesting that m^{5}C might affect translation for some methylated mRNAs. Methylation correlated with translation state in opposite ways for these examples, suggesting that such a role of m^{5}C would have to be context-dependent.

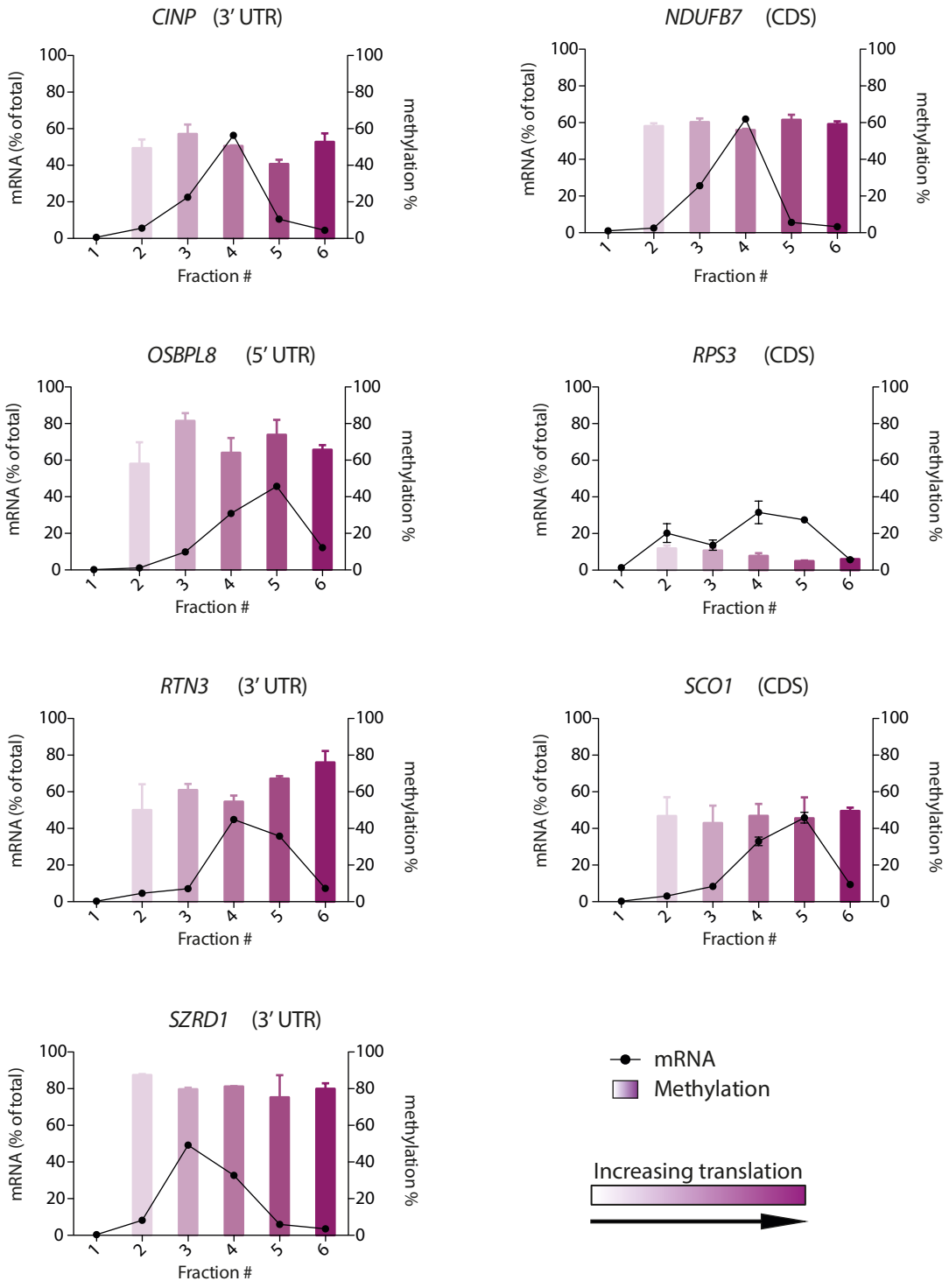

CINP (3′ UTR)
NDUFB7 (CDS)
OSBPL8 (5′ UTR)
RPS3 (CDS)
RTN3 (3′ UTR)
SCO1 (CDS)
SZRD1 (3′ UTR)
mRNA (% of total)
methylation %
Fraction #
mRNA
Methylation
Increasing translation

Figure 5.16. Gradient distribution and methylation profiles of mRNAs validated to contain m^{5}C sites mediated by NSUN2 at different translation states.

For the gradient distribution (line graph) RT-qPCR was performed on reverse transcribed RNA extracted from the polysome gradients and the data normalised to the R-Luc 3xB pA *in vitro* transcript that was spiked into each fraction prior to RNA extraction. The data is expressed as the percentage of total amount of each mRNA extracted from the polysome gradient. For the methylation profiles (bar graphs), MiSeq amplicon sequencing was performed on fractions 2 to 6 of HeLa polysome gradients. The location of the m^{5}C site in the mRNA is indicated above the graphs. The MiSeq amplicon sequencing coverage ranged from 15,000-360,000 mappable sequencing reads. Data is shown as mean with SEM (n=2).

5.3 Discussion

In this chapter, I have described the first study to investigate the molecular functions of m^{5}C in mRNA. I have analysed the bsRNA-seq dataset in order to gain insight into the role of m^{5}C in post-transcriptional gene regulation. To follow this up further, RT-qPCR and sucrose density gradient centrifugation were performed using candidates that were validated to contain m^{5}C sites mediated by NSUN2.

The GSEA and GO analysis of candidate m^{5}C sites in HeLa cells did not reveal any enrichment of complexes or pathways. The GO terms that were enriched were diverse and included mRNA metabolism, in particular all three phases of translation. Thus, while m^{5}C is present in transcripts that encode proteins of diverse cellular functions, it is also present in transcripts that encode proteins involved in mRNA metabolism, suggesting that m^{5}C regulates diverse processes involved in post-transcriptional gene regulation.

5.3.1 Ago2 and miRNA analyses

An investigation was initially made into the relationship of m^{5}C with Ago2 and miRNAs. Previously, we integrated our HeLa bsRNA-seq data with a publicly available HEK293 Ago PAR-CLIP dataset, where an enrichment of Ago footprints was evident upstream of candidate m^{5}C sites in the CDS and 3’ UTR of mRNAs [58]. Examination of another RNA binding protein, PUM2, revealed no enrichment within the 1000 nt

flanking regions of candidate m^{5}C sites. To follow this observation up further, a similar analysis was performed in this study, where HeLa Ago2 HITS-CLIP data was integrated with the candidate m^{5}C sites detected in the 3' UTR and CDS of mRNAs in HeLa cells based on the newly generated HeLa methyltransferase knockdown bsRNA-seq data. Concurring with our previous study, enrichment of Ago2 footprints was evident upstream within 25 nt of candidate m^{5}C sites in the 3' UTR and CDS. Examination of another RBP, HuR, showed no enrichment around candidate m^{5}C sites.

To investigate this relationship further, the density of 7-mer miRNA targets was determined in regions surrounding candidate m^{5}C sites in the 3' UTR and CDS. Interestingly, depletion in the density of 7-mer miRNA targets was evident at candidate m^{5}C sites in the 3' UTR and CDS. This depletion was confirmed to not be a result of uneven base distribution in these regions, such as lower frequency of cytosines.

These findings suggest that m^{5}C sites are devoid in regions of mRNA that are bound by RISC but are enriched downstream of this complex. m^{5}C may either mediate or block the action of RISC, highlighting a potential role for m^{5}C in post-transcriptional gene regulation. The mechanisms in which this could possibly occur through include regulating mRNA stability or translation, as miRNA/RISC are known to affect both processes.

5.3.2 mRNA STEADY-STATE LEVEL ANALYSES

To investigate the molecular functions of m^{5}C in mRNA further, an investigation was also made into the role of m^{5}C in regulating mRNA steady-state levels. An initial analysis involved determining whether depletion of m^{5}C in NSUN2-targeted mRNAs in the NSUN2 knockdown sample relative to the TRDMT1 knockdown and NTC samples affected the overall expression levels of these mRNAs compared to unmethylated mRNAs using the bsRNA-seq data. Although bisulfite conversion results in fragmentation of the RNA, fragmentation is also required for standard RNA-seq indicating that the expression levels determined from the bsRNA-seq data were likely true representations of the expression levels. This analysis revealed a 11.9% decrease in average steady-state expression of NSUN2-targeted mRNAs in the NSUN2 knockdown sample compared to unmethylated mRNAs. This suggests that m^{5}C may function to

stabilise mRNA, potentially through preventing the binding of RISC and other regulatory factors.

To investigate the role of m^{5}C in regulating mRNA steady-state levels further, a candidate approach using RT-qPCR was implemented, focusing on the mRNAs that were validated to contain m^{5}C sites mediated by NSUN2. Although a subtle decrease in the steady-state level over many mRNA targets of NSUN2 was observed in the bsRNA-seq data, it was possible that there may be a few candidates that exhibited a stronger change in levels. Six mRNAs that were validated to contain m^{5}C sites mediated by NSUN2 were investigated, with no changes in the steady-state levels of these transcripts evident in the NSUN2 knockdown sample. The overall steady-state level of all six mRNAs was also investigated in the NSUN2 knockdown sample, with no significant changes observed. A likely explanation for this is the limited sensitivity of the RT-qPCR assay in addition to the small number of candidates interrogated; it is possible that the subtle decrease in mRNA steady-state levels observed in the bsRNA-seq data may only be evident over many NSUN2-targeted mRNAs.

It must also be noted that RNA-seq and RT-qPCR cannot differentiate between expression changes due to changes in the rate of transcription or RNA decay. However, the possibility of the global rate of transcription being affected upon NSUN2 knockdown is unlikely; the global analysis performed using the bsRNA-seq data showed a minimal effect on the steady-state level of transcripts with no detectable m^{5}C sites upon NSUN2 knockdown. However, there is the possibility of the rate of transcription of individual genes being affected, such as the housekeeping genes used for this assay, which may have negated the subtle decrease in mRNA steady-state levels, resulting in the inability to detect these changes.

Interestingly, in some instances, the expression levels of the NSUN2-targeted mRNAs in the TRDMT1 knockdown sample were significantly higher compared to both the NSUN2 knockdown and NTC samples, although the reasoning behind this remains elusive. It is possible that the knockdown of TRDMT1 may have resulted in indirect changes in steady-state levels. As mentioned in Chapter 1, TRDMT1 was initially thought to be a DNA methyltransferase due to its strong homology with DNA

methyltransferases, although its role in methylating DNA remains controversial. Several published studies reported conflicting results for TRDMT1 methylating DNA, with some studies reporting low levels of DNA methylation in non-specific contexts in various species, including *D. melanogaster* [164, 311-313]. Furthermore, a study that investigated DNA methylation by human TRDMT1 using *in vitro* assays reported low levels of DNA methylation in a CpG context [314]. Although this needs to be explored further, it is possible that human TRDMT1 may have residual DNA methyltransferase activity that could result in gene repression. Thus, the knocking down of TRDMT1 may have resulted in an indirect increase in the expression levels of these transcripts.

5.3.3 mRNA TRANSLATION ANALYSIS

To investigate the role of m^{5}C in regulating translation, a candidate approach was used focusing on mRNAs that were validated to contain m^{5}C sites in Chapter 4. It was hypothesised that if m^{5}C in mRNA is involved in regulating translation, then the proportion of methylation in the mRNA would differ within each translation state.

The reasoning behind the candidate approach was to ensure that the m^{5}C sites in the mRNAs investigated were genuine as determined by MiSeq amplicon sequencing in Chapter 4. A drawback with this approach, however, was the possibility of PCR amplification biases of low abundant transcripts, although this has not been previously investigated. Examination of the proportion of methylation of previously validated transcripts in technical replicates revealed that some of the mRNAs exhibited variability in methylation. Therefore, those mRNAs that exhibited variability in the proportion of methylation between technical replicates, such as *MCFD2* and *NAPRT1*, were removed from further analysis.

Seven mRNAs exhibiting validated m^{5}C sites were investigated further to determine the relationship between m^{5}C and translation state using sucrose density gradient centrifugation. Examination of the gradient distribution of these transcripts revealed that the majority were associated with polysomes and were therefore actively translated. While the majority of *RPS3* transcripts were associated with polysomes, a substantial proportion of the transcript was associated with a single ribosome. *RPS3* is a component of the 40S subunit, which forms part of the domain associated with translation initiation

[315]. Translation of ribosomal proteins is largely dependent on cell growth, where they are actively translated in fast growing cells in order to contribute to the active translation of other mRNAs. The presence of multiple peaks in the gradient distribution of *RPS3* was most likely an indication of differences in the cell growth within the population of cells, resulting in different translation efficiencies of *RPS3*. This is also the most likely explanation for the slight differences in the gradient distribution of *RPS3* between the biological replicates.

MiSeq amplicon sequencing of the seven mRNAs demonstrated variability in the proportion of methylation between the mRNAs, with no consistent trend observed. Several transcripts, including *CINP*, *NDUFB7*, *SCO1* and *SZRD1* exhibited similar proportions of methylation at all translation states, while *OSBPL8* exhibited slightly more variability with no discernible trend. Two mRNAs showed interesting trends in the proportion of methylation with translation state. *RPS3* exhibited decreasing methylation with increasing translation. *RTN3* exhibited increasing methylation with increasing translation. While each trend was only present in a single mRNA, it is suggestive that m^{5}C can affect translation and that the role of m^{5}C in mRNA may be context-dependent. The m^{5}C site in the CDS of *RPS3* is located 10 nt downstream of a HeLa Ago2 footprint, consistent with the global finding of Ago2 enrichment upstream of candidate m^{5}C sites in mRNA. This suggests that m^{5}C in this transcript may mediate the binding of RISC to inhibit translation. In contrast, the m^{5}C site in the 3' UTR of *RTN3* was not near any putative or known miRNA binding sites, or HeLa Ago2 footprints, suggesting that this m^{5}C site affects translation through a different mechanism than that of the miRNA pathway, such as recruitment or blocking of other RBPs.

5.3.4 Conclusions

In this chapter, I performed the first study to investigate the molecular functions of m^{5}C in mRNA. The HeLa methyltransferase bsRNA-seq data showed an enrichment of Ago2 footprints upstream of candidate m^{5}C sites in the CDS and 3' UTR of mRNAs and depletion of miRNA target sites at candidate m^{5}C sites in these regions. This suggests a regulatory interaction between m^{5}C and RISC. Global analysis of mRNA steady-state levels using the bsRNA-seq data indicated a subtle decrease in the mean

steady-state levels of NSUN2-targeted mRNAs in the NSUN2 knockdown sample, however the subtle changes observed could not be confirmed using a candidate approach by RT-qPCR. I investigated the role of m^5C in mRNA translation, which revealed two mRNAs that exhibited trends in the proportion of methylation with translation state. *RPS3*, which contains a m^5C site in the CDS, demonstrated a decreasing trend in methylation with increasing translation. *RTN3*, which contains a m^5C site in the 3' UTR, showed an increasing trend in methylation with translation state. These findings suggest that the molecular functions of m^5C in mRNA may be context-dependent.

5.4 CHAPTER HIGHLIGHTS

- mRNAs exhibiting candidate m^{5}C sites were enriched in a broad range of GO terms, which prominently included those related to the translation processes.
- Correlating matched HeLa cell datasets revealed that Ago2 footprints were enriched upstream of candidate m^{5}C sites in both the 3' UTR and CDS of mRNAs. Conversely, m^{5}C sites were depleted at the portion of conserved miRNA target sites that match the miRNA seed, again in both the 3' UTR and CDS.
- This confirms and extends our previously published findings and indicates regulatory interactions between m^{5}C and RISC.
- A statistically significant 11.9% decrease in mean steady-state expression level of NSUN2-targeted mRNAs was seen in NSUN2 knockdown cells, compared to unmethylated mRNAs. RT-qPCR of six chosen NSUN2 target mRNAs failed to validate this effect.
- This suggests a subtle role for m^{5}C formation by NSUN2 in stabilising a subset of its target mRNAs that is revealed only when averaging over many target mRNAs.
- The relationship between m^{5}C status and polysome association was assessed for seven NSUN2-targeted mRNAs (one site in the 5' UTR, three sites each in the CDS and 3' UTR).
- Two examples displayed a correlation between m^{5}C level and translation state: with increasing polysome density *RPS3* showed a decreasing proportion of m^{5}C at a site in the CDS, while *RTN3* exhibited increasing proportion of m^{5}C at a site in the 3' UTR.
- The m^{5}C site in *RPS3* was 10 nt downstream of an Ago2 footprint, while Ago2 footprints were not evident near the m^{5}C site in *RTN3*.
- This suggests that m^{5}C formation by NSUN2 affects translation of target mRNAs in different ways and in a context-dependent manner.
- Overall, these findings provide several correlative links between the presence of m^{5}C in mRNAs and several aspects of post-transcriptional gene regulation. mRNA stability and translation emerge as the processes most likely to be affected. The underlying mechanisms remain to be explored.

CHAPTER SIX

COMPARISON OF THE RNA m^5C PROFILES OF PREC, LNCAP & HELA CELLS

6.1 INTRODUCTION

NSUN2 is involved in a number of biological processes, including Myc-induced proliferation of skin cells and mitotic spindle stability [99, 211, 213]. Recently, GWAS have identified a number of mutations in *NSUN2*, including missense, nonsense and splicing mutations that cause intellectual disabilities [147, 209, 210]. These studies indicate a role for NSUN2 in normal cognitive development.

Several studies have reported that m^5C in tRNA promotes mRNA translation through stabilisation of tRNAs [149, 150]. A recent study investigated the role of tRNA methylation in neuro-developmental disorders and linked this to cellular stress [150]. Loss of NSUN2 in patient fibroblasts and mice resulted in decreased tRNA methylation, leading to angiogenin-dependent endonucleolytic cleavage of tRNAs and an accumulation of 5' tRNA-derived fragments. The biological effect of this was decreased mRNA translation and induction of the cellular stress response that ultimately resulted in decreased cell size and increased apoptosis of neurons. This is the first study to suggest a link between tRNA m^5C methylation and neuro-degenerative disorders.

NSUN2 shows low expression in most tissues, with higher expression in the testis [99, 207, 316]. However, it is activated by the oncogene, Myc, and is overexpressed in a variety of tumours including breast, colorectal and oral cancers [99, 207, 317]. Investigation of several breast cancer cell lines revealed >1.5-fold increase in *NSUN2* mRNA compared to normal breast tissue. Immunofluorescence staining also indicated an increase in NSUN2 protein, although this was not quantified and the increase was variable between cell lines [317]. Examination of NSUN2 expression in prostate tumour samples by immunohistochemistry also revealed an increase in NSUN2 compared to matched normal tissues [207].

The role of RNA m^5C in cancer has not yet been explored. The potential function of m^5C in tRNA promoting mRNA translation and the fact that deregulation of translation can contribute to the development of various cancers, warrants investigation into the role of m^5C in RNA in cancer [318, 319]. In this thesis, I have found that m^5C is present across multiple transcript biotypes and m^5C in mRNA mediated by NSUN2 has a subtle and context-dependent effect on post-transcriptional gene regulation.

Thus, the first aim of this chapter was to determine the differences in the RNA m^{5}C profiles of normal prostate cells and prostate cancer cells using bsRNA-seq. The cell lines used for this study were primary prostate epithelial cells (PrEC) and a prostate adenocarcinoma cell line derived from the left supraclavicular lymph node metastasis from a 50 year old Caucasian male (LNCaP). These cell lines were an attractive choice for an exploration of whether m^{5}C in RNA functions in cancer due to the availability of epigenomic and transcriptomic data of these cell lines from Professor Susan Clark. Findings from bsRNA-seq, particularly any differential RNA methylation sites between normal and cancer cells could be related back to these other available data sets. However, due to the lack of validatable differences in the RNA m^{5}C profiles, the investigation shifted focus towards the candidate m^{5}C sites in common between PrEC and LNCaP cells. An integrative comparison with the HeLa methyltransferase knockdown dataset was also performed to confirm patterns of methylation and validate individual candidate m^{5}C sites across multiple cell lines.

6.2 Results

The following sections document the comparative analysis of candidate m^{5}C sites in RNA in PrEC and LNCaP cells, as well as an integrative comparative analysis with HeLa cells. Dr. Brian J. Parker mapped the bsRNA-seq data and MiSeq amplicon sequencing data and I performed all experimental work and comparative analyses between the cell lines.

6.2.1 Expression level of NSUN2 in PrEC and LNCaP cells

As NSUN2 was reported as overexpressed in prostate tumours when compared to normal prostate tissues, the mRNA and protein levels of NSUN2 were determined in PrEC and LNCaP cells. RT-qPCR revealed a 1.4-fold increase in *NSUN2* in LNCaP cells compared to PrEC cells, although this did not reach statistical significance (Figure 6.1A). Examination of NSUN2 protein levels by western blotting showed more than two-fold higher NSUN2 protein levels in LNCaP cells compared to PrEC cells, and this

was consistent between two biological replicates (Figure 6.1B). Therefore, the pair of cell lines broadly replicates the documented tendency towards overexpression in cancer and we hypothesised that an increase in NSUN2 in LNCaP cells may result in additional m^{5}C sites in RNA, or an increase in the stoichiometry of methylation at a particular m^{5}C site.

As discussed in Chapter 4, knockdown of *NSUN2* in HeLa cells resulted in more than five-fold depletion of *NSUN2* mRNA and almost complete depletion of NSUN2 protein, causing substantial depletion of m^{5}C at NSUN2-mediated sites. Therefore, the two-fold overexpression of NSUN2 in LNCaP cells was expected to cause more subtle increases in methylation, indicating that a screen by bsRNA-seq should feature at least some level of biological replication.

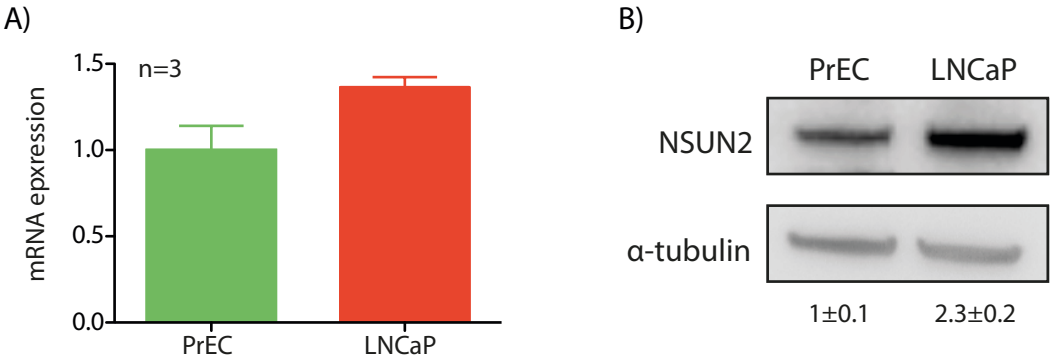


Figure 6.1. NSUN2 levels in PrEC and LNCaP cells based on RT-qPCR and western blotting.

A) *NSUN2* mRNA expression levels are relative to the geometric mean of *B2M*, *SDHA* and *TBP* mRNA expression levels. Data is shown as mean with SEM (n=3). There was no significant difference in the expression of *NSUN2* in LNCaP cells compared to PrEC cells (p=0.0755, Student's t-test). B) Representative western blot of PrEC and LNCaP whole cell lysate against NSUN2. α-tubulin was used as a loading control. Numbers below the blot represent the protein levels relative to PrEC cells based on densitometry with SEM (n=2).

6.2.2 KNOCKDOWN OF *NSUN2* IN LNCAP CELLS

In this thesis, NSUN2 was found to mediate m^5C sites in a broad range of transcript biotypes in HeLa cells. *NSUN2* was also knocked down in LNCaP cells to determine if candidate m^5C sites detected by bsRNAseq in LNCaP cells were mediated by NSUN2 using MiSeq amplicon sequencing. Several attempts were also made to knock down *NSUN2* in PrEC cells, but the cells did not survive the process.

The RNAi-mediated knockdown was performed in triplicate, and a single sample was used for the validation of candidate m^5C sites by MiSeq amplicon sequencing. Confirmation of *NSUN2* knockdown was assessed by RT-qPCR, western blotting and MiSeq amplicon sequencing of the controls. >90% knockdown of *NSUN2* mRNA was evident in LNCaP cells, with almost complete protein depletion (Figure 6.2A and B). Depletion of m^5C at the known NSUN2 target sites was evident in $tRNA^{Asp(GUC)}$ and $tRNA^{Leu(CAA)}$. ~90% methylation was detected in the LNCaP NTC sample at C48 and C49 in $tRNA^{Asp(GUC)}$, which was depleted to ~67% and ~47% in the LNCaP NSUN2 knockdown sample, respectively (Figure 6.2C). Depletion of m^5C in the NSUN2 knockdown sample was confirmed at C34 in spliced and unspliced $tRNA^{Leu(CAA)}$ (Figure 6.2D). ~35% methylation in the unspliced tRNA was evident in the LNCaP NTC sample compared to only ~5% in the LNCaP NSUN2 knockdown sample. For spliced $tRNA^{Leu(CAA)}$, ~70% methylation was evident in the LNCaP NTC sample compared to only ~13% in the NSUN2 knockdown sample. Thus, this confirms depletion of NSUN2 in LNCaP cells. Efficient bisulfite conversion was also evident in a segment of the non-humanised *in vitro* transcript for both samples (Figure 6.2E).

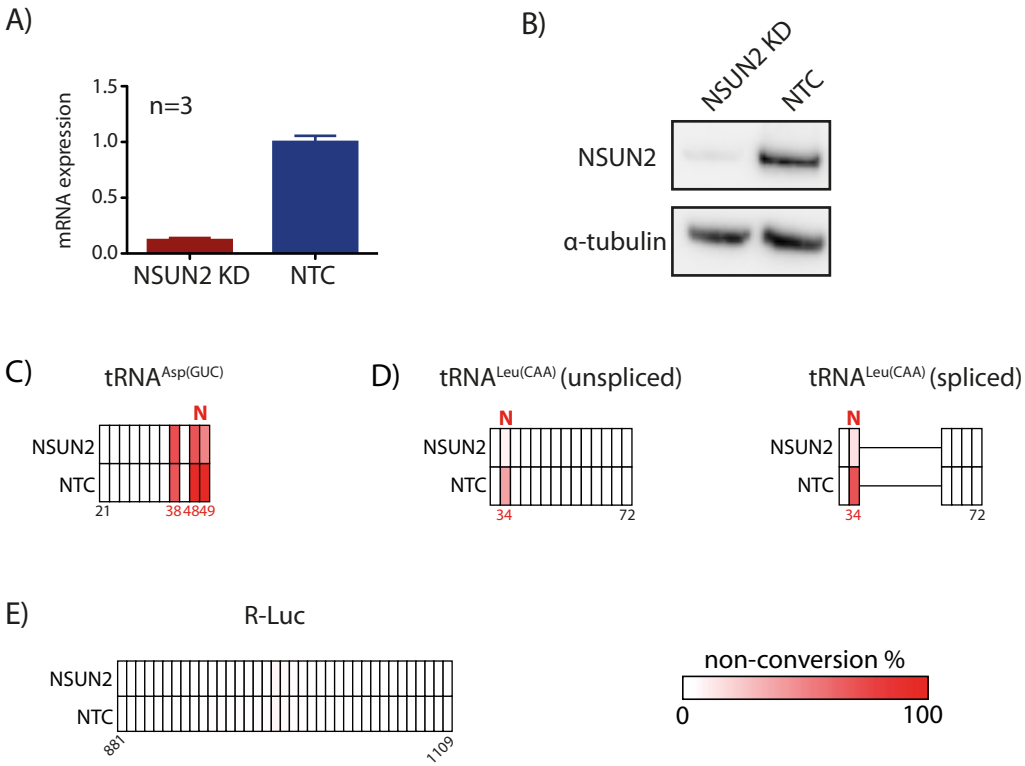


Figure 6.2. Confirmation of RNAi-mediated knockdown of *NSUN2* in LNCaP cells and verification of bisulfite conversion efficiency.

A) *NSUN2* mRNA expression levels in the NSUN2 knockdown and NTC samples relative to the geometric mean of *MRPL9*, *H2AFV* and *TCF25* mRNA expression levels. Data shown as mean fold-change relative to NTC with SEM (n=3). Depletion of *NSUN2* mRNA in the NSUN2 knockdown sample is evident. B) Western blot showing NSUN2 protein levels in the NSUN2 knockdown and NTC samples. α-tubulin was used as a loading control. Depletion of NSUN2 protein is evident in the NSUN2 knockdown sample. C) MiSeq amplicon sequencing of tRNA$^{Asp(GUC)}$ in the NSUN2 knockdown and NTC samples. Depletion of m^{5}C is evident in the NSUN2 knockdown sample at C48 and C49. D) MiSeq amplicon sequencing of tRNA$^{Leu(CAA)}$ in the NSUN2 knockdown and NTC samples. Depletion of m^{5}C is evident in the NSUN2 knockdown sample at C34. The line in the heatmap of the spliced tRNA$^{Leu(CAA)}$ indicates the region spliced. The tRNA structural positions with the known m^{5}C sites are indicated in red below the heatmap. N= NSUN2 target sites. E) MiSeq amplicon sequencing of a segment of the non-humanised R-Luc *in vitro* transcript confirming efficient bisulfite conversion in both samples. The MiSeq coverage ranged from 1,000-75,000 mappable sequencing reads.

6.2.3 Differential m^{5}C sites

6.2.3.A Global distribution of differential m^{5}C sites

As detailed in Chapter 3, bsRNA-seq was performed on two biological replicates of PrEC and LNCaP cells. Additionally, bsRNA-seq, MiSeq amplicon sequencing and Sanger bisulfite sequencing of the R-Luc *in vitro* transcripts, tRNA$^{Asp(GUC)}$ and tRNA$^{Leu(CAA)}$ indicated that all samples were efficiently bisulfite converted.

Using the proportion statistic and appropriate controls, the number and distribution of candidate differential m^{5}C sites in RNA detected in PrEC and LNCaP cells were determined. The criteria for the detection of these candidate sites were chosen as ≥5 reads in all samples and ≥10% mean methylation across the control samples. As only few candidate differential m^{5}C sites were detectable when requiring stringent statistical criteria, a FDR≤40% was used. This yielded 426 candidate m^{5}C sites that were hypermethylated in PrEC cells and 389 candidate m^{5}C sites that were hypermethylated in LNCaP cells (Figure 6.3).

The majority of the candidate differential m^{5}C sites were detected in mRNAs, although a small number of candidate sites were detected in various ncRNA classes. While the global mapping of the tRNAs revealed five candidate differential m^{5}C sites hypermethylated in LNCaP cells, the local mapping revealed no candidate differential m^{5}C sites in tRNAs. The complete lists of candidate differential m^{5}C sites in PrEC and LNCaP cells are located in Appendix 8.3.

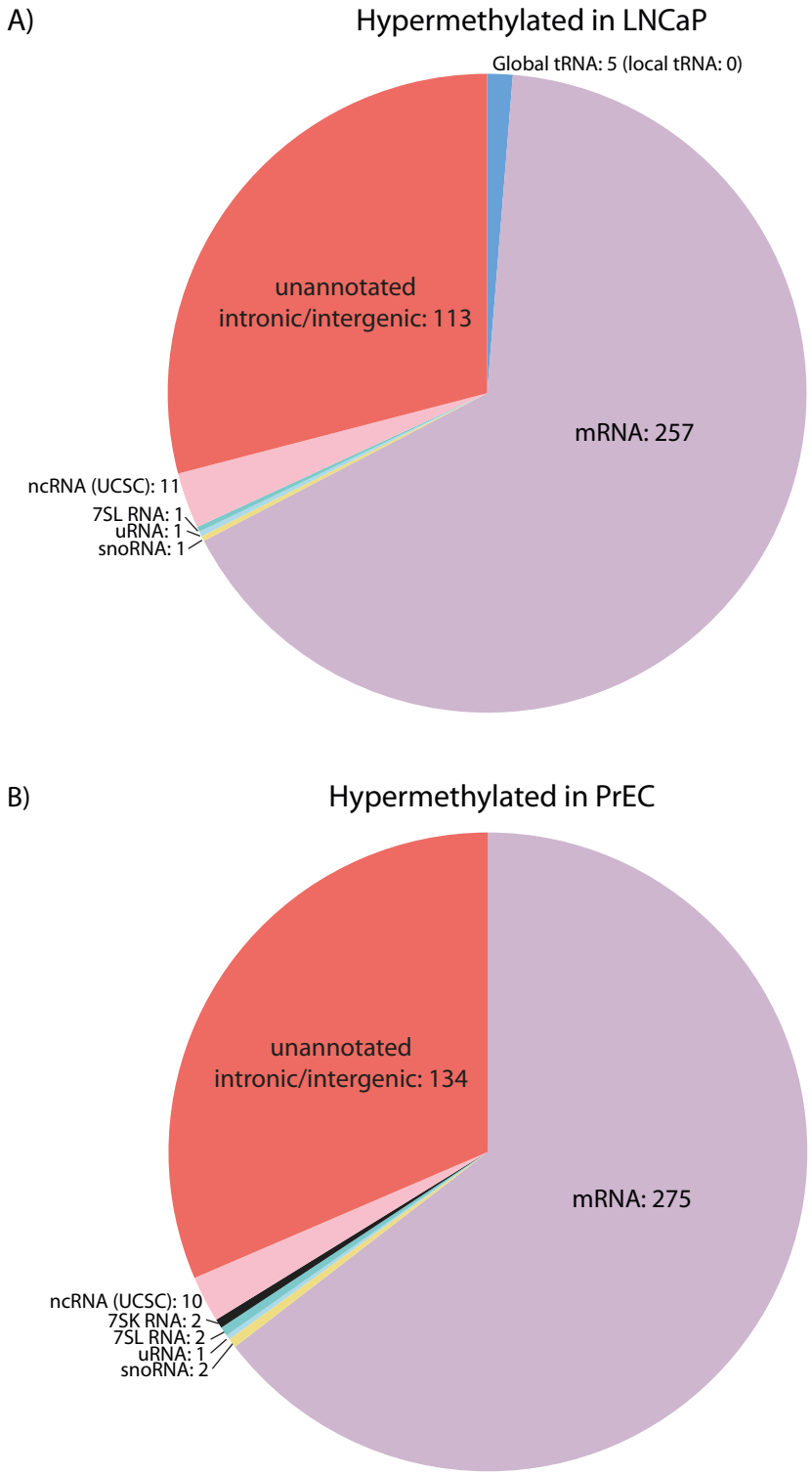


Figure 6.3. The distribution of candidate differential m^{5}C sites across different transcript biotypes in PrEC and LNCaP cells.

≥5 reads in all samples and ≥10% mean methylation across the control samples was required for the detection of candidate differential m^{5}C sites. FDR≤40% was estimated by the Benjamini-Hochberg method. The number of candidate differential m^{5}C sites detected in each transcript biotype is indicated.

6.2.3.B DIFFERENTIAL M^{5}C SITES IN MRNA

Further investigation into mRNAs was motivated by the preferential detection of candidate differential m^{5}C sites in this transcript biotype. Although the candidate differential m^{5}C sites were hypermethylated in either PrEC or LNCaP cells, it was curious that these hypermethylated candidate sites were almost exclusively detected in one cell line and not the other. In addition to candidates exhibiting high coverage and reasonable amounts of methylation, sequencing reads were inspected in IGV prior to selection for validation (see section 3.2.5.A). It was noted that the majority of candidate differential m^{5}C sites were called based on sequencing reads with poor quality values or were otherwise potentially compromised (see below).

Nevertheless, three mRNAs exhibiting candidate differential m^{5}C sites that were either hypermethylated in PrEC or LNCaP cells were selected for validation to investigate whether these were artefacts. In the case that they did validate, a focus was placed on transcripts that encoded proteins with functional importance in cell growth and in the cancer process. This included *SOX4*, *FTSJ3* and another candidate m^{5}C site in *SZRD1*. The candidate m^{5}C site in *SOX4* was hypermethylated in LNCaP cells and this mRNA encodes a transcription factor that is overexpressed in a number of cancers, including prostate cancer, contributing to cellular transformation, cell survival and metastasis [320-324]. Recently, SOX4 was reported as essential for epithelial to mesenchymal transition (EMT), cell survival *in vitro* and for primary tumour growth and metastasis *in vivo* [325]. The candidate m^{5}C site in *FTSJ3* was hypermethylated in PrEC cells and this mRNA encodes a protein that is a putative homologue of Spb1 protein in yeast, which is involved in the methylation of G2922 within the catalytic centre of the ribosome [326]. FTSJ3 has been reported to function in cell proliferation and pre-rRNA processing [327]. The candidate m^{5}C site in *SZRD1* that was hypermethylated in PrEC cells was interesting in that it was an additional candidate m^{5}C site detected upstream of that validated in both PrEC and LNCaP cells (see section 6.2.9). Multiple m^{5}C sites in a single mRNA have not yet been validated.

Examination of the sequencing reads spanning the candidate m^{5}C site in *SOX4* indicated that they were of high sequencing quality. However, MiSeq amplicon sequencing

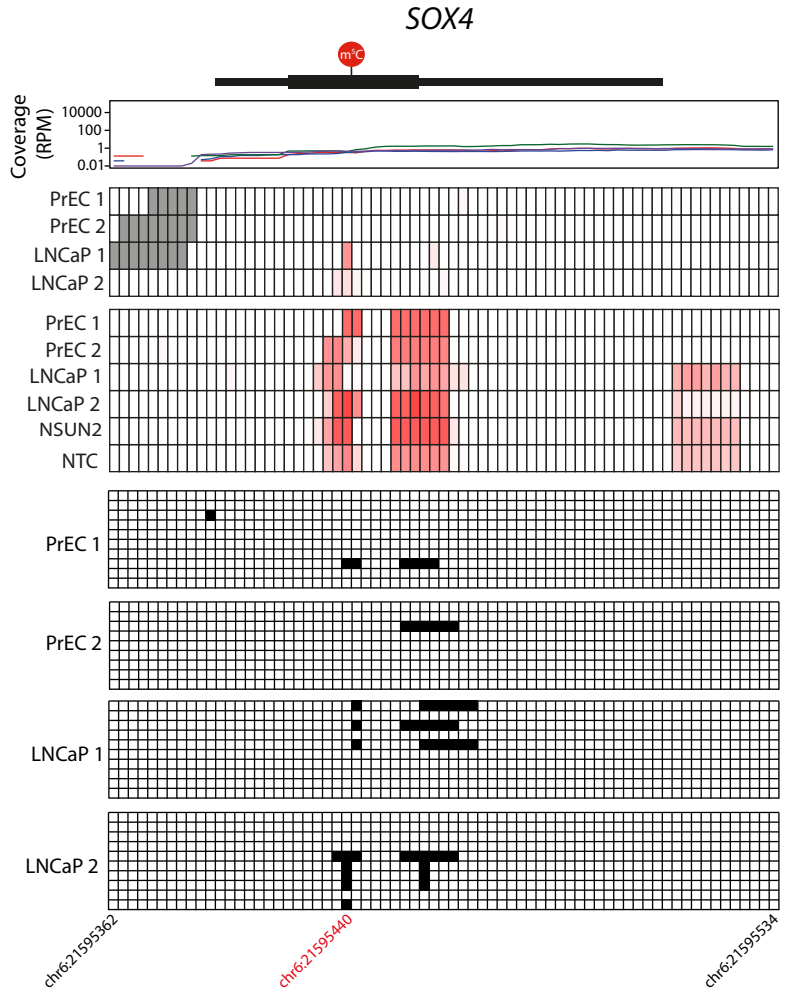
revealed clusters of non-converted cytosines primarily in a CpG context that were not apparent by bsRNA-seq, and this was also confirmed by Sanger bisulfite sequencing (Figure 6.4A). It is interesting that bsRNA-seq did not detect the clusters of non-conversion in *SOX4* despite the alignment of the sequencing reads globally to non-CpG *in silico* bisulfite converted references. However, we have previously found that clusters of non-converted cytosines were evident in longer PCR amplicons derived from highly structured rRNA compared to shorter amplicons (see section 3.2.1.C). The insert size used for SOLiD™ 5500XL sequencing was 50-120 nt in contrast to the amplicons generated for MiSeq amplicon sequencing and Sanger bisulfite sequencing which were ~200 bp. Thus, the shorter fragments used for bsRNA-seq were most likely bisulfite converted more efficiently compared to the longer fragments amplified for MiSeq amplicon sequencing and Sanger bisulfite sequencing. This possibly resulted in the inability to detect the clusters of non-conversion by bsRNA-seq while clearly being detected by MiSeq amplicon sequencing and Sanger bisulfite sequencing. Interestingly, non-conversion was not detected in PrEC cells by bsRNA-seq. It is possible that the detection of clusters of non-conversion may also depend on the sample preparation, which is suggested by the different levels of non-conversion exhibited in *SOX4* in each LNCaP replicate by bsRNA-seq. This is also highlighted by the variations in levels of non-conversion exhibited in each sample by MiSeq amplicon sequencing. Overall, the detection of this candidate m^{5}C site in *SOX4* by bsRNA-seq was most likely a false positive due to RNA secondary structure.

The candidate m^{5}C sites detected in *FTSJ3* and *SZRD1* in PrEC cells could also not be validated (Figure 6.4B and C). Examination of the bsRNA-seq sequencing reads in IGV revealed that the candidate m^{5}C site detected in *FTSJ3* in PrEC cells was downstream and adjacent to a G-to-A SNP (Figure 6.5A). The Single Nucleotide Polymorphism Database (dbSNP), which catalogues annotated SNPs generated by the National Centre for Biotechnology Information (NCBI), indicated that this SNP is annotated (rs9910549) [328]. Furthermore, PrEC and LNCaP RNA-seq data provided by Professor Susan Clark confirmed the presence of this SNP in 76% of the sequencing reads (data not shown). Interestingly, 13-34% of the bsRNA-seq sequencing reads mapping to this region contained the SNP, which exhibited a high quality score. A large proportion of sequencing reads that did not contain the SNP exhibited a low quality

score at that base. Due to the low quality calling of the non-SNP base, the bases adjacent to this were of low quality, causing the incorrect calling of a candidate m^{5}C site. In contrast, the high quality score at that base resulted in the correct calling of a converted cytosine. LNCaP cells do not exhibit a SNP at this position and all bases were of high quality, resulting in the correct calling of a converted cytosine. Similarly, the candidate m^{5}C site in *SZRD1* was of low quality due to the low quality calling of the adjacent base, resulting in the incorrect calling of a candidate m^{5}C site (Figure 6.5B).

In summary, many of the candidate differential m^{5}C sites in mRNA were called based on sequencing reads with poor quality values or were otherwise potentially compromised, which was highlighted by the inability to validate a subset of these sites.

A)



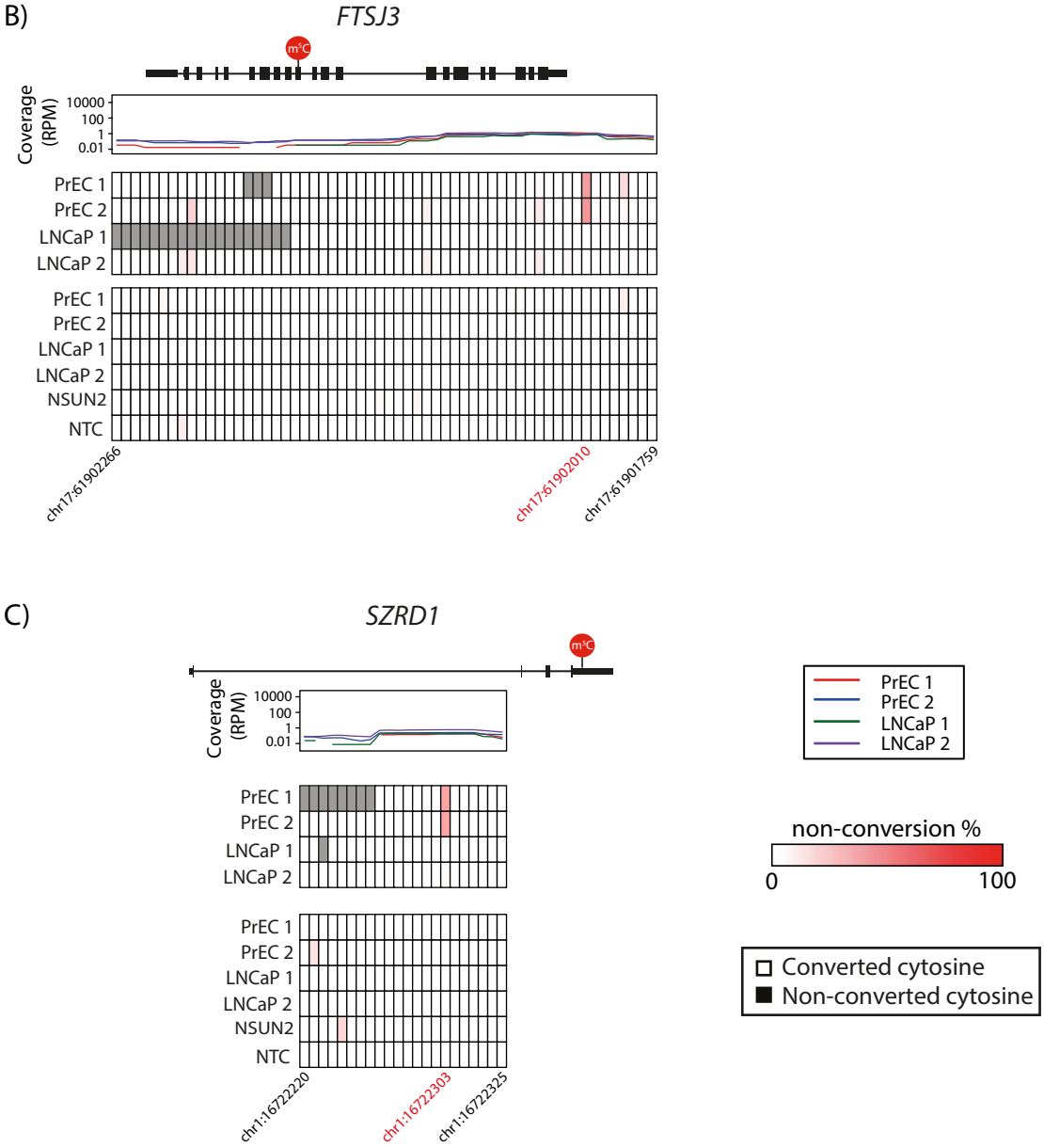
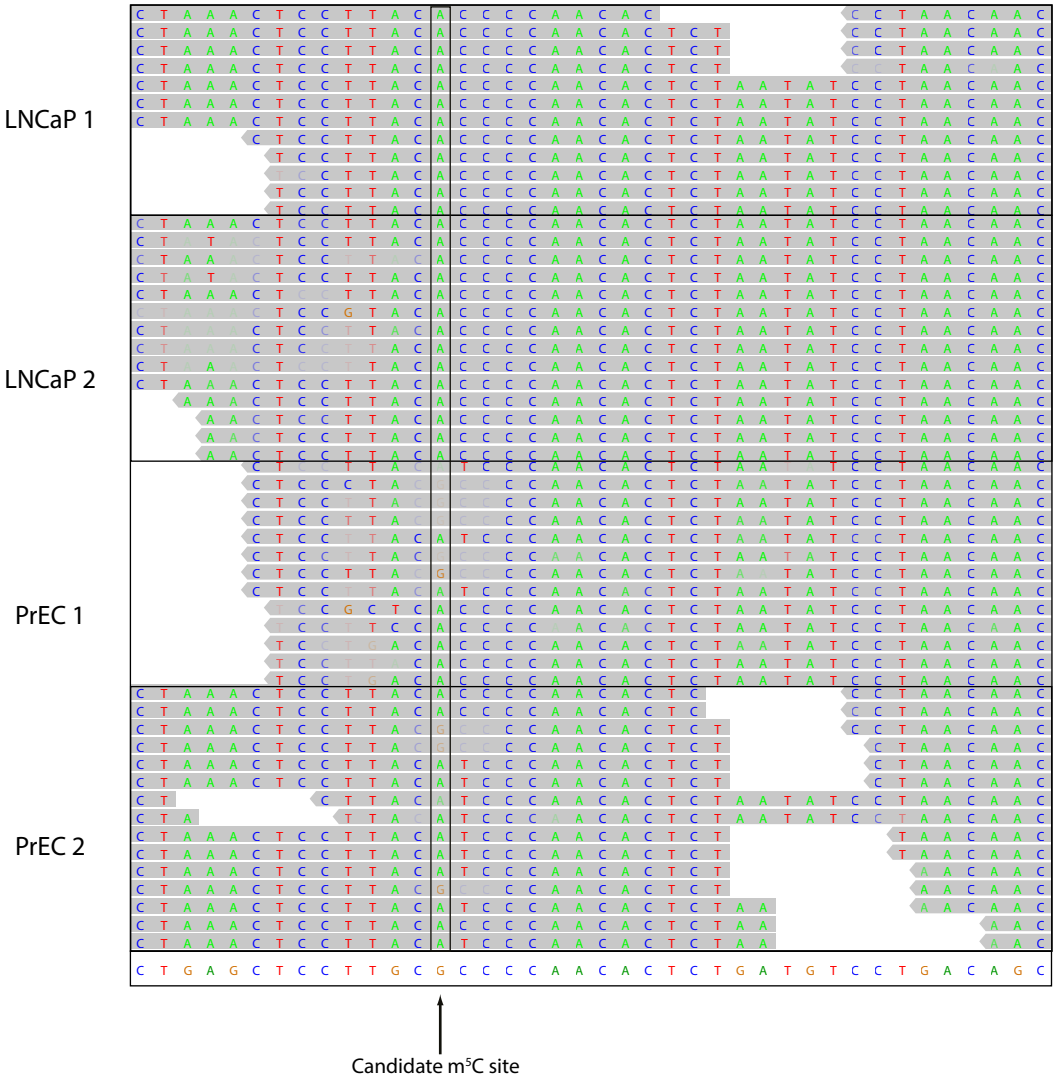


Figure 6.4. Lack of validation of candidate differential m^{5}C sites in mRNA by MiSeq amplicon sequencing in PrEC and LNCaP cells.

A) *SOX4*, B) *FTSJ3*, C) *SZRD1*. The location of the m^{5}C site relative to the regions in the mRNA is indicated at the top. The top panel displays the coverage in RPM based on bsRNA-seq, the middle panel displays the proportion of non-conversion based on bsRNA-seq, and the bottom panel displays the proportion of non-conversion based on MiSeq amplicon sequencing. The shades of grey in the middle panel indicate regions of the transcript where coverage was <5 reads based on bsRNA-seq. Chromosomal locations of the candidate m^{5}C sites are indicated in red below the heatmap. The MiSeq coverage ranged from 29,000-56,000 mappable sequencing reads.

FTSJ3


LNCaP 1
LNCaP 2
PrEC 1
PrEC 2
C T G A G C T C C T T G C G C C C C A A C A C T C T G A T G T C C T G A C A G C
Candidate m5C site

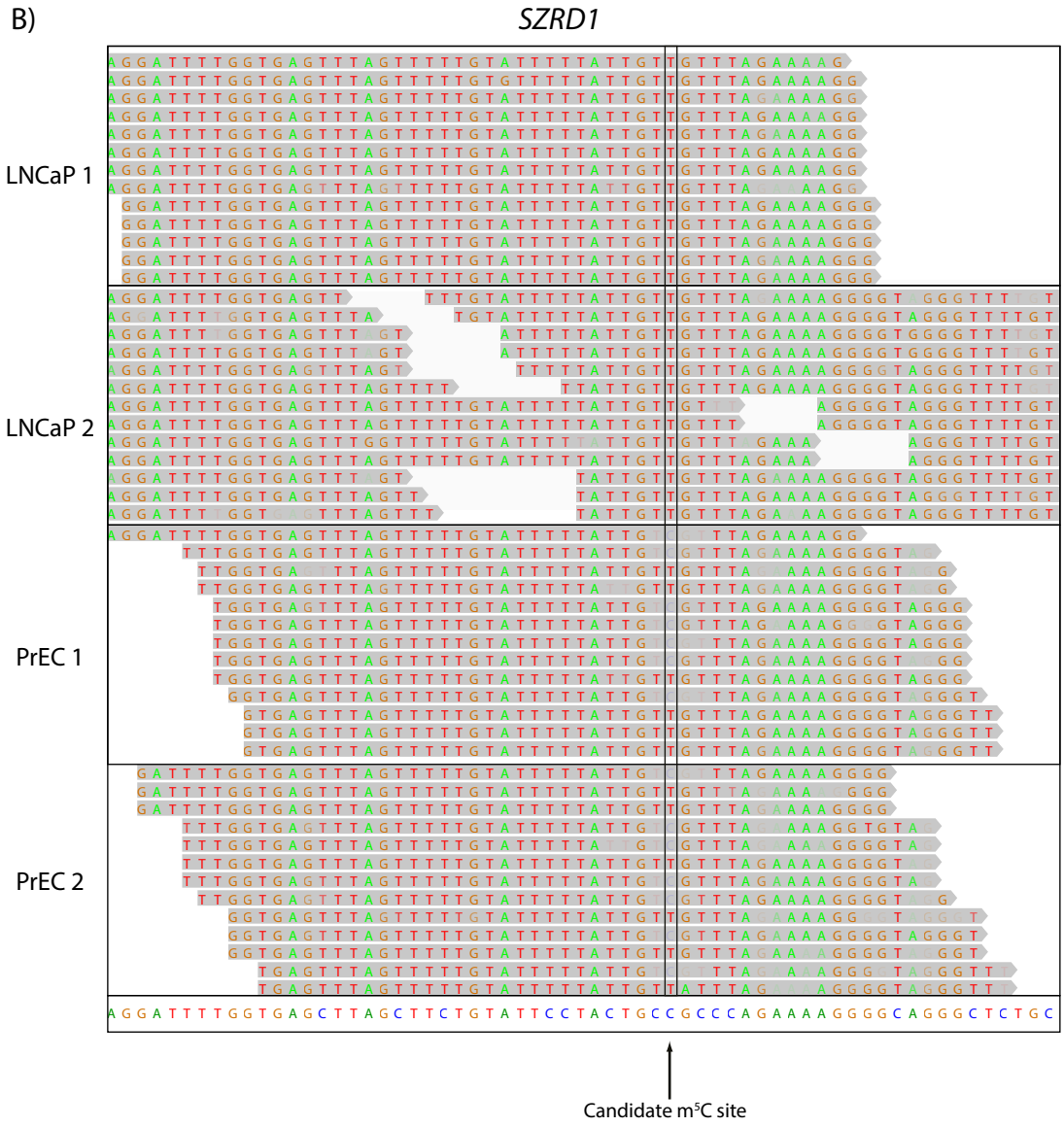


Figure 6.5. Representative snapshots from IGV depicting a subset of sequencing reads containing candidate m^{5}C sites that could not be validated in PrEC cells.

The grey bars represent 50 nt sequencing reads mapped to part of the locus. A) *FTSJ3*. Unconverted guanines represent the false positive site (chr17:61,901,997-61,902,036, - strand), B) *SZRD1*. Unconverted cytosines represent the false positive site (chr1:16,722,266-16,722,328, + strand). Low quality bases are transparent.

6.2.4 STOICHIOMETRY OF M^{5}C SITES IN COMMON BETWEEN PREC AND LNCAP CELLS

Instead of leading to sites of pronounce differential methylation, another plausible consequence of NSUN2 overexpression in cancer would be that the stoichiometry of methylation is altered more subtly across many or all NSUN2 target sites. Thus, this was determined for candidate m^{5}C sites that were in common between PrEC and LNCaP cells. Details of the number and distribution of candidate m^{5}C sites in common between PrEC and LNCaP cells will be addressed in the following section.

The mean proportion of methylation was calculated for each transcript biotype that contained candidate m^{5}C sites in each cell line. In almost all cases, the proportion of methylation was approximately the same in PrEC and LNCaP cells (Figure 6.6). Interestingly, the single candidate m^{5}C site detected in the vtRNA, vtRNA1-3, exhibited a significantly higher proportion of m^{5}C in LNCaP cells compared to PrEC cells, with LNCaP cells having ~35% methylation and PrEC cells ~10% methylation. Examination of the sequencing reads mapping to vtRNA1-3 indicated high quality detection of the candidate m^{5}C site in both cell lines, indicating that this site was not compromised by issues exhibited by candidate differential m^{5}C sites (Figure 6.7). This candidate m^{5}C site was also detected in HeLa cells as a NSUN2 target in this thesis and the Aza-IP study, as well as in HEK293 cells and human fibroblasts by miCLIP (see section 4.3.3.B) [77, 78]. Due to the lack of time, this candidate m^{5}C site was not followed up.

These findings have demonstrated that a moderate increase in NSUN2 in LNCaP cells did not result in a substantial increase in the stoichiometry of methylation across most transcript biotypes, with the exception of an individual vtRNA.

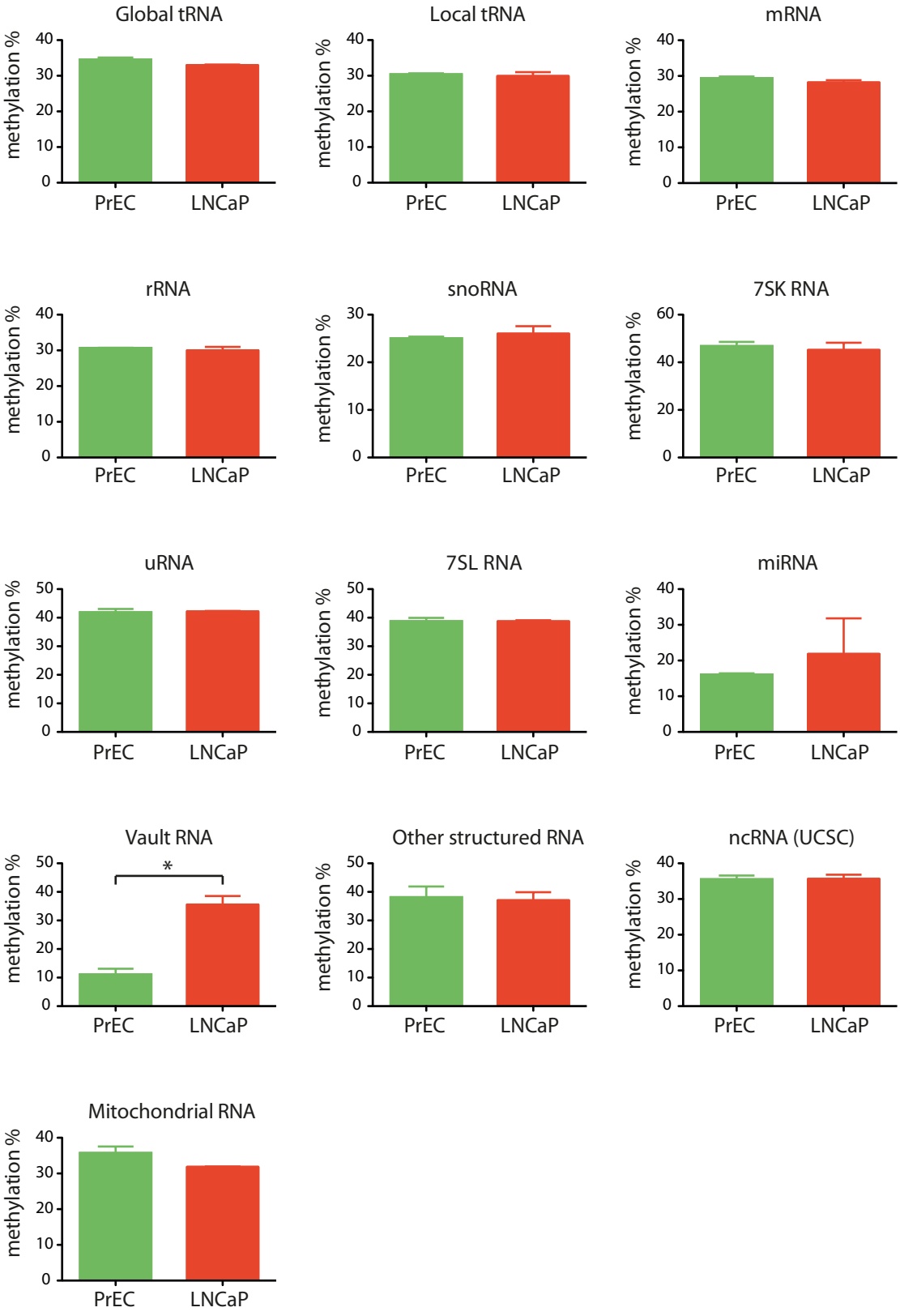

Global tRNA
Local tRNA
mRNA
rRNA
snoRNA
7SK RNA
uRNA
7SL RNA
miRNA
Vault RNA
Other structured RNA
ncRNA (UCSC)
Mitochondrial RNA
methylation %
PrEC
LNCaP
*

Figure 6.6. Mean methylation detected in each sample for all candidate m^{5}C sites in common between PrEC and LNCaP cells based on bsRNA-seq.

The bar charts display the mean methylation at candidate m^{5}C sites in common between PrEC and LNCaP cells within each transcript biotype. A single vtRNA, vtRNA1-3, exhibited a significantly higher proportion of methylation in LNCaP cells compared to PrEC cells (* $p<0.05$, Student's t-test). Data represented as mean with SEM (n=2).

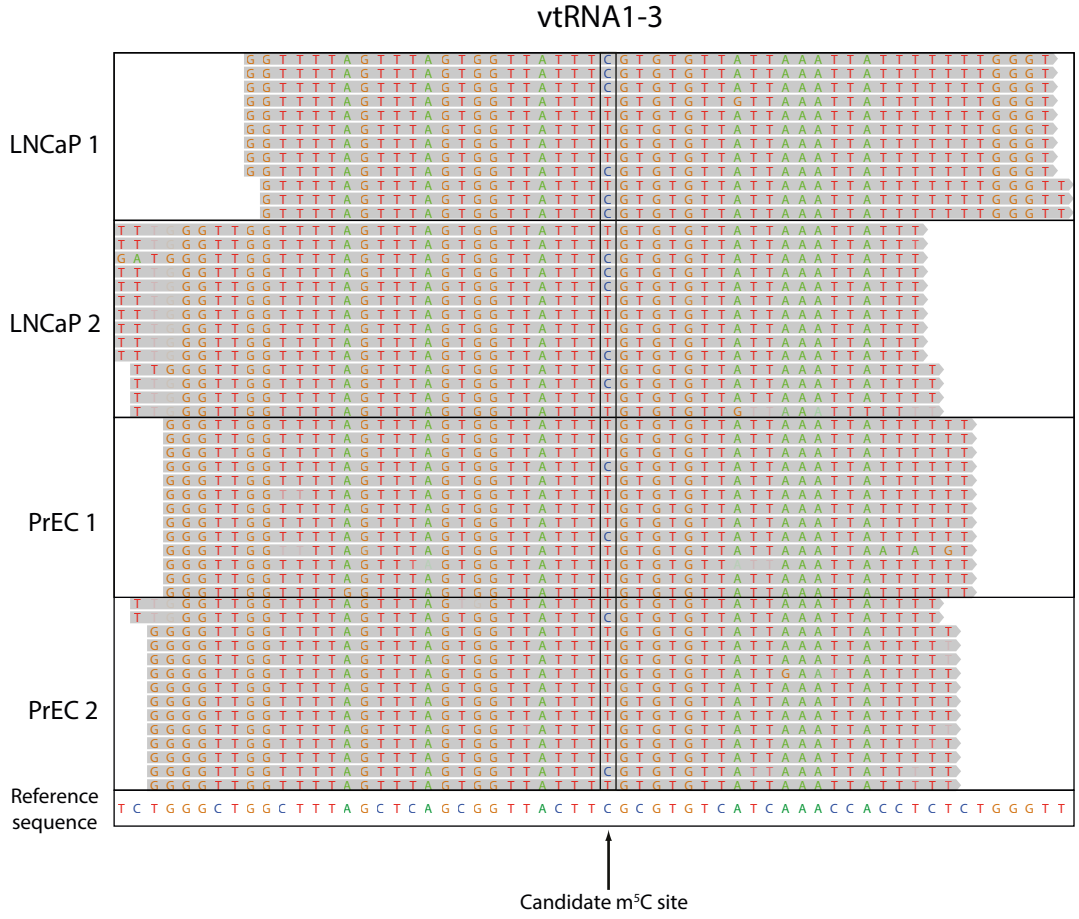


Figure 6.7. Representative snapshot from IGV depicting a subset of sequencing reads mapping to vtRNA1-3 surrounding the candidate m^{5}C site.

The grey bars represent 50 nt sequencing reads mapped to part of the locus. Unconverted cytosines represent the candidate m^{5}C site. The sequencing reads span chr5:140,105,740-140,105,798 (+ strand).

6.2.5 m^{5}C SITES IN COMMON BETWEEN PREC AND LNCAP CELLS

The original aim of this investigation was to perform a comparative screen of differential m^{5}C sites in PrEC and LNCaP cells and to relate the findings back to the available epigenomic and transcriptomic data reported in these cell lines by Professor Susan Clark, particularly to sites of differential DNA methylation in the genome. Due to the difficulty with validating candidate differential m^{5}C sites, this aim was abandoned and an in-depth analysis of candidate m^{5}C sites in common between PrEC and LNCaP cells and an integrative comparison with the HeLa methyltransferase knockdown dataset instead emerged as promising lines of investigation. This was of particular interest to confirm patterns of methylation and validate individual candidate m^{5}C sites across multiple cell lines.

Using the proportion statistic and appropriate controls, the total number of candidate m^{5}C sites in common between PrEC and LNCaP cells was determined. The criteria for the detection of common candidate m^{5}C sites were defined as ≥5 reads in all samples and ≥10% mean methylation across all samples, as well as a FDR≤1%. This yielded a total of 5,653 candidate m^{5}C sites in common between PrEC and LNCaP cells in a wide variety of transcript biotypes, including tRNAs, other ncRNAs and mRNAs (Figure 6.8). 438 candidate m^{5}C sites were detected in tRNAs from the global mapping, and 180 candidate m^{5}C sites were detected in tRNAs from the local mapping. Additionally, 559 candidate m^{5}C sites were detected in other ncRNAs and were distributed across various transcript biotypes. Interestingly, excluding unannotated intronic and intergenic regions, the majority of the candidate m^{5}C sites were detected in mRNAs, with 2,223 candidate m^{5}C sites called. The complete list of candidate m^{5}C sites in common between PrEC and LNCaP cells is located in Appendix 8.3.

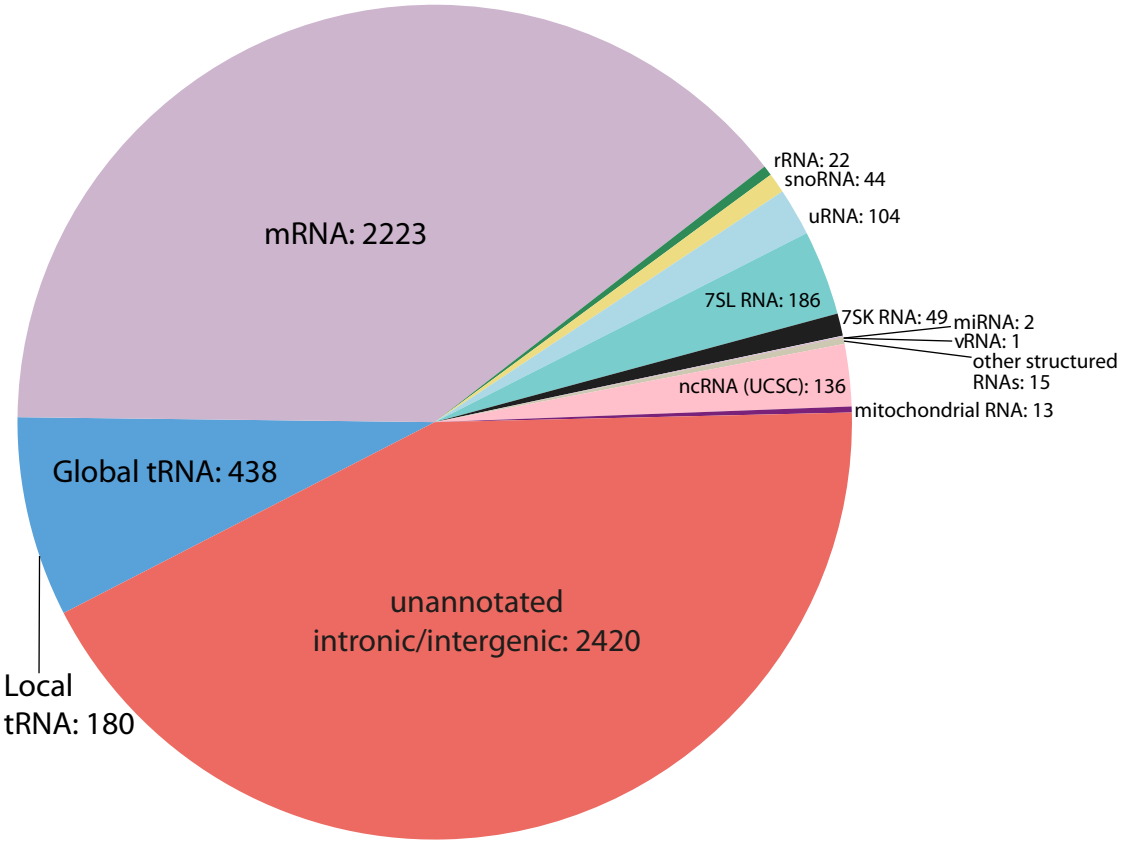


Figure 6.8. The distribution of candidate m^{5}C sites across different transcript biotypes in PrEC and LNCaP cells.

≥5 reads in all samples and ≥10% mean methylation across the control samples was required for the detection of candidate m^{5}C sites in common between PrEC and LNCaP cells. FDR≤1% was estimated by the Benjamini-Hochberg method. The number of candidate m^{5}C sites detected in each transcript biotype is indicated.

6.2.6 Global comparison with HeLa cells

In order to compare the RNA m^{5}C profiles of PrEC, LNCaP and HeLa cells, an integrative comparison of the PrEC and LNCaP (prostate) cell dataset and HeLa methyltransferase knockdown dataset was performed. This section will detail the global comparison of candidate m^{5}C sites in PrEC and LNCaP cells with those in HeLa cells. It should be noted that the total number of candidate m^{5}C sites in PrEC and LNCaP cells could not be determined based on the statistical model used to detect candidate m^{5}C sites. As a result, the global comparisons made in this section were between all candidate m^{5}C sites in HeLa cells and candidate m^{5}C sites in common between PrEC and LNCaP cells.

A direct comparison between the HeLa and prostate candidate m^{5}C sites was not trivial due to the differences in the number of controls used for each dataset (see section 3.2.3.B). To make the datasets more comparable, RPM expression thresholds were used for transcripts exhibiting candidate m^{5}C sites (Figure 6.9). This was particularly important to ensure that transcripts with candidate m^{5}C sites had relatively high expression in both datasets. The thresholds selected for trial ranged from 0.1-100 RPM (Table 6.1). Interestingly, as the expression threshold was increased, the proportion of overlapping candidate m^{5}C sites in HeLa cells and prostate cells also increased, which was a result of an increasing restriction to the highly expressed tRNAs. The overlap of candidate m^{5}C sites for each transcript biotype between all three cell lines is listed in Appendix 8.10

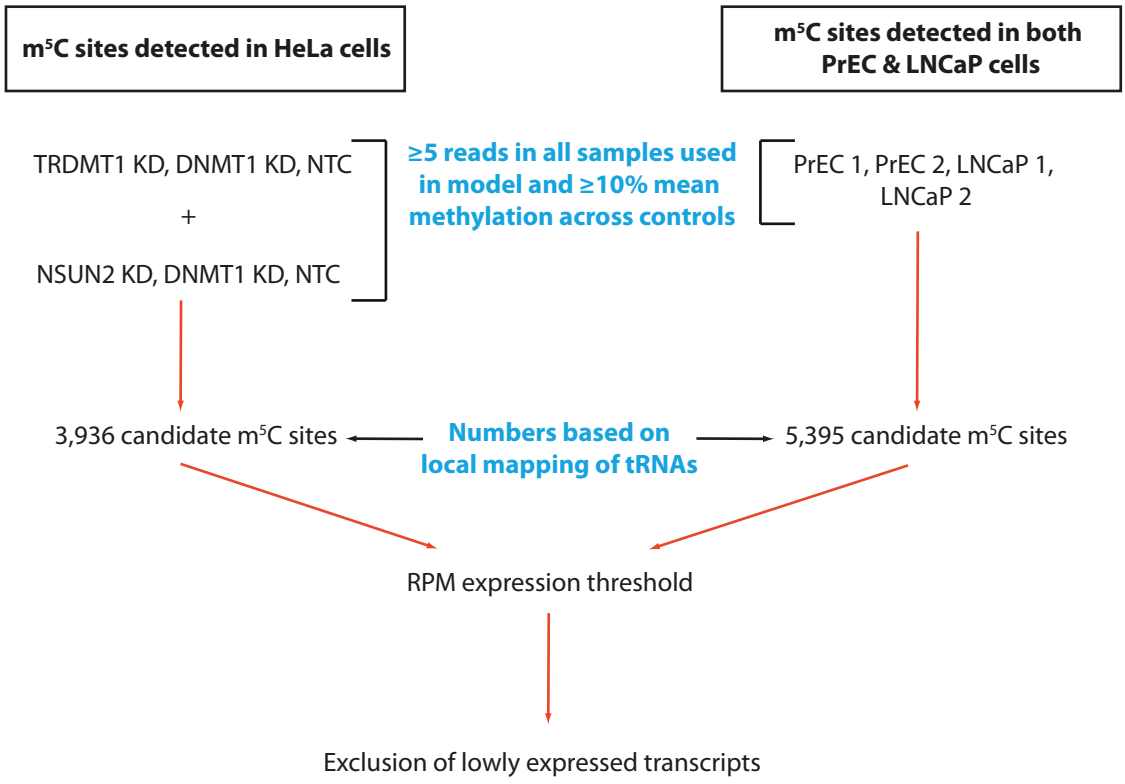


Figure 6.9. Schematic demonstrating the steps taken to make the HeLa methyltransferase knockdown and prostate datasets comparable.

The controls to detect candidate m^{5}C sites in HeLa cells and candidate m^{5}C sites in common between PrEC and LNCaP cells were different. To make the datasets more comparable and to remove lowly expressed transcripts, RPM expression thresholds were applied. Comparisons for tRNAs were based on the local mapping.

Table 6.1. Overlap of candidate m^{5}C sites detected in HeLa cells and sites in common between PrEC and LNCaP cells at different RPM expression thresholds.

Candidate m^{5}C sites in tRNAs were based on the local mapping.

RPM cut-off	Number of m^{5}C sites in HeLa cells	Number of m^{5}C sites in prostate cells	Number of overlapping m^{5}C sites	% overlapping m^{5}C sites
0.1	3933	5345	2250	53.1
0.5	2552	2223	1457	59.2
1	1569	1201	987	70.5
5	570	408	437	81.2
10	375	283	298	80.3
20	253	185	1922	82.1
50	142	95	120	81.6
100	86	68	78	83.0

In order to examine the overlap between HeLa cells and prostate cells in more detail, an expression threshold of 1 RPM was selected. This threshold was selected to focus on transcripts with sufficient coverage to quantify methylation and eliminate lowly expressed transcripts that may only be expressed in one dataset. At this threshold, mRNAs were still well-represented, whereas at higher expression thresholds the number of mRNAs was relatively low. For tRNAs, 88% of the candidate m^{5}C sites detected in the prostate cells overlapped with HeLa cells, indicating high concordance between each dataset (Figure 6.10A). Interestingly, 82% of the candidate m^{5}C sites detected in other ncRNAs in prostate cells overlapped with HeLa cells (Figure 6.10B). The overlap of candidate m^{5}C sites in mRNAs between prostate cells and HeLa cells was still substantial at 46% (Figure 6.10C).

In summary, the global comparison between PrEC and LNCaP cells and HeLa cells showed high concordance between the datasets, indicating that methylation patterns are conserved across multiple cell lines.

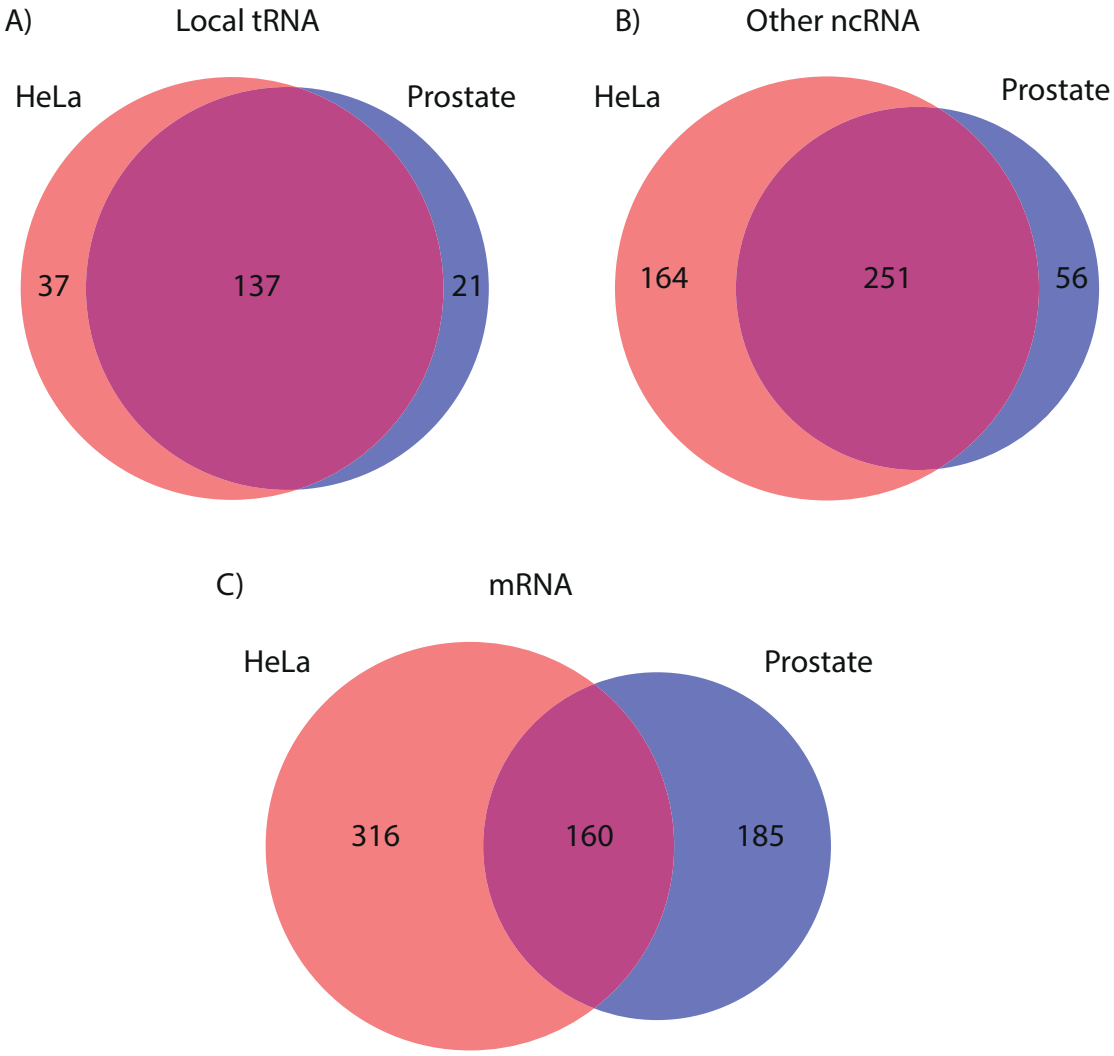


Figure 6.10. Comparison of candidate m^{5}C sites in different transcript biotypes in prostate cells and HeLa cells.

An expression threshold of 1 RPM was used for the comparison between the prostate dataset and HeLa methyltransferase knockdown dataset. A) Candidate m^{5}C sites detected in tRNAs from the local mapping, B) Candidate m^{5}C sites detected in other ncRNA, C) Candidate m^{5}C sites detected in mRNAs.

6.2.7 m^{5}C SITES IN COMMON IN tRNA

A total of 180 candidate m^{5}C sites in tRNAs were in common between PrEC and LNCaP cells from the local mapping. These numbers are comparable to the candidate m^{5}C sites detected in tRNAs in HeLa cells and the comparison revealed that 88% of candidate m^{5}C sites in tRNAs overlapped between all three cell lines. Therefore, it was of interest to document the candidate m^{5}C sites detected in tRNA structural positions, isotypes and isoacceptors in PrEC and LNCaP cells, and to compare this with those detected in HeLa cells. As bsRNA-seq was not performed on NSUN2 and TRDMT1 knockdown samples in PrEC and LNCaP cells, comparisons were made with all candidate m^{5}C sites detected in tRNAs in HeLa cells in this thesis. Due to the high expression of tRNAs, a RPM expression threshold was not applied. As with HeLa cells in Chapter 4, the analyses in this section primarily focused on candidate m^{5}C sites detected from the local mapping.

6.2.7.A tRNA STRUCTURAL POSITIONS AND ISOTYPES EXHIBITING m^{5}C SITES

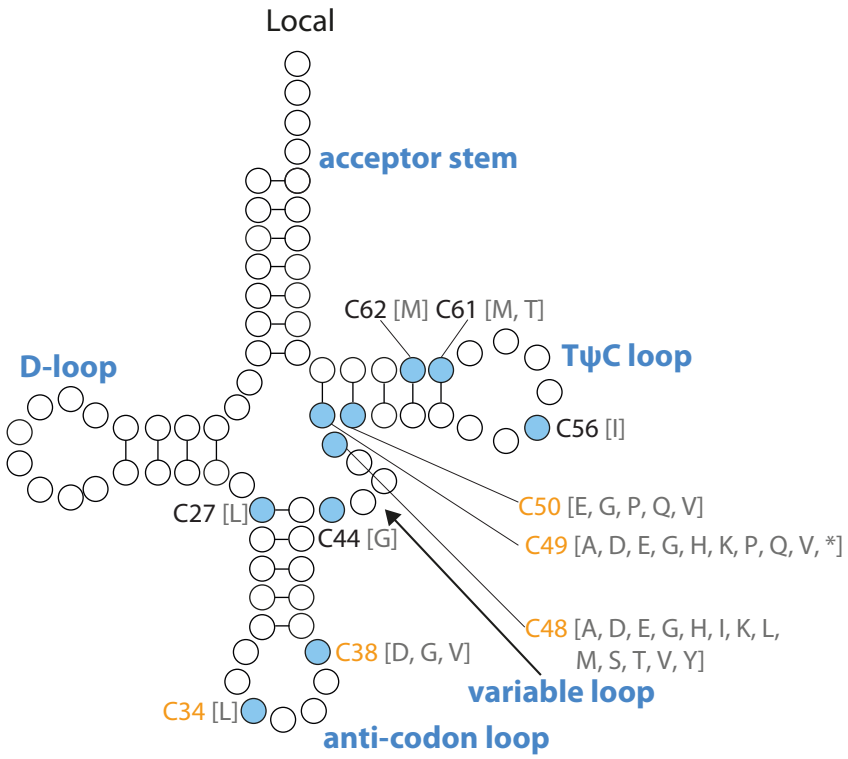
As with HeLa cells, the PrEC and LNCaP bsRNA-seq data was mined to determine the tRNA structural positions and isotypes in those structural positions that exhibited candidate m^{5}C sites in PrEC and LNCaP cells. A tRNA cloverleaf structure was then generated to visualise this (Figure 6.11).

Candidate m^{5}C sites were detected in most known structural positions, except C40 and C72, in PrEC and LNCaP cells (Figure 6.11A). Candidate m^{5}C sites detected at C34 were exclusively in tRNA^{Leu}, and at C38 were in tRNA^{Asp}, tRNA^{Gly} and tRNA^{Val}. Candidate m^{5}C sites at C48 and C49 were widespread across 10 and 13 different tRNA isotypes, respectively, and 7 isotypes overlapped between the two positions. In contrast, only 5 tRNA isotypes exhibited candidate m^{5}C sites at C50. Five non-canonical structural positions exhibited candidate m^{5}C sites in PrEC and LNCaP cells, including C27, C44, C56, C61 and C62. The candidate m^{5}C site at C27 was detected exclusively in tRNA^{Leu}, at C44 exclusively in tRNA^{Gly}, and at C56 exclusively in tRNA^{Ile}. The candidate m^{5}C sites at C61 were detected in tRNA^{Met} and tRNA^{Thr}, and at C62 in tRNA^{Met}.

The global mapping of tRNAs revealed candidate m^5C sites in all known structural positions (Figure 6.11B). The inability to detected candidate m^5C sites at C40 by the local mapping may be a result of identical tRNA sequences being mapped to a single representative sequence, causing dilution of the candidate m^5C site so that it was not detectable by the local mapping. As with HeLa cells, coverage towards the 3' end of the tRNA when locally mapped possibly resulted in the inability to detect m^5C at C72. A total of 11 non-canonical structural positions exhibited candidate m^5C sites based on the global mapping. Three of these non-canonical structural positions were detected in the local mapping, including C27, C44 and C56, although the number of isotypes exhibiting candidate m^5C sites at these positions was higher in the global mapping.

In summary, 96% of candidate m^5C sites in tRNAs that were in common between PrEC and LNCaP cells were in the anticipated structural positions.

A)


Local
acceptor stem
C62 [M] C61 [M, T]
TψC loop
D-loop
C56 [I]
C50 [E, G, P, Q, V]
C27 [L]
C44 [G]
C49 [A, D, E, G, H, K, P, Q, V, *]
C48 [A, D, E, G, H, I, K, L, M, S, T, V, Y]
C38 [D, G, V]
variable loop
C34 [L]
anti-codon loop

B)

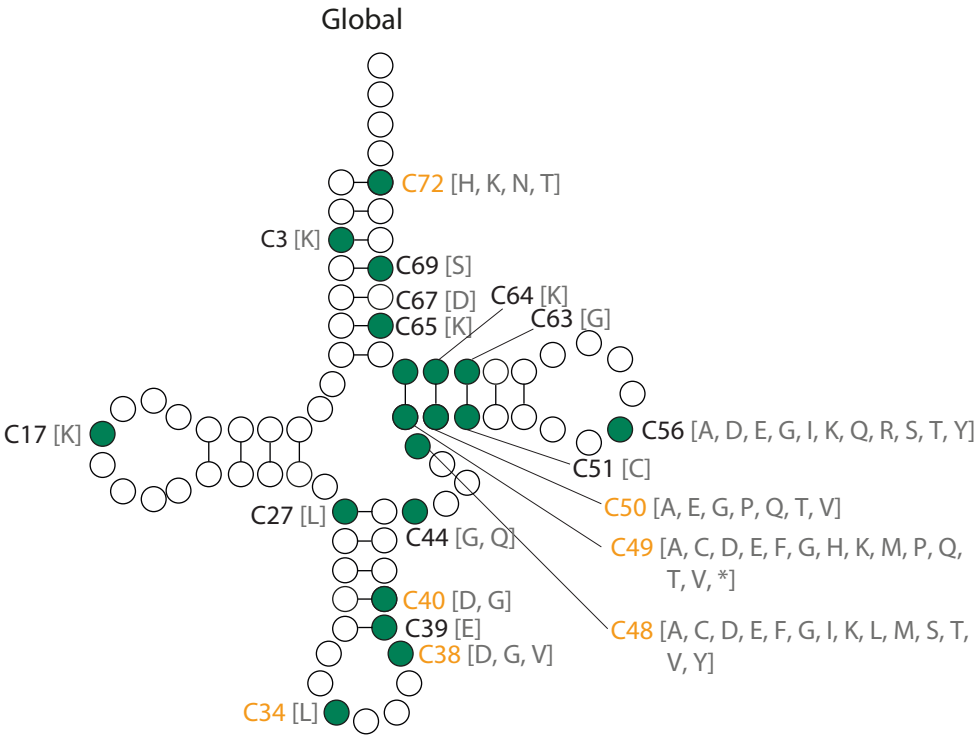

Global
C72 [H, K, N, T]
C3 [K]
C69 [S]
C67 [D]
C64 [K]
C65 [K]
C63 [G]
C17 [K]
C56 [A, D, E, G, I, K, Q, R, S, T, Y]
C51 [C]
C50 [A, E, G, P, Q, T, V]
C27 [L]
C44 [G, Q]
C49 [A, C, D, E, F, G, H, K, M, P, Q, T, V, *]
C40 [D, G]
C39 [E]
C48 [A, C, D, E, F, G, I, K, L, M, S, T, V, Y]
C38 [D, G, V]
C34 [L]

Figure 6.11. tRNA structural positions and isotypes containing candidate m^{5}C sites in PrEC and LNCaP cells.

A) The blue circles represent tRNA structural positions exhibiting candidate m^{5}C sites from the local tRNA mapping. B) The green circles represent tRNA structural positions exhibiting candidate m^{5}C sites from the global tRNA mapping. tRNA isotypes containing m^{5}C sites in these positions are indicated by the amino acid abbreviations. * indicates a suppressor tRNA. The different regions of the tRNA secondary structure are indicated. The seven structural positions known to exhibit m^{5}C sites are indicated in orange.

Comparison with HeLa cells

For the comparison of the tRNA structural positions exhibiting candidate m^{5}C sites between prostate cells and HeLa cells, a tRNA cloverleaf structure was generated indicating the positions in which candidate m^{5}C sites were detected in prostate cells and HeLa cells from the local mapping. In addition, the bsRNA-seq data was mined to determine the tRNA isotypes methylated at these positions that were in common between all three cell lines.

This yielded a total of 8 out of 10 structural positions exhibiting candidate m^{5}C sites in common, with 5 of these being in the canonical structural positions (Figure 6.12A). In addition, the tRNA isotypes targeted at C34, C38, C48, C49 and C50 in PrEC and LNCaP cells were in complete agreement with HeLa cells.

Candidate m^{5}C sites were detected at the non-canonical structural positions at C44, C56 and C61 in all three cell lines. As with the known structural positions, the tRNA isotypes at the non-canonical structural positions exhibiting candidate m^{5}C sites in PrEC and LNCaP cells were in complete agreement with HeLa cells. Interestingly, the candidate m^{5}C sites detected at C27 and C62 in PrEC and LNCaP cells were not detected in HeLa cells, and the candidate m^{5}C site detected at C30 in HeLa cells was not detected in PrEC and LNCaP cells.

The bsRNA-seq data was also mined to compare the tRNA isotypes that contained candidate m^{5}C sites between prostate cells and HeLa cells, regardless of the structural positions. The tRNA isotypes that exhibited candidate m^{5}C sites in PrEC and LNCaP

cells were broad, with 16 of 22 tRNA isotypes methylated (Figure 6.12B). In addition, 14 out of 16 tRNA isotypes overlapped with HeLa cells, indicating high concordance in tRNA isotypes containing candidate m^{5}C sites.

From this analysis, 80% of tRNA structural positions and 88% of tRNA isotypes that exhibited candidate m^{5}C sites in PrEC and LNCaP cells were in agreement HeLa cells.

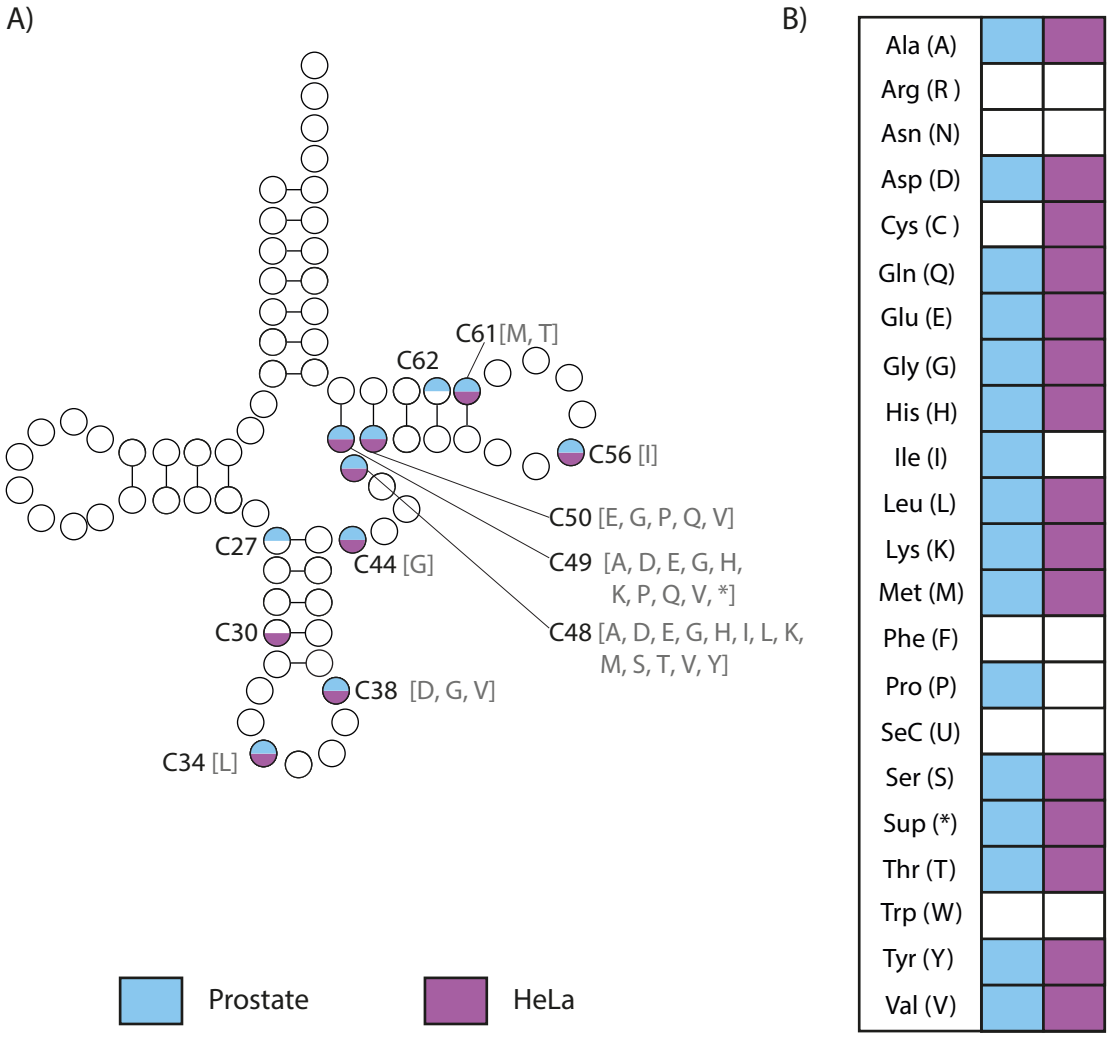


Figure 6.12. Comparison of tRNA structural positions and isotypes exhibiting candidate m^{5}C sites detected by bsRNA-seq in PrEC and LNCaP cells with HeLa cells.
Comparisons are based on the local mapping of tRNAs. A) tRNA structural positions exhibiting candidate m^{5}C sites in PrEC and LNCaP cells (blue) and HeLa cells (purple). Amino acid abbreviations indicate tRNA isotypes in agreement between all three cell lines at each structural position. B) tRNA isotypes exhibiting candidate m^{5}C sites in PrEC and LNCaP cells (blue) and HeLa cells (purple).

6.2.7.B tRNA ISOACCEPTORS EXHIBITING m^{5}C SITES

The PrEC and LNCaP bsRNA-seq data was also mined to determine the tRNA isoacceptors exhibiting candidate m^{5}C sites in both PrEC and LNCaP cells, and a table was generated for the visualisation of this data (Figure 6.13). The local mapping of tRNA data yielded a total of 35 tRNA isoacceptors containing candidate m^{5}C sites in both PrEC and LNCaP cells. These isoacceptors were broad, where multiple isoacceptors of a number of isotypes contained m^{5}C sites. Global mapping of the bsRNA-seq data yielded a total of 38 isoacceptors that contained candidate m^{5}C sites in both PrEC and LNCaP cells. Interestingly, 83% of the tRNA isoacceptors exhibiting candidate m^{5}C sites from the local mapping were in agreement with the global mapping.

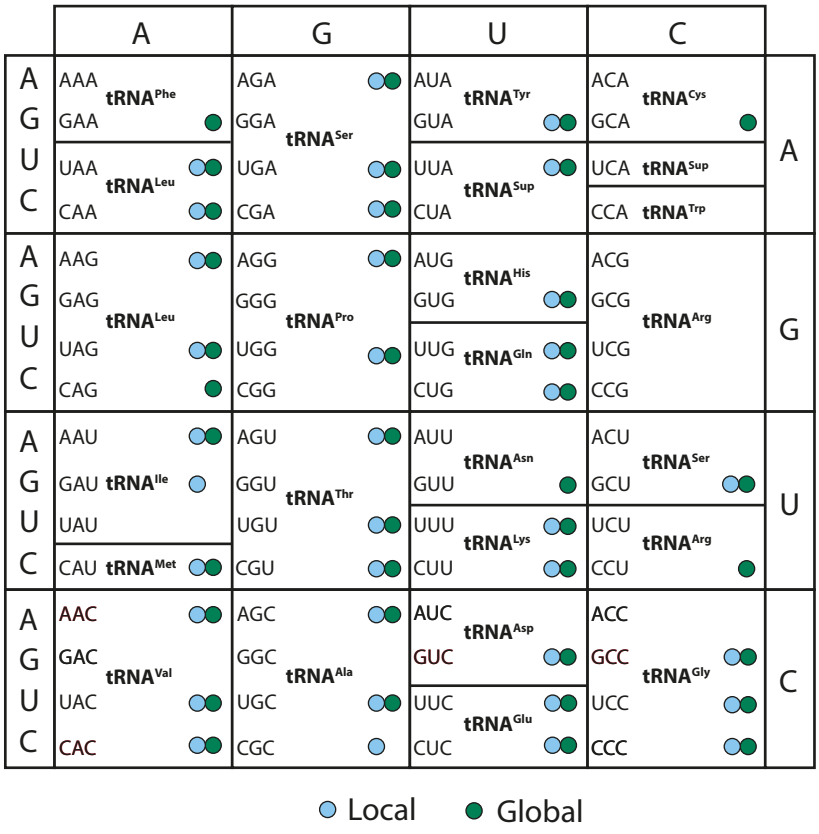


Figure 6.13. tRNA isoacceptors exhibiting candidate m^{5}C sites in PrEC and LNCaP cells. The blue and green circles represent tRNA isoacceptors exhibiting candidate m^{5}C sites based on the local and global mapping, respectively.

Comparison with HeLa cells

A comparison of the tRNA isoacceptors containing candidate m^{5}C sites in PrEC and LNCaP cells was made with those in HeLa cells. Based on the local mapping, 33 out of 35 tRNA isoacceptors exhibiting candidate m^{5}C sites overlapped with those detected in HeLa cells (Figure 6.14). Thus, as with the tRNA structural positions and isotypes, >90% tRNA isoacceptors exhibiting candidate m^{5}C sites in PrEC and LNCaP cells were in agreement with HeLa cells.

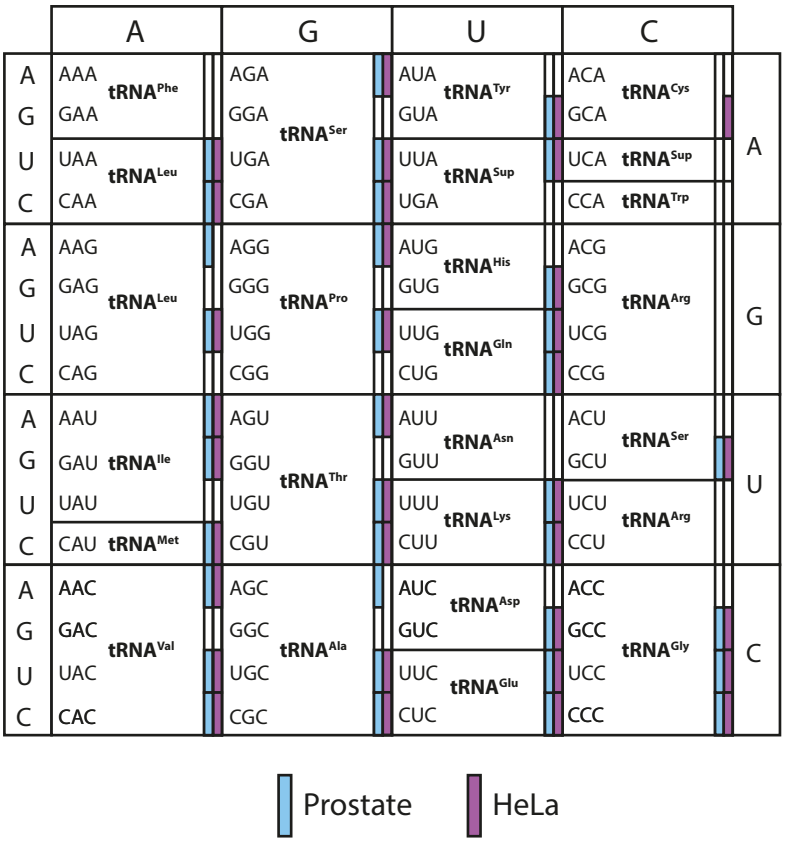


Figure 6.14. Comparison of tRNA isoacceptors exhibiting candidate m^{5}C sites in PrEC and LNCaP cells with HeLa cells.

The comparison was based on all m^{5}C sites detected in HeLa cells regardless of the methyltransferase that mediates the m^{5}C site.

6.2.7.C Validation of m^{5}C sites in tRNA

In order to confirm patterns of m^{5}C in tRNAs across multiple cell lines, four tRNAs exhibiting candidate m^{5}C sites in PrEC and LNCaP cells that were also validated in HeLa cells were selected for validation by MiSeq amplicon sequencing. This included tRNA$^{Glu(CUC)}$, tRNA$^{Gly(GCC)}$, tRNA$^{Thr(UGU)}$ and tRNA$^{Val(CAC)}$. As with HeLa cells, the coverage of the tRNAs selected were high based on the bsRNA-seq, with all tRNAs exhibiting declining coverage towards the 3' end (Figure 6.15; top panel).

As with the m^{5}C sites in tRNAs in HeLa cells, there was an underrepresentation of multiple m^{5}C sites in tRNA$^{Glu(CUC)}$ and tRNA$^{Gly(GCC)}$ detected by bsRNA-seq in PrEC and LNCaP cells. Candidate m^{5}C sites were detected at C38 and C50 of tRNA$^{Gly(GCC)}$ from the bsRNA-seq, but the validation revealed additional m^{5}C sites at C48 and C49 (Figure 6.15; middle and bottom panels). Also, the candidate m^{5}C sites detected at C49 and C50 of tRNA$^{Glu(CUC)}$ from the bsRNA-seq appeared to be underrepresented compared to MiSeq amplicon sequencing.

In all cases, the candidate m^{5}C sites detected by bsRNA-seq were validated, including at C49 and C50 in tRNA$^{Glu(CUC)}$, C48 in tRNA$^{Thr(UGU)}$ and C48 and C49 in tRNA$^{Val(CAC)}$. Depletion of m^{5}C at C48-50 was evident in the NSUN2 knockdown sample, indicating that NSUN2 mediates m^{5}C at these positions in LNCaP cells in addition to HeLa cells. Interestingly, mild depletion of m^{5}C was evident at C38 of tRNA$^{Gly(GCC)}$ in the NSUN2 knockdown sample by MiSeq amplicon sequencing, although this was validated as a TRDMT1-mediated site in HeLa cells in this thesis and by previous studies [73, 77].

In summary, it is clear that m^{5}C sites in at least 4 tRNAs are present across multiple cell lines, and these sites are mediated NSUN2 in both LNCaP and HeLa cells.

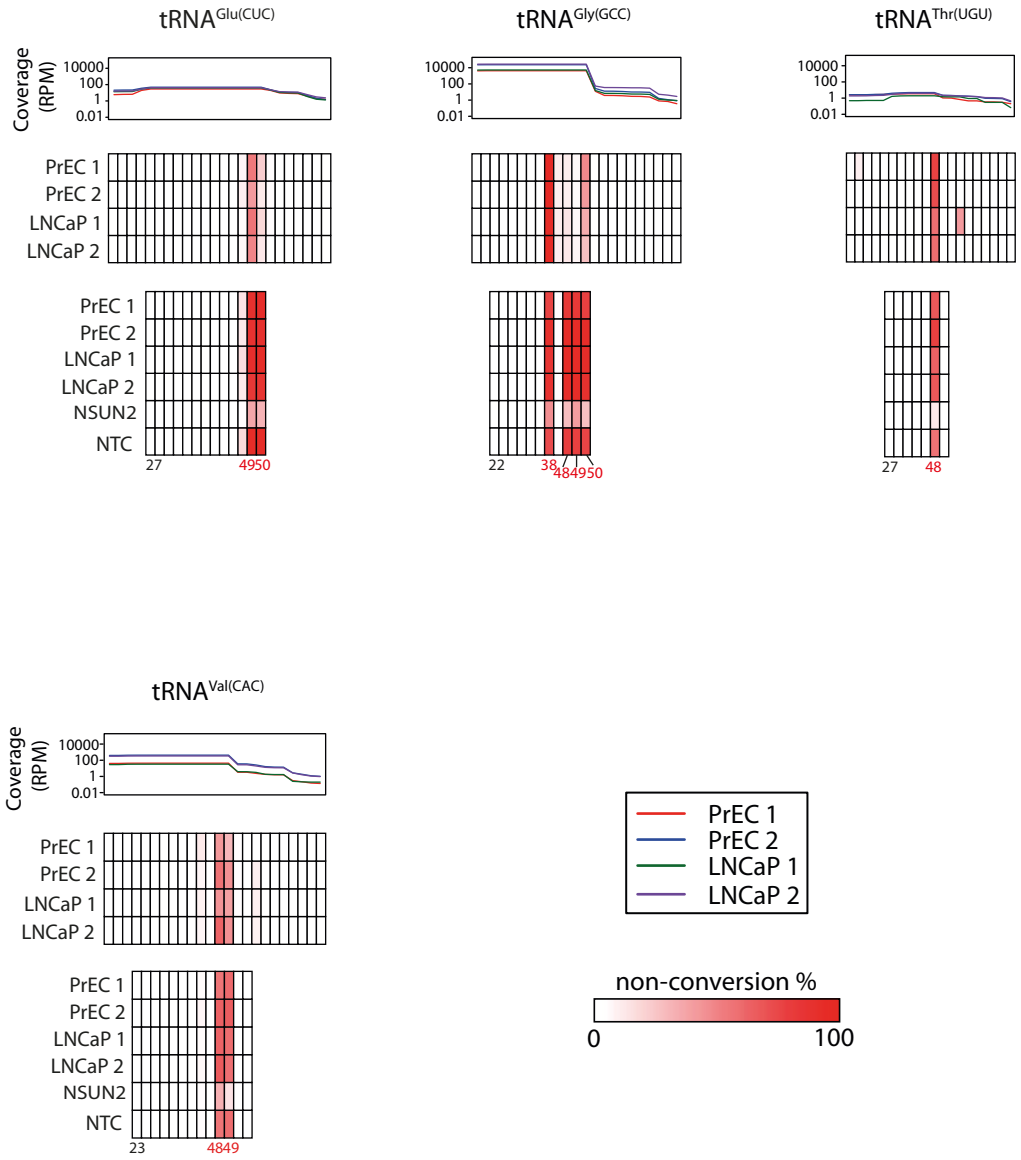


Figure 6.15. Validation of candidate m^{5}C sites in tRNAs in PrEC and LNCaP cells by MiSeq amplicon sequencing.

The top panel displays the coverage throughout the transcript in RPM based on bsRNA-seq, the middle panel displays the proportion of non-conversion based on bsRNA-seq, and the bottom panel displays the proportion of non-conversion in a segment of the transcript based on MiSeq amplicon sequencing. The tRNA structural positions with candidate m^{5}C sites are indicated in red below the heatmap. The bsRNA-seq data is based on the local mapping of tRNAs. The MiSeq coverage ranged from 800-72,000 mappable sequencing reads.

6.2.8 m^{5}C SITES IN COMMON IN OTHER ncRNA

The bsRNA-seq yielded 559 candidate m^{5}C sites in other ncRNA that were in common between PrEC and LNCaP cells. Furthermore, 82% of candidate m^{5}C sites detected in ncRNA in PrEC and LNCaP cells were in concordance with HeLa cells. As such, two candidate m^{5}C sites were selected for validation to confirm individual sites across multiple cell lines. Interestingly, the NSUN2-mediated m^{5}C sites detected in *SCARNA2* and *SNORD62* in HeLa cells were also detected in PrEC and LNCaP cells by bsRNA-seq. Both m^{5}C sites were validated by MiSeq amplicon sequencing in PrEC and LNCaP cells (Figure 6.16). Strong depletion of methylation was observed at both m^{5}C sites in the LNCaP NSUN2 knockdown sample, with the m^{5}C site in *SCARNA2* depleting to 7% and the m^{5}C site in *SNORD62* depleting to 8%, indicating that NSUN2 mediates these sites in LNCaP cells.

In summary, it is clear that m^{5}C sites in several ncRNAs can be present across multiple cell lines, and these sites are mediated NSUN2 in both LNCaP and HeLa cells.

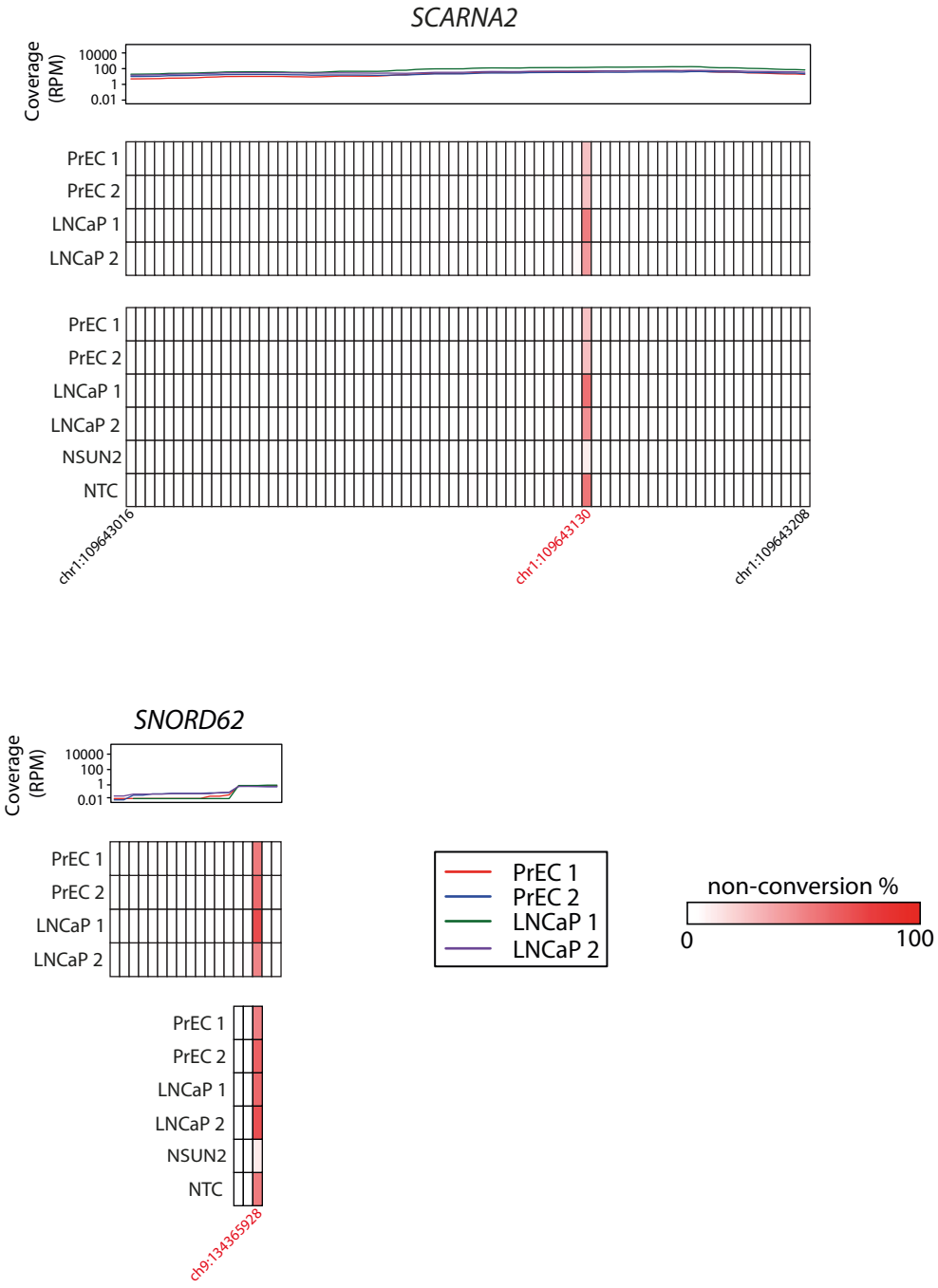


Figure 6.16. Validation of candidate m^{5}C sites in other ncRNA in PrEC and LNCaP cells by MiSeq amplicon sequencing.

The top panel displays the coverage in RPM based on bsRNA-seq, the middle panel displays the proportion of non-conversion based on bsRNA-seq, and the bottom panel displays the proportion of non-conversion based on MiSeq amplicon sequencing. Chromosomal locations of the candidate m^{5}C sites are indicated in red below the heatmap. The MiSeq coverage ranged from 18,000-70,000 mappable sequencing reads.

6.2.9 m^{5}C SITES IN COMMON IN mRNA

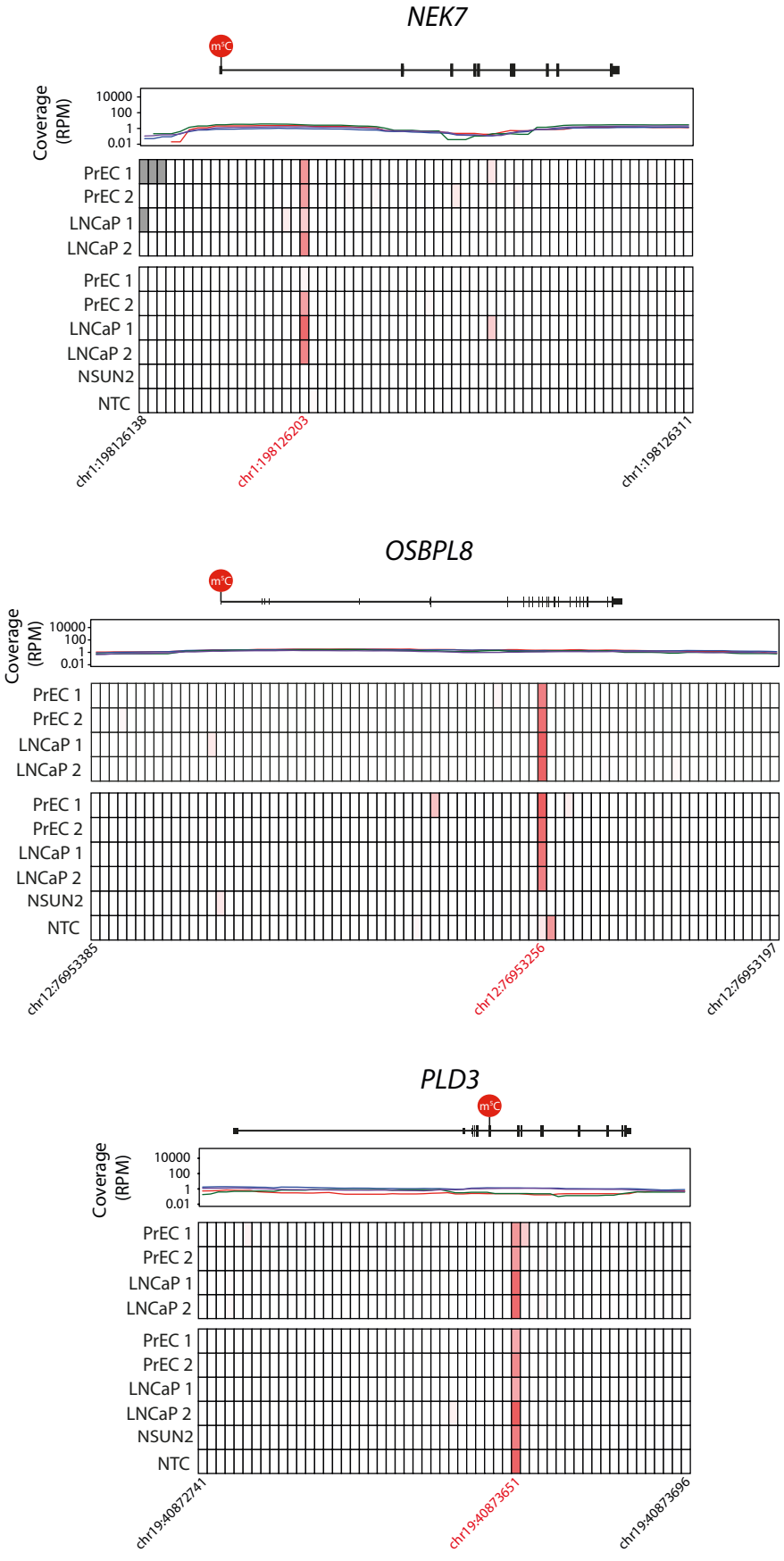
The majority of candidate m^{5}C sites in common between PrEC and LNCaP cells were in mRNAs and the comparison with HeLa cells also indicated 46% of candidate m^{5}C sites in mRNAs were in common between all three cell lines. To confirm these patterns, candidate m^{5}C sites detected and validated in mRNA in HeLa cells that were also detected in PrEC and LNCaP cells were selected for validation by MiSeq amplicon sequencing. These included the m^{5}C sites in *OSBPL8*, *RTN3* and *SZRD1* that were validated as NSUN2 targets in HeLa cells (see section 4.2.4.B). An additional four novel candidate m^{5}C sites detected in PrEC and LNCaP cells were also selected for validation to confirm patterns across the two cell lines and to determine if these sites were mediated by NSUN2 in LNCaP cells. This included candidate m^{5}C sites in *NEK7*, *PLD3*, *PWP2* and *SRRT*.

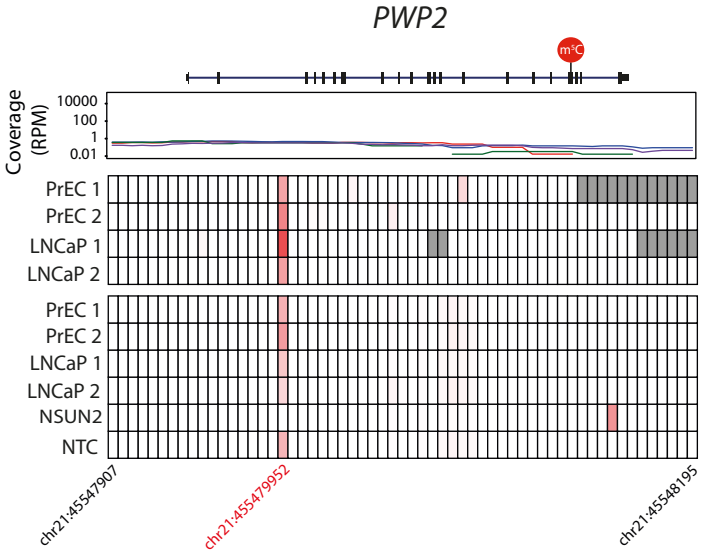
The m^{5}C sites in *OSBPL8*, *RTN3* and *SZRD1* were validated in both PrEC and LNCaP cells (Figure 6.17). Depletion of m^{5}C was evident in the LNCaP NSUN2 knockdown sample in *RTN3* and *SZRD1*, where m^{5}C levels depleted to 8% and <1%, respectively, indicating that these sites are also mediated by NSUN2 in LNCaP cells. Although *OSBPL8* was validated as a NSUN2 target in HeLa cells, little methylation was detected in the LNCaP NTC sample in addition to the LNCaP NSUN2 knockdown. Interestingly, an unconverted cytosine was instead evident one base downstream in the NTC sample. Thus, it could not be concluded that NSUN2 mediated this m^{5}C site in LNCaP cells in addition to HeLa cells.

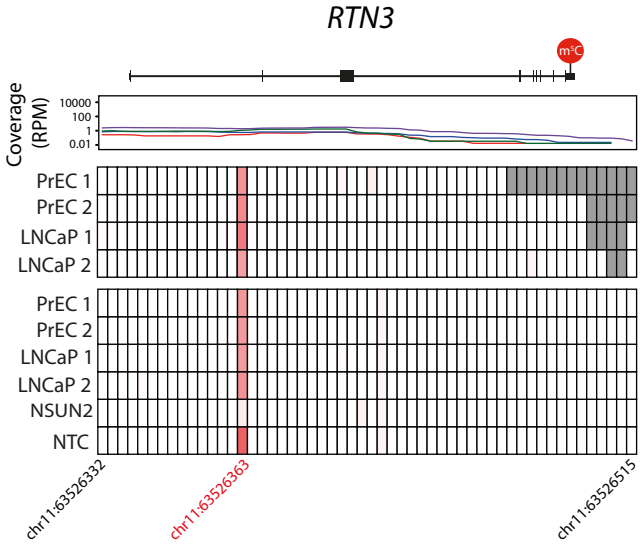
The novel candidate m^{5}C sites detected in *NEK7*, *PLD3*, *PWP2* and *SRRT* were also validated by MiSeq amplicon sequencing. Interestingly, the level of m^{5}C in *NEK7* in PrEC 1 detected by MiSeq amplicon sequencing was substantially lower than that detected by bsRNA-seq. bsRNA-seq detected ~40% methylation at this site, whilst MiSeq amplicon sequencing detected ~5% methylation in this sample. Similarly, <1% methylation was detected in the LNCaP NSUN2 knockdown and NTC samples based on MiSeq amplicon sequencing. In contrast, PrEC 2, LNCaP 1 and LNCaP 2 exhibited ~40-50% methylation. It is possible that the variability in the proportion of methylation may be attributed to PCR amplification biases (see section 5.2.5.A). Although this m^{5}C site was validated, the site was not a convincing NSUN2 target due to the lack of

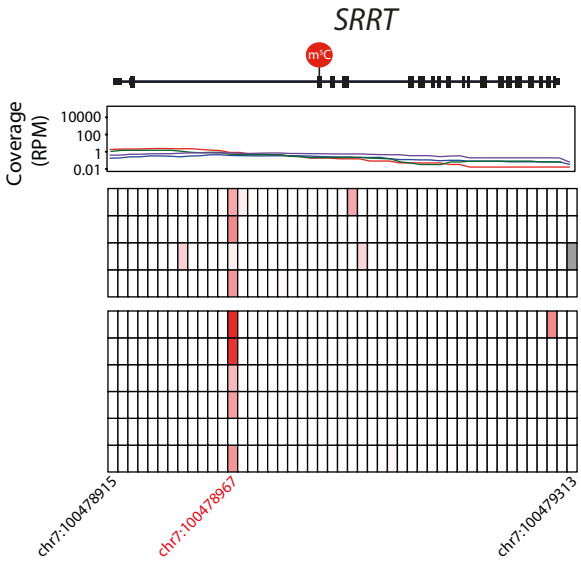
methylation in both the LNCaP NSUN2 knockdown and NTC samples. In contrast, depletion of methylation in the LNCaP NSUN2 knockdown sample was evident in *PWP2* and *SRRT*, indicating that NSUN2 mediates m^{5}C in these transcripts in LNCaP cells. The candidate m^{5}C site validated in *PLD3* in PrEC and LNCaP cells was not convincingly depleted upon NSUN2 knockdown in LNCaP cells. Therefore, this was not a convincing NSUN2-mediated site in LNCaP cells.

These findings have provided evidence that NSUN2-mediated m^{5}C sites in at least seven mRNAs are conserved across multiple cell lines.


NEK7
m5C
Coverage (RPM)
10000
100
1
0.01
PrEC 1
PrEC 2
LNCaP 1
LNCaP 2
PrEC 1
PrEC 2
LNCaP 1
LNCaP 2
NSUN2
NTC
chr1:198126138
chr1:198126203
chr1:198126311
OSBPL8
m5C
Coverage (RPM)
10000
100
1
0.01
PrEC 1
PrEC 2
LNCaP 1
LNCaP 2
PrEC 1
PrEC 2
LNCaP 1
LNCaP 2
NSUN2
NTC
chr12:76953385
chr12:76953256
chr12:76953197
PLD3
m5C
Coverage (RPM)
10000
100
1
0.01
PrEC 1
PrEC 2
LNCaP 1
LNCaP 2
PrEC 1
PrEC 2
LNCaP 1
LNCaP 2
NSUN2
NTC
chr19:40872741
chr19:40873651
chr19:40873696


PWP2
m5C
Coverage (RPM)
10000
100
1
0.01
PrEC 1
PrEC 2
LNCaP 1
LNCaP 2
PrEC 1
PrEC 2
LNCaP 1
LNCaP 2
NSUN2
NTC
chr21:45547907
chr21:45547952
chr21:45548195


RTN3
m5C
Coverage (RPM)
10000
100
1
0.01
PrEC 1
PrEC 2
LNCaP 1
LNCaP 2
PrEC 1
PrEC 2
LNCaP 1
LNCaP 2
NSUN2
NTC
chr11:63526332
chr11:63526363
chr11:63526515


SRRT
m5C
Coverage (RPM)
10000
100
1
0.01
chr7:100478915
chr7:100478967
chr7:100479313

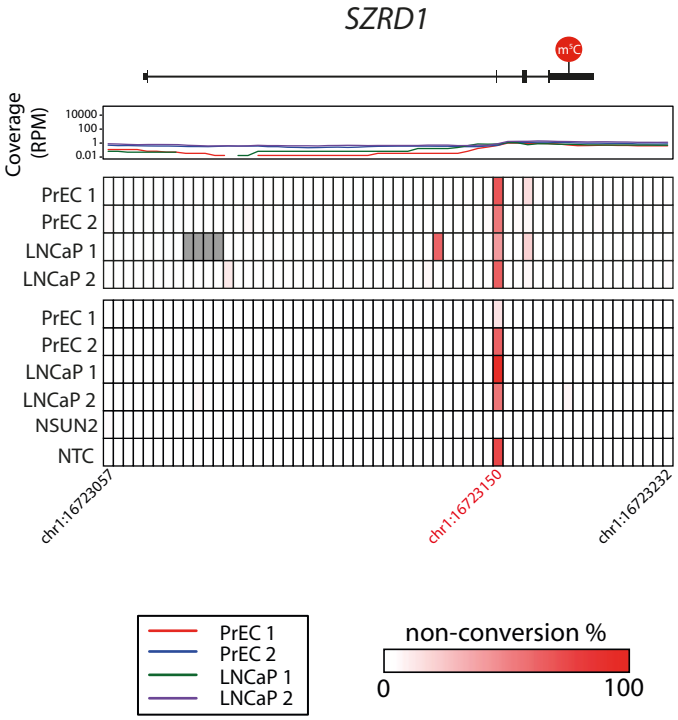


Figure 6.17. Validation of candidate m^{5}C sites in mRNA in PrEC and LNCaP cells by MiSeq amplicon sequencing.

The location of the m^{5}C site relative to the regions in the mRNA is indicated at the top. The top panel displays the coverage in RPM based on bsRNA-seq, the middle panel displays the proportion of non-conversion based on bsRNA-seq, and the bottom panel displays the proportion of non-conversion based on MiSeq amplicon sequencing. The shades of grey in the middle panel indicate regions of the transcript where coverage was <5 reads based on bsRNA-seq. Chromosomal locations of the candidate m^{5}C sites are indicated in red below the heatmap. The MiSeq coverage ranged from 10,000-70,000 mappable sequencing reads.

6.3 Discussion

The original aim of this investigation was to perform a comparative screen of candidate differential m^{5}C sites in PrEC and LNCaP cells and to relate the findings back to the available epigenomic and transcriptomic data reported in these cell lines. Previous studies have reported increased levels of NSUN2 in a number of cancers and it was observed in this thesis that NSUN2 was overexpressed more than two-fold in LNCaP cells compared to PrEC cells. As such, it was originally hypothesised that an increase in NSUN2 may result in either an increased number of m^{5}C sites in LNCaP cells compared to PrEC cells, or an increase in the stoichiometry of methylation at m^{5}C sites detected in both cell lines. However, candidate differential m^{5}C sites could not be validated and as such, I used the data to compare candidate m^{5}C sites across multiple cell lines and to validate individual candidates across transcript biotypes. Here, I will discuss the outcomes of these findings.

6.3.1 Differential m^{5}C sites in PrEC and LNCaP cells

6.3.1.A ncRNAs

The local mapping of tRNAs revealed no candidate differential m^{5}C sites in either cell line or increased stoichiometry of methylation in LNCaP cells. Only a small number of candidate differential m^{5}C sites were detected in other ncRNA and as a result, these candidates were not investigated in great detail. Investigation into the stoichiometry of methylation of candidate m^{5}C sites in common between PrEC and LNCaP cells in different ncRNAs indicated little difference between each cell line. A single vtRNA, vtRNA1-3, showed significantly more methylation in LNCaP cells compared to PrEC cells. Although this site was not independently validated in this thesis, it was detected as a NSUN2 target in HeLa cells in this thesis and two other studies [77, 78], and Hussain et al. reported that loss of NSUN2-mediated methylation in vtRNA1-1 caused decreased processing into svRNAs that can associate with Ago2 and Ago3 and potentially induce silencing of mRNA (see section 4.3.3.B) [78]. Although this was only detected in vtRNA1-1, it is possible that the difference in stoichiometry of m^{5}C in vtRNA1-3 in PrEC and LNCaP cells may result in changes in the processing of this transcript into

smaller fragments. This may contribute to variations in gene expression in PrEC and LNCaP cells that may result in the differences associated with normal and cancer cells.

6.3.1.B mRNAs

While m^{5}C in tRNAs was reported to promote mRNA translation through stabilisation, this does not disregard other factors contributing to the regulation of translation. The thesis has provided several correlative links between the presence of m^{5}C in mRNAs and several aspects of post-transcriptional gene regulation, including mRNA stability and translation. Thus, it was originally of interest to investigate the differences in m^{5}C in mRNA in PrEC and LNCaP cells that could potentially contribute to the deregulation of translation commonly seen in cancer.

It was curious that the majority of candidate differential m^{5}C sites detected by bsRNA-seq were almost exclusively detected in one cell line and not the other, rather than being hypermethylated. Additionally, examination of the sequencing reads mapping to the differential mRNA candidates in IGV revealed that the majority of candidate m^{5}C sites exhibited low quality scores, suggesting that these may be artefacts. To determine if this was the case, three mRNAs containing candidate differential m^{5}C sites were selected for further interrogation by MiSeq amplicon sequencing. Analysis of the candidate m^{5}C site in *SOX4* by MiSeq amplicon sequencing and Sanger bisulfite sequencing revealed that this site was most likely an artefact due to incomplete denaturation of the RNA. This issue is not trivial to overcome. For the bisulfite conversion of RNA, a compromise was made between efficient denaturation and limiting fragmentation of the RNA, suggesting that some transcripts may still exhibit these clusters (see section 3.2.1.B). While reducing the amplicon size would limit the detection of these clusters, the short insert size used for bsRNA-seq resulted in the detection of a single non-converted cytosine in *SOX4* as a convincing candidate m^{5}C site. In the case of 28S rRNA, the bsRNA-seq revealed clusters of non-converted cytosines despite the short insert size possibly due to the substantially strong secondary structure (see section 4.2.3.A). Increasing the insert or amplicon size may increase the probability of detecting clusters of non-conversion, however these clusters would be identifiable by the examination of the sequencing reads that surround these regions. Ultimately, optimisation of a m^{5}C antibody for the

detection of regions exhibiting candidate m^{5}C sites by MeRIP, which was attempted in this thesis, would be of benefit as clusters of non-conversion due to inefficient denaturation would not be detected. Coupling this technique with bsRNA-seq would enable detection of candidate m^{5}C sites at single nucleotide resolution in addition to removal of secondary structure artefacts.

The candidate differential m^{5}C sites in *FTSJ3* and *SZRD1* failed to validate for different reasons to that of *SOX4*. The candidate m^{5}C site in *FTSJ3* was only detected in PrEC cells and an annotated G-to-A SNP was located upstream and adjacent to the candidate m^{5}C site. The SNP was of high quality after examination in IGV, however a small subset of sequencing reads that did not contain the SNP exhibited low quality calling of the base. The low quality calling of this base resulted in the low quality calling of adjacent bases including the candidate m^{5}C site. This ultimately resulted in the detection of a false positive site. This SNP was not detected in LNCaP cells and the base at this location was of high quality, resulting in the correct calling of a converted cytosine. Similarly, the candidate m^{5}C site in *SZRD1* was of low quality, resulting in a false positive.

Taken together, the small number of candidate differential m^{5}C sites detected and the inability to validate a subset of these sites in PrEC and LNCaP cells suggests that the RNA m^{5}C profiles of each cell line are largely similar. This indicates that increased NSUN2 in LNCaP cells did not result in an obvious increase number of m^{5}C sites or increase in the stoichiometry of m^{5}C in this thesis. The similar RNA m^{5}C profiles is not unexpected, as this study was performed as an initial screen and in combination with the moderate overexpression of NSUN2 in LNCaP cells compared to the strong depletion of methylation observed in NSUN2 knockdown cells suggests that any differences detected would be subtle. Therefore, it is possible that the level of overexpression exhibited in LNCaP cells was not sufficient enough to confer hypermethylation of RNA detectable in this screen.

6.3.2 m^{5}C SITES IN COMMON BETWEEN PrEC AND LNCaP CELLS

An investigation into the m^{5}C sites in common between PrEC and LNCaP cells was made. A comparative analysis with HeLa cells was also performed in order to confirm patterns of methylation and validate individual m^{5}C sites across multiple cell lines.

The bsRNA-seq and subsequent bioinformatics pipeline used yielded a total of 5,653 candidate m^{5}C sites in common between PrEC and LNCaP cells. 2,223 candidate m^{5}C sites were detected in mRNAs, 180 candidate m^{5}C sites in tRNAs from the local mapping and 559 candidate m^{5}C sites distributed over other ncRNAs. The wide distribution of candidate m^{5}C sites in various transcript biotypes among three cell lines indicates that this modification is not just limited to tRNAs and highlights a functional importance in other transcript biotypes.

6.3.2.A GLOBAL COMPARISON WITH HELA CELLS

As a different number of control samples were used for the detection of candidate m^{5}C sites in the prostate dataset and HeLa dataset, they were not directly comparable. To make the datasets more comparable, an expression threshold for transcripts containing m^{5}C sites detected in each dataset was used. An expression threshold of 1 RPM enabled detection of tRNAs as well other ncRNA and mRNAs. This revealed high concordance between prostate cells and HeLa cells for all transcript biotypes. 88% of the candidate m^{5}C sites in tRNAs in common between PrEC and LNCaP cells were detected in HeLa cells at this expression threshold. The large overlap of candidate m^{5}C sites in tRNAs between all cell lines suggests that m^{5}C in tRNAs plays an important functional role in all cellular contexts, which has been highlighted by several previous studies that reported increased cleavage of tRNAs devoid of m^{5}C sites [73, 149, 150]. Interestingly, a study that disrupted NSUN2 and TRDMT1 in mice not only reported ablation of all m^{5}C sites in tRNAs, but also reduced steady-state levels of the tRNAs [149]. Additionally, reduced rate of protein synthesis was observed in *Nsun2* and *Trdmt1* double knockout mouse embryonic fibroblasts (MEFs), although other factors such as mRNA methylation may also contribute to this [149]. This finding highlights the importance of m^{5}C in tRNA fidelity, regardless of the cellular context, and also

suggests that a slight deregulation in this modification in tRNAs may compromise the cell.

The candidate m^{5}C sites detected in other ncRNAs were also highly concordant between all cell lines, with 82% of candidate m^{5}C sites in common between PrEC, LNCaP and HeLa cells. The candidate m^{5}C sites in mRNAs were also highly concordant between all cell lines, although this was slightly lower at 46%, suggesting that m^{5}C in mRNAs is more dynamic and may largely depend on the cellular context. The high concordance of candidate m^{5}C sites in other ncRNAs and mRNAs between three different cell lines indicates that this modification is not randomly distributed, suggesting important regulatory functions for this modification in mRNA and ncRNA.

6.3.2.B m^{5}C sites in tRNAs

The local mapping of tRNAs revealed that all candidate m^{5}C sites detected in tRNAs were in common between PrEC and LNCaP cells, highlighting the importance of this modification for the fidelity of tRNAs. As with Chapter 4, candidate m^{5}C sites in tRNA structural positions, isotypes and isoacceptors were investigated and these were compared with all candidate m^{5}C sites detected in HeLa cells. In this section, I will discuss the outcomes of these findings. As absence of NSUN2 and TRDMT1 in mice was reported to result in complete ablation of m^{5}C in tRNAs [149], the tRNA structural positions, isotypes and isoacceptors exhibiting candidate m^{5}C sites in PrEC and LNCaP cells were also juxtaposed with the NSUN2 and TRDMT1 Aza-IP study [77]. As the Aza-IP study mapped tRNAs globally, comparisons were made between candidate m^{5}C sites detected from both the local and global mapping of tRNAs with the Aza-IP study.

The local mapping of tRNAs revealed that 96% of candidate m^{5}C sites were located at five of the seven known structural positions in PrEC and LNCaP cells. Interestingly, candidate m^{5}C sites were not detected at C40, which may be a result of identical tRNA sequences being mapped to a single representative sequence, causing dilution of the candidate m^{5}C site so that it was not detectable in the local mapping. This is highlighted by the fact that the level of methylation detected at this positions in tRNA$^{Gly(GCC)}$ in HeLa cells in this thesis was low (see section 4.2.2.C). In the case of C72, declining

coverage towards the 3' end of the tRNA when mapped locally possibly resulted in the inability to detect m^5C at this position from the local mapping.

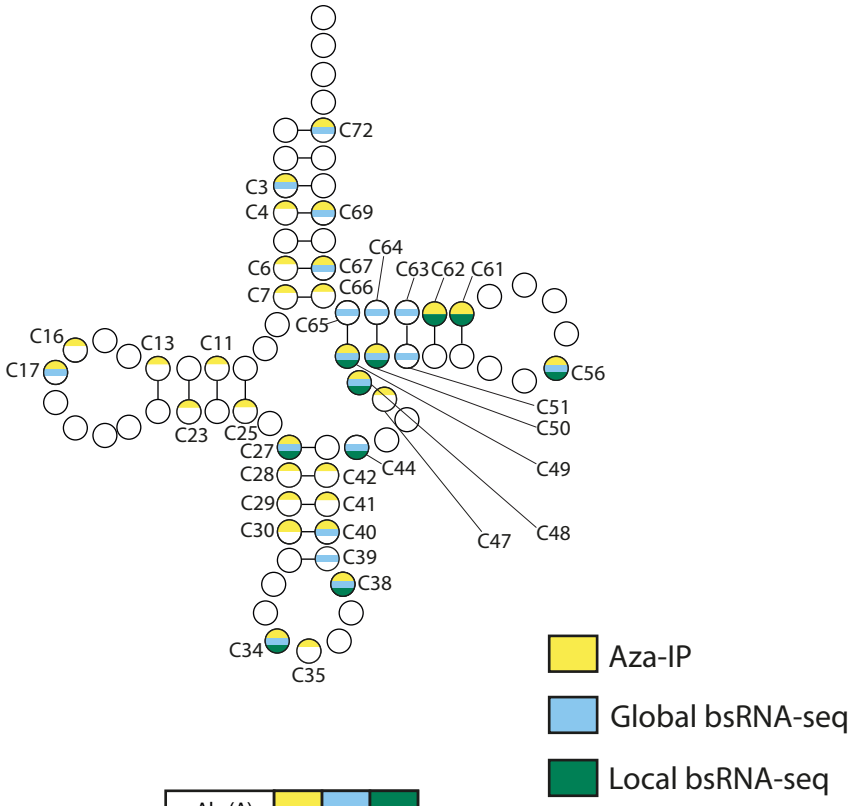
Candidate m^5C sites detected at the anticipated structural positions and the tRNA isotypes in these positions in prostate cells were largely in agreement with those detected in HeLa cells in this thesis and the Aza-IP study, in addition to previous literature (Figure 6.18A). This includes m^5C at C34 in $tRNA^{Leu(CAA)}$ [100], C38 in $tRNA^{Asp(GUC)}$ [58, 59, 77, 159], $tRNA^{Gly(GCC)}$ [73, 77], $tRNA^{Val(AAC)}$ [73, 77] and $tRNA^{Val(CAC)}$ [77], and C48-50 in a broad range of tRNA isotypes [58, 98, 148, 149]. There was high concordance in the tRNA isotypes targeted for methylation at C48-50 with the Aza-IP study, with >85% tRNA isotypes exhibiting candidate m^5C sites at each of these positions in PrEC and LNCaP cells overlapping with the Aza-IP study (data not shown).

Candidate m^5C sites were located in five non-canonical structural positions in PrEC and LNCaP cells, including C27, C44, C56, C61 and C62. The candidate m^5C sites detected at C44, C56 and C62 were also detected in HeLa cells, and the tRNA isotypes at these positions were in complete agreement. Interestingly, four of the five non-canonical structural positions exhibiting candidate m^5C sites in PrEC and LNCaP cells were also detected by Aza-IP, including C27, C56, C61 and C62. The tRNA isotypes methylated at these positions differed between each study, with substantially more isotypes detected by Aza-IP (data not shown). However, there was some overlap including m^5C at C27 in $tRNA^{Leu}$ and m^5C at C61 in $tRNA^{Met}$. As expected, the global mapping of tRNAs in PrEC and LNCaP cells showed a stronger overlap, with candidate m^5C sites detected at C56 in $tRNA^{Ala}$, $tRNA^{Arg}$ and $tRNA^{Gln}$ overlapping with the Aza-IP study. The high concordance of the non-canonical tRNA structural positions exhibiting candidate m^5C sites between PrEC and LNCaP cells, HeLa cells in this thesis and by Aza-IP, highlight that these positions are more commonly modified than previously thought.

The tRNA isotypes exhibiting candidate m^5C sites in PrEC and LNCaP cells were broad, with the local mapping revealing 16 isotypes containing candidate m^5C sites, and the global mapping revealing 21 isotypes containing candidate m^5C sites. As with the structural positions, there was high concordance between both mapping methods, with

all 16 tRNA isotypes in the local mapping overlapping with the global mapping. In addition to being in agreement with HeLa cells in this thesis, the data was largely in agreement with the Aza-IP study, with 18 candidate m^{5}C sites from the global mapping and 15 from the local mapping overlapping with the Aza-IP study (Figure 6.18B).

A)



B)

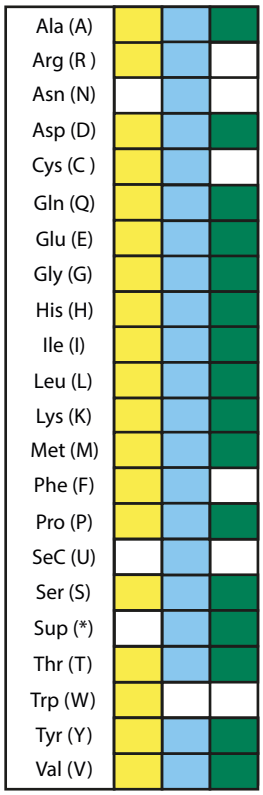


Figure 6.18. Comparison of tRNA structural positions and isotypes exhibiting candidate m^{5}C sites in PrEC and LNCaP cells with the Aza-IP study [77].

All candidate m^{5}C sites in common between PrEC and LNCaP cells were compared with NSUN2 and TRDMT1 targeted tRNAs. A) tRNA structural positions, B) tRNA isotypes.

The local and global mapping of the tRNAs revealed a wide variety of tRNA isoacceptors exhibiting candidate m^{5}C sites in PrEC and LNCaP cells; 35 isoacceptors contained at least one candidate m^{5}C site based on the local mapping, and 38 isoacceptors based on the global mapping. Both methods of mapping were largely in agreement, with 33 isoacceptors overlapping. Furthermore, ~95% of the tRNA isoacceptors containing candidate m^{5}C sites in PrEC and LNCaP cells were in agreement with HeLa cells in this thesis and >90% were in agreement with the Aza-IP study (Figure 6.19).

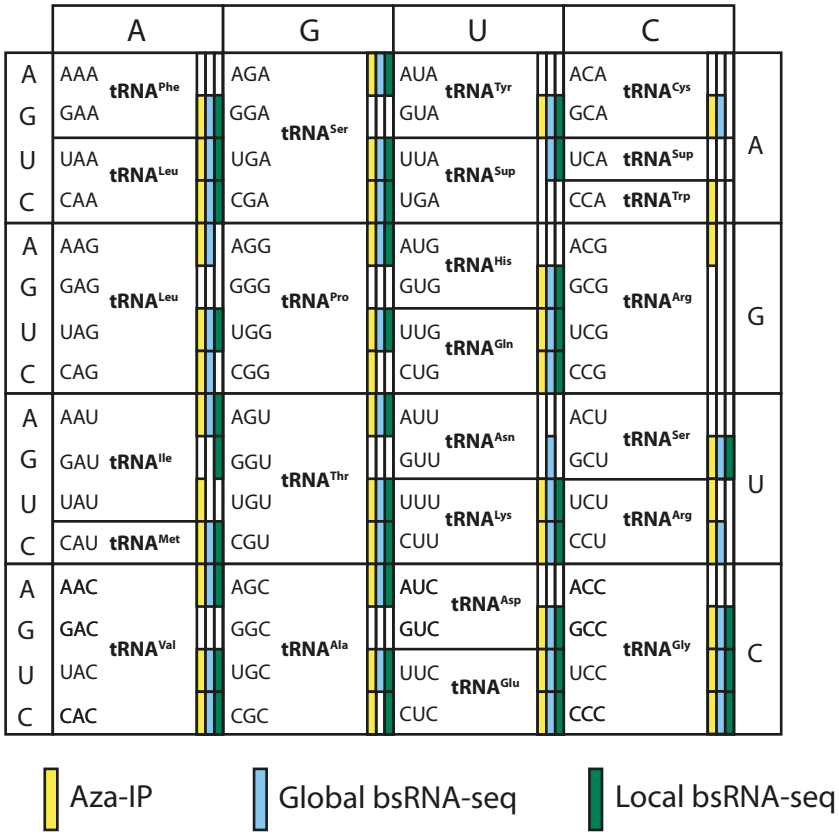


Figure 6.19. Comparison of tRNA isoacceptors exhibiting candidate m^{5}C sites in PrEC and LNCaP cells with the Aza-IP study [77].

All candidate m^{5}C sites in common between PrEC and LNCaP cells were compared with NSUN2 and TRDMT1 targeted tRNAs.

Analysis of the candidate m^{5}C sites in tRNAs that were in common between PrEC and LNCaP cells and the comparison with HeLa cells has revealed high concordance between all three cell lines on three different levels; tRNA structural positions, isotypes and isoacceptors. Several of the tRNAs containing candidate m^{5}C sites in all three cell lines were also validated in PrEC and LNCaP cells, and NSUN2 mediated m^{5}C sites at C34 and C48-50 in LNCaP cells in addition to HeLa cells. Comparison of the tRNA structural positions, isotypes and isoacceptors with the NSUN2 and TRDMT1 Aza-IP study also revealed strong agreement with PrEC and LNCaP cells. This suggests that tRNAs are consistently and highly modified under different cellular contexts, highlighting the importance of this modification in tRNA fidelity.

6.3.2.C m^{5}C sites in other ncRNA

559 candidate m^{5}C sites detected in various classes of ncRNA were in common between PrEC and LNCaP, and the integrative comparison with HeLa cells performed in this chapter also revealed high concordance with HeLa cells. To confirm these patterns, the m^{5}C sites detected and validated in *SCARNA2* and *SNORD62* in HeLa cells, which were also detected in PrEC and LNCaP cells, were selected for validation in PrEC and LNCaP cells. The m^{5}C sites in both ncRNAs were validated in both cell lines and were also confirmed as NSUN2 targets in LNCaP cells. Detection and validation of the same NSUN2-mediated m^{5}C sites in the selected ncRNAs in multiple cell lines highlights that m^{5}C in ncRNAs is not random and can exist in various cellular contexts. This suggests that m^{5}C in other ncRNA may play a regulatory role. While m^{5}C in vtRNAs has been reported to regulate their processing into smaller fragments [78], the role of m^{5}C in other ncRNA transcript biotypes remains to be explored. Possible molecular functions of m^{5}C in ncRNA include regulating stability and association with proteins, although this remains to be explored in further detail.

Comparison of candidate m^{5}C sites with the NSUN2 and TRDMT1 Aza-IP study showed little overlap, with only m^{5}C sites in *RPPH1*, *SCARNA2* and vtRNA1-3 in common with PrEC and LNCaP cells [77]. It is possible that many m^{5}C sites detected in ncRNAs in PrEC and LNCaP cells may be mediated by other enzymes besides NSUN2 and TRDMT1. Another plausible explanation for the differences, which was discussed

in Chapter 4, is the difference in the sensitivity of the techniques. Aza-IP may enrich tRNAs due to their high abundance and high density of m^{5}C, which may have resulted in the depletion of other ncRNAs and mRNAs containing m^{5}C sites (see section 4.3.1.C). The fact that there was high concordance in m^{5}C sites detected in PrEC and LNCaP cells with HeLa cells in this thesis suggests that the differences were not necessarily attributed to different cellular contexts, but more likely a result of the sensitivity of the different techniques.

6.3.2.D m^{5}C sites in mRNA

With the exception of candidate m^{5}C sites in unannotated and intronic regions, the majority of the candidate m^{5}C sites in common between PrEC and LNCaP cells were in mRNA. As with tRNAs and other ncRNAs, there was high concordance in the candidate m^{5}C sites in mRNAs between PrEC, LNCaP and HeLa cells, with 46% of candidate m^{5}C sites in agreement between all three cell lines. To confirm patterns of methylation in mRNA across multiple cell lines, a subset of candidate m^{5}C sites in mRNAs was validated by MiSeq amplicon sequencing. The mRNAs selected for validation encoded proteins of various cellular processes, and several of these transcripts were detected and validated as NSUN2-mediated m^{5}C sites in HeLa cells, including *OSBPL8*, *RTN3* and *SZRD1*. The m^{5}C sites in all three mRNAs were validated in PrEC and LNCaP cells, and the sites in *RTN3* and *SZRD1* were also mediated by NSUN2 in LNCaP cells. The m^{5}C site in *OSBPL8* was not detected in either the NSUN2 knockdown or NTC samples in LNCaP cells, although it is highly likely that this site is also mediated by NSUN2 in LNCaP cells.

Four novel m^{5}C sites in mRNAs validated in PrEC and LNCaP cells were not detected in HeLa cells, and included *NEK7*, *PLD3*, *PWP2* and *SRRT*. The proteins encoded by these transcripts have broad biological functions, suggesting that m^{5}C in mRNA may regulate stability or translation of transcripts encoding proteins that are involved in broad biological processes. NEK7 is part of the NIMA-related family of serine/threonine kinases involved in the mitotic signalling cascade [329]. PLD3 is a non-classical and poorly characterised member of the PLD family of phospholipases, with members of the family involved in catalysing the hydrolysis of

phosphatidylcholine to phosphatidic acid [330]. Members of the PLD family are activated during a number of different cellular events, including vesicle trafficking, cell migration and mitosis, and were reported to be deregulated in a number of cancers [330-333]. Recently, GWAS of late-onset Alzheimer's disease families reported a rare coding variant of PLD3 (Val232Met) that increased the risk of the development of Alzheimer's disease [334]. The m^{5}C site in *PLD3* was located ~400 nt upstream of this variant, suggesting that it would not contribute to the variant. PWP2 is homologous to the yeast Pwp2 and is a nucleolar protein reported to be involved in pre-rRNA processing [335, 336]. SRRT is a nuclear cap-binding protein involved in cell proliferation and studies suggest that it is required for miRNA-mediated silencing [337-339]. The m^{5}C sites in *PWP2* and *SRRT* were mediated by NSUN2 in LNCaP cells, whereas depletion of methylation in *PLD3* was not observed in the NSUN2 knockdown sample, suggesting that either residual NSUN2 after knockdown may have still methylated this site or another methyltransferase may mediate this modification. The m^{5}C site in *NEK7* was not detected in the NSUN2 knockdown and NTC samples in LNCaP cells and therefore it could not be concluded whether NSUN2 mediates this site.

As with tRNAs and other ncRNAs, it is clear that patterns of m^{5}C in mRNAs are conserved across multiple cell lines and can be mediated by the same enzyme, indicating the non-random distribution of this modification in RNA.

6.3.3 Conclusions

This was the first study to compare the RNA m^{5}C profiles of normal and cancer cells. Several hundred differential m^{5}C sites were detected in PrEC and LNCaP cells, with the majority of sites present in mRNAs. Examination of the sequencing reads of the bsRNA-seq data indicated that a large proportion of the candidate m^{5}C sites were of low quality and most likely artefacts, which was highlighted by the inability to validate a subset of these sites. The m^{5}C sites that were in common between PrEC and LNCaP cells showed little difference in the stoichiometry of methylation, with only a single vtRNA showing higher proportion of methylation in LNCaP cells compared to PrEC cells. This suggests that the RNA m^{5}C profiles of PrEC and LNCaP cells were largely similar, and that the higher expression of NSUN2 in LNCaP cells did not result in an increase in the number of m^{5}C sites or stoichiometry at a particular m^{5}C site.

I compared the RNA m^{5}C profiles between PrEC and LNCaP cells and successfully detected 5,653 candidate m^{5}C sites across various transcript biotypes that were in common between PrEC and LNCaP cells, with several successfully validated. In addition to being in common between PrEC and LNCaP cells, a large number of candidate m^{5}C sites were in agreement with HeLa cells. A small subset of these m^{5}C sites in tRNAs, other ncRNAs and mRNAs were validated in PrEC and LNCaP cells, and were mediated by NSUN2 in LNCaP cells in addition to HeLa cells. There was also high concordance in the tRNA structural positions, isotypes and isoacceptors containing m^{5}C sites in all three cell lines and the literature, highlighting the important role of m^{5}C in tRNA stability. Taken together, this indicates that m^{5}C patterns are conserved across multiple cell lines, suggesting a regulatory role for this modification in a number of transcript biotypes.

6.4 Chapter Highlights

- NSUN2 was overexpressed more than two-fold in LNCaP cells compared to PrEC cells.

6.4.1 Differential m^{5}C sites

- Candidate m^{5}C sites that were differentially methylated between PrEC and LNCaP cells were difficult to detect; reducing the FDR criteria to ≤40% allowed the listing of 426 candidate sites as hypermethylated in PrEC cells and 389 candidate m^{5}C sites that were hypermethylated in LNCaP cells, with the majority of candidate sites in mRNAs.
- Visual inspection revealed that many of these candidate differential m^{5}C sites were called based on sequencing reads with poor quality values or were otherwise potentially compromised.
- Three mRNA candidates were analysed by MiSeq amplicon sequencing and failed to validate. The site in *SOX4* was likely called as a result of incomplete denaturation of the RNA, while the sites in *FTSJ3* and *SZRD1* were called based on reads with poor quality values.

- The stoichiometry of m^{5}C for candidate m^{5}C sites in common between PrEC and LNCaP cells were indistinguishable for most transcript biotypes. A single vtRNA, vtRNA1-3, showed significantly higher methylation in LNCaP cells compared to PrEC cells.

6.4.2 m^{5}C SITES IN COMMON

- 5,653 candidate m^{5}C sites were in common between PrEC and LNCaP cells.
- Using a 1 RPM expression threshold, there was high concordance in the candidate m^{5}C sites detected in prostate cells with HeLa cells; 88% of candidate m^{5}C sites in tRNAs, 82% in other ncRNAs and 46% in mRNAs were in common between PrEC, LNCaP and HeLa cells.
- All candidate m^{5}C sites in tRNAs were in common between PrEC and LNCaP cells. There was also high concordance with those detected in HeLa cells. This suggests that m^{5}C in tRNA is important for fidelity.
- The m^{5}C sites validated in the ncRNAs *SCARNA2* and *SNORD62* in HeLa cells were also validated in PrEC and LNCaP cells. Both m^{5}C sites were mediated by NSUN2 in LNCaP cells.
- The m^{5}C sites validated in the mRNAs *OSBPL8*, *RTN3* and *SZRD1* were also validated in PrEC and LNCaP cells. The m^{5}C sites in *RTN3* and *SZRD1* were also mediated by NSUN2 in LNCaP cells.
- Novel m^{5}C sites were validated in *NEK7*, *PLD3*, *PWP2* and *SRRT* in PrEC and LNCaP cells. The m^{5}C sites in *PWP2* and *SRRT* were mediated by NSUN2 in LNCaP cells. Convincing depletion of m^{5}C in *PLD3* in the NSUN2 knockdown sample was not evident.
- These findings indicate that patterns of m^{5}C in tRNAs, other ncRNAs and mRNA are conserved across different cellular contexts and are mediated by the same enzyme.

CHAPTER SEVEN

GENERAL DISCUSSION & FUTURE DIRECTIONS

7.1 DISCUSSION

In this chapter, I will discuss the outcomes of this thesis and the importance of this study for the understanding of m^5C in RNA.

There are currently >100 nucleoside modifications known in RNA, which were identified over several decades mostly by studying the ncRNAs of the translational machinery (tRNA, rRNA). Notable exceptions were m^6A and m^5C, which were also found at low levels in cellular and viral mRNAs [20-39, 47, 48]. Recent next-generation sequencing studies in the emerging field of "epitranscriptomics" have mapped sites of m^5C [58], m^6A [56, 57] and now also ψ [53-55] throughout the transcriptome including mRNAs, hinting at currently hidden regulatory functions. This raises many questions, including whether several or perhaps many of the other nucleoside modifications may also be distributed throughout the transcriptome. Investigations into m^5C, m^6A and ψ all faced the challenge of identifying the writers, readers and erasers of these modifications and deciphering the molecular functions, particularly in mRNA. m^6A currently leads the way, with an emerging consensus that it is involved in mRNA stability and alternative splicing [57, 69, 105, 107, 120]. The discovery of ψ in mRNA in the last year has meant that limited investigations into its molecular function have been performed, although initial analysis suggests a stabilising role in mRNA [54]. Another challenge is whether these modifications are static or dynamic, with indications for m^6A and ψ that many of the sites are preserved across tissues and species, although there is also evidence that a subset of sites respond to changed conditions such as stress or meiosis in yeast [53-57, 67].

While next-generation sequencing has revealed the abundance and distribution of m^5C throughout the transcriptome, the understanding of the enzymology, molecular functions and biological implications of this modification in mRNA remain to be explored in greater detail. This thesis has greatly extended what was previously known about the substrate range of NSUN2 and TRDMT1, revealing multiple novel target sites in mRNAs and some ncRNAs, while consolidating our knowledge of their range of sites in tRNAs. Previous studies have primarily focused on m^5C sites mediated by NSUN2 and TRDMT1 in tRNAs and other ncRNAs, with limited detection or exploration of

m^5C sites in mRNAs [77, 78]. Thus, this was also the first study to explore candidate m^5C sites in mRNAs in detail and explore the molecular functions of m^5C in mRNA.

NSUN2 has been reported by several studies to be overexpressed in cancer, however there is a lack of studies investigating the RNA m^5C profiles across normal and cancer cells [99, 207, 317]. This was the first study to perform a comparative analysis of the RNA m^5C profiles of normal and cancer cells to investigate differential m^5C sites that may contribute to cancer. Little difference in the RNA m^5C profiles was evident between PrEC and LNCaP cells and as such this motivated a comparative analysis of candidate m^5C sites in common between PrEC and LNCaP cells with candidate m^5C sites detected in HeLa cells to investigate the presence of m^5C in different cellular contexts.

To summarise, this is the first study to extensively interrogate the substrate range of NSUN2 and TRDMT1, the specificity of NSUN2 in a range of transcript biotypes and the molecular functions of m^5C in mRNA. Additionally, this is the first study to perform a comparative analysis of RNA m^5C profiles between normal and cancer cells.

7.1.1 Substrate range of NSUN2 and TRDMT1

One of the major aims of this thesis was to investigate the substrate range and specificity of NSUN2 and TRDMT1 by coupling RNAi-mediated knockdown of the enzymes with bsRNA-seq. In addition to detecting thousands of candidate m^5C sites in a wide variety of transcript biotypes in HeLa cells, NSUN2 and TRDMT1 were found to mediate m^5C sites in tRNAs, other ncRNAs and mRNAs. I found that the majority of candidate m^5C sites that were called as NSUN2 or TRDMT1 targets were in mRNAs, substantially extending what was previously known about the substrate range of these enzymes.

Concurring with recent complementary transcriptome-wide studies, 87% of candidate m^5C sites detected in tRNAs were in the anticipated structural positions, with 99% of NSUN2 targets at C34 and C48-50 in a broad range of tRNA isotypes and isoacceptors, and 93% of TRDMT1 targets at C38 exclusively in $tRNA^{Asp(GUC)}$, $tRNA^{Gly(GCC)}$, $tRNA^{Val(AAC)}$ and $tRNA^{Val(CAC)}$. As with tRNAs, NSUN2 was found to be multisite-

specific in other ncRNAs and mRNAs, mediating m^5C within stem-loop structures. Several of the structures surrounding the m^5C sites in ncRNAs and mRNAs were reminiscent of those in tRNAs, although there were other structural features that were targeted by this enzyme. As with the structures, NSUN2 was found to target different sequences in ncRNA, and a computational analysis of candidate m^5C sites in mRNA called as NSUN2 targets revealed a slight preference for CpA and CpG and an abundance of Gs 2-4 bases downstream of the targeted cytosine. It is unclear whether these features are sufficient to mediate m^5C by NSUN2 or if there are additional structural or sequence requirements that were not detected in this thesis to mediate m^5C. It is also possible that other features besides the sequence and structure are required for selective methylation of RNA by NSUN2. One possible scenario is the presence of proteins or RNA that may help guide NSUN2 to mediate m^5C in transcripts.

These findings have extended the previously known substrate specificity of NSUN2, highlighting its multisite specificity. While the role of TRDMT1 in modifying ncRNA and mRNA are yet to be explored in detail, I have provided evidence in this thesis that NSUN2 is an important mRNA modifying enzyme, in addition to modifying tRNAs and other ncRNAs.

7.1.2 m^5C prior to processing of transcripts

In this thesis, I also examined methylation of tRNAs prior to end processing and splicing and have provided evidence of methylation occurring prior to both. I have provided evidence in this thesis of m^5C in $tRNA^{Leu(CAA)}$ occurring prior to splicing, confirming previous reports [100, 148]. In addition, the global tRNA mapping data was analysed to determine if m^5C can occur prior to end processing of tRNAs, with evidence that it does. To this end, it is not known whether end processing precedes splicing of tRNAs in humans, with conflicting evidence in other species. In *S. cerevisiae*, end processing generally occurs prior to splicing [269, 340, 341], however when grown under high temperatures some tRNAs undergo splicing first, suggesting that the order of processing in yeast is dependent on growth conditions [269]. Similarly, based on *Xenopus* oocyte injection experiments, the amount of pre-$tRNA^{Tyr}$ introduced into the nucleus influenced which process occurred first [270]. Injection of low amounts of pre-$tRNA^{Tyr}$ into the nucleus resulted in splicing occurring first, however in the case of high

amounts of pre-tRNA^{Tyr}, end processing preceded splicing. Thus, the findings in this thesis are consistent with both scenarios and indicate that m^{5}C in tRNA occurs early during biogenesis, most likely in the nucleolus, and suggests that this may be the case for other transcript biotypes (Figure 7.1). Although the bsRNA-seq library was not enriched for pre-mRNAs, candidate m^{5}C sites were detected in introns, suggestive of m^{5}C in mRNAs occurring prior to splicing. Similarly, the transcriptome-wide studies of m^{6}A also detected m^{6}A in introns [56, 57]. Thus, it is possible that many of the nucleoside modifications in RNA occur early during RNA biogenesis.

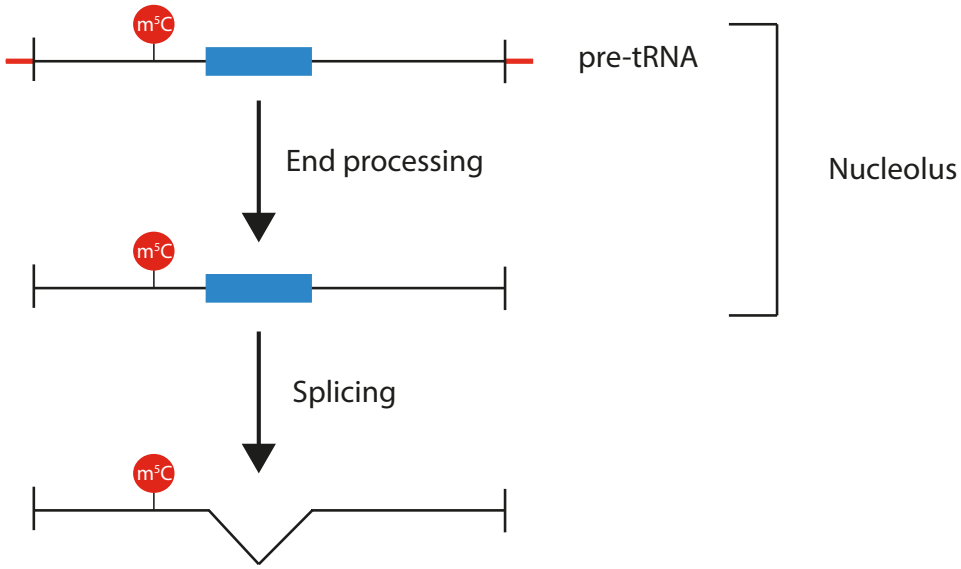


Figure 7.1. Schematic demonstrating m^{5}C in tRNA prior to end processing and splicing. Transcription and end processing of pre-tRNAs occurs in the nucleolus. As shown in this thesis, methylation of tRNAs can occur prior to end processing and splicing.

7.1.3 Differential RNA m^{5}C profiles between normal and cancer cells

Another major goal of this thesis was to examine the differences in the RNA m^{5}C profiles of PrEC and LNCaP cells and relate any findings of this back to the available epigenomic and transcriptomic data of these cell lines. I found a two-fold increase in NSUN2 in LNCaP cells compared to PrEC cells and hypothesised a subtle increase in either the number of m^{5}C sites or an increase in the stoichiometry of m^{5}C in LNCaP cells. In contrast to the number of candidate m^{5}C sites in common between PrEC and

LNCaP cells, only a small number of candidate m^5C sites were differentially methylated. Furthermore, examination of the sequencing reads revealed that a substantial number of candidate differential m^5C sites called were a result of low quality bases and this was confirmed by MiSeq amplicon sequencing. In addition, the stoichiometry of m^5C of candidate m^5C sites in common between PrEC and LNCaP cells were relatively similar across transcript biotypes, although vtRNA1-3 exhibited a significantly higher proportion of methylation in LNCaP cells compared to PrEC cells. As m^5C in vtRNAs has been reported to regulate its processing into svRNAs, this suggests that an increase the proportion of m^5C in vtRNA1-3 may increase its processing into smaller fragments to induce gene silencing in LNCaP cells. In turn, this may contribute to the differences in gene expression profiles between the two cell lines. With the exception of vtRNA1-3, these findings indicate that the two-fold increase in NSUN2 did not result in detectable hypermethylation in LNCaP cells and as a result the RNA m^5C profiles of PrEC and LNCaP cells were largely similar. It is possible that the higher level of NSUN2 in LNCaP cells may function elsewhere, given its broad biological role in various cellular processes. This could potentially be in mitosis where NSUN2 is involved in mitotic spindle stability independent of its methyltransferase activity [211], and where mitotic deregulation is a hallmark of cancer cells. NSUN2 is also phosphorylated at Ser139 by Aurora-B during disassembly of the nucleolus, suppressing its methyltransferase activity (see section 1.4.4) [208]. It is possible that the overexpressed NSUN2 in LNCaP cells may in fact be phosphorylated, resulting in little difference in the RNA m^5C profiles between PrEC and LNCaP cells. Thus, the overexpressed NSUN2 in LNCaP cells may function to stabilise the mitotic spindle necessary for fast-dividing cells such as cancer cells.

7.1.4 Similar m^5C patterns are present across multiple cell lines

The vast majority of the candidate m^5C sites detected in PrEC and LNCaP cells were in common. This motivated a comparative analysis between PrEC, LNCaP and HeLa cells. Using a RPM threshold between the two datasets generated in this thesis to eliminate lowly expressed transcripts that may be present in only one dataset, >70% candidate m^5C sites detected were in common between all three cell lines; >80% candidate m^5C

sites detected in tRNAs and other ncRNAs, and 46% of candidate m^5C sites in mRNAs were in common between all three cell lines. These patterns of methylation were confirmed in four tRNAs, two ncRNAs and several mRNAs as NSUN2 targets in HeLa cells and LNCaP cells. Interestingly, similar patterns of m^6A were reported between normal and cancer tissues as well as different stress conditions, although a subset of dynamically modulated sites were also detected across different conditions [57]. Thus, as with m^6A, the findings in this thesis demonstrate that RNA m^5C patterns are highly conserved across multiple human cell lines, further pointing towards a functional importance of m^5C not only in tRNA, but other ncRNA and mRNA.

7.1.5 Molecular function of m^5C in mRNA

One of the major goals of this thesis was to explore the molecular functions of m^5C in mRNA. Due to the limited detection of m^5C in mRNA in previous studies, this is the first study to investigate this in some detail. In this thesis, m^5C sites were significantly enriched in the 3' UTR, which is a hotspot for regulatory motifs, suggesting a role for m^5C in post-transcriptional gene control. Consistent with this finding, enrichment of Ago2 footprints was evident upstream of m^5C sites and a depletion of miRNA target sites was evident at m^5C sites, indicating that m^5C sites are enriched downstream of RISC and depleted at RISC binding sites in mRNA, suggesting regulatory interactions between m^5C and RISC. In agreement with a role in post-transcriptional gene regulation, a subtle decrease in the mean steady-state level of NSUN2-targeted mRNAs in the NSUN2 knockdown sample based on bsRNA-seq was evident, highlighting a subtle role for NSUN2-mediated m^5C sites in mRNA stability. Thus, this suggests antagonistic interactions between m^5C and Ago2 and miRNAs.

In addition, a correlation of m^5C with translation state was observed which was found to be context-dependent. The m^5C site in the CDS of *RPS3* resulted in a subtle decrease in methylation with translation state, indicating that m^5C in this transcript is negatively correlated with its translation. This m^5C site was 10 nt downstream of an Ago2 footprint, consistent with our global findings, suggesting that at least for this transcript m^5C may mediate binding of RISC to inhibit translation. In contrast, the proportion of m^5C in *RTN3* increased with translation state and in addition was not near any Ago2

footprints or putative miRNA binding sites, suggesting that m^5C in this transcript may interact with other RBPs to mediate translation. While we observed an enrichment of Ago2 footprints upstream of candidate m^5C sites, this does not preclude m^5C from interacting with other RBPs to mediate post-transcriptional gene regulation. In particular, secondary structures such as stem-loops within the 3' UTR can promote or block binding of RBPs to mediate gene expression. The finding in this thesis that m^5C is enriched in stem-loop structures in mRNAs suggests that this modification may regulate the binding of RBPs to the 3' UTR, in addition to Ago2.

The findings in this thesis have provided correlative links between m^5C and mRNA stability and translation, although the underlying mechanisms need to be explored in further detail. Furthermore, the molecular functions reported in this thesis for mRNA are consistent with those reported for rRNA [92], tRNA [73, 149, 150, 304] and vtRNA [78], where m^5C has been demonstrated to regulate stability and translation. Thus, there is an emerging picture that m^5C in RNA largely functions in regulating stability and translation (Figure 7.2).

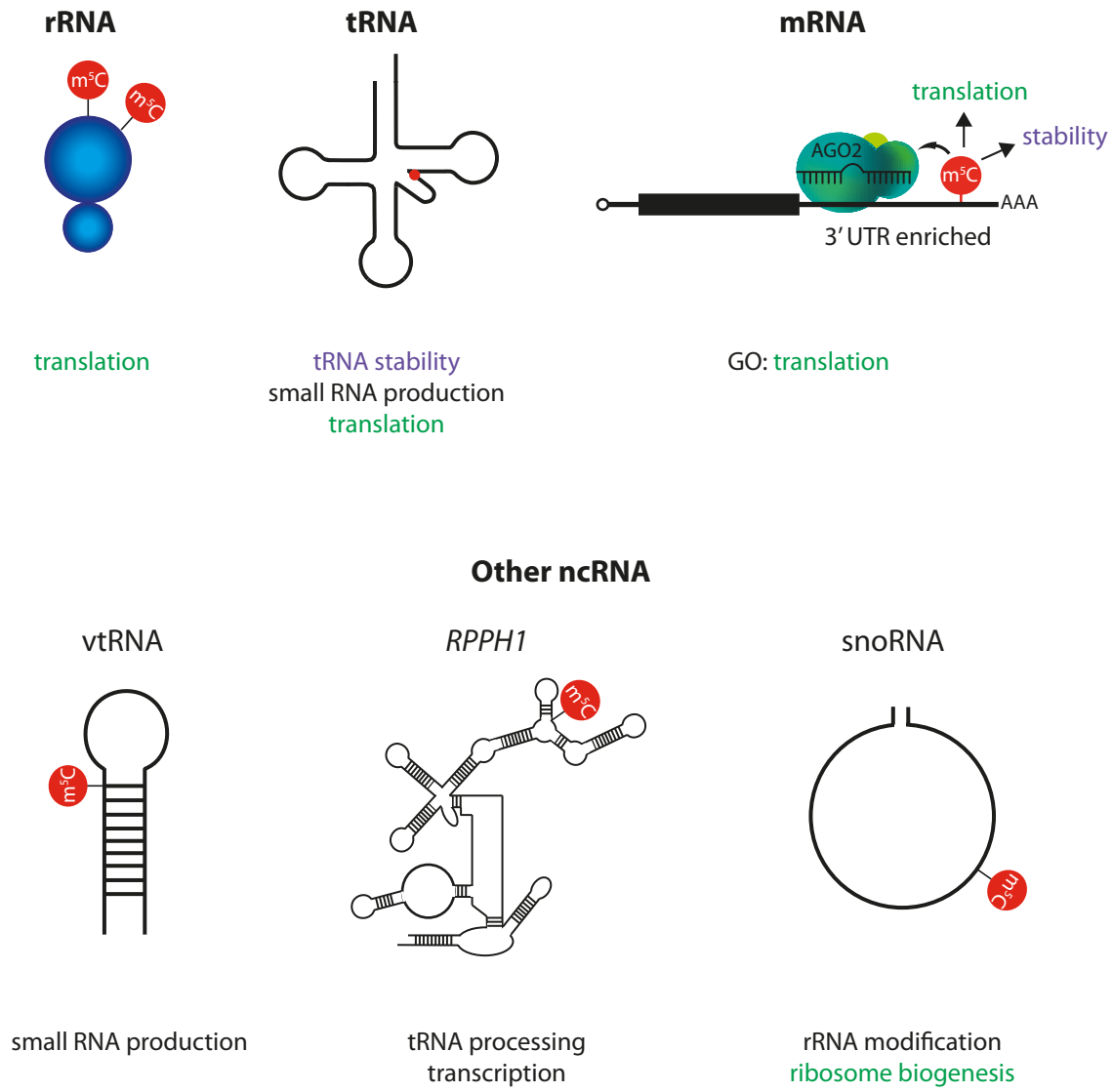


Figure 7.2. Confirmed and speculative molecular functions of m^{5}C in RNA.

Studies have provided evidence for the molecular functions of m^{5}C in rRNA [92], tRNA [73, 149, 150, 304] and vtRNA [78]. This thesis has provided evidence for a correlative link between m^{5}C and stability and translation. Molecular functions in *RPPH1* and snoRNA are speculative. Those molecular functions related to translation are indicated in green, while those related to stability are indicated in purple.

7.2 FUTURE DIRECTIONS

Following from this thesis, future work would focus more heavily on the substrate range and specificity of TRDMT1. A focus on TRDMT1 mediating m^5C in other transcript biotypes is lacking. The Aza-IP study has revealed a single m^5C site in the mRNA *KRT18* as a TRDMT1 target, but this was not validated [77]. This thesis has provided evidence of TRDMT1 mediating m^5C in a range of transcript biotypes and thus warrants validation and further investigation. In addition, this thesis has provided evidence of other RNA:m^5C methyltransferases mediating m^5C sites in rRNA, such as NSUN1, NSUN4 and NSUN5A, which are yet to be confirmed as active methyltransferases in humans. Exploring the substrate range of these enzymes, possibly through deletion cell lines, would substantially extend the known enzymology of m^5C in RNA.

In this thesis, I have provided several correlative links between m^5C in mRNA and post-transcriptional gene regulation. Extending what has been discovered in this thesis warrants further investigation. The steady-state level analyses performed in this thesis could not differentiate between changes in expression as a result of changes in the rate of transcription or stability of the mRNA. As the presence of m^5C in mRNA is not likely to affect the rate of transcription, investigating the role of m^5C specifically in mRNA stability is enticing. This could be achieved by knocking down *NSUN2* and *TRDMT1* (or use of deletion cell lines) followed by treatment with the transcription inhibitor, Actinomycin-D, over a time course. The extracted RNA can then be subjected to RNA-seq to investigate the stability of transcripts exhibiting m^5C sites mediated by NSUN2 or TRDMT1. The benefit of this approach is that this investigation is not just limited to mRNAs. Several studies have used a similar approach to investigate the role of m^6A in mRNA stability, which revealed an inverse correlation of m^6A with stability [105, 120]. Similarly, an extension of the candidate approach used in this thesis to explore the role of m^5C in translation should be investigated by coupling sucrose density gradient centrifugation of at least three biological replicates with RNA-seq, in addition to retaining the resolution of the polysome fractions collected to investigate gradual changes in the proportion of methylation with translation state. Both

transcriptome-wide approaches would provide a wealth of data for the role of m^5C in mRNA in stability and translation.

Another tantalising direction for the understanding of m^5C in RNA is the identification of m^5C-binding proteins, including writers, erasers and readers. Erasers and readers of m^5C have not yet been discovered, and there is the possibility of other writers mediating m^5C, as described above. FTO and ALKBH5 have been recently discovered as m^6A erasers, and FTO mediates m^6A removal by initially oxidising it to N6-hydroxymethyladenosine (hm^6A) and N6-formyladenosine (f^6A) [342]. Interestingly, a similar approach is employed by the TET hydroxylases for oxidation of m^5C to hm^5C and f^5C in DNA [262, 264, 265, 343], and a recent study has reported that TET hydroxylases can mediate hm^5C formation in RNA *in vitro* (see section 3.3.2.E) [263]. Therefore, the role of TET hydroxylases in oxidising and possibly removing m^5C in RNA warrants further investigation. The discovery of m^5C-binding proteins could be investigated by generating RNA baits containing m^5C and coupling this with mass spectrometry. A similar approach was employed for the discovery of the m^6A reader, YTHDF2 [57], which led to a subsequent study that has provided substantial evidence for a role of m^6A in mRNA destabilisation [120]. Thus, identification of m^5C readers using this procedure could also elucidate the molecular functions of m^5C in RNA.

Investigation into the possibility of proteins or RNAs guiding NSUN2 or TRDMT1 to target transcripts for methylation warrants further investigation in order to understand the mechanism of methylation. A possible experiment to identify guide proteins is immunoprecipitation of NSUN2 or TRDMT1 followed by mass spectrometry. Following identification of possible guide protein(s), the level of m^5C can be investigated at known target sites of NSUN2 or TRDMT1 with and without knockdown of the interacting protein to determine if it does function in guiding either methyltransferase. For the investigation of guide RNAs, small RNA-seq could be performed to enrich for small RNAs. Following this, sequence complementarity between small RNAs and the sequence surrounding m^5C sites targeted by NSUN2 or TRDMT1 could be investigated in a similar manner to the detection of snoRNAs and their targets. However, this would be quite a difficult experiment to perform, as it is not known what type of transcripts would help guide m^5C methylation.

Although this thesis suggests antagonistic interactions between m^5C and Ago2 and miRNAs, there is also the possibility of Ago/miRNAs recruiting NSUN2 to promote methylation. Co-immunoprecipitation experiments with NSUN2 and Ago could be used to investigate this further. Additionally, following from this and the work performed in this thesis, the relationship between m^5C sites and Ago could be studied in greater detail by coupling immunoprecipitation of Ago2-bound RNA with bisulfite sequencing to examine the levels of m^5C surrounding Ago2 binding sites. The RNA that was not immunoprecipitated by Ago2 can also be examined by bisulfite sequencing as a comparison. Furthermore, in addition to investigating the stability of NSUN2-targeted transcripts upon NSUN2 knockdown compared to the NTC as mentioned earlier in this chapter, this can be followed up by examining the stability of NSUN2-targeted transcripts that are also miRNA controlled.

Following from this, the biological role of m^5C in RNA should be further investigated. A possible approach is performing transcriptome-wide screens of m^5C in different tissues and comparing the RNA m^5C profiles in a similar way to this thesis. The understanding of the molecular functions of m^5C in RNA in addition to the RNA m^5C profiles provided from these screens could possibly shed light on the biological role of m^5C in RNA. In particular, investigating the RNA m^5C profiles from brains of patients that exhibited intellectual disabilities caused by mutations in *NSUN2* is an enticing direction. A recent study has reported a link between depletion of NSUN2-mediated m^5C sites in tRNAs with accumulation of 5' tRNA-derived fragments, decreased mRNA translation, activation of stress pathways and increased apoptosis of neurons, although m^5C in mRNA or other ncRNAs was not investigated [150]. Understanding the biological role of m^5C in RNA could provide further insight into their role in human diseases and may be a therapeutic target in the future.

7.3 Concluding remarks

The overall aims of this thesis were to improve and establish methods for the detection of m^5C in RNA, to determine the RNA targets of NSUN2 and TRDMT1 and the specificity of NSUN2, investigate the molecular functions of m^5C in mRNA and compare the RNA m^5C profiles across multiple cell lines, including normal and cancer

cells. This was the first study to perform these investigations and thus this thesis has achieved these aims.

CHAPTER EIGHT

APPENDIX

8.1 Sequences for the R-Luc *in vitro* transcripts

Humanised R-Luc *in vitro* transcript

TCCAAGGTGTACGACCCCGAGCAACGCAAACGCATGATCACTGGGCCTCAGTGGTGGGCTCGCT
GCAAGCAAATGAACGTGCTGGACTCCTTCATCAACTACTATGATTCCGAGAAGCACGCCGAGAA
CGCCGTGATTTTTCTGCATGGTAACGCTGCCTCCAGCTACCTGTGGAGGCACGTCGTGCCTCAC
ATCGAGCCCGTGGCTAGATGCATCATCCCTGATCTGATCGGAATGGGTAAGTCCGGCAAGAGCG
GGAATGGCTCATATCGCCTCCTGGATCACTACAAGTACCTCACCGCTTGGTTCGAGCTGCTGAA
CCTTCCAAAGAAAATCATCTTTGTGGGCCACGACTGGGGGGCTTGTCTGGCCTTTCACTACTCC
TACGAGCACCAAGACAAGATCAAGGCCATCGTCCATGCTGAGAGTGTCGTGGACGTGATCGAGT
CCTGGGACGAGTGGCCTGACATCGAGGAGGATATCGCCCTGATCAAGAGCGAAGAGGGCGAGAA
AATGGTGCTTGAGAATAACTTCTTCGTCGAGACCATGTCCCAAGCAAGATCATGCGGAAACTGG
AGCCTGAGGAGTTCGCTGCCTACCTGGAGCCATTCAAGGAGAAGGGCGAGGTTAGACGGCCTAC
CCTCTCCTGGCCTCGCGAGATCCCTCTCGTTAAGGGAGGCAAGCCCGACGTCGTCCAGATTGTC
CGCAACTACAACGCCTACCTTCGGGCCAGCGACGATCTGCCTAAGATGTTCATCGAGTCCGACC
CTGGGTTCTTTTCCAACGCTATTGTCGAGGGAGCTAAGAAGTTCCCTAACACCGAGTTCGTGAA
GGTGAAGGGCCTCCACTTCAGCCAGGAGGACGCTCCAGATGAAATGGGTAAGTACATCAAGAGC
TTCGTGGAGCGCGTGCTGAAGAACGAGCAGTAACGAGCAGTAA

Non-humanised R-Luc *in vitro* transcript

ATGATCCAGAACAAAGGAAACGGATGATAACTGGTCCGCAGTGGTGGGCCAGATGTAAACAAAT
GAATGTTCTTGATTCATTTATTAATTATTATGATTCAGAAAAACATGCAGAAAATGCTGTTATT
TTTTTACATGGTAACGCGGCCTCTTCTTATTTATGGCGACATGTTGTGCCACATATTGAGCCAG
TAGCGCGGTGTATTATACCAGACCTTATTGGTATGGGCAAATCAGGCAAATCTGGTAATGGTTC
TTATAGGTTACTTGATCATTACAAATATCTTACTGCATGGTTTGAACTTCTTAATTTACCAAAG
AAGATCATTTTTGTCGGCCATGATTGGGGTGCTTGTTTGGCATTTCATTATAGCTATGAGCATC
AAGATAAGATCAAAGCAATAGTTCACGCTGAAAGTGTAGTAGATGTGATTGAATCATGGGATGA
ATGGCCTGATATTGAAGAAGATATTGCGTTGATCAAATCTGAAGAAGGAGAAAAAATGGTTTTG
GAGAATAACTTCTTCGTGGAAACCATGTTGCCATCAAAAATCATGAGAAAGTTAGAACCAGAAG
AATTTGCAGCATATCTTGAACCATTCAAAGAGAAAGGTGAAGTTCGTCGTCCAACATTATCATG
GCCTCGTGAAATCCCGTTAGTAAAAGGTGGTAAACCTGACGTTGTACAAATTGTTAGGAATTAT
AATGCTTATCTACGTGCAAGTGATGATTTACCAAAAATGTTTATTGAATCGGACCCAGGATTCT
TTTCCAATGCTATTGTTGAAGGTGCCAAGAAGTTTCCTAATACTGAATTTGTCAAAGTAAAAGG
TCTTCATTTTTCGCAAGAAGATGCACCTGATGAAATGGGAAAATATATCAAATCGTTCGTTGAG
CGAGTTCTCAAAAATGAACAATAATTCTAGAGCGGCCGCTTCGAGCAGACATGATAAGATACAT
TGATGAGTTTGGACAAACCACAACTAGAATGCAGTGAAAAAAATGCTTTATTTGTGAAATTTGT
GATGCTATTGCTTTATTTGTAACCATTATAAGCTGCAATAAACAAGTTAACAACAACAATTGCA

TTCATTTTATGTTTCAGGTTCAGGGGGAGGTGTGGGAGGTTTTTTAAAGCAAGTAAAACCTCTA
CAAATGTGGTAAAATCGATAAGGATCCAAAAAAAAAAAAAAAAAAAAAA

Due to the large size, the following appendix files have been included digitally. Please refer to the attached CD.

8.2 R SCRIPT FOR STRUCTURAL MOTIF ANALYSES

The following .sh file details the shell script used for the structural analyses.

8.3 CANDIDATE m^5C SITES IN HELA CELLS, PREC AND LNCAP CELLS

The following Excel files list the candidate m^5C sites detected in HeLa cells, and PrEC and LNCaP cells.

HeLa: The following Excel files contain the full list of candidate m^5C sites detected in HeLa cells from the global and local mapping. Each file contains three sheets with a list of the total number of candidate m^5C sites (HeLa), candidate m^5C sites called as NSUN2 targets (NSUN2) and candidate m^5C sites called as TRDMT1 targets (TRDMT1).

Prostate: The following Excel files contain three sheets with the full list of candidate m^5C sites in common between PrEC and LNCaP cells (Prostate), candidate m^5C sites hypermethylated in PrEC cells (PrEC) and candidate m^5C sites hypermethylated in LNCaP cells (LNCaP) from the global and local mapping. On the following page are brief descriptions of the headings in these files.

genomic_location	genomic location of candidate m^5C site
reads	total number of reads
TCreads	number of reads that are T or C only
ratio	non-conversion %
genes	names of transcripts with candidate m^5C site
genomic_region	genomic region of candidate m^5C site
P.Value	p-value
Adj.P.Val	adjusted p-value (FDR)
struRNAs:	type of ncRNA
tRNA_type	tRNA isotype/isoacceptors

Additional local mapping headings

chr	coordinates of transcripts with candidate m^5C sites
chr_start/end	location of candidate m^5C site
desc	description of transcripts with candidate m^5C sites
genomic_coord	genomic location of transcripts with candidate m^5C sites

8.4 Global and local tRNA alignments

The following Excel files list the candidate m^5C sites in tRNAs based on the global and local mapping, whether they were called as NSUN2 or TRDMT1 targets, and the structural position of the candidate m^5C sites. These files were used to determine the tRNA structural positions, isotypes and isoacceptors containing candidate m^5C sites.

8.5 Rfam structural alignment of tRNAs exhibiting candidate m^5C sites from the local mapping

The following .pdf file displays the tRNA Rfam structural alignment generated from the local mapping and the location of candidate m^5C sites within the tRNAs.

8.6 MiSeq amplicon sequencing of transcripts exhibiting m^{5}C sites

The following .bed files list the coverage and methylation % of cytosines in segments of the transcripts subjected for m^{5}C site validation by MiSeq amplicon sequencing.

8.7 GSEA and GO analysis for candidate m^{5}C sites in HeLa cells

The following .txt files display the outcome of the GSEA. The following Excel files contain the full list of enriched GO terms of transcripts containing candidate m^{5}C sites. These analyses were performed for all candidate m^{5}C sites in mRNA in HeLa cells (HeLa), NSUN2-targeted mRNA (NSUN2) and TRDMT1-targeted mRNA (TRDMT1). There are three sheets per file showing biological processes (BP), molecular functions (MF) and cellular components (CC).

8.8 Complete list of housekeeping genes for RT-qPCR

The following Excel file contains the list of housekeeping genes from [309] that did not contain detectable levels of m^{5}C by bsRNA-seq.

8.9 Statistical analysis for RT-qPCR gene expression analysis

The following .rtf files display the output from GenStat for the analysis of RT-qPCR data in Chapter 5.

8.10 Overlap of candidate m^{5}C sites in PrEC, LNCaP and HeLa cells

The following Excel file displays the overlap of candidate m^{5}C sites in common between PrEC and LNCaP cells, and HeLa cells for each transcript biotype from 0.1-100 RPM.

REFERENCES

1. Sibbritt T., Patel H.R., and Preiss T., Mapping and significance of the mRNA methylome. *WIREs RNA*, 2013, **4**: 397-422.
2. Cantara W.A., Crain P.F., Rozenski J., McCloskey J.A., Harris K.A., et al., The RNA Modification Database, RNAMDB: 2011 update. *Nucleic Acids Res.*, 2011, **39**: D195-201.
3. Machnicka M.A., Milanowska K., Osman Oglou O., Purta E., Kurkowska M., et al., MODOMICS: a database of RNA modification pathways--2012 update. *Nucleic Acids Res.*, 2012, **41**: D262-267.
4. Belanger F., Stepinski J., Darzynkiewicz E., and Pelletier J., Characterization of hMTr1, a human Cap1 2'-O-ribose methyltransferase. *J. Biol. Chem.*, 2010, **285**: 33037-33044.
5. Werner M., Purta E., Kaminska K.H., Cymerman I.A., Campbell D.A., et al., 2'-O-ribose methylation of cap2 in human: function and evolution in a horizontally mobile family. *Nucleic Acids Res.*, 2011, **39**: 4756-4768.
6. Topisirovic I., Svitkin Y.V., Sonenberg N., and Shatkin A.J., Cap and cap-binding proteins in the control of gene expression. *WIREs RNA*, 2011, **2**: 277-298.
7. Gott J.M. and Emeson R.B., Functions and mechanisms of RNA editing. *Annu. Rev. Genet.*, 2000, **34**: 499-531.
8. Wulff B.E. and Nishikura K., Substitutional A-to-I RNA editing. *WIREs RNA*, 2010, **1**: 90-101.
9. Hall K.B. and McLaughlin L.W., Properties of a U1/mRNA 5' splice site duplex containing pseudouridine as measured by thermodynamic and NMR methods. *Biochemistry (Mosc.)*, 1991, **30**: 1795-1801.
10. Davis D.R. and Poulter C.D., 1H-15N NMR studies of Escherichia coli tRNA(Phe) from hisT mutants: a structural role for pseudouridine. *Biochemistry (Mosc.)*, 1991, **30**: 4223-4231.
11. Arnez J.G. and Steitz T.A., Crystal structure of unmodified tRNA(Gln) complexed with glutaminyl-tRNA synthetase and ATP suggests a possible role for pseudo-uridines in stabilization of RNA structure. *Biochemistry (Mosc.)*, 1994, **33**: 7560-7567.
12. Davis D.R., Veltri C.A., and Nielsen L., An RNA model system for investigation of pseudouridine stabilization of the codon-anticodon interaction in tRNALys, tRNAHis and tRNATyr. *J. Biomol. Struct. Dyn.*, 1998, **15**: 1121-1132.
13. Yarian C.S., Basti M.M., Cain R.J., Ansari G., Guenther R.H., et al., Structural and functional roles of the N1- and N3-protons of Ψ at tRNA’s position 39. *Nucleic Acids Res.*, 1999, **27**: 3543-3549.
14. Charette M. and Gray M.W., Pseudouridine in RNA: what, where, how, and why. *IUBMB Life*, 2000, **49**: 341-351.
15. Hudson G.A., Bloomingdale R.J., and Znosko B.M., Thermodynamic contribution and nearest-neighbor parameters of pseudouridine-adenosine base pairs in oligoribonucleotides. *RNA*, 2013, **19**: 1474-1482.
16. Fernandez I.S., Ng C.L., Kelley A.C., Wu G., Yu Y.T., et al., Unusual base pairing during the decoding of a stop codon by the ribosome. *Nature*, 2013, **500**: 107-110.

17. Karijolich J. and Yu Y.T., Converting nonsense codons into sense codons by targeted pseudouridylation. *Nature*, 2011, **474**: 395-398.
18. Dai Q., Fong R., Saikia M., Stephenson D., Yu Y.T., et al., Identification of recognition residues for ligation-based detection and quantitation of pseudouridine and N6-methyladenosine. *Nucleic Acids Res.*, 2007, **35**: 6322-6329.
19. Furuichi Y. and Shatkin A.J., Viral and cellular mRNA capping: past and prospects. *Adv. Virus Res.*, 2000, **55**: 135-184.
20. Desrosiers R., Friderici K., and Rottman F., Identification of methylated nucleosides in messenger RNA from Novikoff hepatoma cells. *Proc. Natl. Acad. Sci. U. S. A.*, 1974, **71**: 3971-3975.
21. Desrosiers R.C., Friderici K.H., and Rottman F.M., Characterization of Novikoff hepatoma mRNA methylation and heterogeneity in the methylated 5' terminus. *Biochemistry (Mosc.)*, 1975, **14**: 4367-4374.
22. Perry R.P., Kelley D.E., Friderici K., and Rottman F., The methylated constituents of L cell messenger RNA: evidence for an unusual cluster at the 5' terminus. *Cell*, 1975, **4**: 387-394.
23. Wei C.M., Gershowitz A., and Moss B., Methylated nucleotides block 5' terminus of HeLa cell messenger RNA. *Cell*, 1975, **4**: 379-386.
24. Wei C.M., Gershowitz A., and Moss B., 5'-Terminal and internal methylated nucleotide sequences in HeLa cell mRNA. *Biochemistry (Mosc.)*, 1976, **15**: 397-401.
25. Adams J.M. and Cory S., Modified nucleosides and bizarre 5'-termini in mouse myeloma mRNA. *Nature*, 1975, **255**: 28-33.
26. Dubin D.T. and Taylor R.H., The methylation state of poly A-containing messenger RNA from cultured hamster cells. *Nucleic Acids Res.*, 1975, **2**: 1653-1668.
27. Salditt-Georgieff M., Jelinek W., Darnell J.E., Furuichi Y., Morgan M., et al., Methyl labeling of HeLa cell hnRNA: a comparison with mRNA. *Cell*, 1976, **7**: 227-237.
28. Furuichi Y., Morgan M., Shatkin A.J., Jelinek W., Salditt-Georgieff M., et al., Methylated, blocked 5 termini in HeLa cell mRNA. *Proc. Natl. Acad. Sci. U. S. A.*, 1975, **72**: 1904-1908.
29. Lavi S. and Shatkin A.J., Methylated simian virus 40-specific RNA from nuclei and cytoplasm of infected BSC-1 cells. *Proc. Natl. Acad. Sci. U. S. A.*, 1975, **72**: 2012-2016.
30. Sommer S., Salditt-Georgieff M., Bachenheimer S., Darnell J.E., Furuichi Y., et al., The methylation of adenovirus-specific nuclear and cytoplasmic RNA. *Nucleic Acids Res.*, 1976, **3**: 749-765.
31. Wei C.M. and Moss B., Nucleotide sequences at the N6-methyladenosine sites of HeLa cell messenger ribonucleic acid. *Biochemistry (Mosc.)*, 1977, **16**: 1672-1676.
32. Nichols J.L., N6-methyladenosine in maize poly (A)-containing RNA. *Plant Science Letters*, 1979, **15**: 357-361.
33. Beemon K. and Keith J., Localization of N6-methyladenosine in the Rous sarcoma virus genome. *J. Mol. Biol.*, 1977, **113**: 165-179.
34. Levis R. and Penman S., 5'-terminal structures of poly(A)+ cytoplasmic messenger RNA and of poly(A)+ and poly(A)- heterogeneous nuclear RNA of cells of the dipteran Drosophila melanogaster. *J. Mol. Biol.*, 1978, **120**: 487-515.

35. Aloni Y., Dhar R., and Khoury G., Methylation of nuclear simian virus 40 RNAs. *J. Virol.*, 1979, **32**: 52-60.
36. Haugland R.A. and Cline M.G., Post-transcriptional modifications of oat coleoptile ribonucleic acids. 5'-Terminal capping and methylation of internal nucleosides in poly(A)-rich RNA. *Eur. J. Biochem.*, 1980, **104**: 271-277.
37. Kennedy T.D. and Lane B.G., Wheat embryo ribonucleates. XIII. Methyl-substituted nucleoside constituents and 5'-terminal dinucleotide sequences in bulk poly(AR)-rich RNA from imbibing wheat embryos. *Can. J. Biochem.*, 1979, **57**: 927-931.
38. Tuck M.T., The formation of internal 6-methyladenine residues in eucaryotic messenger RNA. *Int. J. Biochem.*, 1992, **24**: 379-386.
39. Schibler U. and Perry R.P., Characterization of the 5' termini of hn RNA in mouse L cells: implications for processing and cap formation. *Cell*, 1976, **9**: 121-130.
40. Bokar J.A., Shambaugh M.E., Polayes D., Matera A.G., and Rottman F.M., Purification and cDNA cloning of the AdoMet-binding subunit of the human mRNA (N6-adenosine)-methyltransferase. *RNA*, 1997, **3**: 1233-1247.
41. Zhong S., Li H., Bodi Z., Button J., Vespa L., et al., MTA is an Arabidopsis messenger RNA adenosine methylase and interacts with a homolog of a sex-specific splicing factor. *Plant Cell*, 2008, **20**: 1278-1288.
42. Clancy M.J., Shambaugh M.E., Timpte C.S., and Bokar J.A., Induction of sporulation in Saccharomyces cerevisiae leads to the formation of N6-methyladenosine in mRNA: a potential mechanism for the activity of the IME4 gene. *Nucleic Acids Res.*, 2002, **30**: 4509-4518.
43. Schibler U., Kelley D.E., and Perry R.P., Comparison of methylated sequences in messenger RNA and heterogeneous nuclear RNA from mouse L cells. *J. Mol. Biol.*, 1977, **115**: 695-714.
44. Bodi Z., Button J.D., Grierson D., and Fray R.G., Yeast targets for mRNA methylation. *Nucleic Acids Res.*, 2010, **38**: 5327-5335.
45. Horowitz S., Horowitz A., Nilsen T.W., Munns T.W., and Rottman F.M., Mapping of N6-methyladenosine residues in bovine prolactin mRNA. *Proc. Natl. Acad. Sci. U. S. A.*, 1984, **81**: 5667-5671.
46. Kane S.E. and Beemon K., Precise localization of m6A in Rous sarcoma virus RNA reveals clustering of methylation sites: implications for RNA processing. *Mol. Cell. Biol.*, 1985, **5**: 2298-2306.
47. Dubin D.T., Stollar V., HsuChen C.C., Timko K., and Guild G.M., Sindbis virus messenger RNA: the 5'-termini and methylated residues of 26 and 42 S RNA. *Virology*, 1977, **77**: 457-470.
48. Dubin D.T. and Stollar V., Methylation of Sindbis virus "26S" messenger RNA. *Biochem. Biophys. Res. Commun.*, 1975, **66**: 1373-1379.
49. Paul M.S. and Bass B.L., Inosine exists in mRNA at tissue-specific levels and is most abundant in brain mRNA. *The EMBO journal*, 1998, **17**: 1120-1127.
50. Squires J.E. and Preiss T., Function and detection of 5-methylcytosine in eukaryotic RNA. *Epigenomics*, 2010, **2**: 709-715.
51. Motorin Y., Lyko F., and Helm M., 5-methylcytosine in RNA: detection, enzymatic formation and biological functions. *Nucleic Acids Res.*, 2010, **38**: 1415-1430.
52. Motorin Y. and Helm M., RNA nucleotide methylation. *WIREs RNA*, 2011, **2**: 611-631.

53. Carlile T.M., Rojas-Duran M.F., Zinshteyn B., Shin H., Bartoli K.M., et al., Pseudouridine profiling reveals regulated mRNA pseudouridylation in yeast and human cells. *Nature*, 2014, **515**: 143-146.
54. Schwartz S., Bernstein D.A., Mumbach M.R., Jovanovic M., Herbst R.H., et al., Transcriptome-wide mapping reveals widespread dynamic-regulated pseudouridylation of ncRNA and mRNA. *Cell*, 2014, **159**: 148-162.
55. Lovejoy A.F., Riordan D.P., and Brown P.O., Transcriptome-wide mapping of pseudouridines: pseudouridine synthases modify specific mRNAs in S. cerevisiae. *PLoS ONE*, 2014, **9**: e110799.
56. Meyer K.D., Saletore Y., Zumbo P., Elemento O., Mason C.E., et al., Comprehensive analysis of mRNA methylation reveals enrichment in 3' UTRs and near stop codons. *Cell*, 2012, **149**: 1635-1646.
57. Dominissini D., Moshitch-Moshkovitz S., Schwartz S., Salmon-Divon M., Ungar L., et al., Topology of the human and mouse m6A RNA methylomes revealed by m6A-seq. *Nature*, 20121-8.
58. Squires J.E., Patel H.R., Nousch M., Sibbritt T., Humphreys D.T., et al., Widespread occurrence of 5-methylcytosine in human coding and non-coding RNA. *Nucleic Acids Res.*, 2012, **40**: 5023-5033.
59. Schaefer M., Pollex T., Hanna K., and Lyko F., RNA cytosine methylation analysis by bisulfite sequencing. *Nucleic Acids Res.*, 2009, **37**: e12.
60. Edmonds M., A history of poly A sequences: from formation to factors to function. *Prog. Nucleic Acid Res. Mol. Biol.*, 2002, **71**: 285-389.
61. Bibikova M. and Fan J.-B., Genome-wide DNA methylation profiling. *WIREs Syst. Biol. Med.*, 2010, **2**: 210-223.
62. Harris R.A., Wang T., Coarfa C., Nagarajan R.P., Hong C., et al., Comparison of sequencing-based methods to profile DNA methylation and identification of monoallelic epigenetic modifications. *Nat. Biotechnol.*, 2010, **28**: 1097-1105.
63. Bakin A. and Ofengand J., Four newly located pseudouridylate residues in Escherichia coli 23S ribosomal RNA are all at the peptidyltransferase center: analysis by the application of a new sequencing technique. *Biochemistry (Mosc.)*, 1993, **32**: 9754-9762.
64. Bakin A. and Ofengand J., Mapping of the 13 pseudouridine residues in Saccharomyces cerevisiae small subunit ribosomal RNA to nucleotide resolution. *Nucleic Acids Res.*, 1995, **23**: 3290-3294.
65. Bakin A.V. and Ofengand J., Mapping of Pseudouridine Residues in RNA to Nucleotide Resolution. *Methods Mol. Biol.*, 1998, **77**: 297-309.
66. Saletore Y., Meyer K., Korlach J., Vilfan I.D., Jaffrey S., et al., The birth of the Epitranscriptome: deciphering the function of RNA modifications. *Genome Biol.*, 2012, **13**: 175.
67. Schwartz S., Agarwala S.D., Mumbach M.R., Jovanovic M., Mertins P., et al., High-resolution mapping reveals a conserved, widespread, dynamic mRNA methylation program in yeast meiosis. *Cell*, 2013, **155**: 1409-1421.
68. Liu J., Yue Y., Han D., Wang X., Fu Y., et al., A METTL3-METTL14 complex mediates mammalian nuclear RNA N6-adenosine methylation. *Nat. Chem. Biol.*, 2014, **10**: 93-95.
69. Ping X.L., Sun B.F., Wang L., Xiao W., Yang X., et al., Mammalian WTAP is a regulatory subunit of the RNA N6-methyladenosine methyltransferase. *Cell Res.*, 2014, **24**: 177-189.

70. Frommer M., McDonald L.E., Millar D.S., Collis C.M., Watt F., et al., A genomic sequencing protocol that yields a positive display of 5-methylcytosine residues in individual DNA strands. *Proc. Natl. Acad. Sci. U. S. A.*, 1992, **89**: 1827-1831.
71. Clark S.J., Statham A., Stirzaker C., Molloy P.L., and Frommer M., DNA methylation: bisulphite modification and analysis. *Nat. Protoc.*, 2006, **1**: 2353-2364.
72. Schaefer M., Hagemann S., Hanna K., and Lyko F., Azacytidine inhibits RNA methylation at DNMT2 target sites in human cancer cell lines. *Cancer Res.*, 2009, **69**: 8127-8132.
73. Schaefer M., Pollex T., Hanna K., Tuorto F., Meusburger M., et al., RNA methylation by Dnmt2 protects transfer RNAs against stress-induced cleavage. *Genes Dev.*, 2010, **24**: 1590-1595.
74. Gu W., Hurto R.L., Hopper A.K., Grayhack E.J., and Phizicky E.M., Depletion of Saccharomyces cerevisiae tRNA(His) guanylyltransferase Thg1p leads to uncharged tRNAHis with additional m(5)C. *Mol. Cell. Biol.*, 2005, **25**: 8191-8201.
75. Ondov B.D., Cochran C., Landers M., Meredith G.D., Dudas M., et al., An alignment algorithm for bisulfite sequencing using the Applied Biosystems SOLiD System. *Bioinformatics*, 2010, **26**: 1901-1902.
76. Edelheit S., Schwartz S., Mumbach M.R., Wurtzel O., and Sorek R., Transcriptome-wide mapping of 5-methylcytidine RNA modifications in bacteria, archaea, and yeast reveals m5C within archaeal mRNAs. *PLoS Genet.*, 2013, **9**: e1003602.
77. Khoddami V. and Cairns B.R., Identification of direct targets and modified bases of RNA cytosine methyltransferases. *Nat. Biotechnol.*, 2013, **31**: 458-464.
78. Hussain S., Sajini A.A., Blanco S., Dietmann S., Lombard P., et al., NSun2-mediated cytosine-5 methylation of vault noncoding RNA determines its processing into regulatory small RNAs. *Cell reports*, 2013, **4**: 255-261.
79. Petrossian T.C. and Clarke S.G., Uncovering the human methyltransferasome. *Molecular Cellular Proteomics*, 2011, **10**: M110.000976.
80. Wlodarski T., Kutner J., Towpik J., Knizewski L., Rychlewski L., et al., Comprehensive structural and substrate specificity classification of the Saccharomyces cerevisiae methyltransferome. *PLoS ONE*, 2011, **6**: e23168.
81. Kozbial P.Z. and Mushegian A.R., Natural history of S-adenosylmethionine-binding proteins. *BMC Struct. Biol.*, 2005, **5**: 19.
82. Bujnicki J.M., Feder M., Ayres C.L., and Redman K.L., Sequence-structure-function studies of tRNA:m5C methyltransferase Trm4p and its relationship to DNA:m5C and RNA:m5U methyltransferases. *Nucleic Acids Res.*, 2004, **32**: 2453-2463.
83. Young B.D., Weiss D.I., Zurita-Lopez C.I., Webb K.J., Clarke S.G., et al., Identification of methylated proteins in the yeast small ribosomal subunit: a role for SPOUT methyltransferases in protein arginine methylation. *Biochemistry (Mosc.)*, 2012, **51**: 5091-5104.
84. Dillon S.C., Zhang X., Trievel R.C., and Cheng X., The SET-domain protein superfamily: protein lysine methyltransferases. *Genome Biol.*, 2005, **6**: 227.
85. Czerwoniec A., Kasprzak J.M., Kaminska K.H., Rother K., Purta E., et al., *Folds and Functions of Domains in RNA Modification Enzymes*, in *DNA and*

RNA Modification Enzymes: Structure, Mechanism, Function and Evolution, H. Grosjean, Editor. 2009, Landes Bioscience: Austin, Texas USA. p. 289-302.
86. Pavlopoulou A. and Kossida S., Phylogenetic analysis of the eukaryotic RNA (cytosine-5)-methyltransferases. *Genomics*, 2009, **93**: 350-357.
87. Bujnicki J.M., Feder M., Radlinska M., and Blumenthal R.M., Structure prediction and phylogenetic analysis of a functionally diverse family of proteins homologous to the MT-A70 subunit of the human mRNA:m(6)A methyltransferase. *J. Mol. Evol.*, 2002, **55**: 431-444.
88. Sunita S., Tkaczuk K.L., Purta E., Kasprzak J.M., Douthwaite S., et al., Crystal structure of the Escherichia coli 23S rRNA:m5C methyltransferase RlmI (YccW) reveals evolutionary links between RNA modification enzymes. *J. Mol. Biol.*, 2008, **383**: 652-666.
89. Reid R., Greene P.J., and Santi D.V., Exposition of a family of RNA m(5)C methyltransferases from searching genomic and proteomic sequences. *Nucleic Acids Res.*, 1999, **27**: 3138-3145.
90. Gustafson W.C., Taylor C.W., Valdez B.C., Henning D., Phippard A., et al., Nucleolar protein p120 contains an arginine-rich domain that binds to ribosomal RNA. *Biochem. J.*, 1998, **331 (Pt 2)**: 387-393.
91. Hong B., Brockenbrough J.S., Wu P., and Aris J.P., Nop2p is required for pre-rRNA processing and 60S ribosome subunit synthesis in yeast. *Mol. Cell. Biol.*, 1997, **17**: 378-388.
92. Sharma S., Yang J., Watzinger P., Kotter P., and Entian K.D., Yeast Nop2 and Rcm1 methylate C2870 and C2278 of the 25S rRNA, respectively. *Nucleic Acids Res.*, 2013, **41**: 9062-9076.
93. Zanchin N.I.T., Roberts P., DeSilva A., Sherman F., and Goldfarb D.S., Saccharomyces cerevisiae Nip7p Is Required for Efficient 60S Ribosome Subunit Biogenesis. *Mol. Cell. Biol.*, 1997, **17**: 5001-5015.
94. Zanchin N.I.T. and Goldfarb D.S., Nip7p Interacts with Nop8p, an Essential Nucleolar Protein Required for 60S Ribosome Biogenesis, and the Exosome Subunit Rrp43p. *Mol. Cell. Biol.*, 1999, **19**: 1518-1525.
95. Camara Y., Asin-Cayuela J., Park C.B., Metodiev M.D., Shi Y., et al., MTERF4 regulates translation by targeting the methyltransferase NSUN4 to the mammalian mitochondrial ribosome. *Cell Metab.*, 2011, **13**: 527-539.
96. Santos M.C., Goldfeder M.B., Zanchin N.I., and Oliveira C.C., The essential nucleolar yeast protein Nop8p controls the exosome function during 60S ribosomal subunit maturation. *PLoS ONE*, 2011, **6**: e21686.
97. Metodiev M.D., Spahr H., Loguercio Polosa P., Meharg C., Becker C., et al., NSUN4 is a dual function mitochondrial protein required for both methylation of 12S rRNA and coordination of mitoribosomal assembly. *PLoS Genet.*, 2014, **10**: e1004110.
98. Motorin Y. and Grosjean H., Multisite-specific tRNA-m5C-methyltransferase (Trm4) in yeast Saccharomyces cerevisiae- identification of the gene and substrate specificity of the enzyme. *RNA*, 1999, **5**: 1105-1118.
99. Frye M. and Watt F.M., The RNA methyltransferase Misu (NSun2) mediates Myc-induced proliferation and is upregulated in tumors. *Curr. Biol.*, 2006, **16**: 971-981.
100. Brzezicha B., Schmidt M., Makalowska I., Jarmolowski A., Pienkowska J., et al., Identification of human tRNA:m5C methyltransferase catalysing intron-

dependent m5C formation in the first position of the anticodon of the pre-tRNA Leu (CAA). *Nucleic Acids Res.*, 2006, **34**: 6034-6043.
101. Forterre P. and Grosjean H., *The Interplay between RNA and DNA Modifications: Back to the RNA World*, in *DNA and RNA Modification Enzymes: Structure, Mechanism, Function and Evolution*, H. Grosjean, Editor. 2009, Landes Bioscience: Austin, Texas USA. p. 259-274.
102. Bokar J.A., Rath-Shambaugh M.E., Ludwiczak R., Narayan P., and Rottman F., Characterization and partial purification of mRNA N6-adenosine methyltransferase from HeLa cell nuclei. *J. Biol. Chem.*, 1994, **262**: 17696-17704.
103. Narayan P. and Rottman F.M., An in vitro system for accurate methylation of internal adenosine residues in messenger RNA. *Science*, 1988, **242**: 1159-1162.
104. Agarwala S.D., Blitzblau H.G., Hochwagen A., and Fink G.R., RNA methylation by the MIS complex regulates a cell fate decision in yeast. *PLoS Genet.*, 2012, **8**: e1002732.
105. Wang Y., Li Y., Toth J.I., Petroski M.D., Zhang Z., et al., N6-methyladenosine modification destabilizes developmental regulators in embryonic stem cells. *Nat. Cell Biol.*, 2014, **16**: 191-198.
106. Bujnicki J.M., Sequence permutations in the molecular evolution of DNA methyltransferases. *BMC Evol. Biol.*, 2002, **2**: 3.
107. Schwartz S., Mumbach M.R., Jovanovic M., Wang T., Maciag K., et al., Perturbation of m6A writers reveals two distinct classes of mRNA methylation at internal and 5' sites. *Cell reports*, 2014, **8**: 284-296.
108. Harper J.E., Miceli S.M., Roberts R.J., and Manley J.L., Sequence specificity of the human mRNA N6-adenosine methylase in vitro. *Nucleic Acids Res.*, 1990, **18**: 5735-5741.
109. Nichols J.L. and Welder L., Nucleotides adjacent to N6-methyladenosine in maize poly(A)-containing RNA. *Plant Science Letters*, 1981, **21**: 75-81.
110. Shimba S., Bokar J.A., Rottman F., and Reddy R., Accurate and efficient N-6-adenosine methylation in spliceosomal U6 small nuclear RNA by HeLa cell extract in vitro. *Nucleic Acids Res.*, 1995, **23**: 2421-2426.
111. Kane S.E. and Beemon K., Inhibition of methylation at two internal N6-methyladenosine sites caused by GAC to GAU mutations. *J. Biol. Chem.*, 1987, **262**: 3422-3427.
112. Csepany T., Lin A., Baldick C.J., and Beemon K., Sequence specificity of mRNA N6-adenosine methyltransferase. *J. Biol. Chem.*, 1990, **265**: 20117-20122.
113. Bringmann P. and Luhrmann R., Antibodies specific for N6-methyladenosine react with intact snRNPs U2 and U4/U6. *FEBS Lett.*, 1987, **213**: 309-315.
114. Bodi Z., Zhong S., Mehra S., Song J., Graham N., et al., Adenosine Methylation in Arabidopsis mRNA is Associated with the 3' End and Reduced Levels Cause Developmental Defects. *Front. Plant Sci.*, 2012, **3**: 48.
115. Stoilov P., Rafalska I., and Stamm S., YTH: a new domain in nuclear proteins. *Trends Biochem. Sci.*, 2002, **27**: 495-497.
116. Zhang Z., Theler D., Kaminska K.H., Hiller M., de la Grange P., et al., The YTH domain is a novel RNA binding domain. *J. Biol. Chem.*, 2010, **285**: 14701-14710.
117. Hartmann A.M., Nayler O., Schwaiger F.W., Obermeier A., and Stamm S., The Interaction and Colocalization of Sam68 with the Splicing-associated Factor

YT521-B in Nuclear Dots Is Regulated by the Src Family Kinase p59fyn. *Mol. Biol. Cell*, 1999, **10**: 3909-3926.
118. Harigaya Y., Tanaka H., Yamanaka S., Tanaka K., Watanabe Y., et al., Selective elimination of messenger RNA prevents an incidence of untimely meiosis. *Nature*, 2006, **442**: 45-50.
119. McPheeters D.S., Cremona N., Sunder S., Chen H.M., Averbeck N., et al., A complex gene regulatory mechanism that operates at the nexus of multiple RNA processing decisions. *Nat. Struct. Mol. Biol.*, 2009, **16**: 255-264.
120. Wang X., Lu Z., Gomez A., Hon G.C., Yue Y., et al., N6-methyladenosine-dependent regulation of messenger RNA stability. *Nature*, 2014, **505**: 117-120.
121. Reijns M.A., Alexander R.D., Spiller M.P., and Beggs J.D., A role for Q/N-rich aggregation-prone regions in P-body localization. *J. Cell Sci.*, 2008, **121**: 2463-2472.
122. Li F., Zhao D., Wu J., and Shi Y., Structure of the YTH domain of human YTHDF2 in complex with an m(6)A mononucleotide reveals an aromatic cage for m(6)A recognition. *Cell Res.*, 2014, **24**: 1490-1492.
123. Zhu T., Roundtree I.A., Wang P., Wang X., Wang L., et al., Crystal structure of the YTH domain of YTHDF2 reveals mechanism for recognition of N6-methyladenosine. *Cell Res.*, 2014, **24**: 1493-1496.
124. Jia G., Fu Y., Zhao X., Dai Q., Zheng G., et al., N6-methyladenosine in nuclear RNA is a major substrate of the obesity-associated FTO. *Nat. Chem. Biol.*, 2011, **7**: 885-887.
125. Zheng G., Dahl J.A., Niu Y., Fedorcsak P., Huang C.-M., et al., ALKBH5 Is a Mammalian RNA Demethylase that Impacts RNA Metabolism and Mouse Fertility. *Mol. Cell*, 2013, **49**: 18-29.
126. Peters T., Ausmeier K., and Ruther U., Cloning of Fatso (Fto), a novel gene deleted by the Fused toes (Ft) mouse mutation. *Mamm. Genome*, 1999, **10**: 983-986.
127. Dina C., Meyre D., Gallina S., Durand E., Körner A., et al., Variation in FTO contributes to childhood obesity and severe adult obesity. *Nat. Genet.*, 2007, **39**: 724-726.
128. Frayling T.M., Genome-wide association studies provide new insights into type 2 diabetes aetiology. *Nat. Rev. Genet.*, 2007, **8**: 657-662.
129. Frayling T.M., Timpson N.J., Weedon M.N., Zeggini E., Freathy R.M., et al., A common variant in the FTO gene is associated with body mass index and predisposes to childhood and adult obesity. *Science*, 2007, **316**: 889-894.
130. Hinney A., Nguyen T.T., Scherag A., Friedel S., Brönner G., et al., Genome wide association (GWA) study for early onset extreme obesity supports the role of fat mass and obesity associated gene (FTO) variants. *PLoS ONE*, 2007, **2**: e1361.
131. Scuteri A., Sanna S., Chen W.-M., Uda M., Albai G., et al., Genome-wide association scan shows genetic variants in the FTO gene are associated with obesity-related traits. *PLoS Genet.*, 2007, **3**: e115.
132. Gerken T., Girard C.A., Tung Y.-C.L., Webby C.J., Saudek V., et al., The obesity-associated FTO gene encodes a 2-oxoglutarate-dependent nucleic acid demethylase. *Science*, 2007, **318**: 1469-1472.
133. Sanchez-Pulido L. and Andrade-Navarro M.A., The FTO (fat mass and obesity associated) gene codes for a novel member of the non-heme dioxygenase superfamily. *BMC biochemistry*, 2007, **8**: 23.

134. Kurowski M.A., Bhagwat A.S., Papaj G., and Bujnicki J.M., Phylogenomic identification of five new human homologs of the DNA repair enzyme AlkB. *BMC Genomics*, 2003, **4**: 48.
135. Trewick S.C., Henshaw T.F., Hausinger R.P., Lindahl T., and Sedgwick B., Oxidative demethylation by Escherichia coli AlkB directly reverts DNA base damage. *Nature*, 2002, **419**: 174-178.
136. Han Z., Niu T., Chang J., Lei X., Zhao M., et al., Crystal structure of the FTO protein reveals basis for its substrate specificity. *Nature*, 2010, **464**: 1205-1209.
137. Larder R., Cheung M.K.M., Tung Y.C.L., Yeo G.S.H., and Coll A.P., Where to go with FTO? *Trends Endocrinol. Metabol.*, 2011, **22**: 53-59.
138. Jia G., Yang C.-G., Yang S., Jian X., Yi C., et al., Oxidative demethylation of 3-methylthymine and 3-methyluracil in single-stranded DNA and RNA by mouse and human FTO. *FEBS Lett.*, 2008, **582**: 3313-3319.
139. Baltz A.G., Munschauer M., Schwanhäusser B., Vasile A., Murakawa Y., et al., The mRNA-Bound Proteome and Its Global Occupancy Profile on Protein-Coding Transcripts. *Mol. Cell*, 2012, **46**: 674-690.
140. Castelló A., Fischer B., Eichelbaum K., Horos R., Beckmann B.M., et al., Insights into RNA biology from an atlas of mammalian mRNA-binding proteins. *Cell*, 2012, **149**: 1393-1406.
141. Robbens S., Rouzé P., Cock J.M., Spring J., Worden A.Z., et al., The FTO gene, implicated in human obesity, is found only in vertebrates and marine algae. *J. Mol. Evol.*, 2008, **66**: 80-84.
142. Wu P., Brockenbrough J.S., Paddy M.R., and Aris J.P., NCL1, a novel gene for a non-essential nuclear protein in Saccharomyces cerevisiae. *Gene*, 1998, **220**: 109-117.
143. Walbott H., Auxilien S., Grosjean H., and Golinelli-Pimpaneau B., The carboxyl-terminal extension of yeast tRNA m5C methyltransferase enhances the catalytic efficiency of the amino-terminal domain. *J. Biol. Chem.*, 2007, **282**: 23663-23671.
144. Walbott H., Husson C., Auxilien S., and Golinelli-Pimpaneau B., Cysteine of sequence motif VI is essential for nucleophilic catalysis by yeast tRNA m5C methyltransferase. *RNA*, 2007, **13**: 967-973.
145. King M.Y. and Redman K.L., RNA methyltransferases utilize two cysteine residues in the formation of 5-methylcytosine. *Biochemistry (Mosc.)*, 2002, **41**: 11218-11225.
146. Redman K.L., Assembly of protein-RNA complexes using natural RNA and mutant forms of an RNA cytosine methyltransferase. *Biomacromolecules*, 2006, **7**: 3321-3326.
147. Martinez F.J., Lee J.H., Lee J.E., Blanco S., Nickerson E., et al., Whole exome sequencing identifies a splicing mutation in NSUN2 as a cause of a Dubowitz-like syndrome. *J. Med. Genet.*, 2012, **49**: 380-385.
148. Auxilien S., Guérineau V., Szweykowska-Kulinska Z., and Golinelli-Pimpaneau B., The Human tRNA m5C methyltransferase Misu is multisite-specific. *RNA Biol.*, 2012, **9**: 1331-1338.
149. Tuorto F., Liebers R., Musch T., Schaefer M., Hofmann S., et al., RNA cytosine methylation by Dnmt2 and NSun2 promotes tRNA stability and protein synthesis. *Nat. Struct. Mol. Biol.*, 2012, **19**: 900-905.

150. Blanco S., Dietmann S., Flores J.V., Hussain S., Kutter C., et al., Aberrant methylation of tRNAs links cellular stress to neuro-developmental disorders. *EMBO J.*, 2014, **33**: 2020-2039.
151. Zhang X., Liu Z., Yi J., Tang H., Xing J., et al., The tRNA methyltransferase NSun2 stabilizes p16(INK4) mRNA by methylating the 3'-untranslated region of p16. *Nat. Commun.*, 2012, **3**: 712.
152. Sugimoto Y., König J., Hussain S., Zupan B., Curk T., et al., Analysis of CLIP and iCLIP methods for nucleotide-resolution studies of protein-RNA interactions. *Genome Biol.*, 2012, **13**: R67.
153. Wilkinson C.R.M., Bartlett R., Nurse P., and Bird A.P., The fission yeast gene pmt1+ encodes a DNA methyltransferase homologue. *Nucleic Acids Res.*, 1995, **23**: 203-210.
154. Van den Wyngaert I., Sprengel J., Kass S.U., and Luyten W.H., Cloning and analysis of a novel human putative DNA methyltransferase. *FEBS Lett.*, 1998, **426**: 283-289.
155. Yoder J.A. and Bestor T.H., A candidate mammalian DNA methyltransferase related to pmt1p of fission yeast. *Hum. Mol. Genet.*, 1998, **7**: 279-284.
156. Jurkowski T.P. and Jeltsch A., On the evolutionary origin of eukaryotic DNA methyltransferases and Dnmt2. *PLoS ONE*, 2011, **6**: e28104.
157. Jeltsch A., Nellen W., and Lyko F., Two substrates are better than one: dual specificities for Dnmt2 methyltransferases. *Trends Biochem. Sci.*, 2006, **31**: 306-308.
158. Schaefer M. and Lyko F., Solving the Dnmt2 enigma. *Chromosoma*, 2010, **119**: 35-40.
159. Goll M.G., Kirpekar F., Maggert K.A., Yoder J.A., Hsieh C.L., et al., Methylation of tRNAAsp by the DNA methyltransferase homolog Dnmt2. *Science*, 2006, **311**: 395-398.
160. Jurkowski T.P., Meusburger M., Phalke S., Helm M., Nellen W., et al., Human DNMT2 methylates tRNA(Asp) molecules using a DNA methyltransferase-like catalytic mechanism. *RNA*, 2008, **14**: 1663-1670.
161. Tovy A., Siman Tov R., Gaentzsch R., Helm M., and Ankri S., A new nuclear function of the Entamoeba histolytica glycolytic enzyme enolase: the metabolic regulation of cytosine-5 methyltransferase 2 (Dnmt2) activity. *PLoS Pathog.*, 2010, **6**: e1000775.
162. Dong A., Yoder J.A., Zhang X., Zhou L., Bestor T.H., et al., Structure of human DNMT2, an enigmatic DNA methyltransferase homolog that displays denaturant-resistant binding to DNA. *Nucleic Acids Res.*, 2001, **29**: 439-448.
163. Jurkowski T.P., Shanmugam R., Helm M., and Jeltsch A., Mapping the tRNA binding site on the surface of human DNMT2 methyltransferase. *Biochemistry (Mosc.)*, 2012, **51**: 4438-4444.
164. Kuhlmann M., Borisova B.E., Kaller M., Larsson P., Stach D., et al., Silencing of retrotransposons in Dictyostelium by DNA methylation and RNAi. *Nucleic Acids Res.*, 2005, **33**: 6405-6417.
165. Schaefer M., Steringer J.P., and Lyko F., The Drosophila Cytosine-5 Methyltransferase Dnmt2 Is Associated with the Nuclear Matrix and Can Access DNA during Mitosis. *PLoS ONE*, 2008, **3**: e1414.
166. Thiagarajan D., Dev R.R., and Khosla S., The DNA methyltranferase Dnmt2 participates in RNA processing during cellular stress. *Epigenetics*, 2011, **6**: 103-113.

167. Shah J.C. and Clancy M.J., IME4, a gene that mediates MAT and nutritional control of meiosis in Saccharomyces cerevisiae. *Mol. Cell. Biol.*, 1992, **12**: 1078-1086.
168. He C., Grand challenge commentary: RNA epigenetics? *Nat. Chem. Biol.*, 2010, **6**: 863-865.
169. Rowles J., Wong M., Powers R., and Olsen M., FTO, RNA epigenetics and epilepsy. *Epigenetics*, 2012, **7**: 1094-1097.
170. Ransohoff R.M., Narayan P., Ayers D.F., Rottman F.M., and Nilsen T.W., Priming of influenza mRNA transcription is inhibited in CHO cells treated with the methylation inhibitor, neplanocin A. *Antiviral Res.*, 1987, **7**: 317-327.
171. Amalric F., Bachellerie J.P., and Caboche M., RNA methylation and control of eukaryotic RNA biosynthesis: processing and utilization of undermethylated tRNAs in CHO cells. *Nucleic Acids Res.*, 1977, **4**: 4357-4370.
172. Bachellerie J.P., Amalric F., and Caboche M., Biosynthesis and utilization of extensively undermethylated poly(A)+ RNA in CHO cells during a cycloleucine treatment. *Nucleic Acids Res.*, 1978, **5**: 2927-2943.
173. Caboche M. and Bachellerie J.P., RNA methylation and control of eukaryotic RNA biosynthesis. Effects of cycloleucine, a specific inhibitor of methylation, on ribosomal RNA maturation. *Eur. J. Biochem.*, 1977, **74**: 19-29.
174. Caboche M. and La Bonnardiere C., Vesicular stomatitis virus mRNA methylation in vivo: effect of cycloleucine, an inhibitor of S-adenosylmethionine biosynthesis, on viral transcription and translation. *Virology*, 1979, **93**: 547-557.
175. Stoltzfus C.M. and Dane R.W., Accumulation of spliced avian retrovirus mRNA is inhibited in S-adenosylmethionine-depleted chicken embryo fibroblasts. *J. Virol.*, 1982, **42**: 918-931.
176. Finkel D. and Groner Y., Methylations of adenosine residues (m6A) in pre-mRNA are important for formation of late simian virus 40 mRNAs. *Virology*, 1983, **131**: 409-425.
177. Tuck M.T., Wiehl P.E., and Pan T., Inhibition of 6-methyladenine formation decreases the translation efficiency of dihydrofolate reductase transcripts. *Int. J. Biochem. Cell Biol.*, 1999, **31**: 837-851.
178. Leach R.A. and Tuck M.T., Methionine depletion induces transcription of the mRNA (N6-adenosine)methyltransferase. *Int. J. Biochem. Cell Biol.*, 2001, **33**: 1116-1128.
179. Lu L.W., Chiang G.H., Medina D., and Randerath K., Drug effects on nucleic acid modification. I. A specific effect of 5-azacytidine on mammalian transfer RNA methylation in vivo. *Biochem. Biophys. Res. Commun.*, 1976, **68**: 1094-1101.
180. Lu L.J., Chiang G.H., and Randerath K., Effects of 5-azacytidine on transfer RNA modification: comparative study on normal and malignant tissues. *Life Sci.*, 1980, **27**: 577-584.
181. Lu L.J. and Randerath K., Mechanism of 5-azacytidine-induced transfer RNA cytosine-5-methyltransferase deficiency. *Cancer Res.*, 1980, **40**: 2701-2705.
182. Fustin J.M., Doi M., Yamaguchi Y., Hida H., Nishimura S., et al., RNA-methylation-dependent RNA processing controls the speed of the circadian clock. *Cell*, 2013, **155**: 793-806.

183. Leach R.R. and Tuck M.T., Expression of the mRNA (N6-adenosine)-methyltransferase S-adenosyl-L-methionine binding subunit mRNA in cultured cells. *Int. J. Biochem. Cell Biol.*, 2001, **33**: 984-999.
184. Tuck M.T., James C.B., Kelder B., and Kopchick J.J., Elevation of internal 6-methyladenine mRNA methyltransferase activity after cellular transformation. *Cancer Lett.*, 1996, **103**: 107-113.
185. Hongay C.F. and Orr-Weaver T.L., Drosophila Inducer of Meiosis 4 (IME4) is required for Notch signaling during oogenesis. *Proc. Natl. Acad. Sci. U. S. A.*, 2011, **108**: 14855-14860.
186. Nguyen T.T., Ma L.N., Slovak M.L., Bangs C.D., Cherry A.M., et al., Identification of novel Runx1 (AML1) translocation partner genes SH3D19, YTHDf2, and ZNF687 in acute myeloid leukemia. *Genes Chromosomes Cancer*, 2006, **45**: 918-932.
187. Okuda T., Nishimura M., Nakao M., and Fujitaa Y., RUNX1/AML1: a central player in hematopoiesis. *Int. J. Hematol.*, 2001, **74**: 252-257.
188. Cardelli M., Marchegiani F., Cavallone L., Olivieri F., Giovagnetti S., et al., A polymorphism of the YTHDF2 gene (1p35) located in an Alu-Rich genomic domain is associated with human longevity. *J. Gerontol.*, 2006, **61A**: 547-556.
189. Cheung M.K. and Yeo G.S.H., FTO biology and obesity: why do a billion of us weigh 3 kg more? *Front. Endocrin.*, 2011, **2**: 4.
190. Choquet H. and Meyre D., Genetics of obesity: What have we learned? *Curr. Genomics*, 2011, **12**: 169-179.
191. Fawcett K.A. and Barroso I., The genetics of obesity: FTO leads the way. *Trends Genet.*, 2010, **26**: 266-274.
192. Frayling T.M. and Ong K., Piecing together the FTO jigsaw. *Genome Biol.*, 2011, **12**: 104.
193. Rivera M., Cohen-Woods S., Kapur K., Breen G., Ng M.Y., et al., Depressive disorder moderates the effect of the FTO gene on body mass index. *Mol. Psychiatry*, 2012, **17**: 604-611.
194. Ho A.J., Stein J.L., Hua X., Lee S., Hibar D.P., et al., A commonly carried allele of the obesity-related FTO gene is associated with reduced brain volume in the healthy elderly. *Proc. Natl. Acad. Sci. U. S. A.*, 2010, **107**: 8404-8409.
195. Keller L., Xu W., Wang H.-X., Winblad B., Fratiglioni L., et al., The obesity related gene, FTO, interacts with APOE, and is associated with Alzheimer's disease risk: a prospective cohort study. *J. Alzheimers Dis.*, 2011, **23**: 461-469.
196. Boissel S., Reish O., Proulx K., Kawagoe-Takaki H., Sedgwick B., et al., Loss-of-function mutation in the dioxygenase-encoding FTO gene causes severe growth retardation and multiple malformations. *Am. J. Hum. Genet.*, 2009, **85**: 106-111.
197. Peters T., Ausmeier K., Dildrop R., and Ruther U., The mouse Fused toes (Ft) mutation is the result of a 1.6-Mb deletion including the entire Iroquois B gene cluster. *Mamm. Genome*, 2002, **13**: 186-188.
198. Fischer J., Koch L., Emmerling C., Vierkotten J., Peters T., et al., Inactivation of the Fto gene protects from obesity. *Nature*, 2009, **458**: 894-898.
199. Fredriksson R., Hagglund M., Olszewski P.K., Stephansson O., Jacobsson J.A., et al., The obesity gene, FTO, is of ancient origin, up-regulated during food deprivation and expressed in neurons of feeding-related nuclei of the brain. *Endocrinology*, 2008, **149**: 2062-2071.

200. McMurray F., Moir L., and Cox R.D., From mice to humans. *Curr. Diab. Rep.*, 2012, **12**: 651-658.
201. Church C., Lee S., Bagg E.A.L., McTaggart J.S., Deacon R., et al., A mouse model for the metabolic effects of the human fat mass and obesity associated FTO gene. *PLoS Genet.*, 2009, **5**: e1000599.
202. Church C., Moir L., McMurray F., Girard C., Banks G.T., et al., Overexpression of Fto leads to increased food intake and results in obesity. *Nat. Genet.*, 2010, **42**: 1086-1092.
203. Gao X., Shin Y.-H., Li M., Wang F., Tong Q., et al., The fat mass and obesity associated gene FTO functions in the brain to regulate postnatal growth in mice. *PLoS ONE*, 2010, **5**: e14005.
204. Smemo S., Tena J.J., Kim K.H., Gamazon E.R., Sakabe N.J., et al., Obesity-associated variants within FTO form long-range functional connections with IRX3. *Nature*, 2014, **507**: 371-375.
205. McMurray F., Church C.D., Larder R., Nicholson G., Wells S., et al., Adult onset global loss of the fto gene alters body composition and metabolism in the mouse. *PLoS Genet.*, 2013, **9**: e1003166.
206. Gulati P., Cheung M.K., Antrobus R., Church C.D., Harding H.P., et al., Role for the obesity-related FTO gene in the cellular sensing of amino acids. *Proc. Natl. Acad. Sci. U. S. A.*, 2013, **110**: 2557-2562.
207. Okamoto M., Hirata S., Sato S., Koga S., Fujii M., et al., Frequent Increased Gene Copy Number and High Protein Expression of tRNA (Cytosine-5-)-Methyltransferase (NSUN2) in Human Cancers. *DNA Cell Biol.*, 2011, **31**: 660-671.
208. Sakita-Suto S., Kanda A., Suzuki F., Sato S., Takata T., et al., Aurora-B regulates RNA methyltransferase NSUN2. *Mol. Biol. Cell*, 2007, **18**: 1107-1117.
209. Abbasi-Moheb L., Mertel S., Gonsior M., Nouri-Vahid L., Kahrizi K., et al., Mutations in NSUN2 cause autosomal-recessive intellectual disability. *Am. J. Hum. Genet.*, 2012, **90**: 847-855.
210. Khan M.A., Rafiq M.A., Noor A., Hussain S., Flores J.V., et al., Mutation in NSUN2, which encodes an RNA methyltransferase, causes autosomal-recessive intellectual disability. *Am. J. Hum. Genet.*, 2012, **90**: 856-863.
211. Hussain S., Benavente S.B., Nascimento E., Dragoni I., Kurowski A., et al., The nucleolar RNA methyltransferase Misu (NSun2) is required for mitotic spindle stability. *J. Cell Biol.*, 2009, **186**: 27-40.
212. Hussain S., Tuorto F., Menon S., Blanco S., Cox C., et al., The mouse cytosine-5 RNA methyltransferase NSun2 is a component of the chromatoid body and required for testis differentiation. *Mol. Cell. Biol.*, 2013, **33**: 1561-1570.
213. Blanco S., Kurowski A., Nichols J., Watt F.M., Benitah S.A., et al., The RNA-Methyltransferase Misu (NSun2) Poises Epidermal Stem Cells to Differentiate. *PLoS Genet.*, 2011, **7**: e1002403.
214. Freude K., Hoffmann K., Jensen L.-R., Delatycki M.B., des Portes V., et al., Mutations in the FTSJ1 gene coding for a novel S-adenosylmethionine-binding protein cause nonsyndromic X-linked mental retardation. *Am. J. Hum. Genet.*, 2004, **75**: 305-309.
215. Ramser J., Winnepenninckx B., Lenski C., Errijgers V., Platzer M., et al., A splice site mutation in the methyltransferase gene FTSJ1 in Xp11.23 is associated with non-syndromic mental retardation in a large Belgian family (MRX9). *J. Med. Genet.*, 2004, **41**: 679-683.

216. Najmabadi H., Hu H., Garshasbi M., Zemojtel T., Abedini S.S., et al., Deep sequencing reveals 50 novel genes for recessive cognitive disorders. *Nature*, 2011, **478**: 57-63.
217. Fisher O., Siman Tov R., and Ankri S., Pleiotropic phenotype in Entamoeba histolytica overexpressing DNA methyltransferase (Ehmeth). *Mol. Biochem. Parasitol.*, 2006, **147**: 48-54.
218. Lin M.-J., Tang L.-Y., Reddy M.N., and Shen C.-K.J., DNA methyltransferase gene dDnmt2 and longevity of Drosophila. *J. Biol. Chem.*, 2005, **280**: 861-864.
219. Rai K., Chidester S., Zavala C.V., Manos E.J., James S.R., et al., Dnmt2 functions in the cytoplasm to promote liver, brain, and retina development in zebrafish. *Genes Dev.*, 2007, **21**: 261-266.
220. Scott L.J., Mohlke K.L., Bonnycastle L.L., Willer C.J., Li Y., et al., A genome-wide association study of type 2 diabetes in Finns detects multiple susceptibility variants. *Science*, 2007, **316**: 1341-1345.
221. Hornung V., Ellegast J., Kim S., Brzozka K., Jung A., et al., 5'-Triphosphate RNA is the ligand for RIG-I. *Science*, 2006, **314**: 994-997.
222. Züst R., Cervantes-Barragan L., Habjan M., Maier R., Neuman B.W., et al., Ribose 2'-O-methylation provides a molecular signature for the distinction of self and non-self mRNA dependent on the RNA sensor Mda5. *Nat. Immunol.*, 2011, **12**: 137-143.
223. Karikó K., Buckstein M., Ni H., and Weissman D., Suppression of RNA recognition by Toll-like receptors: the impact of nucleoside modification and the evolutionary origin of RNA. *Immunity*, 2005, **23**: 165-175.
224. Karikó K., Muramatsu H., Welsh F.A., Ludwig J., Kato H., et al., Incorporation of pseudouridine into mRNA yields superior nonimmunogenic vector with increased translational capacity and biological stability. *Mol. Ther.*, 2008, **16**: 1833-1840.
225. Warren L., Manos P.D., Ahfeldt T., Loh Y.-H., Li H., et al., Highly efficient reprogramming to pluripotency and directed differentiation of human cells with synthetic modified mRNA. *Cell Stem Cell*, 2010, **7**: 618-630.
226. Kruse S., Zhong S., Bodi Z., Button J., Alcocer M.J.C., et al., A novel synthesis and detection method for cap-associated adenosine modifications in mouse mRNA. *Scientific reports*, 2011, **1**: 126.
227. Narayan P., Ludwiczak R.L., Goodwin E.C., and Rottman F.M., Context effects on N6-adenosine methylation sites in prolactin mRNA. *Nucleic Acids Res.*, 1994, **22**: 419-426.
228. Carroll S.M., Narayan P., and Rottman F.M., N6-methyladenosine residues in an intron-specific region of prolactin pre-mRNA. *Mol. Cell. Biol.*, 1990, **10**: 4456-4465.
229. Canaani D., Kahana C., Lavi S., and Groner Y., Identification and mapping of N6-methyladenosine containing sequences in simian virus 40 RNA. *Nucleic Acids Res.*, 1979, **6**: 2879-2899.
230. Camper S.A., Albers R.J., Coward J.K., and Rottman F.M., Effect of undermethylation on mRNA cytoplasmic appearance and half-life. *Mol. Cell. Biol.*, 1984, **4**: 538-543.
231. Zhao X., Yang Y., Sun B.F., Shi Y., Yang X., et al., FTO-dependent demethylation of N6-methyladenosine regulates mRNA splicing and is required for adipogenesis. *Cell Res.*, 2014, **24**: 1403-1419.

232. Parker B.J., Moltke I., Roth A., Washietl S., Wen J., et al., New families of human regulatory RNA structures identified by comparative analysis of vertebrate genomes. *Genome Res.*, 2011, **21**: 1929-1943.
233. Hafner M., Landthaler M., Burger L., Khorshid M., Hausser J., et al., Transcriptome-wide identification of RNA-binding protein and microRNA target sites by PAR-CLIP. *Cell*, 2010, **141**: 129-141.
234. Bhattacharyya S.N., Habermacher R., Martine U., Closs E.I., and Filipowicz W., Relief of microRNA-mediated translational repression in human cells subjected to stress. *Cell*, 2006, **125**: 1111-1124.
235. Untergasser A., Cutcutache I., Koressaar T., Ye J., Faircloth B.C., et al., Primer3--new capabilities and interfaces. *Nucleic Acids Res.*, 2012, **40**: e115.
236. Koressaar T. and Remm M., Enhancements and modifications of primer design program Primer3. *Bioinformatics*, 2007, **23**: 1289-1291.
237. Kreck B., Marnellos G., Richter J., Krueger F., Siebert R., et al., B-SOLANA: an approach for the analysis of two-base encoding bisulfite sequencing data. *Bioinformatics*, 2012, **28**: 428-429.
238. Trapnell C., Pachter L., and Salzberg S.L., TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics*, 2009, **25**: 1105-1111.
239. Snedecor G.W. and Cochran W.G., *Statistical methods*. 7th ed. 1980: Iowa State University Press Ames, Iowa.
240. Smyth G.K., *limma: Linear Models for Microarray Data*, in *Bioinformatics and Computational Biology Solutions Using R and Bioconductor*, R. Gentleman, et al., Editors. 2005, Springer New York. p. 397-420.
241. Yao Z., Weinberg Z., and Ruzzo W.L., CMfinder--a covariance model based RNA motif finding algorithm. *Bioinformatics*, 2006, **22**: 445-452.
242. Bembom O., *seqLogo: Sequence logos for DNA sequence alignments*, R package version 1.30.0.
243. Shannon C.E., A Mathematical Theory of Communication. *The Bell System Technical Journal*, 1948, **27**: 379-423, 623-656.
244. Lebedeva S., Jens M., Theil K., Schwanhausser B., Selbach M., et al., Transcriptome-wide analysis of regulatory interactions of the RNA-binding protein HuR. *Mol. Cell*, 2011, **43**: 340-352.
245. Pollard K.S., Hubisz M.J., Rosenbloom K.R., and Siepel A., Detection of nonneutral substitution rates on mammalian phylogenies. *Genome Res.*, 2010, **20**: 110-121.
246. Luo W., Friedman M.S., Shedden K., Hankenson K.D., and Woolf P.J., GAGE: generally applicable gene set enrichment for pathway analysis. *BMC Bioinformatics*, 2009, **10**: 161.
247. Alexa A. and Rahnenfuhrer J., *topGO: Enrichment analysis for Gene Ontology*, 2010, R package version 2.6.0.
248. Bolger A.M., Lohse M., and Usadel B., Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*, 2014, **30**: 2114-2120.
249. Krueger F. and Andrews S.R., Bismark: a flexible aligner and methylation caller for Bisulfite-Seq applications. *Bioinformatics*, 2011, **27**: 1571-1572.
250. Tusnady G.E., Simon I., Varadi A., and Aranyi T., BiSearch: primer-design and search tool for PCR on bisulfite-treated genomes. *Nucleic Acids Res.*, 2005, **33**: e9.

251. Nwagwu M. and Nana M., Ribonucleic acid synthesis in embryonic chick muscle, rates of synthesis and half-lives of transfer and ribosomal RNA species. *J. Embryol. Exp. Morphol.*, 1980, **56**: 253-267.
252. Karnahl U. and Wasternack C., Half-life of cytoplasmic rRNA and tRNA, of plastid rRNA and of uridine nucleotides in heterotrophically and photoorganotrophically grown cells of Euglena gracilis and its apoplastic mutant W3BUL. *Int. J. Biochem.*, 1992, **24**: 493-497.
253. Kanerva P.A. and Mäenpää P.H., Codon-specific serine transfer ribonucleic acid degradation in avian liver during vitellogenin induction. *Acta Chem. Scand.*, 1981, **35**: 379-385.
254. Archer S.K., Shirokikh N.E., and Preiss T., Selective and flexible depletion of problematic sequences from RNA-seq libraries at the cDNA stage. *BMC Genomics*, 2014, **15**: 401.
255. Baer R.J. and Dubin D.T., Methylated regions of hamster mitochondrial ribosomal RNA: structural and functional correlates. *Nucleic Acids Res.*, 1981, **9**: 323-337.
256. Iwanami Y. and Brown G.M., Methylated bases of ribosomal ribonucleic acid from HeLa cells. *Arch. Biochem. Biophys.*, 1968, **126**: 8-15.
257. Olson M.V., Page G.S., Sentenac A., Piper P.W., Worthington M., et al., Only one of two closely related yeast suppressor tRNA genes contains an intervening sequence. *Nature*, 1981, **291**: 464-469.
258. Weissenbach J., Kiraly I., and Dirheimer G., Primary structure of tRNA Thr 1a and b from brewer's yeast. *Biochimie*, 1977, **59**: 381-391.
259. Ryvkin P., Leung Y.Y., Silverman I.M., Childress M., Valladares O., et al., HAMR: high-throughput annotation of modified ribonucleotides. *RNA*, 2013, **19**: 1684-1692.
260. Huang Y., Pastor W.A., Shen Y., Tahiliani M., Liu D.R., et al., The behaviour of 5-hydroxymethylcytosine in bisulfite sequencing. *PLoS ONE*, 2010, **5**: e8888.
261. Jin S.G., Kadam S., and Pfeifer G.P., Examination of the specificity of DNA methylation profiling techniques towards 5-methylcytosine and 5-hydroxymethylcytosine. *Nucleic Acids Res.*, 2010, **38**: e125.
262. Tahiliani M., Koh K.P., Shen Y., Pastor W.A., Bandukwala H., et al., Conversion of 5-methylcytosine to 5-hydroxymethylcytosine in mammalian DNA by MLL partner TET1. *Science*, 2009, **324**: 930-935.
263. Fu L., Guerrero C.R., Zhong N., Amato N.J., Liu Y., et al., Tet-mediated formation of 5-hydroxymethylcytosine in RNA. *J. Am. Chem. Soc.*, 2014, **136**: 11582-11585.
264. Ito S., Shen L., Dai Q., Wu S.C., Collins L.B., et al., Tet proteins can convert 5-methylcytosine to 5-formylcytosine and 5-carboxylcytosine. *Science*, 2011, **333**: 1300-1303.
265. Pfaffeneder T., Hackner B., Truss M., Munzel M., Muller M., et al., The discovery of 5-formylcytosine in embryonic stem cell DNA. *Angew. Chem. Int. Ed. Engl.*, 2011, **50**: 7008-7012.
266. Booth M.J., Branco M.R., Ficz G., Oxley D., Krueger F., et al., Quantitative sequencing of 5-methylcytosine and 5-hydroxymethylcytosine at single-base resolution. *Science*, 2012, **336**: 934-937.
267. Quail M.A., Kozarewa I., Smith F., Scally A., Stephens P.J., et al., A large genome center's improvements to the Illumina sequencing system. *Nat. Methods*, 2008, **5**: 1005-1010.

268. Wolin S.L. and Matera A.G., The trials and travels of tRNA. *Genes Dev.*, 1999, **13**: 1-10.
269. O'Connor J.P. and Peebles C.L., In vivo pre-tRNA processing in Saccharomyces cerevisiae. *Mol. Cell. Biol.*, 1991, **11**: 425-439.
270. Lund E. and Dahlberg J.E., Proofreading and Aminoacylation of tRNAs Before Export from the Nucleus. *Science*, 1998, **282**: 1082-2085.
271. Chan P.P. and Lowe T.M., GtRNAdb: a database of transfer RNA genes detected in genomic sequence. *Nucleic Acids Res.*, 2009, **37**: D93-97.
272. Johnson P.F. and Abelson J., The yeast tRNATyr gene intron is essential for correct modification of its tRNA product. *Nature*, 1983, **302**: 681-687.
273. Dubin D.T. and HsuChen C.C., The 3'-terminal region of mosquito mitochondrial small ribosomal subunit RNA: sequence and localization of methylated residues. *Plasmid*, 1983, **9**: 307-320.
274. Reiner R., Ben-Asouli Y., Krilovetzky I., and Jarrous N., A role for the catalytic ribonucleoprotein RNase P in RNA polymerase III transcription. *Genes Dev.*, 2006, **20**: 1621-1635.
275. Esakova O. and Krasilnikov A.S., Of proteins and RNA: the RNase P/MRP family. *RNA*, 2010, **16**: 1725-1747.
276. Tycowski K.T., Aab A., and Steitz J.A., Guide RNAs with 5' caps and novel box C/D snoRNA-like domains for modification of snRNAs in metazoa. *Curr. Biol.*, 2004, **14**: 1985-1995.
277. Tycowski K.T., Smith C.M., Shu M.-D., and Steitz J.A., A small nucleolar RNA requirement for site-specific ribose methylation of rRNA in Xenopus. *Proc. Natl. Acad. Sci. U. S. A.*, 1996, **93**: 14480-14485.
278. Lestrade L. and Weber M.J., snoRNA-LBME-db, a comprehensive database of human H/ACA and C/D box snoRNAs. *Nucleic Acids Res.*, 2006, **34**: D158-162.
279. Brown J.W., The ribonuclease P database. *Nucleic Acids Res.*, 1997, **25**: 263-264.
280. Burge S.W., Daub J., Eberhardt R., Tate J., Barquist L., et al., Rfam 11.0: 10 years of RNA families. *Nucleic Acids Res.*, 2013, **41**: D226-232.
281. Andronescu M., Bereg V., Hoos H.H., and Condon A., RNA STRAND: the RNA secondary structure and statistical analysis database. *BMC Bioinformatics*, 2008, **9**: 340.
282. Stadler P.F., Chen J.J., Hackermuller J., Hoffmann S., Horn F., et al., Evolution of vault RNAs. *Mol. Biol. Evol.*, 2009, **26**: 1975-1991.
283. Beier H. and Grimm M., Misreading of termination codons in eukaryotes by natural nonsense suppressor tRNAs. *Nucleic Acids Res.*, 2001, **29**: 4767-4782.
284. Li J., Esberg B., Curran J.F., and Bjork G.R., Three Modified Nucleosides Present in the Anticodon Stem and Loop Influence the in vivo aa-tRNA Selection in a tRNA-dependent Manner. *J. Mol. Biol.*, 1997, **271**: 209-221.
285. Pütz J., Florentz C., Benseler F., and Giegé R., A single methyl group prevents the mischarging of a tRNA. *Nat. Struct. Biol.*, 1994, **1**: 580-582.
286. Frank D.N. and Pace N.R., Ribonuclease P- unity and diversity in a tRNA processing ribozyme. *Annu. Rev. Biochem.*, 1998, **67**: 153-180.
287. Matera A.G., Frey M.R., Margelot K., and Wolin S.L., A perinucleolar compartment contains several RNA polymerase III transcripts as well as the polypyrimidine tract-binding protein, hnRNP I. *J. Cell Biol.*, 1995, **129**: 1181-1193.

288. Lee B., Matera A.G., Ward D.C., and Craft J., Association of RNase mitochondrial RNA processing enzyme with ribonuclease P in higher ordered structures in the nucleolus- a possible coordinate role in ribosome. *Proc. Natl. Acad. Sci. U. S. A.*, 1996, **93**: 11471-11476.
289. Green C.J., Sohel I., and Vold B.S., The discovery of new intron containing human tRNA genes using the polymerase chain reaction. *J. Biol. Chem.*, 1990, **21**: 12139-12142.
290. Spåhr H., Habermann B., Gustafsson C.M., Larsson N.-G.G., and Hallberg B.M., Structure of the human MTERF4-NSUN4 protein complex that regulates mitochondrial ribosome biogenesis. *Proc. Natl. Acad. Sci. U. S. A.*, 2012, **109**: 15253-15258.
291. Yakubovskaya E., Guja K.E., Mejia E., Castano S., Hambardjieva E., et al., Structure of the essential MTERF4:NSUN4 protein complex reveals how an MTERF protein collaborates to facilitate rRNA modification. *Structure*, 2012, **20**: 1940-1947.
292. Hu S., Wu J., Chen L., and Shan G., Signals from noncoding RNAs: unconventional roles for conventional pol III transcripts. *Int. J. Biochem. Cell Biol.*, 2012, **44**: 1847-1851.
293. Gerard M.A., Myslinski E., Chylak N., Baudrey S., Krol A., et al., The scaRNA2 is produced by an independent transcription unit and its processing is directed by the encoding region. *Nucleic Acids Res.*, 2010, **38**: 370-381.
294. Pannucci J.A., Haas E.S., Hall T.A., Harris K.J., and Brown J.W., RNase P RNAs from some Archaea are catalytically active. *Proc. Natl. Acad. Sci. U. S. A.*, 1999, **96**: 7803-4808.
295. Kikovska E., Svard S.G., and Kirsebom L.A., Eukaryotic RNase P RNA mediates cleavage in the absence of protein. *Proc. Natl. Acad. Sci. U. S. A.*, 2007, **104**: 2062-2067.
296. Brannvall M., Kikovska E., and Kirsebom L.A., Cross talk between the +73/294 interaction and the cleavage site in RNase P RNA mediated cleavage. *Nucleic Acids Res.*, 2004, **32**: 5418-5429.
297. Kirsebom L.A. and Svärd S.G., Base pairing between Escherichia coli RNase P RNA and its substrate. *EMBO J.*, 1994, **13**: 4870-4876.
298. Busch S., Kirsebom L.A., Notbohm H., and Hartmann R.K., Differential role of the intermolecular base-pairs G292-C(75) and G293-C(74) in the reaction catalyzed by Escherichia coli RNase P RNA. *J. Mol. Biol.*, 2000, **299**: 941-951.
299. Persson H., Kvist A., Vallon-Christersson J., Medstrand P., Borg A., et al., The non-coding RNA of the multidrug resistance-linked vault particle encodes multiple regulatory small RNAs. *Nat. Cell Biol.*, 2009, **11**: 1268-1271.
300. Doll A. and Grzeschik K.H., Characterization of two novel genes, WBSCR20 and WBSCR22, deleted in Williams-Beuren syndrome. *Cytogenet. Cell Genet.*, 2001, **95**: 20-27.
301. Lovejoy C.A., Xu X., Bansbach C.E., Glick G.G., Zhao R., et al., Functional genomic screens identify CINP as a genome maintenance protein. *Proc. Natl. Acad. Sci. U. S. A.*, 2009, **106**: 19304-19309.
302. Zhang B., Cunningham M.A., Nichols W.C., Bernat J.A., Seligsohn U., et al., Bleeding due to disruption of a cargo-specific ER-to- Golgi transport complex. *Nat. Genet.*, 2003, **34**: 220-225.
303. He W., Shi Q., Hu X., and Yan R., The membrane topology of RTN3 and its effect on binding of RTN3 to BACE1. *J. Biol. Chem.*, 2007, **282**: 29144-29151.

304. Yamasaki S., Ivanov P., Hu G.F., and Anderson P., Angiogenin cleaves tRNA and promotes stress-induced translational repression. *J. Cell Biol.*, 2009, **185**: 35-42.
305. Ashburner M., Ball C.A., Blake J.A., Botstein D., Butler H., et al., Gene Ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.*, 2000, **25**: 25-29.
306. Kanehisa M. and Goto S., KEGG: Kyoto Encylopedia of Genes and Genomes. *Nucleic Acids Res.*, 2000, **28**: 27-30.
307. Kanehisa M., Goto S., Sato Y., Kawashima M., Furumichi M., et al., Data, information, knowledge and principle: back to metabolism in KEGG. *Nucleic Acids Res.*, 2014, **42**: D199-205.
308. Chi S.W., Zang J.B., Mele A., and Darnell R.B., Argonaute HITS-CLIP decodes microRNA-mRNA interaction maps. *Nature*, 2009, **460**: 479-486.
309. Eisenberg E. and Levanon E.Y., Human housekeeping genes are compact. *Trends Genet.*, 2003, **19**: 362-365.
310. Lewis B.P., Burge C.B., and Bartel D.P., Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets. *Cell*, 2005, **120**: 15-20.
311. Lyko F., Ramsahoy B.H., and Jaenisch R., DNA methylation in Drosophila melanogaster. *Nature*, 2000, **408**: 538-540.
312. Phalke S., Nickel O., Walluscheck D., Hortig F., Onorati M.C., et al., Retrotransposon silencing and telomere integrity in somatic cells of Drosophila depends on the cytosine-5 methyltransferase DNMT2. *Nat. Genet.*, 2009, **41**: 696-702.
313. Raddatz G., Guzzardo P.M., Olova N., Fantappie M.R., Rampp M., et al., Dnmt2-dependent methylomes lack defined DNA methylation patterns. *Proc. Natl. Acad. Sci. U. S. A.*, 2013, **110**: 8627-8631.
314. Hermann A., Schmitt S., and Jeltsch A., The human Dnmt2 has residual DNA-(cytosine-C5) methyltransferase activity. *J. Biol. Chem.*, 2003, **278**: 31717-31721.
315. Schafer T., Maco B., Petfalski E., Tollervey D., Bottcher B., et al., Hrr25-dependent phosphorylation state regulates organization of the pre-40S subunit. *Nature*, 2006, **441**: 651-655.
316. Tuorto F., Liebers R., Musch T., Schaefer M., Hofmann S., et al., RNA cytosine methylation by Dnmt2 and NSun2 promotes tRNA stability and protein synthesis. *Nat. Struct. Mol. Biol.*, 2012, **19**: 900-905.
317. Frye M., Dragoni I., Chin S.-F., Spiteri I., Kurowski A., et al., Genomic gain of 5p15 leads to over-expression of Misu (NSUN2) in breast cancer. *Cancer Lett.*, 2010, **289**: 71-80.
318. Ruggero D. and Pandolfi P.P., Does the ribosome translate cancer? *Nat. Rev. Cancer*, 2003, **3**: 179-192.
319. Pandolfi P.P., Aberrant mRNA translation in cancer pathogenesis: an old concept revisited comes finally of age. *Oncogene*, 2004, **23**: 3134-3137.
320. Liu P., Ramachandran S., Ali Seyed M., Scharer C.D., Laycock N., et al., Sex-determining region Y box 4 is a transforming oncogene in human prostate cancer cells. *Cancer Res.*, 2006, **66**: 4011-4019.
321. Pramoonjago P., Baras A.S., and Moskaluk C.A., Knockdown of Sox4 expression by RNAi induces apoptosis in ACC3 cells. *Oncogene*, 2006, **25**: 5626-5639.

322. Liao Y.L., Sun Y.M., Chau G.Y., Chau Y.P., Lai T.C., et al., Identification of SOX4 target genes using phylogenetic footprinting-based prediction from expression microarrays suggests that overexpression of SOX4 potentiates metastasis in hepatocellular carcinoma. *Oncogene*, 2008, **27**: 5578-5589.
323. Tavazoie S.F., Alarcon C., Oskarsson T., Padua D., Wang Q., et al., Endogenous human microRNAs that suppress breast cancer metastasis. *Nature*, 2008, **451**: 147-152.
324. Aaboe M., Birkenkamp-Demtroder K., Wiuf C., Sorensen F.B., Thykjaer T., et al., SOX4 expression in bladder carcinoma: clinical aspects and in vitro functional characterization. *Cancer Res.*, 2006, **66**: 3434-3442.
325. Tiwari N., Tiwari V.K., Waldmeier L., Balwierz P.J., Arnold P., et al., Sox4 is a master regulator of epithelial-mesenchymal transition by controlling Ezh2 expression and epigenetic reprogramming. *Cancer Cell*, 2013, **23**: 768-783.
326. Lapeyre B. and Purushothaman S.K., Spb1p-directed formation of Gm2922 in the ribosome catalytic center occurs at a late processing stage. *Mol. Cell*, 2004, **16**: 663-669.
327. Morello L.G., Coltri P.P., Quaresma A.J., Simabuco F.M., Silva T.C., et al., The human nucleolar protein FTSJ3 associates with NIP7 and functions in pre-rRNA processing. *PLoS ONE*, 2011, **6**: e29174.
328. Sherry S.T., Ward M.-H., Kholodov M., Baker J., Phan L., et al., dbSNP- the NCBI database of genetic variation. *Nucleic Acids Res.*, 2001, **29**: 308-311.
329. Belham C., Roig J., Caldwell J.A., Aoyama Y., Kemp B.E., et al., A mitotic cascade of NIMA family kinases. Nercc1/Nek9 activates the Nek6 and Nek7 kinases. *J. Biol. Chem.*, 2003, **278**: 34897-34909.
330. Foster D.A. and Xu L., Phospholipase D in Cell Proliferation and Cancer. *Mol. Cancer Res.*, 2003, **1**: 789-800.
331. Lu Z., Hornia A., Joseph T., Sukezane T., Frankel P., et al., Phospholipase D and RalA cooperate with the epidermal growth factor receptor to transform 3Y1 rat fibroblasts. *Mol. Cell. Biol.*, 2000, **20**: 462-467.
332. Joseph T., Wooden R., Bryant A., Zhong M., Lu Z., et al., Transformation of cells overexpressing a tyrosine kinase by phospholipase D1 and D2. *Biochem. Biophys. Res. Commun.*, 2001, **289**: 1019-1024.
333. Zhong M., Shen Y., Zheng Y., Joseph T., Jackson D., et al., Phospholipase D prevents apoptosis in v-Src-transformed rat fibroblasts and MDA-MB-231 breast cancer cells. *Biochem. Biophys. Res. Commun.*, 2003, **302**: 615-619.
334. Cruchaga C., Karch C.M., Jin S.C., Benitez B.A., Cai Y., et al., Rare coding variants in the phospholipase D3 gene confer risk for Alzheimer's disease. *Nature*, 2014, **505**: 550-554.
335. Lalioti M.D., Chen H., Rossier C., Shafaatian R., Reid J.D., et al., Cloning the cDNA of human PWP2, which encodes a protein with WD repeats and maps to 21q22.3. *Genomics*, 1996, **35**: 321-327.
336. Dosil M. and Bustelo X.R., Functional characterization of Pwp2, a WD family protein essential for the assembly of the 90 S pre-ribosomal particle. *J. Biol. Chem.*, 2004, **279**: 37385-37397.
337. Gruber J.J., Zatechka D.S., Sabin L.R., Yong J., Lum J.J., et al., Ars2 links the nuclear cap-binding complex to RNA interference and cell proliferation. *Cell*, 2009, **138**: 328-339.

338. Sabin L.R., Zhou R., Gruber J.J., Lukinova N., Bambina S., et al., Ars2 regulates both miRNA- and siRNA- dependent silencing and suppresses RNA virus infection in Drosophila. *Cell*, 2009, **138**: 340-351.
339. Laubinger S., Sachsenberg T., Zeller G., Busch W., Lohmann J.U., et al., Dual roles of the nuclear cap-binding complex and SERRATE in pre-mRNA splicing and microRNA processing in Arabidopsis thaliana. *Proc. Natl. Acad. Sci. U. S. A.*, 2008, **105**: 8795-8800.
340. Lee J.Y., Rohlman C.E., Molony L.A., and Engelke D.R., Characterization of RPR1, an essential gene encoding the RNA component of Saccharomyces cerevisiae nuclear RNase P. *Mol. Cell. Biol.*, 1991, **11**: 721-730.
341. Yoo C.J. and Wolin S.L., The Yeast La Protein Is Required for the 3' Endonucleolytic Cleavage That Matures tRNA Precursors. *Cell*, 1997, **89**: 393-402.
342. Fu Y., Jia G., Pang X., Wang R.N., Wang X., et al., FTO-mediated formation of N6-hydroxymethyladenosine and N6-formyladenosine in mammalian RNA. *Nat. Commun.*, 2013, **4**: 1798.
343. He Y.F., Li B.Z., Li Z., Liu P., Wang Y., et al., Tet-Mediated Formation of 5-Carboxylcytosine and Its Excision by TDG in Mammalian DNA. *Science*, 2011, **333**: 1303-1307.